Artificial Intelligence Review (2025) 58:381
https://doi.org/10.1007/s10462-025-11368-7

A survey on deep learning fundamentals

Chunwei Tian[1] · Tongtong Cheng[2] · Zhe Peng[3] · Wangmeng Zuo[1] · Yonglin Tian[4] · Qingfu Zhang[5] · Fei-Yue Wang[4] · David Zhang[2,6]

Accepted: 15 August 2025 / Published online: 17 October 2025

Abstract
Deep learning, driven by big data and graphic processing units, has garnered significant attention across various domains. The flexibility of network architectures, combined with their diverse components, has allowed deep learning techniques to be applied across a wide range of domains, expanding from low- and high-level computer vision tasks to encompass video processing, natural language processing (NLP), and 3D data processing. However, there has been relatively little effort to systematically summarise these works from principles to applications in terms of deep learning fundamentals. The present study aims to address this gap in the literature by presenting components of deep networks for image applications, and describing several classical deep networks for image applications. The study then introduces principles, relations, ranges, and applications of deep networks across an expanded scope, covering low-level vision tasks, high-level vision tasks, video processing, NLP, and 3D data processing. The study then compares the performance of different networks across these diverse tasks. Finally, it summarises potential focuses and challenges of deep learning research for these applications with concluding remarks.

Keywords Deep learning · Vision tasks · Natural language processing · 3D convolutional neural networks · Artificial intelligence

1 Introduction

Deep learning has achieved remarkable progress in recent years, demonstrating significant potential across a wide range of applications. Its core strength lies in the ability to automatically learn complex features from data, providing efficient solutions for various tasks. Deep learning for image application has been extensively utilized in fields such as aeronautics and space, biomedical engineering, communication engineering, industrial automation, military and public security, robotics, and e-commerce, making it a focal point of research for many scholars. Deep learning has diverse applications across multiple domains, including image segmentation, classification, object detection, and restoration, as well as video processing tasks like video tracking and action recognition, natural language processing

Extended author information available on the last page of the article

tasks such as text generation and cross-modal integration, and 3D data processing tasks like 3D object recognition and reconstruction. Notably, deep-learning-based image denoising has become a crucial area. In the past 20 years, its development exemplifies the progress of deep-learning applications in the image domain. Buades et al. (2005) proposed a non-local algorithm method to deal with image denoising. Lan et al. (2006) fused belief propagation inference method and Markov Random Fields (MRFs) to address image denoising. Dabov et al. (2007) presented to transform grouping similar two-dimensional image fragments into three-dimensional data arrays to improve sparsity for image denoising. Although, these selection and extraction methods have amazing performance for image denoising, most of these traditional methods have two challenges as follows (Zhang et al. 2017). First, these methods are non-convex, which need to manually set parameters. Second, these methods refer a complex optimization problem for the test stage, which results in high computational costs.

However, research have illustrated that deep learning technologies can reply to deeper architectures to automatically learn and find more suitable image features rather than manual setting parameters, which effectively addresses drawbacks of traditional methods above (Krizhevsky et al. 2012). Big data and GPU are also essential for deep learning technologies to improve learning ability (Li et al. 2019). The learning ability of deep learning is finished by a model (also referred to as a network) and that is made up of many layers, i.e., convolutional layer, pooling layer, batch normalization layer and full connection layer. That is, deep learning technologies can convert input data (e.g., images, speech and video) into outputs (e.g., object category, password unlocking and traffic information) by a deep model (Litjens et al. 2017). Specifically, convolutional neural networks (CNNs) stand out as the most typical and successful deep-learning networks when it comes to applications in the realm of images (Lawrence et al. 1997). CNNs originated LeNet in 1998 and it was successfully used in hand-written digit recognition, which obtained excellent performance (LeCun et al. 1989). Although, CNNs have obtained initial success for hand-written, they have not been widely used in other real applications until arise of GPU and big data. In other words, the real success of CNNs attributed to ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012), where a new CNN was proposed, named AlexNet and became world champion in this ILSVRC 2012 (Krizhevsky et al. 2012).

In the years that followed, deeper neural networks gained significant popularity. These networks have demonstrated remarkable efficacy in various image-related tasks (Simonyan and Zisserman 2014). Simonyan and Zisserman (2014) increased the depth of neural networks to 16–19 weighted layers and convolution filter size of each layer is set to 3×3 to improve performance for image recognition. Christian Szegedy et al. (2015) used sparsely connected layers instead of fully connected layers to increase width and depth in neural networks for image classification which is named Inception V1. Inception V1 effectively prevented to overfitting from enlarged size (width) of network and reduced the computing resource from increased depth of network. Ioffe (2015) used two 3 convolutions to take place of a convolution with 5 in speech recognition, which reduced the number of parameters and established more nonlinear transformation to improve learning ability for extraction features. Besides, batch normalization (BN) was proposed in this network to address internal covariate shift in image denoising which was called as to Inception V2. Inception V3 based on Inception V2 was proposed by Szegedy et al. (2016), which decomposed a two-dimension convolution into two one-dimension convolutions to deal with image clas-

sification, i.e., a two-dimension convolution of 7×7 was divided into two one-dimensions of 1×7 and 7×1. Thus Inception V3 not only accelerate computing speed of network, but also increased the depth of network to enhance the non-linearity of network. Despite deep networks have obtained successfully applications for image classification (Sermanet et al. 2013), they can generate vanishing gradient when network depth, which makes network hamper convergence. Also deeper neural networks get to converge, networks are saturated and degrade quickly with increasing depth of networks. The appearance of residual network effectively deals with problems above for image recognition (He et al. 2016). Additionally, the idea of training two models jointly was proposed and named the Generative Adversarial Network (GAN) for image-related applications (Goodfellow 2016). A GAN was trained together. One model was named as to a generator, which was used to generate the same distribution as the training data. Another model is regarded as a discriminator, which was exploited to determine if the data was real data or counterfeit (Pan et al. 2017). GAN method has a lot of applications for image single image super-resolution (Ledig et al. 2017), image to image translation (Isola et al. 2017), image editing (Brock et al. 2016) and text to image (Reed et al. 2016). Besides, there are other effective methods for image applications. For instance, attention methods are good methods to deal with object recognition (Mnih et al. 2014) and image generation (Mansimov et al. 2015). The ResNet method has been tested to be very effectively for image classification (Xie et al. 2017). Residual Dense Network (RDN) is also effective tool for image super-resolution (Zhang et al. 2018). DiracNets (Zagoruyko and Komodakis 2017), IndRNN (Li et al. 2018) and variational U-Net (Esser et al. 2018) are also very good technologies for image applications.

In recent years, two potent deep-learning methods have been extensively applied to images: Transformers (Vaswani 2017) and Diffusion Models (Nichol and Dhariwal 2021). Transformer models, originally designed for natural language processing tasks, have been adapted for images due to their ability to capture long-range dependencies in data. Unlike traditional CNNs, which are limited by their local receptive fields, transformer models leverage self-attention mechanisms to capture global dependencies within images. This global context awareness makes them particularly effective for image applications that require understanding of the entire image structure. Vision Transformers (ViTs) (Dosovitskiy 2020) and other transformer-based models (i.e., Shifted Window Transformer (Swin Transformer) (Liu et al. 2021) and Convolutional Vision Transformer(CvT) (Wu et al. 2021)) have demonstrated impressive results in various image applications, including image classification, object detection and image inpainting. Beyond image applications, Transformers have also revolutionized natural language processing and cross-modal tasks, while Diffusion Models are increasingly applied to video generation and 3D reconstruction, demonstrating their versatility across the expanded scope of deep learning applications. On the other hand, Diffusion models, such as Denoising Diffusion Probabilistic Models (DDPM) (Ho et al. 2020), were proposed to address the limitations of previous generative models in capturing complex data distributions. These models simulate a gradual denoising process. The key advantage of diffusion models lies in their stepwise denoising approach, which allows for finer recovery of image details, significantly enhancing restoration quality (Croitoru et al. 2023). This iterative process enables the model to learn and reverse the noise addition process, resulting in high-quality image generation and restoration.

Although the above research reveals that deep-learning technologies have achieved remarkable success in image-related applications, there is a lack of a comprehensive survey

specifically focused on deep-learning technologies for handling images. In the context of image-related tasks, i.e., image segmentation, image classification, object detection, image restoration and action recognition have big difference, however, deep learning technologies refer to properties of different tasks to propose wise solution methods, which are embedded in multiple hidden layers with end-end connection to better deal with different applications. Therefore, a survey is important and necessary to review the principles, performance, difference, merits, shortcomings and technical potential for image different tasks (e.g., image segmentation, image classification, object detection, image restoration and action recognition). Additionally, some basic plug-ins, i.e., convolutional layer, pooling layer, active function, fully connected layer and BN layer in deep neural networks are illustrated in this survey, which make readers know the implementations of neural networks. The implementation environment: software and hardware are revealed. The concepts behind typical deep-learning techniques, along with the reasons for their success in image-related applications, are presented. To better show the robustness of deep learning technologies, the performance of deep learning for different image tasks are shown. Finally, we give challenges and trends of deep learning technologies in the future are also offered in this survey. An outline of this survey is shown in Fig. 1.

This overview covers more than 500 papers about deep learning for image application in recent year. Differing from previous research, our research provides a comprehensive report on classical techniques across multiple domains, including low- and high-level vision tasks, video processing, natural language processing, and 3D data processing, rather than focusing

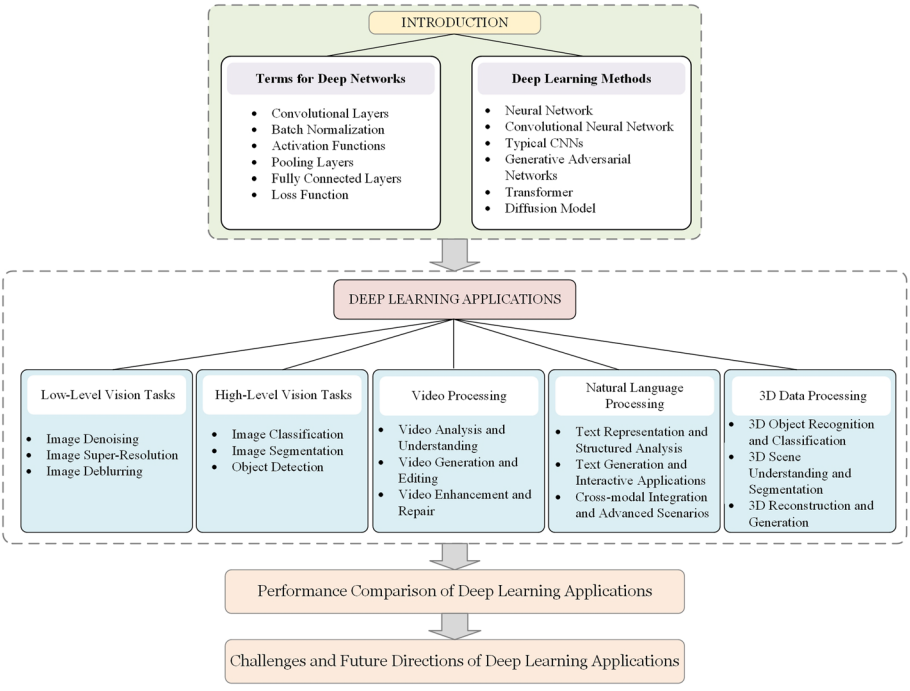


Fig. 1 Outline of the survey. It consists of nine parts, including definitions of terms, overview of deep learning methods, low-level tasks, high-level tasks, video processing, natural language processing, 3D data processing, performance comparison, challenges and potential directions

on a single application area, which is highly beneficial for starters, engineers, and scholars in diverse research fields. Besides, we also analyze relations and differences of these techniques to make scholars convenient do research of interdisciplinary individuals to facilitate development of difference industry, i.e., traditional chemistry, medical, etc.

The main contributions in this paper can be summarized as follows.

- The overview illustrates the effects of deep learning methods for diverse applications, including image processing, video analysis, natural language processing, and 3D data processing.
- The overview offers an extensive review of deep learning applications across both low-level and high-level vision tasks, as well as video, language, and 3D data processing. It systematically examines various deep learning models and their effectiveness in improving image quality for tasks such as denoising, super-resolution, and deblurring, while also addressing video analysis, text processing, and 3D scene understanding. Additionally, it explores the latest techniques in image classification, object detection, and segmentation, offering a comprehensive perspective on how deep learning is advancing diverse fields, from image and video processing to language understanding and 3D data analysis.
- The overview points out some potential challenges and directions for deep learning across diverse applications, including image processing, video analysis, natural language processing, and 3D data processing.

The rest of this overview is organized as follows.

Section 2 explores the fundamental terminology related to deep neural networks, laying a solid conceptual foundation. Section 3 provides an overview of various deep learning methodologies, highlighting their core principles and functionalities. Section 4 then illustrates the broad applications of deep learning in addressing low-level vision tasks, emphasizing its effectiveness in enhancing image quality. In contrast, Sect. 5 focuses on high-level vision tasks, detailing how deep learning facilitates a deeper understanding and interpretation of image content. Section 6 explores deep learning's role in video processing, addressing video analysis and understanding, generation and editing, and enhancement and repair techniques for dynamic visual data. Section 7 investigates deep learning applications in natural language processing, focusing on text representation and structured analysis, text generation and interactive applications, and cross-modal integration for advanced scenarios. Section 8 delves into deep learning for 3D data processing, covering 3D object recognition and classification, scene understanding and segmentation, and reconstruction and generation for spatial data applications. Section 9 presents a rigorous performance analysis, comparing and contrasting the effectiveness of various deep learning approaches. Section 10 explores the untapped potential of deep learning in image applications, outlining promising research directions and offering a visionary perspective on its future trajectory. Finally, Sect. 11 summarizes the key findings and contributions of this paper, while last section presents a comprehensive list of references, underscoring the breadth and depth of the research conducted.

2 Definitions of terms for deep networks

In this section, we show the general technical terms in the deep networks, explain ideas for deep networks, which is significant to better know the implementations of deep networks. Basic components of deep networks are provided as follows.

1. Convolutional layers are core components of CNNs. They apply filters (or kernels) to local input regions (receptive fields) via convolution, producing initial feature maps. This process typically involves convolving the input with filter weights, adding a bias, and applying an activation function (LeCun et al. 1998) for nonlinearity. For example, a convolutional layer with a Tanh activation can be expressed as $z = a(W * i + b)$, where i is the input, W is the filter weights, b is the bias, and $a(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. And detailed information of mentioned activation functions can be shown in later.

2. Batch Normalization (BN) layers, typically applied after convolutional layers, address the covariate shift problem by normalizing layer inputs to a standard distribution (mean 0, variance 1), followed by a learnable scale and shift (Ioffe 2015). Covariate shift occurs when the input distribution to a layer changes during training (e.g., after $i' = W * i + b$, where W is the filter weights and b is the bias), violating the independent and identically distributed (IID) hypothesis question (Clauset 2011) that training and test data share the same distribution. In deep networks, this shift pushes distributions toward the saturation regions of nonlinear activation functions, causing vanishing gradients in lower layers and slowing convergence. BN mitigates this by maintaining larger gradients and accelerating training. However, BN's effectiveness depends on batch size: small batches yield inaccurate mean and variance estimates, while large batches strain GPU memory (Wu and He 2018). Alternatives like Group Normalization (GN) (Wu and He 2018) and Layer Normalization (LN) (Ba et al. 2016) overcome these limitations by normalizing across groups or layers, stabilizing training without batch size dependency.
3. If a neural network includes a linear convolution and fully connection operation, where this neural network only represents linear mapping. However, most data is nonlinear in the real world, which makes neural network be hard to deal with nonlinear data. An activation function is embedded into a neural network to improve nonlinear modeling ability of neural network, which can effectively data of nonlinear distribution in real applications. Ans it is after BN layer. Common activation functions include Sigmoid (DasGupta and Schnitger 1992), Tanh (Jarrett et al. 2009) and ReLU (Nair and Hinton 2010).

   Sigmoid, a nonlinear activation function, maps inputs to (0, 1) with the formula $f(x) = \frac{1}{1+e^{-x}}$, where x is input and $f(x)$ represents an output of an activation function. However, it suffers from vanishing gradients for large or small inputs, slowing convergence and complicating deep network training. Additionally, its non-zero-centered output can hinder gradient descent efficiency.

   Tanh, defined as $a(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, where x and $a(x)$ represent input and output of this activation function, is a nonlinear activation function with zero-centered outputs, improving over Sigmoid. Yet, it still faces vanishing gradient issues for large inputs, limiting its effectiveness in deep networks.

ReLU is actually a piecewise function and its equation is shown in Eq. (1), where output of f with input x stand for a standard model. This accelerates convergence and induces sparsity in neurons, aiding feature extraction and reducing overfitting. More deformation of ReLU is also very popular for neural networks, which can be obtained at He et al. (2015).

$$f(x) = \text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{1}$$

4. The pooling layer, typically applied after the convolutional and activation layers, reduces the spatial dimensions of feature maps from a convolutional layer. This improves computational efficiency and mitigates the risk of overfitting (Yu et al. 2012). It divides the feature map into regions, representing each with a single value (e.g., maximum or average) to decrease dimensionality and computational complexity. Common pooling methods include general pooling (e.g., max and average pooling (Lee et al. 2016)), overlapping pooling (Krizhevsky et al. 2012), and spatial pyramid pooling (SPP) (He et al. 2015), which are detailed below.
   General pooling methods extract features from non-overlapping regions of an image. The window size (*sizeX*) equals the stride, ensuring that adjacent pooling windows do not overlap. Common techniques include average pooling, which computes the mean of a region, and max pooling, which uses the maximum value of a region.
   Overlapping pooling uses a window size larger than the stride ($sizeX > stride$), allowing adjacent windows to overlap. This approach reduces overfitting and enhances performance, lowering top-1 and top-5 error rates by 0.4 and 0.3%, respectively, on ImageNet LSVRC-2012.
   Fully connected layers in CNNs require fixed-size inputs, while convolutional operations can process variable-size images. In practice, input sizes vary, so SPP (He et al. 2015) addresses this by transforming convolutional features of arbitrary sizes into fixed-length vectors for fully connected layers. This preserves important information, avoiding losses from cropping or warping.
   Advanced pooling techniques enhance CNN flexibility and performance. Adaptive feature pooling (Liu et al. 2018) dynamically adjusts output sizes, while global pooling (e.g., global average pooling (Lin 2013)) summarizes feature maps for classification. Methods like fractional (Graham 2014), mixed (Yu et al. 2014), and attention-based pooling (Lee et al. 2019) improve detail retention and adaptability. For specialized tasks, graph pooling (e.g., graph coarsening (Cai et al. 2021) and hierarchical pooling (Zhang et al. 2019)) handles non-Euclidean data, wavelet pooling (Williams and Li 2018) preserves structure, and anti-aliasing pooling (Zhang 2019) reduces artifacts. These innovations boost generalization, efficiency, and adaptability across applications.
5. Fully connection(FC)layer is used after a pooling layer and it maps learned distributed feature representation sample label space (Lin et al. 2013). However, parameters of fully connected layer are redundant. Thus, popular ResNet and GoogLeNet use global average pooling (GAP) (Szegedy et al. 2015) instead of FC to fuse learned features and apply softmax loss function (Liu et al. 2016) as an objective function to guide the learning processing. GAP can be also utilized to reduce the model size and address overfitting.

6. Loss Function is a function to compute difference of the predict value and obtained real value to update parameters (Janocha and Czarnecki 2017). Loss function can be adjusted by gradient descent method. That is, when difference above is less than specified value, parameters of model are considered to confirm. Common loss functions include Euclidean (Wan et al. 2014), Cross Entropy (De Boer et al. 2005) and Mean Square loss (Stewart et al. 2001) and Softmax (Liu et al. 2016). Different loss functions have different applications. For instance, Euclidean is used for regression problem. Softmax is exploited for only one label in class labels, i.e., face recognition and object recognition. When the number of class labels is more than 1, Cross Entropy method is effective for classification task. Mean Square is a good tool for image denoising and resolution.

To provide a concise overview of the key components discussed, Table 1 summarizes the definitions and roles of essential terms in deep networks, serving as a quick reference for understanding their implementations.

3 Overview of deep learning methods

3.1 Neural network

Traditional neural networks form the foundation of most deep learning technologies (Schalkoff 1997). They consist of neurons, weights W, biases b, and an activation function $f(x)$, which introduces non-linearity to the linear combination of inputs X, as shown in Eq. (2):

$$f(X;W;b) = f(W^T X + b) \tag{2}$$

Sigmoid, Tanh and ReLU are typical activations functions for neural networks and they have been shown in Sect. 2.

A multilayer perceptron (MLP) extends this to multiple layers (Gardner and Dorling 1998), as in Eq. (3):

$$f(X;W;b) = f(W^n f(W^{n-1} \dots f(W^0 X + b^0) \dots b^{n-1}) + b^n) \tag{3}$$

Table 1 Definitions of terms for deep networks

Component	Description
Convolutional layers	Extracts high-level features, generates feature maps
Batch normalization	Normalizes inputs, speeds up and stabilizes training
Activation functions	Adds nonlinearity (e.g., Sigmoid, Tanh, ReLU)
Pooling layers	Reduces feature map size, lowers computation (e.g., max, average)
Fully connected layers	Maps features to outputs, used in small or large models
Loss function	Measures prediction error (e.g., cross-entropy, MSE)

Here, W^l denotes weights at layer l, and n is the final layer. Layers between input and output are hidden layers. Networks with over three layers are deep neural networks, with the final layer's activation predicting object classes.

A fully connected network links all neurons between consecutive layers, while partially connected networks link only some. For a two-layer network (hidden and output), with inputs x_1, x_2, x_3 and output o_1, parameters include weights $w_1, \ldots, w_{12}$ and biases $b_1, \ldots, b_4$. Using a Sigmoid activation, the hidden neuron h_1 computes $o(h_1) = \frac{1}{1+e^{-(w_1 x_1 + w_4 x_2 + w_7 x_3 + b_1)}}$, and the output $o(o_1)$ follows similarly. Backpropagation (BP) (Hirose et al. 1991) adjusts parameters by comparing $o(o_1)$ to a target, but Sigmoid's saturation can cause gradient vanishing, complicating training.

Hinton et al. (2006) proposed pre-training with unsupervised methods and fine-tuning with supervision to mitigate this, using models like stacked autoencoders (SAE) and deep belief networks (DBN). However, these are complex and time-intensive (Litjens et al. 2017). Modern approaches favor end-to-end supervised training, with architectures like CNN variants (e.g., AlexNet (Krizhevsky et al. 2012), ResNet (He et al. 2016)) and RNNs (Graves et al. 2013), widely applied to image tasks. These are detailed in Sect. 3.7.

3.2 Convolution neural networks (CNNs)

CNNs are feedforward neural networks trained using backpropagation (BP), achieving significant success in image applications and speech recognition (Ciresan et al. 2011). They extract features hierarchically: low layers detect basic edges, while deeper layers identify distinct objects. Unlike MLPs, CNNs differ in two key ways (Patra and Kot 2002). First, CNNs use weight sharing in convolutional operations, where weights learned from a local image region are applied across the entire image, reducing the parameter count. These convolutional operations resemble those in traditional neural networks. Second, CNNs incorporate pooling layers after each convolutional layer (see Sect. 2), which downsample and refine features. These features are then fed into fully connected layers, mapping them to labels (or classes). Finally, an activation function (e.g., Softmax) in the last layer produces a class distribution, and the model is trained using maximum likelihood (Li et al. 2021).

3.3 Typical convolution neural networks

Due to the popular of CNNs in image applications, we provide typical architectures and show differences of these architectures.

3.3.1 AlexNet

AlexNet, proposed by Hinton et al., won the ILSVRC-2012 competition (Krizhevsky et al. 2012) and is a classic CNN. It employs 5 convolutional layers and 3 fully connected layers to extract image features.

The success of AlexNet in image classification stems from six key factors:

- It uses data augmentation (e.g., rotations at $0°, 90°, 180°, 270°, 360°$, flips, clipping, and color/light adjustments) to enhance learning and prevent overfitting (Krizhevsky et al. 2012).

- Traditional CNNs fused model predictions to reduce test errors (Bell and Koren 2007), but their high computational cost and overfitting risk with limited data were drawbacks. AlexNet's dropout method (Hinton et al. 2012) randomly sets hidden neuron outputs to 0 (probability 0.5), excluding them from forward and backward passes. This creates varied network structures with shared weights per input, breaking neuron dependencies and forcing robust feature learning. During testing, outputs are scaled by 0.5 to compute a geometric mean of predictions (Krizhevsky et al. 2012), also mitigating overfitting (Nielsen 2015).
- Purely linear operations (convolution and full connection) limit a network to linear mapping, inadequate for nonlinear real-world data. Activation functions enhance nonlinearity; while (Jarrett et al. 2009) and Logistic (Zhang et al. 1998) improve expressiveness, they risk gradient divergence. AlexNet's Rectified Linear Units (ReLUs) (Nair and Hinton 2010), detailed in Sect. 2, overcome this issue.
- Local Response Normalization (LRN), akin to Jarrett et al.'s method (Jarrett et al. 2009), normalizes using neighboring data without mean subtraction (unlike "brightness normalization"). It boosts generalization, reducing top-1 and top-5 error rates by 1.4 and 1.2%, respectively, in image classification (Krizhevsky et al. 2012), proving its value for deep learning.
- Pooling reduces data dimensions post-activation using kernel-based neighbor information. Traditional non-overlapping pooling (Hinton et al. 2012) is replaced in AlexNet with overlapping pooling, improving feature precision and reducing overfitting. This cuts top-1 and top-5 error rates by 0.4 and 0.3%, respectively, in ImageNet ILSVRC-2012.
- GPUs are vital for deep learning in image, video, and speech tasks (Chetlur et al. 2014). As data scales, a single GPU struggles with large CNNs. AlexNet uses two parallel GPUs, each handling half the neurons and interacting only at specific layers. Experiments show dual GPUs slightly outperform a single GPU in training time (Krizhevsky et al. 2012), making multi-GPU setups effective for big data.

Despite its success in image recognition, AlexNet lacks flexibility. Removing a convolutional layer can degrade performance, a limitation deeper networks like VGG (Simonyan and Zisserman 2014) address effectively.

3.3.2 Vgg

Vgg respectively obtained the first and second places in image classification and tracks in ImageNet Challenge 2014 (Simonyan and Zisserman 2014). However, it is different from GoogLeNet (Szegedy et al. 2015), which only improves the performance by increasing depth with filter. Vgg-19 denotes 19 layers from Vgg. The reasons of performance of Vgg-19 have three-fold. First, each layer of Vgg-19 uses filter, which effectives the number of parameters and prevents overfitting problem. Second, it uses 3 fully connected layers instead of 3 convolutional layers, which the fully connected network can allow any size of input. Finally, architecture of the Vgg-19 is deeper, can better learn features. Despite, Vgg method has improved the accuracy as 84.0% for single-image action classification in VOC-2012, Vgg of increasing depth may make accuracy saturated then degrades rapidly. He

et al. (2016) proposed ResNet method to address problem above, which obtained excellent performance for image recognition.

3.3.3 GoogLeNet

As deep learning advances, scholars have shifted focus from powerful hardware and large datasets to novel ideas, algorithms, and network architecture improvements. Widening CNNs is a straightforward way to boost performance. GoogLeNet (Szegedy et al. 2015) employs multiple kernels and max pooling, placing a 1×1 convolutional kernel before larger kernels to reduce dimensionality and parameters. With 22 layers, it mitigates the vanishing gradient problem using two auxiliary losses. Its architecture stacks 1×1, 3×3, and 5×5 convolutions alongside a 3×3 max pooling layer, increasing depth and scale adaptability.

Initially termed the Inception Module Native Version (IMNV), this design directly stacked convolutions, resulting in excessive filters at the top layer and high computational costs. Adding pooling units fused their outputs with convolutional layers, inflating the output size and hindering training efficiency. To address this, the refined "Inception module" (also known as GoogLeNet) uses 1×1 convolutions to reduce dimensions before 3×3 and 5×5 convolutions, doubling as rectified linear activations. GoogLeNet excels in image classification and detection. Similarly, VGG, another prominent CNN, enhances image recognition in ILSVRC 2014 (Simonyan and Zisserman 2014) by deepening the network.

3.3.4 ResNet

Deep CNNs have driven significant breakthroughs in image recognition (Zeiler 2014), particularly in image classification (Simonyan and Zisserman 2014), and benefit various visual recognition tasks. However, deeper networks risk vanishing or exploding gradients (Bengio et al. 1994), where increased depth amplifies training errors. ResNet (Wu and He 2018) addresses this by adding the input of every two layers to their outputs, forming residual blocks defined as $f(x) + x$, where x is the input and f is the activation function. ResNet's popularity stems from several strengths: First, it prioritizes depth over width, controlling parameter count and reducing overfitting. Second, it relies less on pooling layers, using more downsampling to enhance transmission efficiency. Third, it employs batch normalization (BN) and average pooling for regularization, speeding up training. Finally, it uses 3×3 filters in convolutional layers, training faster than mixed 3×3 and 1×1 filters. These advantages led ResNet to win ILSVRC 2015, achieving a 3.57% error rate on the ImageNet test set.

Variants of residual networks (He et al. 2016) are widely applied in image classification, denoising, and resolution enhancement. GANs and attention mechanisms have also achieved notable success in image applications (Denton et al. 2015).

3.4 Generative adversarial networks

While CNNs excel at extracting local features from structured data and building hierarchical representations, GANs advance generative modeling by creating samples that mimic a target data distribution. GANs comprise two networks: a generator, which produces syn-

thetic data from random noise (Random Z), and a discriminator, which assesses whether a sample—real (Real Sample) or fake (Fake Sample)—is genuine, outputting a probability. Through adversarial training, the generator refines its outputs to be more realistic, while the discriminator improves at detecting fakes, driving mutual enhancement.

This adversarial paradigm enables GANs to excel in generating high-dimensional data, particularly in image synthesis, producing photorealistic images from noise or partial inputs. They are widely applied in tasks like image inpainting (Liu et al. 2021) and style transfer (Lin et al. 2020), yielding outputs with impressive visual fidelity.

However, GANs face challenges. Training instability arises from the need to balance generator and discriminator performance. Mode collapse, where the generator outputs limited variety, fails to reflect the target distribution's diversity. Convergence issues can also occur, with the generator producing quality data that the discriminator struggles to critique effectively. These problems often necessitate advanced methods like Wasserstein GANs (WGANs) (Gulrajani et al. 2017) or Progressive GANs (Karras 2017) to stabilize training and boost performance.

GANs offer a potent generative modeling framework via adversarial training, excelling in diverse applications. Yet, their training instabilities remain a key research focus, with efforts ongoing to enhance robustness and generalization.

3.5 Transformer

Transformers, initially developed for sequence modeling in NLP, have transformed tasks like machine translation, text generation, and summarization (Lin et al. 2022). Unlike CNNs or RNNs, they use self-attention to capture global dependencies in input data, efficiently modeling both local and long-range interactions. This innovation extends beyond NLP, impacting computer vision via models like the Vision Transformer (ViT) (Yuan et al. 2021).

The Transformer's core is the scaled dot-product attention mechanism, computing relationships across all input elements simultaneously. For an input sequence of length n, it is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$

- Q, K, V: The query, key, and value matrices, derived from the input embeddings through learned linear projections.
- $QK^\top$: The dot product between queries and keys, which quantifies the similarity between input elements.
- $\sqrt{d_k}$: A scaling factor that stabilizes gradients during training.
- V: The value matrix, which represents the contextualized information passed through the attention mechanism.

The output is a weighted sum of values, with weights reflecting each element's relevance.

A typical Transformer features an encoder-decoder structure. The encoder stacks self-attention and feedforward layers to process inputs, while the decoder generates outputs by attending to encoder outputs and its own sequence (Vaswani 2017).

In computer vision, ViT adapts self-attention for images by splitting an image I $H \times W \times C$ (height, width, and channels) into$N$ fixed-size patches (e.g., 16×16), flattening them into vectors, and embedding them as tokens:

$$\mathrm{z}_0 = [\mathrm{p}_1\mathrm{E}; \mathrm{p}_2\mathrm{E}; \ldots; \mathrm{p}_N\mathrm{E}] + \mathrm{E}_{\mathrm{pos}} \tag{5}$$

- E: The learnable linear projection matrix.
- $\mathrm{E}_{\mathrm{pos}}$: Positional embeddings to encode spatial information.

The sequence z_0 is fed into the Transformer encoder, enabling global patch interactions. This allows ViT to outperform CNNs on large-scale datasets by capturing spatial and contextual relationships.

Unlike CNNs, which use localized receptive fields, Transformers' self-attention enables every token to interact globally in one step, excelling in tasks requiring holistic input understanding, such as image classification, object detection, and video analysis.

3.6 Diffusion model

Diffusion models have emerged as a significant advancement in generative modeling, distinguished by their stability and capacity to produce high-quality outputs (Croitoru et al. 2023). Diffusion models are built on the principle of gradually perturbing data through a stochastic noise process and then learning to reverse this process to reconstruct the original data. This approach ensures robust generative capabilities while addressing many of the instability issues associated with adversarial training in GANs.

The core of diffusion models lies in two processes:

- Forward Process (Diffusion): Starting from a clean data sample x_0, Gaussian noise is incrementally added over T time steps to produce a series of noisy samples $\{x_t\}_{t=1}^{T}$. This process can be formalized as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathrm{I}) \tag{6}$$

where $q(x_t \mid x_{t-1})$ represents the probability distribution for transitioning from the noisy sample x_{t-1} at step $t-1$ to the sample x_t at step t. The distribution is Gaussian ($\mathcal{N}$), characterized by a mean of $\sqrt{\alpha_t}x_{t-1}$, where α_t is a time-dependent parameter that scales the original data to control the amount of retained information. The variance term $(1-\alpha_t)\mathrm{I}$ adds isotropic Gaussian noise, with I being the identity matrix, ensuring uniform noise across all dimensions. This process describes how clean data is progressively corrupted with noise as time t increases.

- Reverse Process (Denoising): The reverse process aims to remove noise step by step,

reconstructing the data distribution. This is parameterized as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{7}$$

where $p_\theta(x_{t-1} \mid x_t)$ models the reverse transition, predicting the previous sample x_{t-1} from the current noisy sample x_t. The distribution is Gaussian ($\mathcal{N}$) with a learned mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$, both parameterized by a neural network with parameters θ. The mean $\mu_\theta(x_t, t)$ predicts the denoised data direction, while the variance $\Sigma_\theta(x_t, t)$ accounts for uncertainty in this prediction. Together, these terms guide the step-by-step removal of noise, enabling the reconstruction of the original data from pure noise.
The model is trained to minimize the discrepancy between the true forward process $q(x_{t-1}|x_t)$ and the predicted reverse process $p_\theta(x_{t-1}|x_t)$, often using a variational lower bound (VLB) (Ho et al. 2020) as the optimization objective. Figure 2 is a schematic illustration of the diffusion model's structure, showcasing the forward diffusion process (adding noise) and the reverse denoising process (removing noise).

Diffusion models, once trained, generate data by starting with a Gaussian noise sample x_T and iteratively denoising it using learned reverse transitions. This step-by-step refinement gradually reconstructs the target data distribution, producing high-fidelity outputs with fine detail. The iterative nature of diffusion models ensures precise and coherent data synthesis, making them highly effective across diverse applications.

In high-resolution image generation, diffusion models excel at creating photorealistic images with rich textures and intricate details. Models like Denoising Diffusion Probabilistic Models (DDPM) (Croitoru et al. 2023) and Imagen (Saharia et al. 2022a) achieve top-tier performance, leveraging tailored noise schedules and iterative refinement to preserve key features of the target distribution, ideal for realistic, high-quality image synthesis.

For image super-resolution, diffusion models enhance low-quality images while preserving structure and detail. Models such as Super-Resolution via Repeated Refinements (SR3) (Saharia et al. 2022b) and Latent Diffusion Models (LDMs) (Rombach et al. 2022) showcase this strength. SR3 progressively refines blurry images to recover high-frequency details, while LDMs use latent space representations for efficient, scalable high-resolution synthesis.

The versatility of diffusion models highlights their power in generative tasks. Their iterative refinement and probabilistic framework offer a stable, scalable alternative to traditional methods like GANs, often surpassing them in output quality and training reliability.

3.7 Tools for deep learning methods

GPU is a critical driver of deep learning success in image applications. Unlike CPUs, GPUs leverage numerous computing units, longer pipelines, simpler control logic, and reduced

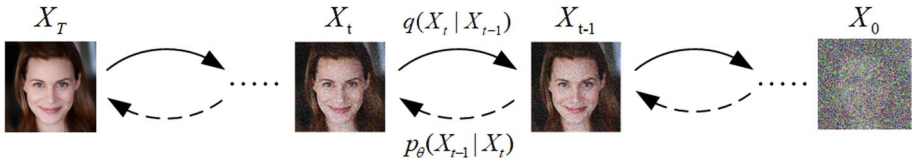


Fig. 2 Forward and reverse processes in a diffusion model

cache, delivering superior computational power. Additionally, GPUs support more threads than CPUs, accelerating processing speeds. Common GPU hardware libraries for deep learning include CUDA (Sanders and Kandrot 2010) and OpenCL (Stone et al. 2010). Modern GPU options have evolved beyond older models, with current standards including NVIDIA's RTX 4090, A100, and H100, optimized for AI workloads. Table 2 summarizing the content about GPU software libraries for deep learning.

GPU software libraries (also called tools) are as vital as hardware libraries for deep learning. The following are popular tools updated to reflect current usage:

- Convolutional Architecture for Fast Feature Embedding (Caffe) (Jia et al. 2014) is an efficient, open-source deep learning framework written in C++, with Python and MATLAB interfaces. It runs on both CPUs and GPUs and excels in object detection. However, it demands strong C++ skills and has seen reduced adoption compared to newer

Table 2 Popular GPU software libraries for deep learning

Tool	Languages	Key features	Applications	Notes
Caffe	C++, Python, MATLAB	Efficient, Open-Source, CPU/GPU	Object Detection	Needs C++ Skills, Less Popular Now
TensorFlow	C++, Python	High-Level, Auto-Differentiation, Portable	Detection, Classification, Denoising, Super-Resolution	Widely Used, Outperforms Theano
Keras	Python (via TensorFlow)	User-Friendly, Multi-GPU	Classification, Detection, Super-Resolution, Denoising, Action Recognition	Now Part of TensorFlow
PyTorch	Python	Intuitive, Research-Oriented	Classification, Detection, Segmentation, Action Recognition, Super-Resolution, Tracking	Top Choice for Research
Theano	Python	NumPy Integration, Efficient	Super-Resolution, Denoising, Classification	Development Ended (2017), Outdated
MatConvNet	MATLAB	MATLAB-Based, Specialized	Classification, Denoising, Super-Resolution, Tracking	Requires MATLAB, Limited Use

tools.

- TensorFlow (Abadi et al. 2016) is a versatile, high-level machine learning library supporting neural network design with automatic differentiation (not just backpropagation). It offers C++ and Python interfaces, flexible portability, and faster compilation than Theano. TensorFlow is widely used for object detection, image classification, denoising, and super-resolution.
- Keras (Ketkar and Ketkar 2017), now integrated into TensorFlow, is a Python-based API for easy neural network implementation. It supports multiple GPUs and is user-friendly, making it ideal for image classification, object detection, super-resolution, denoising, and action recognition.
- PyTorch (Paszke et al. 2017) is a leading neural network library, porting Torch to Python. With its intuitive Python interface, PyTorch dominates in research and applications like image classification, object detection, segmentation, action recognition, super-resolution, and visual tracking.
- Theano (Bastien et al. 2012), a mathematical expression compiler for large-scale neural networks, integrates with NumPy and generates efficient C code dynamically. While historically popular for its low learning curve and stability, its development ceased in 2017, limiting its use today to legacy projects like image resolution, denoising, and classification.
- MatConvNet (Vedaldi and Lenc 2015) provides a MATLAB interface for image classification, denoising, super-resolution, and visual tracking. It requires MATLAB expertise and has niche use due to its complexity and limited community support.

4 Deep learning for low-level vision tasks

Low-Level tasks in computer vision aim to restore high-quality images. Common low-level tasks include image denoising, image super-resolution, and image deblurring. This section reviews the currently proposed deep learning methods in image applications from these three aspects Fig. 3 provides an overview of Sect. 4.

4.1 Deep learning for image denoising

4.1.1 Deep learning for additive white noisy image denoising

Due to the shortage of real noisy images, additive white noisy images (AWNIs) are widely used in training and testing of image denoising models. AWNIs include Gaussian, Poisson, Salt, Pepper and multiplicative noisy images Additive white noise simulates noise. There are deep learning techniques for AWNIs denoising, including methods based on ResNet, GAN and GNN.

A CNN-based image denoising algorithm utilizes neural networks to learn the mapping relationship from noisy images to clear images (Tian et al. 2020). By training on a large dataset containing pairs of noisy and corresponding clear images, a CNN can learn a complex nonlinear mapping from noisy images to clean images (Xie et al. 2012). During the training process, the CNN continually adjusts its weights by using the backpropagation algorithm to minimize the difference between the output image and given clean

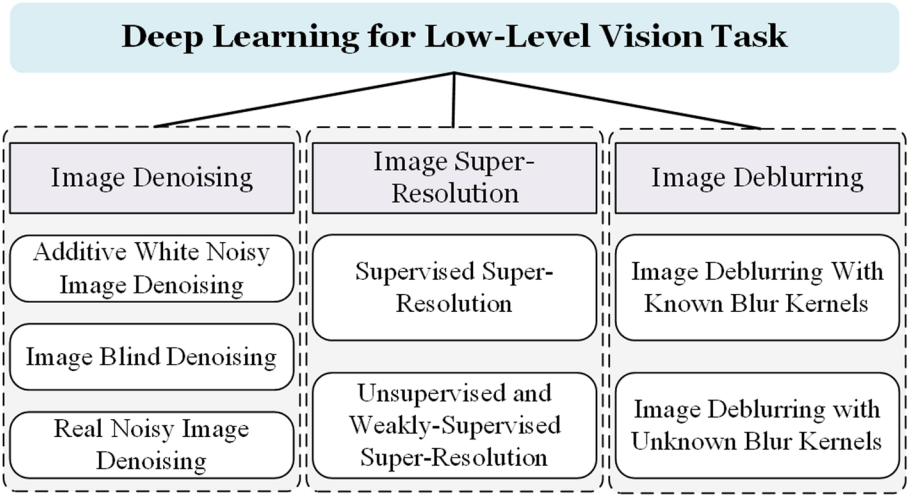


Fig. 3 Overview of low-level vision tasks

image (Singh et al. 2022). Common loss functions include Mean Squared Error (MSE) and Structural Similarity Index (SSIM) (Sara et al. 2019). Zhang et al. (2017) proposed a feed-forward denoising convolutional neural network (DnCNN). This method employs a deeper architecture , residual learning techniques, regularization, and batch normalization to enhance denoising performance. Compared to traditional CNN denoising methods, the advantage of DnCNN lies in its ability to handle Gaussian denoising with unknown noise levels. FFDNet (Zhang et al. 2018) is an upgrade of DnCNN. While maintaining a similar architecture to DnCNN, FFDNet includes a user-controlled parameter in a network input to improve algorithm's adaptability in image denoising. Guo et al. added a user-input noise level parameter σ and incorporated a fully convolutional network to learn this parameter to achieve adaptive denoising model (Guo et al. 2019). Tian et al. (2020) introduced a BRDNet model, which uses a batch renormalization to address the performance degradation of small batch data in training process. BRDNet employs residual learning and dilated convolutions to extract more contextual information and prevent gradient vanishing or explosion. BRDNet demonstrates excellent denoising performance on both synthetic and real noisy images with relatively low computational cost, making it suitable for smartphones and cameras. Zhang et al. (2021) proposed DRNet composed of convolutions, batch normalization, and ReLU activation functions extracted richer features and addressed the vanishing gradient problem. Adding DC-ResBlock into a DRNet to enhance expressive ability of a deep network to improve denoising performance. Gurrola-Ramos et al. (2021) proposed a neural network model based on a Residual Dense Network (RDUNet) for image denoising. This model combines the characteristics of residual learning and dense connection networks to improve denoising performance and reduce computational complexity. RDUNet combines the residual learning and dense connections to enhance network performance. When using deep learning methods for image denoising, a large number of training image sample pairs are usually required, that is, images with noisy and denoised images. However, denoised images are often difficult to obtain in real world, i.e., to overcome this question, Lehtinen et al. (2018) presents a denoising method (Noise2Noise) that does not require a noiseless

image as a label. Mansour and Heckel (2023) improved the N2N method via using a single noisy image to generate a pair of noisy pictures and using pairs to generate a simple two-layer neural network.

GANs are a deep learning framework consisting of two mutually adversarial neural networks: a generator and a discriminator. The generator is responsible for creating artificial images that resemble real samples, while the discriminator's task is to distinguish between real samples and generated images. Through continuous adversarial training, the generator eventually produces high-quality artificial samples that are difficult to distinguish from real ones. Leveraging GAN's powerful generative capabilities, the generator is tasked with producing clean images from noisy ones, while the discriminator assesses whether the generated images are realistically noise-free. By adversarial training, the generator learns the mapping from noisy images to denoised images to obtain high-quality images. Khmag (2023) proposed used a GAN and a semi-soft thresholding idea via two phases to address additive Gaussian noisy image denoising. The first phase uses wavelet transform in a semi-soft thresholding way to remove noise in the high-frequency sub-bands. The second phase utilizes a GAN to further eliminate noise to enhance image quality. Lyu et al. (2020) introduced a GAN-based hybrid noise removal model (DeGAN) to address mixed noise, i.e., additive white Gaussian noise (AWGN) and impulse noise (IN) in noisy images They designed a new joint loss function that incorporated image feature information and human visual perception information into a CNN to simultaneously deal with mixed noise to improve visual denoising effects. Yi and Babyn (2018) used a Conditional Generative Adversarial Network (cGAN) as well as SAGAN via adversarial and sharpness loss to reduce blurring and preserve details in image denoising.

However, CNNs primarily process Euclidean data, such as 2D images and 1D text, which are specific cases of graph data. Consequently, CNNs may not be as well-suited for handling more general graph-structured data (Wu et al. 2020). Graph Neural Networks (GNNs) (Scarselli et al. 2008) can capture dependency in the graph via propagation of information between nodes to better deal with unstructured or complex data. Also, applying GNNs to image denoising can improve the model's processing efficiency and generalization ability. Su et al. (2020) proposed a new image denoising method based on GNNs by using analytical graph filters (GraphBio) as convolutional filters and optimizing graph topology to improve denoising performance. The proposed GNN used predefined filters that don't require training and enhances performance by optimizing graph topology. These filters act like low-pass filters with biorthogonal conditions in signal processing, with the graph spectrum optimized through data training (Sanchez-Lengeling et al. 2021). Graph Convolutional Networks (GCNs) (Kipf and Welling 2016) share the same processing flow as GNNs including aggregation, update, and recurrence. GCNs mathematically constrain the weight parameters in the aggregation step by adding degree normalization to the averaging method. GCNs are ingeniously designed to extract features from graph data, allowing us to use these features for tasks such as node classification, graph classification, link prediction, and graph embedding, making them highly versatile. Valsesia et al. (2020) constructed graph convolutional layers to utilize non-local similarity information to achieve pixel-level adaptive receptive fields and enhanced performance of image denoising . It also introduced a lightweight Edge-Conditioned Convolution (ECC) Simonovsky and Komodakis (2017) to address gradient vanishing and over-parameterization issues. GCNNs have shown excellent performance on both synthetic Gaussian noisy and real noisy image denoising. Chen et al.

(2021) proposed an encoder-decoder structured graph convolutional network (ED-GCN) for CT image denoising. ED-GCN can handle both Gaussian noise and Poisson noise by combining local convolution and graph convolution to manage local and non-local features for enhancing denoising performance. Since GCN treats all neighboring nodes equally during convolutional computing and cannot assign different weights, Graph Attention Networks (GATs) (Veličković et al. 2017) were introduced to address this issue. GATs utilized an attention mechanism to allocate varying weights to distinct nodes, rendering it apt for inductive tasks. Jiang et al. (2023) integrated the graph attention mechanism with an attention network to devise an efficient image denoising framework as well as GAiA-Net. By leveraging the graph attention mechanism, GAiA-Net had a capability to dynamically modify the interconnection weights among different nodes throughout the training, enabling the network to autonomously concentrate on crucial image areas for denoising including edges, textures, and other intricate details, ultimately enhancing the denoising outcome.

Transformer-based models have emerged as powerful tools in image denoising due to their ability to capture long-range dependencies and contextual information. Chen et al. (2021) proposed the Image Transformer, which applies transformer layers to image patches, allowing the model to learn global relationships across the entire image. By fine-tuning on noisy datasets, the Image Transformer effectively suppresses additive white Gaussian noise (AWGN) while preserving intricate details. Similarly, Zamir et al. (2022) introduced the Restormer, a transformer-based model specifically designed for image restoration tasks, including denoising. Restormer leverages self-attention in the frequency domain and incorporates channel-wise processing to efficiently address noise while maintaining computational efficiency. In the context of denoising additive Gaussian noise, Jian et al. (2024) proposed SwinCT, a Swin Transformer-based model for low-dose CT image denoising. By integrating a feature enhancement module into an encoder-decoder framework, SwinCT effectively extracts and enhances high-level features, preserving fine tissue and lesion details while producing high-quality denoised images. Moreover, Wang et al. (2022) developed the Uformer, a U-shaped transformer network, which integrates multi-scale feature extraction with global self-attention to achieve state-of-the-art performance in image denoising.

Recent advancements in denoising have also explored the potential of diffusion models, which offer a probabilistic framework for image generation and restoration. Diffusion models operate by iteratively refining noisy images through a sequence of steps, gradually reducing noise while preserving structural details. For AWGN denoising, these models leverage a reverse diffusion process, starting from a highly noisy image and systematically reconstructing the original clean image. Ho et al. (2020) introduced the Denoising Diffusion Probabilistic Model (DDPM), a generative model that learns the data distribution by modeling a diffusion process. By training on noisy data, DDPM can effectively reverse the diffusion process to denoise images. Similarly, Song et al. (2021) proposed a score-based generative model that uses stochastic differential equations (SDEs) to handle various noise levels, including AWGN, and demonstrated its capacity to reconstruct high-quality images with fine details. These diffusion-based methods have shown promise in achieving superior denoising performance, especially in handling complex noise patterns, by leveraging their strong theoretical foundation and flexibility in modeling diverse noise distributions.

More detail information of deep learning methods for AWNIs denoising can be founded in Table 3.

4.1.2 Deep learning for image blind denoising

Blind image denoising is used to address image denoising with unknown noise. It mainly depends on given noisy images to estimate noise to obtain a denoiser (Pan et al. 2017). Due to the loss of prior knowledge, blind image denoising involves bigger uncertainty and difficulty. That causes that traditional discriminative learning methods, i.e., DnCNN in supervised ways are not suitable for blind denoising problems. To address this problem, blind image denoising based CNNs are developed. For instance, Chen et al. (2018) proposed a two-phase denoising method based a GAN and CNN as well as GCBD. The first phase employs a GAN to estimate noise distribution and generate noise samples to create a paired training dataset. The second phase utilizes a CNN to remove noise. Wu et al. (2020) combined a self-supervised learning and knowledge distillation to overcome unpair images to achieve a blind denoising model. Byun et al. (2021) simplified FBI-Net blind-spot network to quickly estimate noise to achieve a blind denoising model. Table 4 summarizes classical deep learning for blind image denoising.

4.1.3 Deep learning for real noisy image denoising

Deep learning for real noisy image denoising is a challenging and important research area focused on developing neural network-based techniques to remove noise from images affected by real-world noise sources (Zhong et al. 2022). Differing from synthetic noisy image with Gaussian noise, real noisy images can be corrupted by different factors, i.e., sensor noise (Gow et al. 2007), compression artifacts (Dong et al. 2015), and environmental factors (Tian et al. 2020). Due to complex and unknown noise, real noisy image denoising is very difficult to remove noise. Prior knowledge in traditional machine learning is very effective for real noisy image denoising (Xu et al. 2018). Inspired by that, Zhang et al. (2017) embedded iterative learning into a CNN to learn noise mapping with wide range noise levels for address real noisy image denoising. To robustness of an obtained denoiser in the real world, Garibi et al. (2024) utilized a diffusion model in interactive steps to learn respective noisy information to enhance image inversion for real image denoising. Alternatively, Zhang et al. (2022) proposed a self-supervised method via using iterative data refinement to estimate noise for real noisy image denoising. Zou et al. (2023) employed an iterative optimization of to achieve a self-supervised denoiser to remove noise from real noisy image denoising. Besides, to improve relations of key pixels, A dual-branch Transformer network architecture with iterative learning is used to extract salient noise information in real noisy image denoising (Zhang and Zhou 2023). Also, Anwar and Barnes (2019) explored feature attention mechanisms to enhance relations of hierarchical structural information to separate salient noise information from real noisy images. Wang et al. (2023) combined low noise correlation and Transformer mechanism to capture context information to finish real noisy image denoising. Besides, other attention methods, i.e., pyramid mechanisms Mei et al. (2023) also effective for real noisy image denoising. More detail information of deep learning methods for real image denoising can be founded in Table 5.

Table 3 Deep learning for AWNIs denoising

References	Methods	Applications	Key words
Zhang et al. (2017)	CNN	AWNIs denoising	CNN with Residual learning and batch normalization for AWNIs denoising
Zhang et al. (2018)	CNN	AWNIs denoising, Spatially varying nosing	CNN with varying noise level for AWNIs denoising
Guo et al. (2019)	CNN	AWNIs denoising, Real image denoisng	CNN and cameral processing pipeline for AWNIs denoising
Tian et al. (2020)	CNN	AWNIs denoising, Real image denoising	CNN with BRN, RL and dilated convolutions for image denosing
Lehtinen et al. (2018)	CNN	AWNIs denoising	Self-supervised CNN do not require paired clean data for denoising
Mansour and Heckel (2023)	CNN	AWNIs denoising	CNN denoising without clean images as training data
Ma et al. (2021)	CNN	AWNIs denoising, Blind denoising, deblurring	CNN with Residual learning for AWNIs denoising
Zhang et al. (2021)	CNN	AWNIs denoising	CNN with Multi-layer Residual Blocks and Feature Extraction for image denoising
Gurrola-Ramos et al. (2021)	CNN	AWNIs denoising, CT/MRI image denoising	CNN with Residual Dense Network and U-Net Architecture for image denoisng
Zhang et al. (2023)	CNN	AWNIs denoising, Real image denoising	Swin-Conv-UNet architecture for image blind denoising
Khmag 2023	GAN	AWNIs denoising	GAN with Wavelet transform and Semi-soft thresholding for image denoisng
Lyu et al. (2020)	GAN	AWNIs denoising	GAN architecture and joint loss function for hybrid noise removal
Yi and Babyn (2018)	GAN	Low-dose CT denoising, AWNIs denoising	Conditional GAN for LDCT image denoising
Wu et al. (2020)	GNN	AWNIs denoising	GNN with graph filters and topology optimization for image denoising
Valsesia et al. (2020)	GCN	AWNIs denoising	GCN with Edge-Conditioned Convolution(ECC) for image denoising
Chen et al. (2021)	GCN	CT image denoising, AWNIs denoising	GCN with Encoder-decoder architecture for image denoisng
Jiang et al. (2023)	GAT	AWNIs denoising	Graph attention for complex noise reduction
Chen et al. (2021)	Transformer	AWGN denoising	Image Transformer learning global relationships for denoising
Zamir et al. (2022)	Transformer	Image restoration, AWGN denoising	Restormer leveraging frequency domain self-attention
Jian et al. (2024)	Transformer	Low-dose CT denoising, AWGN denoising	SwinCT with feature enhancement module for high-quality denoising
Wang et al. (2022)	Transformer	Image denoising	Uformer with multi-scale feature extraction and global self-attention
Ho et al. (2020)	Diffusion	AWGN denoising	DDPM leveraging reverse diffusion for noise reduction

Table 3 (continued)

References	Methods	Applications	Key words
Song et al. (2021)	Diffusion	AWGN denoising	Score-based generative model using SDEs for high-quality reconstruction

Table 4 Deep learning for blind image denoising

References	Methods	Applications	Key words
Chen et al. (2018)	GAN.CNN	Blind image denoising, real image denoising	GAN for blind image denoising
Soh and Cho (2021)	CNN	Blind image denoising	CNN with Bayesian perspective for blind image denoising
Wu et al. (2020)	CNN	Blind image denoising	CNN with knowledge distillation and Self-Supervised learning for image denoising
Byun et al. (2021)	CNN	Fast blind image denoising	CNN for fast blind image denoising
Liang et al. (2021)	Transformer	Blind image denoising	Vision Transformer for blind image denoising

4.2 Deep learning for image super-resolution

4.2.1 Supervised super-resolution

Existing deep learning-based super-resolution (SR) models are primarily supervised SR models via using low-resolution (LR) images and their corresponding high-resolution (HR) images. These methods are developed by designing new architectures, up-sampling methods, new learning strategies. To make readers' convenient understand difference of different methods, this section summarized these methods from principle to differences as follows. Image super-resolution is an ill-posed problem, which can learn a mapping from LR images to HR images, according to different upsampling ways (Wang et al. 2020). That can be categorized into three kinds: Pre-amplifying resolution, post-amplifying resolution and progressive amplifying resolution.

Pre-amplifying resolution method enlarged a low-resolution image as a same size image with given high-resolution image via upsampling operation, i.e., bicubic interpolation before entering a deep network for image super-resolution. For instance, Dong et al. (2016) designed a 3-layer network by stacking three convolutional layers to achieve a pixel mapping from LR images to HR images, where an input of this network can be obtained by using bicubic operation to amplify LR images (Dong et al. 2016). Although this network has obtained better SR performance, it has poor flexibility for deeper network.

To overcome this problem, VDSR stacked multi convolutional layers to enlarge perception field to capture more structural information in image super-resolution (Kim et al. 2016b). To improve expressive ability of a deep network, Tai et al. used skip connections

Table 5 Deep learning for real image denoising

References	Methods	Applications	Key words
Zhang et al. (2017)	CNN	Real Image Denoising	Deep CNN with residual learning for real image denoisinge
Garibi et al. (2024)	CNN	Real Image Denoisinge	CNN with iterative steps for real image denoising
Zou et al. (2023)	CNN	Real Image Denoisinge	CNN iterative denoising with self-supervised learning
Anwar and Barnes (2019)	CNN	Real Image Denoisinge	CNN image denoising using feature attention mechanisme
Liu et al. (2020)	CNN	Real Image Denoising, image super-resolution	CNN combines residual aggregation and spatial attention mechanism for image restoration
Zhao et al. (2019)	CNN	Real Image Denoisinge	CNN combined residual pyramid structure and channel attention mechanism for image denoising
Li et al. (2024)	CNN	Real Image Denoising, HSI restoration	CNN combines supervised and self-supervised learning strategies for image denoisinge
Mei et al. (2023)	CNN	Real Image Denoising, image restoration	CNN combined residual pyramid structure and channel attention mechanism for image denoising
Wang et al. (2023)	CNN, Transformer	Real Image Denoising	Image denoising is performed using CNN for feature extraction and transformer for long-range dependence modeling capabilities
Zhang and Zhou (2023)	Transformer	Real Image Denoising	Self-supervised Context-aware Transformer for image denoisng

to keep memory of shallow layers to deep layers to enhance strong learning ability of an obtained super-resolution model for image restoration tasks, i.e., image denoising, image super-resolution and JPEG deblocking (Tai et al. 2017). Kim et al. (2016a) used a recursive gate and residual learning operation to improve image super-resolution with fewer parameters. This method is simple for image super-resolution. However, it has higher computational costs. These methods can be summarized in Table 6.

To improve efficiency of image super-resolution, post-amplifying resolution methods amplified resolution of obtained low-frequency mapping in a deep layer of a deep network for image super-resolution. Dong et al. directly used LR images to enter a deep network to obtain low-frequency information and amplify obtained information in a deep layer to construct high-quality images, which can reduce complexity (Dong et al. 2016). Lim et al. (2017) repeatedly used residual learning operations to integrate hierarchical layers to extract more accurate structure information for SISR. Besides, multi-scale idea is embedded into a CNN for image super-resolution. Tong et al. (2017) introduced a Densenet architecture to achieve a SRDenseNet for addressing SR problem, which can effectively alleviate vanishing gradient problem of a deeper network for SISR. Progressive amplifying resolution performs majority of calculations during the low-resolution stage and incrementally raises resolution of images to prevent the loss of details in the up-sampling process. Since the Laplacian pyramid structure can be used to optimize image details layer by layer, Lai et al. (2017) proposed a Laplacian Pyramid Super-Resolution Network (LapSRN) via leveraging a Laplacian pyramid architecture and residual recursive modules for image super-resolution. LapSRN can gradually restore high-resolution images by predicting low-frequency information at each pyramid level and using transposed convolutions to up-sample obtained

Table 6 Deep learning in supervised ways for image super-resolution

References	Methods	Applications	Key words
Dong et al. (2016)	CNN	Image Super-Resolution	CNN for real-time image super-resolution
Kim et al. (2016a)	CNN	Image Super-Resolution	CNN combined with deep recursion for image super-resolution
Kim et al. (2016b)	CNN	Image Super-Resolution	CNN combines deep network structure and residual learning strategy for image super-resolution
Lim et al. (2017)	CNN	Image Super-Resolution	CNN optimized residual network architecture multi-scale super-resolution system
Lai et al. (2017)	CNN	Image Super-Resolution	CNN with pyramid structure for image super-resolution
Lai et al. (2018)	CNN	Image Super-Resolution	CNN with pyramid structure for image super-resolution
Tong et al. (2017)	CNN	Image Super-Resolution	CNN introduces dense skip connections and deconvolution layers for image super-resolution
Haris et al. (2018)	CNN	Image Super-Resolution	CNN combines iterative upsampling and downsampling for image super-resolution
Li et al. (2019)	CNN, RNN	Image Super-Resolution	CNN with feedback blocks and denser skip connections for image super-resolution
Lai et al. (2024)	CNN	HIS Super-Resolution, Remote sensing	CNN combines heterogeneous feature extraction, multi-stage feature alignment, and attention feature fusion for HSI super-resolution
Zhang et al. (2018)	CNN	Image Super-Resolution	CNN with residual channel attention for image super-resolution
Wang et al. (2023)	CNN	Face Image Super-Resolution	Duplex fusing-embedding learning for face super-resolution in low-light environments
Wang et al. (2022)	Transformer, CNN	Face Image Super-Resolution	FaceFormer aggregating global and local representations for face super-resolution
Lu et al. (2022)	CNN	Face Image Super-Resolution	Prior-guided face super-resolution with facial component prior

information to progressively amplify resolution of obtained high-frequency information for SISR. Also, recursive residual modules are utilized to share weights and reduce the number of model parameters to accelerate training speed. To further reduce the number of model parameters, MS-LapSRN (Lai et al. 2018) utilizes same convolutional kernels on different pyramid levels to share weights. By integrating Charbonnier loss, MS-LapSRN significantly enhances stability and convergence speed of a SR model. Alternatively, due to game of a generative adversarial network, Wang et al. (2018) used a forward pyramid technique in a generator and an inverse pyramid technique in a discriminator to achieve a GAN for facilitating high-resolution images.

Table 6 summarizes classical deep learning methods in supervised ways for image super-resolution.

4.2.2 Unsupervised and weakly-supervised super-resolution

Most existing super-resolution (SR) methods focus on supervised learning, which involves learning from paired low-resolution (LR) and high-resolution (HR) images (Ledig et al. 2017). However, high-resolution images have higher requirements for shooting environment and devices, which results in difficult acquirement high-resolution images (Timofte

et al. 2017). Thus, unsupervised and weakly-supervised super-resolution methods are developed (Ulyanov et al. 2018).

Deep Image Prior (DIP) leverages inherent structural properties of convolutional neural networks as a prior to achieve high-quality image reconstruction (Ulyanov et al. 2018). It utilizes a randomly initialized CNN and optimizes parameters to minimize a loss function to approximate the target image rather than require training data. During the optimization process, the network progressively generates the target image. Although it is effective, single-scale DIP may struggle to capture complex details, particularly in images with multi-scale characteristics. To address this limitation, Multi DIP (Wang et al. 2021) extends the basic DIP framework by running DIP in parallel across multiple scales. This multi-scale approach allows the method to better capture both local and global features to enhance the quality of image reconstruction. Another improvement of original DIP is that DIP combines Total Variation Regularization (DIP-TV) (Liu et al. 2019) to reduce noise and artifacts for obtaining clearer, higher-quality reconstructed images. Alternatively, replying on properties of internal images can achieve a weakly-supervised super-resolution model. These methods can use partially labeled or weakly labeled data to reduce requirement of large data to achieve strong performance (Wei et al. 2018). For instance, a cycle consistency loss method can be used to ensure consistency of a model from mapping between low-resolution and high-resolution images. Inspired by that, Zhu et al. (2017) use two generative adversarial networks to enforce this consistency of a SR model. That is, one GAN generates HR images from LR images and the other converts the HR images to LR images to ensure reconstructed LR images closely resembles original images. Yuan et al. (2018) developed a Cycle-in-Cycle GAN (CinCGAN) by using a similar cycle-consistency approach to preserve image details LR-to-HR and HR-to-LR mappings for image super-resolution.

Transformer-based models have recently emerged as a powerful tool for image super-resolution, leveraging the self-attention mechanism to capture long-range dependencies in images. For example, Image Processing Transformer(IPT) (Chen et al. 2021) uses a transformer architecture pre-trained on multiple image processing tasks, including super-resolution, to achieve competitive performance across different resolutions. Swin Transformer for Image Restoration(SwinIR) (Liang et al. 2021) introduces hierarchical Swin Transformer blocks, which improve the ability to process high-resolution image patches and maintain computational efficiency. These methods surpass traditional convolutional methods by effectively modeling global context, especially in cases where structural details need to be reconstructed in high-quality SR outputs.

Diffusion models represent a novel direction for SR by framing the task as a conditional generative process. These models progressively denoise a random latent vector towards the desired high-resolution output. For instance, Super-Resolution via Repeated Refinement (SR3) (Saharia et al. 2022b) uses a cascaded diffusion process to gradually refine LR images into HR images. The process starts with Gaussian noise and iteratively learns a denoising function conditioned on the LR image.

Table 7 summarizes more unsupervised and weakly-supervised SR methods.

4.3 Deep learning for image deblurring

4.3.1 Deep learning for image deblurring with known blur kernels

Image deblurring focuses on recovering a sharp image from a given blurry image. If blur kernel is known, this process is called as non-blind image deblurring. Current deep learning-based deblurring methods fall into two primary categories: deconvolution-based approaches (Zeiler et al. 2010) and prior knowledge-based approaches (Ulyanov et al. 2018).

Deconvolution methods can use reverse process of convolution operations to estimate blur kernels for image deblurring (Joshi et al. 2008). For example, Schuler et al. (2013) introduced a multi-layer perceptron to achieve deconvolution operation to obtain blur kernels. That is, this method first recovers a sharp image via regularized inverse operations in Fourier domain, then uses a neural network to remove artifacts in a sharp image for image deblurring. Similarly, Ren et al. (2018) proposed a convolutional neural network by using generalized low-rank approximation (GLA) to handle various blur kernels, where GLA depends on low-rank properties of pseudo-inverse kernel to achieve good performance in image deblurring. To improve robustness of an obtained deblurring model, Dong et al. (2021) proposed a deep Wiener deconvolution method via using a linear filter to extract

Table 7 Deep learning for unsupervised super-resolution

References	Methods	Applications	Key words
Ulyanov et al. (2018)	CNN	Image super-resolution	No training priors required for image super-resolution
Wang et al. (2021)	CNN	Image super-resolution	Image super-resolution Based on Depth Image Prior (DIP)
Liu et al. (2019)	CNN	Image super-resolution, deblurring	CNN with traditional regularization(TV) for image super-resolution
Zhu et al. (2017)	GAN	Image super-resolution	GAN without paired data for image restoration
Yuan et al. (2018)	GAN	Image super-resolution	GAN without paired data for image restoration
Chen et al. (2021)	Transformer	Image Super-Resolution	Transformer-based network leveraging self-attention for high-quality SR
Liang et al. (2021)	Transformer	Image Super-Resolution	Swin Transformer with hierarchical structure for efficient image super-resolution
Saharia et al. (2022b)	Diffusion	Image Super-Resolution	Diffusion-based generative model using progressive refinement for SR

features from blurred images as inputs of Winer deconvolution to estimate blur kernels for addressing image deblurring. Differing from previous methods, Zhang et al. can use a framework for noisy image deblurring (Zhang et al. 2017). That is, it can use a fully convolutional network to remove noise in gradient domain and depend on iterative deconvolution to address problem of image deblurring. To address image deblurring with limited samples, Arjomand Bigdeli et al. (2017) proposed a GradNet by introducing a mean-shift vector field to smooth the natural image distribution and using gradient descent to minimize Bayes risk in non-blind deblurring. Jin et al. (2017) further enhanced GradNet via integrating various image priors and improving MAP to accelerate deblurring processing. Zhang et al. (2017) incorporated discriminative denoisers within a model-based optimization framework for non-blind deblurring. Although it was effective for image deblurring, it had limitation for artifacts. Besides, some methods, i.e., USRNet (Zhang et al. 2020) can solve non-blind deblurring in the process of image super-resolution. More detailed information of these methods can be generalized in Table 8.

4.3.2 Deep learning for image deblurring with unknown blur kernels

In real-world images, different regions can affect non-uniform nature of the blur to increase difficulty of image deblurring. To address these challenges, a variety of deep learning-based algorithms for blind image deblurring have been developed. This section provides an overview of these approaches.

Single image deblurring methods reply on U-Net architecture with residual learning techniques to obtain unknown blur kernels for image deblurring. Kim et al. (2022) proposed multi-scale-stage network via employing a coarse-to-fine architecture to deal with blur kernels of different scales in image deblurring. Similarly, Tao et al. (2018) analyzed various U-Net and denoising autoencoder architectures to achieve a scale-recurrent network in image deblurring. That is, a U-Net is used to roughly deblur images. Then, another U-Net combines an autoencoder to further deblur images to obtain clearer images. To address image deblurring with unknown blur kernels of limited samples, GANs are used to developed (Kupyn et al. 2018). For instance, Kupyn et al. (2019) used an end-to-end conditional GAN including two strided convolutional blocks, nine residual blocks, and two transposed convolutional blocks in the generator to convert blurry images into sharp images for motion deblurring. Besides, incorporating a relativistic conditional GAN and a dual-scale discriminator with local and global branches can improve performance of image blurring (Kupyn et al. 2019). To address real image blurring, image restoration techniques and image blurring techniques are combined to improve applicability of obtained blurred models. Yang et al. (2021) used GAN and prior to recover face detailed information and achieve face image blurring. Additionally, recent advancements have seen the application of Transformer architectures to image deblurring tasks, achieving remarkable results. Tsai et al. (2022) introduced Stripformer, a Transformer model specifically designed for motion deblurring by effectively extracting and aggregating features from strip regions. Liang et al. (2024) proposed a Transformer-based approach capable of handling both defocus and motion blur, showcasing its versatility across multiple blur types. Kong et al. (2023) developed an efficient Transformer model focused on achieving high-quality motion deblurring. By leveraging the long-range dependency capturing capability of Transformers along

Table 8 Deep learning for image deblurring with known blur kernels

References	Methods	Applications	Key words
Schuler et al. (2013)	MLP	Image deblurring (Gaussian, motion)	MLP with the inverse regularization in the Fourier domain recovers the sharp image
Ren et al. (2018)	CNN	Image deblurring (Gaussian, disk, motion)	CNN adopt generalized low-rank approximation and unified framework to deal with different convolution kernels
Dong et al. (2021)	CNN	Image deblurring (Gaussian, disk, motion)	CNN with wiener convolution deconvolution for image deblurring
Zhang et al. (2017)	CNN	Image deblurring (Gaussian, motion)	CNN with iterative deconvolution in a multi-stage framework for image deblurring
Xu et al. (2014)	CNN	Image deblurring (Gaussian, disk)	CNN combines traditional optimization schemes and neural networks for deblurring
Kruse et al. (2017)	CNN	Image deblurring (Motion)	CNN-based prior with FFT-based deconvolution for image deblurring
Arjomand Bigdeli et al. (2017)	DAE, CNN	Image deblurring (Motion, disk)	CNN with natural image priors and Bayesian estimation for image deblurring
Jin et al. (2017)	CNN	Image deblurring (Motion)	CNN combine with Bayesian framework for image deblurring
Zhang et al. (2020)	CNN	Image deblurring (Motion, Gaussian)	CNN combines model-based and learning-based methods for image deblurring
Li et al. (2019)	CNN	Image deblurring (Motion)	CNN integrates TV-regularized algorithm for image deblurring

with computational efficiency, these methods have significantly advanced the field of image deblurring. Table 9 summarizes discussed methods mentioned in this section.

5 Deep learning for high-level vision tasks

High-Level vision tasks encompass image classification, object detection, and image segmentation in this section, where Fig. 4 summarizes content of Sect. 5.

Table 9 Deep learning for image deblurring with unknown blur kernels

References	Methods	Applications	Key words
Su et al. (2017)	CNN	Image deblurring (motion), video deblurring	An end-to-end data-driven CNN for deblurring
Ren et al. (2020)	CNN	Real world image deblurring	Multi-scale CNN for image deblurring
Kim et al. (2022)	CNN	Image deblurring (motion)	CNN with multi-scale multi-stage architecture for image deblurring
Tao et al. (2018)	CNN	Image deblurring	Deblurring via a scale-recurrent network that shares network weights across scales
Kupyn et al. (2018)	GAN	Image deblurring (motion)	GAN for image deblurring
Kupyn et al. (2019)	GAN	Image deblurring (motion)	GAN and FPN for image deblurring
Yang et al. (2021)	GAN	Blind face restoration (BFR)	GAN prior embedded network for image deblurring
Tsai et al. (2022)	Transformer	Image deblurring (motion)	Transformer for image deblurring
Liang et al. (2024)	Transformer	Image deblurring (defocus, motion)	Transformer for image deblurring
Kong et al. (2023)	Transformer	Image deblurring (motion)	Transformer for High-Quality image deblurring

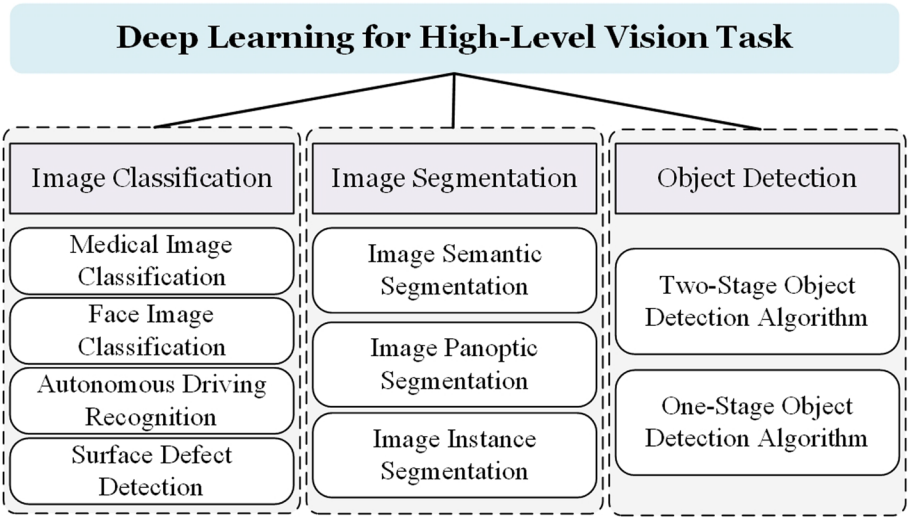


Fig. 4 Overview of high-level vision tasks

5.1 Deep learning for image classification

5.1.1 Deep learning for medical image classification

Due to strong ability of feature extraction, deep learning technique has been widely applied in medical image classification (Yu et al. 2024). For instance, Huang et al. enhanced relations of different layers to facilitate hierarchical information for medical image classification (Huang et al. 2017). Differing from ordinary images, medical images include noise and blurred boundaries. To deal with these challenges, a U-Net architecture employs skip connections to gather low- and high-resolution feature mapping to effectively integrate features of different super-resolution. This characteristic makes U-Net become a benchmark for numerous medical image classification.

CheXNet utilized a 121-layer DenseNet architecture to extract complex features and applied fully connected layers for disease classification, which can provide predicted probabilities for various thoracic conditions (Rajpurkar et al. 2017). To ensure reliability of disease diagnosis, some scholars exploited two-phase methods to deal with mentioned problem. The first phase is used for image segmentation. The second phase is used for image classification. For instance, Oktay et al. can use residual learning strategy and an attention mechanism for image segmentation and fully connected layer for medical image classification (Oktay et al. 2018). Additionally, the progression of brain tumors can lead to changes in their shape, size, and location over time, combining CNNs and RNNs is used for medical image classification. Specifically, CNN is used to extract spatial features from brain tumor images and RNN can capture the temporal sequence of these images, where their combinations can predict brain tumor disease different time periods (Raza et al. 2022). Table 10 summarizes key information of more methods for medical image classification in this section.

5.1.2 Deep learning for face image classification

The primary challenge in facial recognition is to minimize variations in facial images of the same individual while maximizing differences among images of different individuals. To address this problem, Taigman et al. (2014) leveraged deep convolutional neural networks on LFW dataset to obtain high accuracy of face image recognition. Inspired by that, Schroff et al. (2015) utilized a triplet loss to effectively cluster facial feature points of same individual while distinguishing those of others to obtain excellent performance of face image classification. Sun et al. (2014) jointly applied identification and verification signals to learn facial representation for face image classification, where different identities can promote separation of different ID features to aggregate features of same ID for improving classification results of face image classification. To improve robustness of a face recognition model, Parkhi et al. (2015) constructed a large dataset including 2.6 million facial images to enhance rich samples of face images, where this dataset is conducted via a combination of automated processes and manual participation. To further improve recognition results, a loss function is optimized.

Wen et al. (2016) introduced center loss via minimizing intra-class distance among deep features to enhance discriminative power for face image recognition. That is, integrating a center loss and SoftMax to facilitate robust features to ensure both inter-class separability and intra-class compactness. Similarity, Liu et al. (2017) proposed angular SoftMax

Table 10 Deep learning for medical image classification

References	Methods	Applications	Key words
Huang et al. (2017)	DenseNet	Disease Detection in X-ray Images	DenseNet for medical image classification
Rajpurkar et al. (2017)	DenseNet	Image Analysis for Pneumonia Detection	Deep CNN for X-ray image classification
Xu et al. (2018)	LSTM	Drug-drug interaction (DDI)	BR-LSTM for biomedical Resources
Fang et al. (2022)	AlexNet	Extraction of the key ROIs in X-ray images	AlexNet and Two-Class combined model for X-ray images extraction
Oktay et al. (2018)	CNN	Pancreas Segmentation in Medical Scans	CNN with attention gate for CT image classification
Gulshan et al. (2016)	CNN	Ophthalmology Image Analysis	CNN for diabetic retinopathy detection
Raza et al. (2022)	CNN	Brain Tumor Detection and Classification	CNN with GoogleNet for brain tumor image classifications
Yang et al. (2023)	Diffusion	General medical image classification	Diffusion for general medical image classifications

Table 11 Deep learning for face image classification

References	Methods	Applications	Key words
Taigman et al. (2014)	CNN	Face Verification	CNN for 3D face modeling and recognition
Schroff et al. (2015)	CNN	Face Recognition, Clustering	CNN for face recognition
Sun et al. (2014)	CNN	Face Recognition	CNN for face recognition
Parkhi et al. (2015)	CNN	Face Recognition	CNN for face recognition
Wen et al. (2016)	CNN	Face Recognition	CNN with center loss function for face recognition
Liu et al. (2017)	CNN	Face Forgery Analysis	CNN with angular Softmax loss for face recognition
Chen et al. (2024)	Diffusion	Face Recognition	Comprehensive Dataset for Diffusion-Based Face Forgery Analysis

loss to break identified limitations in the traditional SoftMax loss for open-set recognition. Applying discriminative constraints on a hypersphere to learn robust angle-discriminative features to enhance discriminative ability for challenging recognition scenarios. Table 11 summarizes key information of previous introduction methods for face image classification.

5.1.3 Deep learning for autonomous driving recognition

To design suitable autonomous driving model, sophisticate networks be developed via intricate image data. Motivated by that, Huang et al. (2017) introduced densely connected convolutional networks to emphasize importance of densely connected layers to extract effective information for autonomous driving recognition. To reduce computational costs, Howard et al. (2017) utilized point-wise and depth convolutions to separately obtain features and applied a catenation operation to merge obtained features, which can obtain complementary features and reduce computational costs for autonomous driving recognition. Tan and Le (2019) can employ EfficientNet to layer-to-layer increase the number of channels and improve super-resolution to achieve a high recognition result for autonomous driving recognition. To further enhance autonomous driving recognition, recent studies have explored the use of advanced Transformer-based models. Tu et al. (2024) introduced a novel approach by fine-tuning diffusion transformers for autonomous driving tasks, effectively leveraging the strengths of diffusion models in conjunction with Transformer architectures to improve recognition accuracy. Similarly, Ma et al. (2023) proposed a cross-view Transformer network (CVTNet) specifically designed for LiDAR-based place recognition. This method excels in capturing spatial relationships across multiple views, significantly boosting the performance of autonomous driving systems. These approaches demonstrate the growing potential of Transformer models in addressing the complex requirements of autonomous driving recognition tasks. Table 12 summarizes more related research about deep learning for autonomous driving recognition in this section.

5.1.4 Deep learning for surface defect detection

In industrial production processes, existing technologies and working conditions often compromise product quality. Surface defect detection plays a critical role in identifying issues such as spots, scars, and color variations, thereby ensuring product integrity. Ren et al. (2017) proposed a surface defect detection algorithm that utilizes fully convolutional networks to generate heat-maps indicating the probability of defects at various locations, achieving excellent results across multiple datasets. Similarly, Yi et al. (2017) introduced a CNN-based method for rail defect detection that automatically learns image features without the need for manually designed feature extractors, demonstrating strong feature representation and generalization capabilities. Furthermore, Tabernik et al. (2020) developed an integrated approach for surface defect detection that combines segmentation and decision networks. This method first performs pixel-level segmentation on input images to isolate defect areas from the background, followed by the application of DecisionNet for further analysis and evaluation of the segmented results. This two-stage segmentation architecture enables the model to be trained effectively with a limited number of samples while enhancing the accuracy and reliability of surface defect detection. Cui et al. (2021) proposed the SDDNet method, which introduces Feature Retention Blocks (FRB) to preserve texture information that may be lost during downsampling. Additionally, the method incorporates a Skip Dense Connection Module (SDCM) to propagate fine-grained details from low-level feature maps to high-level feature maps, thereby improving the detection accuracy of texture variations and small defect sizes. The strong glossiness of metal surfaces, along with the complexity of various surface defect types, poses significant challenges for surface defect detection. To

Table 12 Deep learning for autonomous driving recognition

References	Methods	Applications	Key words
Krizhevsky et al. (2012)	CNN	Autonomous Driving	CNN with ReLU and Dropout regularization for image classification
LeCun et al. (1989)	CNN	Autonomous Driving	CNN with small convolutional filters for image classification
Szegedy et al. (2015)	CNN	Autonomous Driving	CNN with GooLeNet for image classification
He et al. (2016)	ResNet, CNN	Autonomous Driving	Residual networks for image classification
Huang et al. (2017)	CNN	Autonomous Driving	DenseNet for image classification
Howard et al. (2017)	CNN	Autonomous Driving	CNN with depthwise separable convolution for image classification
Tan and Le (2019)	CNN	Autonomous Driving	CNN with composite scaling method and neural architecture search for image classification
Tu et al. (2024)	Transformer, Diffusion	Autonomous Driving	Fine-tuning diffusion transformers for autonomous driving
Ma et al. (2023)	Transformer	Autonomous Driving	A cross-view transformer network for LiDAR-based place recognition

address these challenges, Tao et al. (2018) employed a Cascade Autoencoder (CASAE) for defect localization and segmentation. This cascaded network converts input defect images into pixel-level prediction masks based on semantic segmentation, followed by classification of the defects using a Convolutional Neural Network (CNN). The method demonstrates high robustness and accuracy in successfully detecting various types of metal defects under industrial conditions. More related work can be summarized in Table 13.

5.2 Deep learning for image segmentation

There are three types for image segmentation: semantic segmentation, instance segmentation, and panoramic segmentation in this section. Due to advance of deep learning, this section introduces deep learning for image segmentation as follows.

5.2.1 Deep learning for image semantic segmentation

Image semantic segmentation aims to assign each pixel in an image to a specific semantic category for pixel-level understanding and analysis (Mottaghi et al. 2014). This process enables machines to differentiate between various object classes and background regions to facilitate context recognition of digital images, which has wide applications in landscapes

Table 13 Deep learning for Surface defect detection

References	Methods	Applications	Key words
Ren et al. (2017)	CNN	Surface defect detection, image classification	CNN for automated surface inspection (ASI)
Yi et al. (2017)	CNN	Steel strip surface defect detection	CNN for steel strip surface defects
Tabernik et al. (2020)	CNN	Surface defect detection	CNN with segmentation network for surface-defect detection
Cui et al. (2021)	CNN	Surface defect detection, real-time processing	CNN with FRB and SDCM for surface defect detection
Tao et al. (2018)	CNN	Surface defect detection	CNN for automated surface inspection (ASI)
Gao et al. (2022)	Transformer	Surface defect detection	A variant swin transformer for surface-defect detection
Shang et al. (2023)	Transformer	Surface defect detection	Transformer for surface-defect detection
Zhu et al. (2023)	Transformer	Steel surface defect detection	Transformer for steel surface-defect detection

(Devereux et al. 2004), portraits (Gallagher and Chen 2009), and medical images (Ma et al. 2024).

With advancements in artificial intelligence techniques, deep learning techniques have become essential for training models for image segmentation. For instance, Long et al. (2015) proposed fully convolutional networks (FCNs) rather than traditional fully connected layers to produce heatmaps for image segmentation. To address issue of inconsistent image sizes caused by convolutional operations and pooling operation, upsampling methods are employed to restore original dimensions. For instance, using skip connections acts between convolutional layers and non-adjacent layers in a U-Net to prevent data loss during process of downsampling and enhance resolution of predicted images (Ronneberger et al. 2015), where U-Net can preserve intricate details for biomedical image segmentation. Google DeepLab model integrates dilated convolutions and fully connected conditional random fields into a CNN to obtain more detailed information and improve effect of image segmentation (Chen et al. 2017). PSPNet (Zhao et al. 2017) gathers encoder-decoder and a pyramid pooling layer to enhance pixel-wise computations for image segmentation. Besides, Luc et al. (2017) trained a standard CNN for semantic segmentation alongside an adversarial network designed to differentiate between true and predicted segmentations, according to GANs. This method can generate outputs distinguish can better actual segmentations for image segmentation. With the rapid evolution of deep learning methods, Transformer-based approaches have also been increasingly adopted for image semantic segmentation, offering significant advancements in pixel-level understanding tasks. Strudel

et al. (2021) introduced the Segmenter, a Transformer-based model specifically designed for semantic segmentation, leveraging the long-range dependency capabilities of Transformers to improve segmentation accuracy and contextual understanding. Furthermore, He et al. (2022) proposed a novel integration of the Swin Transformer with U-Net architecture for remote sensing image semantic segmentation. By embedding Swin Transformer modules into U-Net, their approach effectively captures both local and global features, achieving superior performance in segmenting complex remote sensing images. These Transformer-driven methods underscore their transformative potential in advancing the field of image semantic segmentation across various applications. Table 14 can be used to summarize discussed methods mentioned in this section.

5.2.2 Deep learning for image instance segmentation

Instance segmentation reverses the priority order of semantic segmentation by precisely partitioning individual object instances rather than just predicting the semantic classification of each pixel (Hafiz and Bhat 2020). Differing from semantic segmentation, instance

Table 14 Deep learning for image semantic segmentation

References	Methods	Applications	Key words
Long et al. (2015)	CNN	Image semantic segmentation	Fully convolutional networks and deconvolution layers for image deblurring
Ronneberger et al. (2015)	CNN	Medical image segmentation, biomedical image analysis	U-Net for biomedical images segmentation
Chen et al. (2017)	CNN	Image semantic segmentation	CNN with atrous convolution and CRF for image segmentation
Zhao et al. (2017)	CNN	Image semantic segmentation, remote sensing	CNN with pyramid parsing module for image segmentation
Luc et al. (2017)	CNN	Image semantic segmentation	Autoregressive convolutional neural networks for image segmentation
Soni et al. (2023)	CNN	Arbitrary-Shaped Text segmentation	Arbitrary-Shaped text segmentation in edge-fainted noisy scene images
Strudel et al. (2021)	Transformer	Image semantic segmentation	Transformer for semantic segmentation
He et al. (2022)	Transformer, U-Net	Remote sensing image semantic segmentation	Swin transformer embedding UNet for remote sensing image semantic segmentation

segmentation focuses on differences of different instances of object detection rather than recognizing pixel levels. It can be mainly divided into two categories, i.e., a two-stage and one-stage methods as follows.

Two-stage method first obtain different bounding boxes, according to different instances. Then, it can finish image segmentation in each bounding box, according to pixel levels. Girshick et al. (2014) employed a selective search algorithm to extract approximately 2000 candidate region proposals from an image and applied a CNN to extract information for distinguishing different instances in image instance segmentation. Because each candidate box need be recognized instances, that may cause bigger computational costs and slower detection speeds. To overcome limitation of candidate boxes with fixed sizes, He et al. (2015) used an ROI pooling layer to achieve function of candidate boxes with arbitrary sizes. Besides, this method can use feature mapping from candidate boxes to identify patch for accelerate speed in image instance segmentation. Although mentioned methods can perform well in image instance segmentation, they often suffer from inefficiencies and bigger computational costs caused sliding window strategies. To overcome this drawback, reducing windows are presented. Mask R-CNN (He et al. 2017) based Faster R-CNN (Ren et al. 2015) enhances the instance segmentation landscape by incorporating a mask prediction branch to improve speed in image instance segmentation, where faster R-CNN contains two stages to achieve efficient performance. The first stage can use a region proposal network to generates regions of interest (RoIs) and apply classification and bounding box regression to extract fixed-size features in RoIs. The second stage can use a semantic segmentation method to deal with obtained features for image instance segmentation. This method allows to leverage advancements in object detection, significantly improving instance segmentation accuracy by integrating superior detectors. To make this network lightweight, cascade Mask R-CNN (Cai and Vasconcelos 2018) is developed via introducing mask branches based on Cascade R-CNN, which can break the limitations of sample selection associated with a single IoU threshold. This strategy can enhance the model's ability to differentiate between positive and negative samples during training. Despite good effect of this method, it still faces challenges of rough prediction from edges on large objects.

Differing from two-stage instance segmentation, one-stage instance segmentation Can simultaneously executes object detection and mask generation. For instance, YOLO (Jiang et al. 2022) offered a real-time solution by dividing the instance segmentation task into parallel computations for object detection and mask generation to achieve rapid process speed. Table 15 can be utilized to concluded key information of methods above in this sections.

5.2.3 Deep learning for image panoptic segmentation

Image panoptic segmentation is an integrated approach that combines semantic and instance segmentation. This method not only identifies "stuff" in an image but also distinguishes between individual object instances. Each pixel in the image is assigned both a semantic label and a unique instance identifier (Kirillov et al. 2019). The primary aim of panoptic segmentation is to provide a comprehensive understanding of both objects and backgrounds while simultaneously localizing and segmenting object instances. Detailed introduction of its method can be shown as follows.

Kirillov et al. (2019) merged a fully convolutional network to achieve simultaneously achieve semantic and instance segmentation to further finish image panoptic segmenta-

Table 15 Deep learning for image instance segmentation

References	Methods	Applications	Key words
Girshick et al. (2014)	CNN	Image instance segmentation	R-CNN for image segmentation
He et al. (2015)	CNN	Image instance segmentation	CNN with Spatial Pyramid Pooling(SPP) for image segmentation
He et al. (2017)	CNN	Image instance segmentation	Mask R-CNN for image segmentation
Girshick (2015)	CNN	Image instance segmentation	Fast R-CNN for image segmentation
Ren et al. (2015)	CNN	Image instance segmentation	CNN with Regional Proposal Network (RPN) for image segmentation
Cai and Vasconcelos (2018)	CNN	Image instance segmentation	CNN with multi-cascade object detection architecture for image segmentation
Hu et al. (2021)	Transformer	Image instance segmentation	End-to-end instance segmentation with transformers
Ye et al. (2023)	Transformer	Remote sensing instance segmentation	Remote sensing image instance segmentation

tion. Similarly, Yang et al. (2019) combined sharing encoder-decoder architecture for the two sub-tasks and an Atreus Spatial Pyramid Pooling module at the end of an encoder for image panoptic segmentation. To better collaborative work, unifying image semantic and instance segmentation into a framework is conducted for image panoptic segmentation. For instance, Li et al. (2019) proposed a unified framework via integrating proposal attention module and mask attention module to obtain pixel- and object-level information to improve expressive ability of obtained image panoptic segmentation model. To ensure consistency between instance and stuff segmentation, Li et al. (2018) aligned instance foreground mask mapping and original image's feature and computed their difference via a L2 loss function to improve robustness of an obtained image panoptic segmentation model. Alternatively, Xiong et al. (2019) incorporated a panoptic head to effectively resolve conflicts between instance and semantic segmentation to improve accuracy of instance and class label predictions. Additionally, recent advancements have introduced Transformer and diffusion-based methods to further enhance image panoptic segmentation. Li et al. (2022) proposed a Transformer-based approach that effectively integrates semantic and instance segmentation tasks into a unified framework, leveraging the long-range dependency modeling capabilities of Transformers to improve segmentation accuracy and coherence. Building upon this, Van Gansbeke and De Brabandere (2025) introduced a novel method combining diffusion models and Transformer architectures for panoptic segmentation. Their approach not only

excels in segmenting both "stuff" and object instances but also incorporates mask inpainting techniques to handle occlusions and incomplete object instances. Table 16 can be used to generalize mentioned methods.

5.3 Deep learning for object detection

5.3.1 Two-stage object detection algorithm

Two-stage object detection algorithm firstly can use a region proposal network to obtain region proposals. Then, it utilized CNNs to extract structural information into these region proposals to obtain information of position regression, enabling precise object localization for object detection. Girshick et al. (2014) utilized a high-capacity CNN object object localization and segmentation of bottom-up region proposals. Also, they used supervised pre-training and domain-specific fine-tuning to recognize objects. To overcome offset drawback of object location, He et al. (2015) incorporated a spatial pyramid pooling layer in a R-CNN to address fixed input image sizes and object deformation. Girshick (2015) can optimize R-CNN by introducing a ROI Pooling layer in a single forward pass to improve detection accuracy and efficiency. Ren et al. (2015) further advanced a region proposal network to obtain detection region proposals to improve an object detection algorithm. Alternatively, Dai et al. (2016) can distinguish detection region proposals via scoring to intentionally enhance monitor of important areas to improve effects of object detection. Additionaly,

Table 16 Deep learning for image panoptic segmentation

References	Methods	Applications	Key words
Kirillov et al. (2019)	CNN	Image panoptic segmentation	CNN combined with FPN and RCN for image segmentation
Yang et al. (2019)	CNN	Image panoptic segmentation	FCN for image segmentation
Li et al. (2019)	CNN	Image panoptic segmentation	CNN combines attention mechanism and unified framework for image segmentation
Li et al. (2018)	CNN	Image panoptic segmentation	An end-to-end approach for image segmentation
Xiong et al. (2019)	CNN	Image panoptic segmentation	CNN with variable convolution for image segmentation
Li et al. (2022)	Transformer	Image panoptic segmentation	Image panoptic segmentation with transformers
Van Gansbeke and De Brabandere (2025)	Diffusion	Image panoptic segmentation	Image panoptic segmentation and mask inpainting with transformers

Chen et al. (2023) proposed DiffusionDet, a diffusion-based model tailored for object detection. By incorporating diffusion processes, the model effectively learns object boundaries and improves localization precision in complex scenes. Building on this, Fang et al. (2024) introduced a controllable diffusion model for object detection, allowing for more flexible and adaptive detection by integrating controllable parameters within the diffusion process. These diffusion-based approaches offer innovative solutions to address the challenges of object localization and detection, pushing the boundaries of traditional two-stage object detection methods. Table 17 can be used to summarize important information of more two-stage object detection algorithms in this section.

5.3.2 One-stage object detection algorithm

One-Stage object detection algorithm directly extracts features from the entire image rather than searching region proposals for object detection algorithm. For instance, Sermanet et al. (2013) introduced a unified framework via sharing convolutional layers to achieve adaptive weights and fully connected layers for image classification, object localization object detection. Liu et al. (2016) used forward convolutional neural networks to obtain different scale features for detecting different objects. Although this method can perform well for small object detection, it is not effective for big object detection. To address this problem, DSSD (Fu et al. 2017) employs a Top-Down network structure to fuse high- and low-level features to enhance multi-scale feature maps to increase detection accuracy. To deal with object detection in terms of real time, you only look once (YOLO) (Redmon 2016) is proposed as follows (Table 18).

Joseph Redmon et al. used regression solution to achieve a YOLOv1 to directly detect objects rather than generating region proposals (Redmon 2016). To improve speed, YOLOv2 (Redmon and Farhadi 2017) used batch renormalization technique to act a DARKNet-19 to refine network architecture to accelerate training speed in object detection. To improve performance of object detection, using logistic regression can guide a YOLOv2 to improve expressive ability of an obtained object detection model (Redmon and Farhadi 2018).

Table 17 Two-stage object detection algorithm

References	Methods	Applications	Key words
Girshick et al. (2014)	CNN	Object detection, Semantic segmentation	Supervised pre-training CNN for object detection
He et al. (2015)	CNN	Object detection, Image classification	CNN with spatial pyramid pooling for object detection
Girshick (2015)	CNN	Object detection	Fast R-CNN for object detection
Ren et al. (2015)	CNN	Object detection	Faster R-CNN for object detection
Dai et al. (2016)	CNN	Object detection	Fully convolutional R-CNN for object detection
Chen et al. (2023)	Diffusion	Object detection	Diffusion model for object detection
Fang et al. (2024)	Diffusion	Object detection	Controllable diffusion model for object detection

Table 18 One-stage object detection algorithm

References	Methods	Applications	Key words
Sermanet et al. (2013)	CNN	Image classification, object detection	CNN uses multi-scale and sliding window methods for object detection
Liu et al. (2016)	CNN	Object detection	Single deep neural network for object detection
Fu et al. (2017)	CNN	Object detection	SSD combined with Residual-101 for object detection
Lin (2017)	ResNet	Dense Object Detection tasks	Solution to Class Imbalance in Object Detection
Carion et al. (2020)	Transformer	Object detection	Transformer encoder–decoder with bipartite matching, end-to-end object detection
Beal et al. (2020)	Transformer	Object detection	Transformer for object detection

That is, residual learning operations and feature pyramid architecture are gathered into a YOLOv2 to achieve YOLOv3 (Redmon and Farhadi 2018) to ensure accuracy alongside practicality for object detection. To further improve efficiency, YOLOv3 is combined with an activation function of MISH rather than ReLU as YOLOv4 to enhance detection accuracy and speed (Bochkovskiy et al. 2020). To make a tradeoff detection accuracy and speed, varying network depth and width are proposed to improve YOLOv4 as YOLOv5 (Wu et al. 2021), i.e., CornerNet (Law and Deng 2018), CenterNet (Zhou et al. 2019), FCOS (Tian et al. 2019) and YOLOX (Ge et al. 2021). Anchor-free mechanism is removed from YOLOv5 as YOLOv6 (Li et al. 2022) to further improve training speed for object detection. Increasing network depth in Extended-ELAN to improve YOLOv6 as YOLOv7 (Wang et al. 2023) to improve ability of object detection. To enhance performance and flexibility of object detection, removed anchor-free YOLOv5 is combined with spatial pyramid pooling to address image features of arbitrary sizes as YOLOv8 to reduce computational costs and accelerate training speed for object detection (Li et al. 2023). YOLOv9 (Wang et al. 2024) presents the concept of Programmable Gradient Information (PGI) to provide comprehensive input information to improve adaptive abilities of different deep networks for object detection. Also, it can facilitate reliable gradient information for weight updates. Besides, a general efficient layer aggregation network based on gradient path planning (GELAN) is used to optimize YOLOv9 to confirm excellent results of PGI yields on lightweight models. According to mentioned illustrations, we can see that core principle of the YOLO series, i.e., YOLOv1-YOLOv9 is that it can covert object detection problem to regression problem to deal with the entire image to detect objects. Table 19 summaries key information of classical YOLO algorithms for object detection.

Table 19 Key information of classical YOLO algorithms for object detection

Models	Anchor	Inputs	Backbones	Necks
YOLOv1 (Redmon 2016)	Anchors	Resize (448×448×3)	GoogLeNet	–
YOLOv2 (Redmon and Farhadi 2017)	Anchors	Resize (416×416×3)	Darknet-19	–
YOLOv3 (Redmon and Farhadi 2018)	Anchors	Resize (608×608×3)	Darknet-53 (53×Conv)	FPN
YOLOv4 (Bochkovskiy et al. 2020)	Anchors	Resize (608×608×3)	CSPDarknet53	SPP, PAN
YOLOv5 (Wu et al. 2021)	Anchors	Resize (608×608×3)	CSPDarknet53	SPP, PAN
YOLOX (Ge et al. 2021)	Anchor-free	Resize (608×608×3)	Darknet-53	SPP, FPN
YOLOv6 (Li et al. 2022)	Anchor-free	Resize (640×640×3)	EfficientRep Backbone	Rep-PAN
YOLOv7 (Wang et al. 2023)	Anchors	Resize (640×640×3)	E-ELAN	SPP, PAN
YOLOv8 (Li et al. 2023)	Anchor-free	Resize (640×640×3)	CSPDarknet variant	SPP, PAN
YOLOv9 (Wang et al. 2024)	Anchor-free	Resize (640×640×3)	GELAN	PGI, FPN

6 Deep learning for video processing

This section explores the ways in which deep learning is revolutionizing video processing, covering a wide range of applications from classification to enhancement. It also emphasizes the significant impact that this technology has on accuracy, efficiency and future development.

6.1 Video analysis and understanding

The field of video analysis and understanding has evolved significantly with the advent of deep learning, propelled by innovations in both infrastructure and task-specific methodologies. Early breakthroughs, such as the 3D CNNs introduced by Tran et al. (2015) and the

two-stream networks by Feichtenhofer et al. (2016), pioneered the integration of spatiotemporal feature fusion, effectively capturing dynamic visual and temporal information critical for understanding video content. Subsequent developments, including Temporal Segment Networks by Wang et al. (2016), introduced sparse sampling to model long-term temporal dependencies, while Pseudo-3D Residual Networks by Qiu et al. (2017) enhanced computational efficiency by decoupling spatial and temporal operations. More recently, transformer-based architectures, such as UniFormer by Li et al. (2022) and Multi-Entity Video Transformers by Walmer et al. (2023), have redefined video understanding by leveraging local context aggregation and global attention mechanisms. These advancements effectively balance spatiotemporal redundancy and dependency modeling, achieving remarkable improvements in both accuracy and efficiency for video analysis tasks.

Task-specific innovations have further advanced the capabilities of deep learning models in video analysis and understanding, particularly in areas like video classification, behavior recognition, and object detection. In video classification, techniques such as temporal attention mechanisms by Yang et al. (2020) and hierarchical spatio-temporal transformers by Cai and Cai (2020) have improved performance by prioritizing discriminative frames and actions. Multi-stream models by Kang et al. (2023), which integrate RGB and motion pathways, have also proven effective in earlier frameworks for robust feature extraction. In behavior recognition, time-aligned self-supervised methods, such as those employing local alignment contrastive loss by Oei et al. (2024) and Hadji et al. (2021), have reduced dependency on labeled data while enhancing the recognition of fine-grained actions. Additionally, few-shot action recognition approaches by Zhang et al. (2020) have excelled in low-data scenarios by utilizing spatio-temporal attention. In the realm of object detection and tracking within videos, the Deltaframe method by Han and Roy (2018) optimizes real-time efficiency by processing frame differences, significantly reducing computational overhead. Together, these advancements highlight the rapid evolution and increasing sophistication of deep learning techniques, making them essential for deepening our understanding of video content in diverse applications. Table 20 can be used to summarize key information about the methods mentioned in this section.

6.2 Video generation and editing

Video generation and editing represented a pivotal component within the broader domains of computer vision and artificial intelligence, exhibiting substantial advancements in recent years. Early research in the field of video generation and editing primarily relied on techniques such as GANs, which addressed the challenges of temporal coherence and visual quality in video generation. A notable milestone in this field was achieved with the introduction of MoCoGAN by Tulyakov et al. (2018) The proposed method involved the decomposition of motion and content, leading to the formulation of a framework for video generation that utilised two sub-networks: one responsible for the processing of content information and the other for motion information. These sub-networks were then integrated to generate the video frames. This approach was shown to enhance temporal coherence in generated videos, thus providing a significant foundation for subsequent research in this field.

Wang et al. (2018) proposed that Video-to-Video Synthesis represented a significant expansion of the field, offering a method for mapping from an input source video (e.g. a semantic segmentation mask sequence) to an output realistic video. This method employed

Table 20 Deep learning for video analysis and understanding

References	Methods	Applications	Key words
Tran et al. (2015)	3D CNN	Video classification	Spatiotemporal feature fusion for dynamic visual information
Feichtenhofer et al. (2016)	Two-stream networks	Video classification	Integration of spatial and temporal streams
Wang et al. (2016)	Temporal Segment Networks	Video classification	Sparse sampling for long-term temporal dependencies
Qiu et al. (2017)	Pseudo-3D Residual Networks	Video classification	Decoupling spatial and temporal operations for efficiency
Li et al. (2022)	UniFormer (Transformer)	Video understanding	Local context aggregation and global attention mechanisms
Walmer et al. (2023)	Multi-Entity Video Transformers	Video understanding	Balancing spatiotemporal redundancy and dependency
Yang et al. (2020)	Temporal attention mechanisms	Video classification	Prioritizing discriminative frames and actions
Cai and Cai (2020)	Hierarchical spatio-temporal transformers	Video classification	Enhanced performance through hierarchical modeling
Kang et al. (2023)	Multi-stream models	Video classification	Integration of RGB and motion pathways for robust feature extraction
Oei et al. (2024), Hadji et al. (2021)	Time-aligned self-supervised methods	Behavior recognition	Local alignment contrastive loss, reduced labeled data dependency
Zhang et al. (2020)	Spatio-temporal attention	Few-shot action recognition	Effective in low-data scenarios
Han and Roy (2018)	Deltaframe method	Object detection and tracking	Real-time efficiency through frame difference processing

a temporal-spatial adversarial learning framework, incorporating a meticulously designed generator and discriminator architecture to ensure temporal and spatial consistency in the generated video. The method was particularly well-suited to the task of mapping from sketches or poses to real videos. In contrast, the approach proposed by Kim et al. (2018) focused on the re-animation of portrait videos, generating realistic videos of a target actor

by transferring complete 3D head position, rotation, facial expression, eye gaze, and blink information from the source actor. The core of this method was a generative neural network that combined the synthetic rendering of a parametric face model to achieve high-fidelity video generation.

Siarohin et al. (2019) presented an alternative innovative approach that facilitated the animation of complex movements through self-supervised decoupling of appearance and motion information using a set of learned keypoints and their local affine transformations. This approach was notable for its independence from object-specific annotations, a characteristic that facilitated its generalisability across diverse categories, such as faces or the human body. In a related study, Chan et al. (2019) demonstrated the capacity to transfer dance movements from source videos to target individuals, thereby generating realistic dance videos through the utilisation of pose estimation and image synthesis techniques. These techniques were particularly well-suited for entertainment and virtual reality applications. The seminal contributions of these early works laid the technical foundation for video generation and editing, addressing the challenges of temporal inconsistency and suboptimal visual quality in early video generation.

In recent years, advancements in generative models, such as diffusion models and transformers, further refined the capabilities of video generation and editing, particularly in terms of quality, temporal coherence, and controllability. A seminal contribution in this area was made by Bar-Tal et al. (2024), who proposed Lumiere, a spatiotemporal diffusion model that generated videos directly in the spatiotemporal domain. This model significantly improved the coherence and visual fidelity of the generated videos by combining spatial and temporal information. This approach was shown to excel in generating extended videos and complex scenes, thus providing novel concepts for high-dynamic video generation. Hu (2024) focused on consistent and controllable image-to-video synthesis, especially for character animation applications. The proposed method ensured consistency in terms of the appearance and movements of the characters by enhancing the control over the generated content, and was suitable for animation production and virtual character generation. Qing et al. (2024) proposed a hierarchical spatio-temporal decoupling method, HiGen, for text-to-video generation, which effectively addressed the challenge of generating complex videos from text descriptions by decomposing the generation process into manageable parts. The method demonstrated particular proficiency in processing extensive text inputs and generating a wide variety of videos.

In contrast, Zeng et al. (2024) directed their attention towards the domain of high-dynamic video generation, with the objective of generating videos characterised by intricate motion and substantial dynamic range. The generation of more realistic motion effects was achieved by optimising the training strategy of the diffusion model. Lin et al. (2024) provided an open-source large-scale video generation model, Open-Sora Plan, with the aim of providing the research community with powerful tools to support a variety of generation tasks, such as text-to-video and image-to-video. The project was based on the latest diffusion model and large language model technology, emphasised openness and accessibility, and promoted further innovation in this field. Table 21 can be used to summarize key information about the methods mentioned in this section.

Table 21 Deep learning for video generation and editing

References	Methods	Applications	Key words
Tulyakov et al. (2018)	MoCoGAN	Video generation	Decomposition of motion and content, temporal coherence
Wang et al. (2018)	Video-to-Video Synthesis	Video synthesis	Temporal-spatial adversarial learning, mapping sketches to videos
Kim et al. (2018)	Generative neural network	Portrait video re-animation	Transfer of 3D head position and facial dynamics
Siarohin et al. (2019)	First Order Motion Model	Image animation	Self-supervised decoupling, keypoint-based motion transfer
Chan et al. (2019)	Pose estimation and image synthesis	Dance video generation	Transfer dance movements, entertainment applications
Bar-Tal et al. (2024)	Lumiere (Spatio-temporal Diffusion Model)	Video generation	Spatiotemporal domain generation, high coherence
Hu (2024)	Animate Anyone	Image-to-video synthesis	Consistent and controllable character animation
Qing et al. (2024)	HiGen	Text-to-video generation	Hierarchical spatio-temporal decoupling, complex video generation
Zeng et al. (2024)	Diffusion model optimization	High-dynamic video generation	Realistic motion effects, intricate motion handling
Lin et al. (2024)	Open-Sora Plan	Video generation	Open-source, supports text-to-video and image-to-video tasks

6.3 Video enhancement and repair

The initial advancements in the domains of video enhancement and restoration were predominantly reliant on conventional image processing methodologies. These methods established the foundation for subsequent research in the field. Notable among these classical methods are histogram equalisation and its adaptive variants, including contrast-limited adaptive histogram equalisation (CLAHE) (Reza 2004). These techniques have been extensively applied in the context of video frame processing, with the objective of enhancing image contrast. The Retinex theory (Land and McCann 1971), which models human colour perception, has inspired illumination normalisation and colour correction algorithms such

as the multi-scale Retinex of Jobson et al. (1997), which improves visual quality in low-light conditions by decomposing an image into reflectance and illumination components. For video-specific challenges, spatiotemporal filters such as the non-local mean algorithm of Buades et al. (2005) are used to maintain interframe consistency, especially in denoising tasks. Despite the computational efficiency of these conventional approaches, their efficacy is constrained in complex scenarios, such as low-light conditions or high dynamic range.

The advent of deep learning witnessed the emergence of pioneering methods that incorporated learning frameworks, thereby inaugurating novel avenues for video enhancement and restoration. A notable example is the work of Kappeler et al. (2016) utilised CNNs for the purpose of video super-resolution, marking the first instance of deep learning being applied to the field of video enhancement. Zhang et al. (2017) proposed DnCNN, which provided a residual learning framework for image denoising. Although these initial applications were focused on images, the concept was later expanded to encompass video processing by incorporating temporal information to process frame sequences. Despite the evident superiority of these pioneering deep learning methods in terms of performance when compared to traditional approaches, challenges persist in terms of temporal consistency and computational efficiency when processing extended videos or dynamic scenes.

In recent years, the development of deep learning technology has had a significant impact on the field of video enhancement and restoration, particularly in the areas of low-light conditions, temporal consistency and restoration tasks. Tu et al.'s MAXIM (Tu et al. 2022) employd a multi-axis MLP architecture to capture long-distance dependencies, rendering it suitable for low-level visual tasks such as enhancement and denoising. Guo et al. (2020) introduced the Zero-DCE method, which uses a zero-sample learning method to achieve low-light image enhancement through deep curve estimation and is extended to the video domain. In addition, for video restoration, Deep Video Inpainting by Kim et al. (2019) used a deep network architecture to synthesise unknown regions by collecting and refining information from neighbouring frames to ensure spatial and temporal consistency. The efficacy of these methods in processing complex scenes, such as fast-moving objects or low-light conditions, is well-documented. However, challenges remain with regard to the efficiency of video processing and the diversity of generated content. Table 22 can be used to summarize key information about the methods mentioned in this section.

7 Deep learning for natural language processing

Deep learning has revolutionized natural language processing (NLP) by enabling machines to understand, generate, and interact with human language in unprecedented ways. This section explores the key advancements in text analysis, generation, and cross-modal applications that have driven this transformation.

7.1 Text representation and structured analysis

In the field of natural language processing (NLP), text representation and structured analysis are core tasks, involving the conversion of text into a vector representation that is comprehensible to machines, as well as the parsing of the grammatical and semantic structure of

Table 22 Deep learning for video enhancement and repair

References	Methods	Applications	Key words
Reza (2004)	CLAHE	Video frame enhancement	Contrast-limited adaptive histogram equalisation
Land and McCann (1971)	Retinex theory	Illumination normalisation	Models human colour perception, inspires colour correction
Jobson et al. (1997)	Multi-scale Retinex	Low-light enhancement	Decomposes image into reflectance and illumination
Buades et al. (2005)	Non-local mean algorithm	Video denoising	Spatiotemporal filters, maintains interframe consistency
Kappeler et al. (2016)	CNN	Video super-resolution	First application of deep learning in video enhancement
Zhang et al. (2017)	DnCNN	Image and video denoising	Residual learning framework, extends to video processing
Tu et al. (2022)	MAXIM	Video enhancement and denoising	Multi-axis MLP, captures long-distance dependencies
Guo et al. (2020)	Zero-DCE	Low-light video enhancement	Zero-sample learning, deep curve estimation
Kim et al. (2019)	Deep Video Inpainting	Video restoration	Synthesises unknown regions, ensures spatial-temporal consistency

sentences. These technologies play a key role in machine translation, dialogue systems and information extraction.

Before the popularity of deep learning, NLP mainly relied on traditional methods. Text representations often use one-hot encoding (Rodríguez et al. 2018), bag-of-words (Zhang et al. 2010) or TF-IDF (Aizawa 2003). These methods are simple but cannot capture the semantic relationships between words. For example, bag-of-words ignores word order and context, and cannot distinguish between the different meanings of 'bank' in finance or in a river. Structured analysis tasks such as syntactic parsing and named entity recognition (NER) rely on rule systems or statistical models such as hidden Markov models (HMMs) (Eddy 1996) and conditional random fields (CRFs) (Sutton et al. 2012). These methods

require a lot of feature engineering and have limited generalisation capabilities, especially across languages or in complex domains.

The introduction of deep learning marked a major turning point in NLP, starting with the development of distributed word embeddings. Mikolov et al. (2013) proposed Word-2Vec, which learns word embeddings via skip-gram and CBOW models, capturing semantic similarities such that 'king' and 'queen' are close in the vector space. GloVe by Pennington et al. (2014) generates embeddings using global word co-occurrence statistics, emphasising semantic relationships with linear substructures. These static embeddings provide a basis for subsequent tasks. Subsequently, contextualised embedding models emerged. CNNs and RNNs are also widely used in NLP tasks. Chen (2015) demonstrated the effectiveness of CNNs in sentence classification, extracting local features via convolution and pooling operations. Huang et al. (2015) combined bidirectional LSTM and CRF for sequence labelling tasks such as NER, significantly improving performance.

The emergence of the Transformer model has further promoted the development of this field. Devlin et al. (2019) proposed Bidirectional Encoder Representations from Transformers (BERT), which is pre-trained on large corpora using a masked language model and a next sentence prediction task to generate contextualized representations. Liu et al. (2019) proposed RoBERTa, which optimizes the pre-training process of BERT by adjusting the training data and hyperparameter to further improve performance. These models have excelled in a variety of NLP tasks, including text classification, question answering, and structured prediction.

Transformer models have shown strong capabilities in structured analysis tasks. Kitaev and Klein (2018) used self-attention encoders for composition parsing and demonstrated that Transformers can effectively model the hierarchical structure of language. Tan et al. (2018) applied self-attention to semantic role labeling (SRL) and improved the ability to extract semantic relationships in complex sentences. In addition, recent studies such as Graph-to-Graph Transformer proposed by Mohammadshahi and Henderson (2019) have further explored graph-based parsing methods, especially in dependency parsing of transition systems. Table 23 can be used to summarize key information about the methods mentioned in this section.

7.2 Text generation and interactive applications

In the field of NLP, text generation and interactive applications are core tasks. They involve converting language into a form that machines can understand and generating naturally flowing text to support applications such as dialogue systems, chatbots and information generation. These technologies play a key role in enhancing the human–machine interaction experience and automating content creation.

Before the introduction of deep learning, text generation mainly relied on traditional methods such as rule-based template generation and statistical language models. For example, n-gram models (Brown et al. 1992) generate text by statistically probabilistic word sequences, but fail to capture long-distance dependencies and semantic relationships. Interactive applications such as early dialogue systems relied on finite state machines or scripted responses (Lee and Yannakakis 1996), lacking flexibility and contextual understanding. These methods are simple to implement, but the generated results are often stiff and difficult to adapt to complex scenarios or cross-language needs.

Table 23 Deep learning for natural language processing

References	Methods	Applications	Key words
Rodríguez et al. (2018)	One-hot encoding	Text representation	Simple encoding, no semantic relationships
Zhang et al. (2010)	Bag-of-words	Text representation	Ignores word order and context
Aizawa (2003)	TF-IDF	Text representation	Term frequency-inverse document frequency
Eddy (1996)	Hidden Markov Models (HMMs)	Syntactic parsing, NER	Statistical model, feature engineering
Sutton et al. (2012)	Conditional Random Fields (CRFs)	Structured analysis	Sequence modeling, limited generalization
Mikolov et al. (2013)	Word2Vec	Distributed word embeddings	Skip-gram, CBOW, semantic similarity
Pennington et al. (2014)	GloVe	Word embeddings	Global co-occurrence, linear substructures
Chen (2015)	CNN	Sentence classification	Convolution, pooling, local feature extraction
Huang et al. (2015)	Bidirectional LSTM + CRF	Sequence labeling (e.g., NER)	Contextual features, improved performance
Devlin et al. (2019)	BERT	Contextual embeddings	Masked language model, next sentence prediction
Liu et al. (2019)	RoBERTa	Text classification, QA	Optimized BERT, enhanced pre-training
Kitaev and Klein (2018)	Self-attention (Transformer)	Constituency parsing	Hierarchical structure modeling
Tan et al. (2018)	Self-attention (Transformer)	Semantic role labeling (SRL)	Semantic relationship extraction
Mohammadshahi and Henderson (2019)	Graph-to-Graph Transformer	Dependency parsing	Graph-based parsing, transition systems

The introduction of deep learning marked a major turning point in text generation and interactive applications, and sequence-to-sequence models began to emerge. Bahdanau et al. (2014) proposed an attention mechanism method that improves the quality of generation by dynamically aligning input and output sequences, laying the foundation for subsequent research. Vaswani (2017) proposed the Transformer architecture method, which relies on self-attention mechanisms to generate high-quality text, significantly accelerates train-

ing, and promotes the development of large-scale language models. Radford et al. (2019) proposed the GPT-2 unsupervised multi-task learning method, which generates natural text using large-scale pretraining and is suitable for a variety of interactive scenarios. Subsequently, Brown et al. (2020) proposed the GPT-3 few-shot learning method, which further demonstrated the potential of ultra-large-scale models in dialogue generation.

Deep learning models have demonstrated stronger capabilities in the fields of dialogue systems and interactive applications. Li et al. (2016) proposed a deep reinforcement learning method that enhances the interactive experience by optimizing dialogue consistency and informativeness through a reward mechanism. Zhang et al. (2019) proposed the DIALOGPT dialogue response generation method, which generates fluent dialogue text based on the GPT architecture, which has promoted the development of open-domain chatbots. Bao et al. (2020) proposed the PLATO-2 curriculum learning method, which improves the quality of open-domain dialogues through step-by-step training. Faltings et al. (2023) proposed an interactive text generation method that uses user feedback to enhance model adaptability, which is particularly suitable for real-time interactions. These techniques have together promoted the evolution of text generation from static output to dynamic interaction, although the issues of generation consistency and ethics still need to be further explored. Table 24 can be used to summarize key information about the methods mentioned in this section.

Table 24 Deep Learning for text generation and interactive applications

References	Methods	Applications	Key words
Brown et al. (1992)	N-gram models	Text generation	Probabilistic word sequences, no long-distance dependencies
Lee and Yannakakis (1996)	Finite state machines	Early dialogue systems	Scripted responses, limited flexibility
Bahdanau et al. (2014)	Attention mechanism	Sequence-to-sequence generation	Dynamic alignment, improved generation quality
Vaswani (2017)	Transformer	High-quality text generation	Self-attention, large-scale language models
Radford et al. (2019)	GPT-2	Unsupervised text generation	Multi-task learning, natural text output
Brown et al. (2020)	GPT-3	Dialogue generation	Few-shot learning, ultra-large-scale models
Li et al. (2016)	Deep reinforcement learning	Dialogue systems	Reward mechanism, consistency, informativeness
Zhang et al. (2019)	DIALOGPT	Dialogue response generation	GPT-based, fluent open-domain chatbots
Bao et al. (2020)	PLATO-2	Open-domain dialogue	Curriculum learning, step-by-step training
Faltings et al. (2023)	Interactive text generation	Real-time interaction	User feedback, model adaptability

7.3 Cross-modal integration and advanced scenarios

Before the popularization of deep learning, cross-modal integration usually relied on manually designed features and independent processing pipelines. Fusion occurred at the decision level, making it difficult to capture the complex interactions between modalities. Frome et al. (2013) proposed the DeViSE model, which learns a shared embedding space for images and text and supports zero-shot image classification. This work laid the foundation for cross-modal representation learning. Karpathy and Fei-Fei proposed (Karpathy and Fei-Fei 2015) a method to align image regions with words in sentences, facilitating image caption generation and establishing a correspondence between visual and textual data. Vinyals et al. (2015) developed the 'Show and Tell' model, which combines CNNs and RNNs to generate natural language descriptions of images, demonstrating the potential of end-to-end learning. Antol et al. (2015) introduced the visual question answering (VQA) task, which requires models to answer image-related questions by integrating visual and textual information, driving the diversification of cross-modal tasks.

In recent years, the rise of large-scale pre-trained models has significantly improved the ability of cross-modal integration. Radford et al. (2021) proposed the CLIP model, which was trained with a large-scale image-text pair dataset and supported zero-shot classification and other tasks, demonstrating the versatility of cross-modal representation learning. Ramesh et al. (2021) introduced DALL E, a Transformer-based model that can generate images from text descriptions, demonstrating advanced cross-modal generation capabilities. Zhang et al. (2021) improved visual feature extraction through VinVL, improving the performance of benchmarks such as image captioning and VQA, emphasising the importance of visual representation. Lu et al. (2022) proposed the COTS dual-stream model for cross-modal retrieval, integrating visual and textual information to achieve state-of-the-art results, demonstrating the potential of the dual-stream architecture in retrieval tasks.

In advanced scenarios, cross-modal integration plays a key role in complex applications. In the field of medical imaging, Dunnmon et al. (2020) developed weakly supervised techniques for cross-modal data programming, which can significantly reduce development costs by using auxiliary modalities (such as text reports) to generate training labels, reducing the labeling workload. This is applicable to clinical tasks such as radiology and CT scans. In the field of autonomous driving, Huang et al. (2022) investigated multimodal sensor fusion techniques, emphasising the importance of integrating LiDAR, camera and radar data for accurate perception and safe navigation, and discussed current challenges and future directions. These applications demonstrate the potential of cross-modal integration in processing complex, real-world data. Table 25 can be used to summarize key information about the methods mentioned in this section.

8 Deep learning for 3D data processing

The application of deep learning in the field of 3D data processing is becoming more and more widespread, bringing revolutionary breakthroughs to computer vision and related technologies. This chapter will explore in depth how deep learning can help with the analysis and generation of 3D data, covering multiple key areas from object recognition to scene understanding and then to 3D model reconstruction.

Table 25 Deep learning for cross-modal integration and advanced scenarios

References	Methods	Applications	Key words
Frome et al. (2013)	DeViSE	Zero-shot image classification	Shared embedding space, image-text alignment
Karpathy and Fei-Fei (2015)	Region-word alignment	Image caption generation	Visual-textual correspondence
Vinyals et al. (2015)	Show and Tell (CNN + RNN)	Image description generation	End-to-end learning, natural language output
Antol et al. (2015)	VQA task	Visual question answering	Visual-textual integration, task diversification
Radford et al. (2021)	CLIP	Zero-shot classification	Large-scale pre-training, versatile representation
Ramesh et al. (2021)	DALL E	Text-to-image generation	Transformer-based, cross-modal generation
Zhang et al. (2021)	VinVL	Image captioning, VQA	Enhanced visual features, benchmark improvement
Lu et al. (2022)	COTS (dual-stream)	Cross-modal retrieval	Visual-textual integration, state-of-the-art
Dunnmon et al. (2020)	Weakly supervised learning	Medical imaging (radiology, CT)	Cross-modal data programming, reduced labeling
Huang et al. (2022)	Multimodal sensor fusion	Autonomous driving	LiDAR-camera-radar integration, safe navigation

8.1 3D object recognition and classification

Before the popularisation of deep learning, 3D object recognition mainly relied on hand-designed feature descriptors. Johnson and Hebert (1999) proposed the Spin Images method, which captures the local shape of a point cloud by creating 2D histograms, and is suitable for efficient object recognition in cluttered scenes. Chen and Bhanu (2007) introduced the Local Surface Patches method, which describes local geometry using surface normal and curvature, and is suitable for free-form object recognition in range images. These methods perform poorly on sparse point clouds or occluded/overlapping objects.

Qi et al. (2017a) proposed the PointNet method, which is the first deep learning architecture that directly processes unordered point clouds. It solves the problem of permutation invariance in classification and segmentation tasks and lays the foundation for deep learning of point clouds. In the same year, Qi et al. (2017b) further proposed PointNet++, which captures local structures at different scales through hierarchical feature learning, improving performance in complex scenes. Wang et al. (2019) developed Dynamic Graph CNN (DGCNN), which uses dynamic graphs to model interactions between points, enhancing the flexibility of feature extraction. Thomas et al. (2019) proposed KPConv, a deformable con-

volution operator that allows the network to adapt to local geometric changes, significantly improving the performance of point cloud classification and segmentation tasks.

Transformers have potential in point cloud processing, especially in capturing global context. Zhao et al. (2021) introduced Point Transformer, which applies a self-attention mechanism to capture long-range dependencies in point clouds and achieves significant improvements in semantic segmentation tasks. Yu et al. (2022) developed Point-BERT, which pretrains a point cloud Transformer with mask point modeling to enhance feature learning capabilities and is suitable for a variety of downstream tasks. Table 26 can be used to summarize key information about the methods mentioned in this section.

8.2 3D scene understanding and segmentation

Before the popularisation of deep learning, 3D scene understanding mainly relied on traditional methods such as clustering and region growing based on geometric features. With the development of deep learning, the introduction of large-scale datasets has become crucial. Armeni et al. (2016) proposed the S3DIS dataset, which contains point cloud data of six large indoor areas, with each point annotated with 13 semantic categories, providing a benchmark for indoor scene segmentation. Dai et al. (2017) launched ScanNet, which provides 3D reconstruction and semantic annotation of more than 1500 indoor scenes, covering a wide range of scene types. These datasets have promoted the development and evaluation of subsequent methods.

Qi et al. (2017a) proposed PointNet, which solves the permutation invariance problem in classification and segmentation tasks and lays the foundation for deep learning of point

Table 26 Deep learning for 3D object recognition and classification

References	Methods	Applications	Key words
Johnson and Hebert (1999)	Spin Images	Object recognition in cluttered scenes	2D histograms, local shape capture
Chen and Bhanu (2007)	Local Surface Patches	Free-form object recognition	Surface normal, curvature, range images
Qi et al. (2017a)	PointNet	Point cloud classification, segmentation	Permutation invariance, unordered point clouds
Qi et al. (2017b)	PointNet++	Complex scene recognition	Hierarchical feature learning, local structures
Wang et al. (2019)	Dynamic Graph CNN (DGCNN)	Feature extraction in point clouds	Dynamic graphs, point interactions
Thomas et al. (2019)	KPConv	Point cloud classification, segmentation	Deformable convolution, local geometry adaptation
Zhao et al. (2021)	Point Transformer	Semantic segmentation	Self-attention, long-range dependencies
Yu et al. (2022)	Point-BERT	Downstream point cloud tasks	Pretraining, mask point modeling, feature learning

clouds. PointNet++ (Qi et al. 2017b) improves performance in complex scenes by capturing local structures at different scales through hierarchical feature learning. Tchapmi et al. (2017) proposed SEGCloud, which combines 3D CNNs and CRFs to achieve semantic segmentation of point clouds, achieving an mIoU of 60.3% on the S3DIS dataset.

Thomas et al. (2019) proposed KPConv, a deformable convolution operator that allows the network to adapt to local geometric changes, significantly improving the performance of the point cloud segmentation task. Hu et al. (2020) proposed RandLA-Net, which efficiently processes large-scale point clouds through random sampling and local feature aggregation, and is particularly suitable for outdoor scenes such as Semantic3D (Hackel et al. 2017). Zhao et al. (2021) introduced Point Transformer, which applies a self-attention mechanism to capture long-distance dependencies in point clouds. Qian et al. (2022) re-examined PointNet++ through PointNeXt, integrated improved training strategies and model scaling techniques, setting a new top performance on ScanNet. Rozenberszki et al. (2024) proposed UnScene3D, exploring the application of unsupervised learning in 3D instance segmentation, and improved the performance of indoor scene segmentation through self-supervised pseudo-label generation, especially improving the AP score on ScanNet by 300%. Table 27 can be used to summarize key information about the methods mentioned in this section.

Table 27 Deep learning for 3D scene understanding and segmentation

References	Methods	Applications	Key words
Armeni et al. (2016)	S3DIS dataset	Indoor scene segmentation	Point cloud data, 13 semantic categories
Dai et al. (2017)	ScanNet dataset	3D reconstruction, semantic annotation	1500+ indoor scenes, diverse scene types
Qi et al. (2017a)	PointNet	Point cloud segmentation	Permutation invariance, foundational deep learning
Qi et al. (2017b)	PointNet++	Complex scene segmentation	Hierarchical feature learning, local structures
Tchapmi et al. (2017)	SEGCloud	Semantic segmentation	3D CNNs, CRFs, 60.3% mIoU on S3DIS
Thomas et al. (2019)	KPConv	Point cloud segmentation	Deformable convolution, local geometry adaptation
Hu et al. (2020)	RandLA-Net	Large-scale outdoor segmentation	Random sampling, local feature aggregation
Zhao et al. (2021)	Point Transformer	Semantic segmentation	Self-attention, long-distance dependencies
Qian et al. (2022)	PointNeXt	Scene segmentation	Improved PointNet++, model scaling, top ScanNet performance
Rozenberszki et al. (2024)	UnScene3D	3D instance segmentation	Unsupervised learning, pseudo-labels, 300% AP boost

8.3 3D reconstruction and generation

In the early days, 3D reconstruction mainly relied on traditional computer vision methods, covering techniques such as motion from structure, shape from shadow and surface reconstruction. Longuet-Higgins (1981) proposed an algorithm for reconstructing a scene from two projections, introducing the eight-point algorithm and laying the foundation for multi-view 3D reconstruction. Hoppe et al. (1992) developed a method for reconstructing surfaces from unorganized point clouds, which generates a simplified surface approximation of an unknown manifold by triangulation, and is widely used in the processing of scanned data. Seitz et al. (2006) compared various multi-view stereo algorithms, providing a quantitative evaluation framework and setting a benchmark for subsequent research. Kazhdan et al. (2006) proposed a Poisson surface reconstruction method, which formulates surface reconstruction as a spatial Poisson problem and is robust to noise. Experiments show that it reconstructs a more detailed surface than before on publicly available scanned data.

The application of deep learning techniques has significantly advanced the field of 3D reconstruction and generation. Choy et al. (2016) proposed 3D-R2N2, which uses a recurrent neural network to uniformly process single-view and multi-view 3D object reconstruction, and generates a voxel representation through a 3D convolutional network, achieving 80% IoU on the ShapeNet dataset (Chang et al. 2015). Mescheder et al. (2019) proposed the occupancy network, which learns 3D reconstruction in the function space, representing shape through an implicit function, supporting complex topologies. Park et al. (2019) developed DeepSDF, which represents shapes using a continuous signed distance function, suitable for high-resolution model generation. Mildenhall et al. (2021) proposed NeRF, which represents scenes as neural radiance fields, for high-quality view synthesis, significantly improving the detail and realism of novel view generation. In addition, Wu et al. (2016) proposed 3D GAN, which learns the probability latent space of object shapes through generative adversarial networks, opening up research on 3D generative models. Table 28 can be used to summarize key information about the methods mentioned in this section.

9 Experience results

9.1 Deep learning for image denoising

9.1.1 Deep learning for additive white noise image denoising

To evaluate the performance of the additive white noisy image denoising methods discussed in Sect. 4.1.1, we utilize the BSD68 (Schmidt and Roth 2014), Set12 (Dabov et al. 2007), CBSD68 (Martin et al. 2001), Kodak24 (Franzen 1999), and McMaster (Zhang et al. 2011) datasets to evaluate denoising effect of deep learning techniques. Most denoising methods employ Peak Signal-to-Noise Ratio (PSNR) (Hore and Ziou 2010) as a quantitative metric. Table 29 presents PSNR values of various networks at different noise levels in additive white noisy image denoising. For assessing the denoising capability of different networks on a single grayscale additive white noise image, we conduct test on Set12 dataset and its results displayed in Table 30. Additionally, Table 31 illustrates denoising performance of various methods applied to color additive white noisy images. To qualitatively analyze

Table 28 Deep learning for 3D reconstruction and generation

References	Methods	Applications	Key words
Longuet-Higgins (1981)	Eight-point algorithm	Multi-view 3D reconstruction	Scene reconstruction, two projections
Hoppe et al. (1992)	Surface reconstruction	Processing scanned point clouds	Triangulation, simplified surface approximation
Seitz et al. (2006)	Multi-view stereo comparison	Evaluation framework	Quantitative benchmark, stereo algorithms
Kazhdan et al. (2006)	Poisson surface reconstruction	Noisy data reconstruction	Spatial Poisson problem, detailed surfaces
Choy et al. (2016)	3D-R2N2	Single/multi-view reconstruction	RNN, voxel representation, 80% IoU on ShapeNet
Mescheder et al. (2019)	Occupancy Network	3D shape reconstruction	Implicit function, complex topologies
Park et al. (2019)	DeepSDF	High-resolution shape generation	Signed distance function, continuous representation
Mildenhall et al. (2021)	NeRF	View synthesis	Neural radiance fields, high-quality realism
Wu et al. (2016)	3D GAN	3D generative modeling	Generative adversarial networks, latent space

denoising effects, we provide visual comparisons via comparing clarities of observation areas of obtained visual figures from different methods as shown in Figs. 5, 6 and 7, where a clearer observation area indicates more effective denoising of its corresponding method.

9.1.2 Deep learning for image blind denoising

Blind denoising methods refer to obtained denoisers without any known priors. To evaluate their denoising performance, we conduct experiments via using state-of-the-art denoising methods, i.e., DnCNN (Zhang et al. 2017), FFDNet (Zhang et al. 2018), ADNet (Tian et al. 2020), SCNN (Pan et al. 2018) on public datasets containing BSD68 (Schmidt and Roth 2014) and Set12 (Dabov et al. 2007) datasets. Their denoising PSNR results are shown in Tables 32 and 33.

Table 29 PSNR(dB) of different methods on the BSD68 for different noise levels (i.e., 15, 25 and 50)

Denoising methods	15	25	50
BM3D (Danielyan et al. 2011)	31.07	28.57	25.62
MLP (Burger et al. 2012)	–	28.96	26.03
CSF (Schmidt and Roth 2014)	31.24	28.74	–
TNRD (Chen and Pock 2016)	31.42	28.92	25.97
ECNDNet (Tian et al. 2019)	31.71	29.22	26.23
RED (Mao et al. 2016)	–	–	26.35
DnCNN (Zhang et al. 2017)	31.72	29.23	26.23
DDRN (Wang et al. 2017)	31.68	29.18	26.21
MemNet (Tai et al. 2017)	–	–	26.35
MWCNN (Liu et al. 2018)	31.86	29.41	26.53
MPFE-CNN (Kadimesetty et al. 2018)	31.79	29.31	26.34
IRCNN (Zhang et al. 2017)	31.63	29.15	26.19
FFDNet (Zhang et al. 2018)	31.62	29.19	26.30
BRDNet (Tian et al. 2020)	31.79	29.29	26.36
ADNet (Tian et al. 2020)	31.74	29.25	26.29
DAGL (Mou et al. 2021)	31.93	29.46	26.51
AGP-Net (Jiang et al. 2023)	31.02	29.59	26.71
GAiA-Net (Jiang et al. 2023)	32.09	29.67	26.75
DRUNet (Fang et al. 2021)	31.91	29.48	26.59
EFF-Net (Freeman et al. 2018)	31.92	29.49	26.61
CTNet (Tian et al. 2024)	31.94	29.46	26.49

9.1.3 Deep learning for real image denoising

We selected public datasets, i.e., DND (Plotz and Roth 2017), SIDD (Abdelhamed et al. 2018), Nam (Nam et al. 2016) and cc (Luccioni and Viviano 2021) to verify denoising performance of deep learning techniques on real images. To test applicability of obtained denoisers, we add several traditional denoising methods as comparative denoising methods. As shown in Table 34, DRDN has achieved the best denoising performance on DND and SSID datasets. As shown in Table 35, the AGAN method has achieved the best performance in terms of compressed noisy image denoising. As shown in Table 36, SDNet and BRDNet methods achieve the best performance for real noisy images with different ISO values.

9.2 Deep learning for image super-resolution

9.2.1 Deep learning for supervised image super-resolution

Most of existing super-resolution methods are supervised super-resolution methods, which use paired LR-HR images to obtain super-resolution models. We use PSNR and SSIM (Setiadi 2021) as metrics to evaluate super-resolution performance of different methods. We select five commonly used super-resolution datasets such as Set5 (Bevilacqua et al. 2012), Set14 (Zeyde et al. 2012), BSD100 (Martin et al. 2001), Urban100 (Huang et al. 2015) and Manga109 (Matsui et al. 2017) for testing effects of recovering images from different image super-resolution methods. We test super-resolution performance of different deep learning methods for different scales. As increasement of magnification, super-resolution

Table 30 PSNR(dB) of different methods on the Set12 for different noise levels (i.e., 15, 25 and 50)

Images	C.man	House	Peppers	Starfish	Monarch	Airplane	Parrot	Barbara	Boat	Man	Couple	Average
Noise level 15												
BM3D (Danielyan et al. 2011)	31.91	34.93	32.69	31.14	31.85	31.07	31.37	33.10	32.13	31.92	32.10	32.37
WNNM (Gu et al. 2014)	32.17	35.13	32.99	31.82	32.71	31.39	31.62	33.60	32.27	32.11	32.17	32.70
CSF (Schmidt and Roth 2014)	31.95	34.39	32.85	31.55	32.33	31.33	31.37	31.92	32.01	32.08	31.98	32.32
TNRD (Chen and Pock 2016)	32.19	34.53	33.04	31.75	32.56	31.46	31.63	32.13	32.14	32.23	32.11	32.50
ECNDNet (Tian et al. 2019)	32.56	34.97	33.25	32.17	33.11	31.70	31.82	32.41	32.37	32.39	32.39	32.81
DnCNN (Zhang et al. 2017)	32.61	34.97	33.30	32.20	33.09	31.70	31.83	32.64	32.42	32.46	32.47	32.86
IRCNN (Zhang et al. 2017)	32.55	34.89	33.31	32.02	32.82	31.70	31.84	32.43	32.34	32.40	32.40	32.77
FFDNet (Zhang et al. 2018)	32.43	35.07	33.25	31.99	32.66	31.57	31.81	32.54	32.38	32.41	32.46	32.77
BRDNet (Tian et al. 2020)	32.80	35.27	33.47	32.24	33.35	31.85	32.00	32.93	32.55	32.50	32.62	33.03
ADNet (Tian et al. 2020)	32.81	35.22	33.49	32.17	33.17	31.86	31.96	32.80	32.57	32.47	32.58	32.98
CTNet (Tian et al. 2024)	32.82	35.86	33.69	32.65	33.53	32.07	32.21	33.87	32.75	32.61	32.77	33.31
Noise level 25												
BM3D (Danielyan et al. 2011)	29.45	32.85	30.16	28.56	29.25	28.42	28.93	30.71	29.90	29.61	29.71	29.97
WNNM (Gu et al. 2014)	29.64	33.22	30.42	29.03	29.84	28.69	29.15	31.24	30.03	29.76	29.82	30.26
MLP (Burger et al. 2012)	29.61	32.56	30.30	28.82	29.61	28.82	29.25	29.54	29.97	29.88	29.73	30.03
CSF (Schmidt and Roth 2014)	29.48	32.39	30.32	28.80	29.62	28.72	28.90	29.03	29.76	29.71	29.53	29.84
TNRD (Chen and Pock 2016)	29.72	32.53	30.57	29.02	29.85	28.88	29.18	29.41	29.91	29.87	29.71	30.06
ECNDNet (Tian et al. 2019)	30.11	33.08	30.85	29.43	30.30	29.07	29.38	29.84	30.14	30.03	30.03	30.39
DnCNN (Zhang et al. 2017)	30.18	33.06	30.87	29.41	30.28	29.13	29.43	30.00	30.21	30.10	30.12	30.43
IRCNN (Zhang et al. 2017)	30.08	33.06	30.88	29.27	30.09	29.12	29.47	29.92	30.17	30.04	30.08	30.38
FFDNet (Zhang et al. 2018)	30.10	33.28	30.93	29.32	30.08	29.04	29.44	30.01	30.25	30.11	30.20	30.44
BRDNet (Tian et al. 2020)	31.39	33.41	31.04	29.46	30.50	29.20	29.55	30.34	30.33	30.14	30.28	30.61
ADNet (Tian et al. 2020)	30.34	33.41	31.14	29.41	30.39	29.17	29.49	30.25	30.37	30.08	30.24	30.58
CTNet (Tian et al. 2024)	30.40	33.86	31.33	30.03	30.68	29.50	29.73	31.62	30.54	30.27	30.49	30.94
Noise level 50												
BM3D (Danielyan et al. 2011)	26.13	29.69	26.68	25.04	25.82	25.10	25.90	27.22	26.78	26.81	26.46	26.72
WNNM (Gu et al. 2014)	26.45	30.33	26.95	25.44	26.32	25.42	26.14	27.79	26.97	26.94	26.64	27.05
MLP (Burger et al. 2012)	26.37	29.64	26.68	25.43	26.26	25.56	26.12	25.24	27.03	27.06	26.67	26.78

Table 30 (continued)

Images	C.man	House	Peppers	Starfish	Monarch	Airplane	Parrot	Barbara	Boat	Man	Couple	Average
TNRD (Chen and Pock 2016)	26.62	29.48	27.10	25.42	26.31	25.59	26.16	25.70	26.94	26.98	26.50	26.81
ECNDNet (Tian et al. 2019)	27.07	30.12	27.30	25.72	26.82	25.79	26.32	26.26	27.16	27.11	26.84	27.15
DnCNN (Zhang et al. 2017)	27.03	30.00	27.32	25.70	26.78	25.87	26.48	26.22	27.20	27.24	26.90	27.18
IRCNN (Zhang et al. 2017)	26.88	29.96	27.33	25.57	26.61	25.89	26.55	26.24	27.17	27.17	26.88	27.14
FFDNet (Zhang et al. 2018)	27.05	30.37	27.54	25.75	26.81	25.89	26.57	26.45	27.33	27.29	27.08	27.32
BRDNet (Tian et al. 2020)	27.44	30.53	27.67	25.77	26.97	25.93	26.66	26.85	27.38	27.27	27.17	27.45
ADNet (Tian et al. 2020)	27.31	30.59	27.69	25.70	26.90	25.88	26.56	26.64	27.35	27.17	27.07	27.37
CTNet(Tian et al. 2024)	27.47	30.98	27.92	26.45	27.14	26.28	26.70	28.29	27.52	27.41	27.37	27.79

Table 31 PSNR(dB) of different on the CBSD68, Kodak24 and McMaster for different noise levels (i.e., 15, 25, 35, 50 and 75)

Datasets	Methods	15	25	35	50	75
CBSD68	CBM3D (Dabov et al. 2007)	33.52	30.71	28.89	27.38	25.74
	DnCNN (Zhang et al. 2017)	33.98	31.31	29.65	28.01	–
	DDRN (Wang et al. 2017)	33.93	31.24	–	27.86	–
	BM3D-Net (Yang and Sun 2017)	33.79	30.79	–	27.48	–
	IRCNN (Zhang et al. 2017)	33.86	31.16	29.50	27.86	–
	FFDNet (Zhang et al. 2018)	33.80	31.18	29.57	27.96	26.24
	BRDNet (Tian et al. 2020)	34.10	31.43	29.77	28.16	26.43
	ADNet (Oktay et al. 2018)	33.99	31.31	29.66	28.04	26.33
Kodak24	CBM3D (Dabov et al. 2007)	34.28	31.68	29.90	28.46	26.82
	DnCNN (Zhang et al. 2017)	34.73	32.23	30.64	29.02	–
	IRCNN (Zhang et al. 2017)	34.56	32.03	30.43	28.81	–
	FFDNet (Zhang et al. 2018)	34.55	32.11	30.56	28.99	27.25
	BRDNet (Tian et al. 2020)	34.88	32.41	30.80	29.22	27.49
	ADNet (Oktay et al. 2018)	34.76	32.26	30.68	29.10	27.40
McMaster	CBM3D (Dabov et al. 2007)	34.06	31.66	29.92	28.51	26.79
	DnCNN (Zhang et al. 2017)	34.80	–	30.91	29.21	–
	IRCNN (Zhang et al. 2017)	34.58	32.18	30.59	28.91	–
	FFDNet (Zhang et al. 2018)	34.47	32.25	30.76	29.14	27.29
	BRDNet (Tian et al. 2020)	35.08	32.75	31.15	29.52	27.72
	ADNet (Oktay et al. 2018)	34.93	32.56	31.00	29.36	27.53

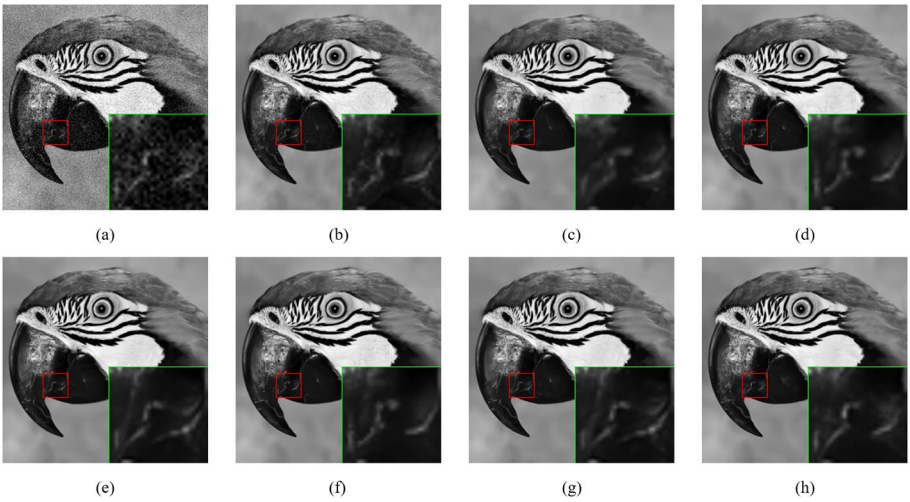


Fig. 5 Denoising results of different methods on one image from Set12 when noise level is 15. **a** Original image **b** Noisy image **c** DnCNN **d** FFDNet **e** ADNet **f** BRDNet **g** ECNDNet **h** RDDCNN

performance of different methods may decrease. Their comparison results are shown in Tables 37, 38 and 39.

To compare super-resolution performance of different deep learning methods, we test different super-resolution methods on B100 and Set5 datasets, respectively. As shown in Figs.

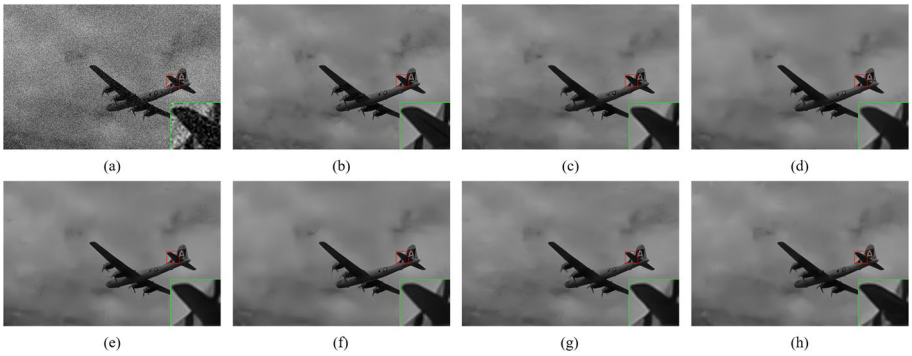


Fig. 6 Denoising results of different methods on one image from BSD68 when noise level 25. **a** Original image **b** Noisy image **c** DnCNN **d** FFDNet **e** ADNet **f** BRDNet **g** ECNDNet **h** RDDCNN

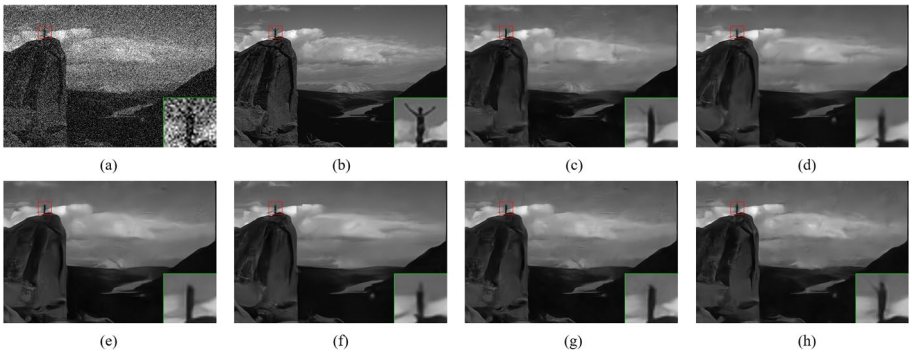


Fig. 7 Denoising results of different methods on one image from BSD68 when noise level is 50. **a** Original image **b** Noisy image **c** DnCNN **d** FFDNet **e** ADNet **f** BRDNet **g** ECNDNet **h** RDDCNN

Table 32 PSNR (dB) of different methods on the BSD68 for different noise levels, i.e., 15, 25 and 50

Methods	15	25	50
DnCNN-B (Zhang et al. 2017)	31.61	29.16	26.23
FFDNet (Zhang et al. 2018)	31.62	29.19	26.30
SCNN (Pan et al. 2018)	31.48	29.03	26.08
ADNet-B (Oktay et al. 2018)	–	29.00	25.95
DnCNN-SURE-T (Soltanayev and Chun 2018)	–	29.20	26.22
DnCNN-MSE-GT (Soltanayev and Chun 2018)	31.55	28.93	25.73

Table 33 Average PSNR(dB) result of different methods on Set12 with noise level of 25 and 50

Images	C.man	House	Peppers	Starfish	Monarch	Airplane	Parrot	Barbara	Boat	Man	Couple	Average
Noise level 25												
DnCNN-B (Zhang et al. 2017)	29.94	33.05	30.84	29.34	30.25	29.09	29.35	29.69	30.20	30.09	30.10	30.36
FFDNet (Zhang et al. 2018)	30.10	33.28	30.93	29.32	30.08	29.04	29.44	30.01	30.25	30.11	30.20	30.44
ADNet-B (Oktay et al. 2018)	29.94	33.38	30.99	29.22	30.38	29.16	29.41	30.05	30.28	30.01	30.15	30.46
DudeNet-B (Tian et al. 2021)	30.01	33.15	30.87	29.39	30.31	29.07	29.40	29.76	30.18	30.03	30.06	30.39
DnCNN-SURE-T (Soltanayev and Chun 2018)	29.86	32.73	30.57	29.11	30.13	28.93	29.26	29.44	29.86	29.91	29.78	30.14
DnCNN-MSE-GT (Soltanayev and Chun 2018)	30.14	33.16	30.84	29.40	30.45	29.11	29.36	29.91	30.11	30.08	30.06	30.42
Noise level 50												
DnCNN-B (Zhang et al. 2017)	27.03	30.02	27.39	25.72	26.83	25.89	26.48	26.38	27.23	27.23	26.91	27.21
FFDNet (Zhang et al. 2018)	27.05	30.37	27.54	25.75	26.81	25.89	26.57	26.45	27.33	27.29	27.08	27.32
ADNet-B (Oktay et al. 2018)	27.22	30.43	27.70	25.63	26.92	26.03	26.56	26.51	27.22	27.19	27.05	27.33
DudeNet-B (Tian et al. 2021)	27.19	30.11	27.50	25.69	26.82	25.85	26.46	26.38	27.20	27.13	26.90	27.22
DnCNN-SURE-T (Soltanayev and Chun 2018)	26.47	29.20	26.78	25.39	26.53	25.65	26.21	25.23	26.79	26.97	26.48	26.71
DnCNN-MSE-GT (Soltanayev and Chun 2018)	27.03	29.92	27.27	25.65	26.95	25.93	26.43	26.17	27.12	27.22	26.94	27.16

8 and 9, the observed area is clearer, the corresponding method has better super-resolution performance.

9.2.2 Deep learning for unsupervised image super-resolution

Unsupervised image denoising methods do not require clean images as references, which are suitable for scenarios. Specifically, labeled data is scarce, i.e., low-light image enhancement, old photo restoration, etc. We tested some unsupervised image denoising methods on NTIRE 2018 Track 2 dataset (Timofte et al. 2017), and their results are shown in Table 40.

9.3 Deep learning for image deblurring

To verify deblurring performance of deep learning methods, we conducted some experiments via using different networks on the Set5, Set14, BSD100 and Urban100 datasets in terms of quantitative and qualitative evaluations. PSNR and SSIM are used to as evaluation metrics to test performance of image deblurring from various deep learning methods, i.e., SRCNN, FSRCNN, VDSR, RDN, D-DBPN, EDSR and GCEDSR. Detailed test results of these methods are presented in Table 41. Visual compressions of typical deblurring methods can be shown in Fig. 10.

9.4 Deep learning for image classification

9.4.1 Deep learning for medical image classification

The challenges of medical image classification come from: (1) the complexity of image data: the quality of images generated by medical imaging modalities has been improved, but the image information captured by medical imaging modalities is not complete and clear. (2) Complexity of imaging model: medical imaging involves different applications on different structures and features of the human body (Miranda et al. 2016). Thus, medical imaging needs to interpret objects based on prior knowledge and identify different tissue structures. Due to the wide range of medical images, we selected the ISIC2018 (Milton 2019) and Covid19 (Yang et al. 2020) datasets to test performance of medical image classification method based deep learning techniques, obtained results of deep networks are shown in Tables 42 and 43.

9.4.2 Deep learning for face image classification

In human social interaction, facial expressions are key non-verbal signals that convey emotions, intentions, and mental states. With the rapid development of artificial intelligence technology, facial expression recognition has become a hot research direction in the field of computer vision, which aims to automatically recognize and understand human facial expressions by analyzing and processing facial images. In the field of facial expression recognition, methods based on convolutional neural networks have significantly improved the accuracy and robustness of recognition by learning complex patterns of facial expressions from a large amount of training data. Table 44 shows test accuracy of deep learning methods

Table 34 PSNR(dB) of different methods on the DND and SIDD for real noise image denoising

Methods	DND	Methods	SIDD
EPLL (Zoran and Weiss 2011)	33.51	CBM3D (Dabov et al. 2007)	25.65
TNRD (Chen and Pock 2016)	33.65	WNNM (Gu et al. 2014)	25.78
NCSR (Dong et al. 2012)	34.05	MLP (Burger et al. 2012)	24.71
MLP (Burger et al. 2012)	34.23	DnCNN-B (Zhang et al. 2017)	23.66
BM3D (Dabov et al. 2007)	34.51	CBDNet (Guo et al. 2019)	33.28
FoE (Roth and Black 2005)	34.62	VDN (Yue et al. 2019)	39.23
WNNM (Gu et al. 2014)	34.67	DRDN (Song et al. 2019)	39.6
KSVD (Aharon et al. 2006)	36.49		
CDnCNN-B (Zhang et al. 2017)	32.43		
FFDNet (Zhang et al. 2018)	34.4		
MCWNNM (Liu et al. 2018)	37.38		
TWSC (Xu et al. 2018)	37.94		
GCBD (Chen et al. 2018)	35.58		
CIMM (Anwar et al. 2017)	36.04		
CBDNet (Guo et al. 2019)	37.72		
VDN (Yue et al. 2019)	39.38		
DRDN (Song et al. 2019)	39.4		
AGAN (Lin et al. 2019)	38.13		

Table 35 PSNR(dB) of different methods on the Nam for real noise image denoising

Methods	TWSC (Xu et al. 2018)	BM3D (Dabov et al. 2007)	NC (Lebrun et al. 2015)	WNNM (Gu et al. 2014)	CDnCNN-B (Zhang et al. 2017)	MC-WNNM (Liu et al. 2018)	CBDNet (Guo et al. 2019)	CBDNet (JPEG) (Guo et al. 2019)	DRDN (Song et al. 2019)
Nam	31.52	37.52	39.84	40.41	41.04	37.49	37.91	41.02	41.31

for facial expression recognition on different datasets (RAF-DB (Li et al. 2017), FER-2013 (Giannopoulos et al. 2018), and JAFFE (Shih et al. 2008)).

9.4.3 Deep learning for autonomous driving recognition

The performance of deep learning methods for autonomous driving image classification is summarized in Table 45, where we classify the models into lightweight and heavyweight models. In Table 45, we show information such as Top-1 accuracy, Top-5 accuracy, number of parameters, number of network layers, number of FLOPs and model size for each model. It can be observed from Table 45 that the heavyweight models achieve better accuracy levels mostly by increasing the layer and parameter numbers without regard for the model size or energy consumption. On the other hand, the lightweight CNNs, aim to reduce the model

Table 36 PSNR (dB) of different methods on the CC for real noising image denoising

Camera settings	CBM3D (Dabov et al. 2007)	MLP (Burger et al. 2012)	TNRD (Chen and Pock 2016)	DnCNN (Zhang et al. 2017)	NC (Lebrun et al. 2015)	WNNM (Gu et al. 2014)	BRDNet (Tian et al. 2020)	SDNet (Zhao et al. 2019)	ADNet (Tian et al. 2020)
Canon 5D ISO = 3200	39.76	39.00	39.51	37.26	38.76	37.51	37.63	39.83	35.96
	36.40	36.34	36.47	34.13	35.69	33.86	37.28	37.25	36.11
	36.37	36.33	36.45	34.09	35.54	31.43	37.75	36.79	34.49
Nikon D600 ISO = 3200	41.8	34.70	34.79	33.62	35.57	33.46	34.55	35.50	33.94
	35.07	36.20	36.37	34.48	36.70	36.09	35.99	37.24	34.33
	34.70	37.13	39.33	39.49	35.41	39.28	39.86	38.62	41.18
Nikon D800 ISO = 1600	36.81	37.95	38.11	35.79	38.01	6.35	9.22	38.77	37.61
	37.76	40.23	40.52	36.08	39.05	39.99	39.67	40.87	38.24
	39.15	37.51	37.94	38.17	35.48	38.20	37.15	39.04	38.86
Nikon D800 ISO = 3200	35.05	37.55	37.69	34.08	38.07	38.60	38.28	39.94	37.20
	36.93	34.07	35.91	35.90	33.70	35.72	36.04	37.18	36.78
	35.67	35.80	34.42	38.15	38.21	33.31	36.76	39.73	38.85
Nikon D800 ISO = 6400	31.13	32.69	2.81	29.83	33.49	33.29	32.75	33.34	32.24
	31.94	31.22	32.33	32.33	30.55	32.79	31.16	33.24	33.29
	32.59	32.51	30.97	32.29	32.29	30.09	32.86	31.98	32.89
Average	35.19	36.46	36.61	33.86	36.43	35.77	36.73	37.51	35.69

sizes by conceding their accuracy performance, i.e., SqueezeNet, MobileNets, ShiftNet, and SqueezeNext.

9.4.4 Deep learning for surface defect detection

To compare performance of different deep learning methods for surface detection, we conduct performance test on NEU-DET (He et al. 2019) and DAGM (Carvalho et al. 2022) datasets and calculate the accuracy, precision, recall, F1 score, IoU and mAP of typical deep learning methods as shown in Table 46. As can be seen from Table 46, AIS-Net has obtained the highest accuracy, precision, recall, F1 score, IoU, and mAP values on NEU-DET dataset, which indicate its best performance for surface detect detection. On DAGM dataset, MF-GAN has the highest accuracy, precision, recall, F1 score, IoU, and mAP value.

9.5 Deep learning for image segmentation

9.5.1 Deep learning for image semantic segmentation

Image segmentation is a crucial task in computer vision. Currently, several commonly used 2D image datasets include PASCAL Visual Object Classes (VOC) (Everingham et al. 2010), PASCAL Context (Mottaghi et al. 2014), Microsoft Common Objects in Context (MS COCO) (Lin et al. 2014), Cityscapes (Cordts et al. 2016), SiftFlow (Liu et al. 2009), and Stanford Background (Gould et al. 2009). Datasets are conducted by scanners, i.e., NYU-D V2 (Silberman et al. 2012), SUN-3D (Xiao et al. 2013), and SUN RGB-D (Song et al. 2015). Besides, 3D datasets contain Stanford 2D-3D (Xiang et al. 2016) and ShapeNet Core (Chang et al. 2015). To evaluate image segmentation performance of various deep learning methods, we employ the mean intersection over union (MIoU) as metric. Table 47 presents the results of image segmentation models tested on the PASCAL VOC dataset and Table 48 displays the results for models based deep learning tested on the Cityscapes dataset.

9.5.2 Deep learning for image instance segmentation

Instance segmentation not only has characteristic of semantic segmentation with pixel-level classification, but also has characteristics of object detection, which needs to locate different instances. To test the performance of one- and two-stage instance segmentation methods, we have conducted test on the COCO dataset (Lin et al. 2014) as shown in Table 49, where we can see that Mask R-CNN (He et al. 2017) has achieved the best accuracy from two-stage instance segmentation methods and TensorMask (Chen et al. 2019) method has achieved the best accuracy from one-stage instance segmentation methods.

9.5.3 Deep learning for image panoptic segmentation

Panorama segmentation is a fusion of instance segmentation and semantic segmentation, which aims to distinguish things in the scene. Differing from instance segmentation and semantic methods, each object is distinguished from others by assigning different colors for image panorama segmentation. Table 50 shows obtained results of several existing panoptic segmentation methods based deep learning techniques referred to metrics of AP and IoU

Table 37 PSNR(dB)/SSIM of different methods on ×2

Methods	Scale	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	×2	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN (Dong et al. 2015)	×2	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
VDSR (Kim et al. 2016b)	×2	37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140	37.22/0.9750
EDSR (Lim et al. 2017)	×2	38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351	39.10/0.9773
RCAN (Zhang et al. 2018)	×2	38.27/0.9614	34.11/0.9216	32.41/0.9026	33.34/0.9385	39.43/0.9786
NLRN (Liu et al. 2018)	×2	38.00/0.9603	33.46/0.9159	32.19/0.8992	31.81/0.9246	–/–
SRFBN (Li et al. 2019)	×2	38.11/0.9609	33.82/0.9196	32.29/0.9010	32.62/0.9328	39.08/0.9779
SAN (Dai et al. 2019)	×2	38.31/0.9620	34.07/0.9213	32.42/0.9028	33.10/0.9370	39.32/0.9792
RDN (Zhang et al. 2018)	×2	38.24/0.9614	34.01/0.9212	32.34/0.9017	32.89/0.9353	39.18/0.9780
USRNet (Zhang et al. 2020)	×2	37.77/0.9592	33.49/0.9156	32.10/0.8981	31.79/0.9255	38.37/0.9760
SRGAT (Yan et al. 2021)	×2	38.20/0.9610	33.93/0.9201	32.34/0.9014	32.90/0.9359	39.30/0.9785
SwinIR (Liang et al. 2021)	×2	38.35/0.9620	34.14/0.9215	32.44/0.9030	33.40/0.9393	39.60/0.9792
RGCN (Schlichtkrull et al. 2018)	×2	38.30/0.9616	34.10/0.9213	32.44/0.9030	33.15/0.9377	39.38/0.9784

metrics on different datasets including Cityscapes (Cordts et al. 2016), COCO (Lin et al. 2014), ADE20K (Zhou et al. 2019), Mapillary Vitas (Neuhold et al. 2017), ADE20K (Zhou et al. 2017), KITTI (Geiger et al. 2013), and Semantic KITTI (Behley et al. 2019).

9.6 Deep learning for object detection

9.6.1 Two-stage object detection algorithms

Table 51 compares the performance of several object detection algorithms: R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, and R-FCN. These algorithms vary in backbone networks, maximum frame rates (FPS), and accuracy across datasets, each suited for different applications. R-CNN, while foundational, lacks FPS data, suggesting slower processing. Fast R-CNN offers 3 FPS with 70.0% accuracy on VOC 2007, making it viable when speed is secondary. Faster R-CNN improves speed (7 FPS) and accuracy (73.2%), while Mask R-CNN achieves 11 FPS and 76.4% accuracy, suitable for real-time tasks. R-FCN demonstrates performance variance, achieving up to 80.5% accuracy with ResNet-101. Users should select algorithms based on specific needs, balancing processing speed and accuracy according to their application scenarios.

Table 38 PSNR(dB)/SSIM of different methods on $\times 3$

Methods	Scale	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	$\times 3$	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN (Dong et al. 2015)	$\times 3$	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
VDSR (Kim et al. 2016b)	$\times 3$	33.67/0.9210	29.78/0.8320	28.83/0.7990	27.14/0.8290	32.01/0.9340
EDSR (Lim et al. 2017)	$\times 3$	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	34.17/0.9476
RCAN (Zhang et al. 2018)	$\times 3$	34.74/0.9299	30.65/0.8482	29.32/0.8111	29.09/0.8702	34.44/0.9499
NLRN (Liu et al. 2018)	$\times 3$	34.27/0.9266	30.16/0.8374	29.06/0.8026	27.93/0.8453	–/–
SRFBN (Li et al. 2019)	$\times 3$	34.70/0.9292	30.51/0.8461	29.24/0.8084	28.73/0.8641	34.18/0.9481
SAN (Dai et al. 2019)	$\times 3$	34.75/0.9300	30.59/0.8476	29.33/0.8112	28.93/0.8671	34.30/0.9494
RDN (Zhang et al. 2018)	$\times 3$	34.71/0.9296	30.57/0.8468	29.26/0.8093	28.80/0.8653	34.13/0.9484
USRNet (Zhang et al. 2020)	$\times 3$	34.43/0.9279	30.51/0.8446	29.18/0.8076	28.38/0.8575	34.05/0.9466
SRGAT (Yan et al. 2021)	$\times 3$	34.75/0.9297	30.63/0.8474	29.29/0.8099	28.90/0.8666	34.42/0.9495
RGCN (Schlichtkrull et al. 2018)	$\times 3$	34.77/0.9301	30.67/0.8486	29.33/0.8114	28.99/0.8679	34.47/0.9501
ESRT (Lu et al. 2022)	$\times 3$	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574	33.95/0.9455
LBNet (Gao et al. 2022)	$\times 3$	34.47/0.9277	30.38/0.8417	29.13/0.8061	28.42/0.8559	33.82/0.9460
SwinIR (Liang et al. 2021)	$\times 3$	34.89/0.9312	30.77/0.8503	29.37/0.8124	29.29/0.8744	34.74/0.9518

9.6.2 One-stage object detection algorithms

Table 52 compares the performance of several one-stage object detectors, including YOLOv1-448, YOLOv2-416, SSD-300, DSSD-321, YOLOV-3416, YOLOv4, and YOLOv5. These detectors vary in maximum frame rate, average accuracy, and the datasets used, and have different advantages for different scenarios. For example, YOLOv1-448 is fast and small, but it has low accuracy and is not sensitive to small targets. YOLOv2-416 improves accuracy, but sacrifices some speed. SSD-300 can deal with irregular targets, but it is not good for small targets. DSSD-321 enhances the detection ability of small targets, but the speed is reduced. YOLOV-3416 improves accuracy by increasing network depth. YOLOv4 uses a series of strengthening measures to greatly improve the accuracy of object detection. YOLOv5 provides more flexibility, but performs slightly worse than YOLOv4. Besides, we give visual figures to test detection performance of different YOLO methods in Fig. 11. Therefore, the user can choose the appropriate detector, according to the specific application scenario and requirements.

Table 39 PSNR(dB)/SSIM of different methods on $\times 4$

Methods	Scale	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	$\times 4$	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN (Dong et al. 2015)	$\times 4$	30.48/0.8628	27.50/0.7513	26.90/0.7101	25.52/0.7221	27.58/0.8555
VDSR (Kim et al. 2016b)	$\times 4$	31.35/0.8830	28.02/0.7680	27.29/0.7260	25.18/0.7540	28.83/0.8870
EDSR (Lim et al. 2017)	$\times 4$	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	31.02/0.9148
RCAN (Zhang et al. 2018)	$\times 4$	32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087	31.22/0.9173
NLRN (Liu et al. 2018)	$\times 4$	31.92/0.8916	28.36/0.7745	27.48/0.7346	25.79/0.7729	–/–
SRFBN (Li et al. 2019)	$\times 4$	32.47/0.8983	28.81/0.7868	27.72/0.7409	26.60/0.8015	31.15/0.9160
SAN (Dai et al. 2019)	$\times 4$	32.64/0.9003	28.92/0.7888	27.78/0.7436	26.79/0.8068	31.18/0.9169
RDN (Zhang et al. 2018)	$\times 4$	32.47/0.8990	28.81/0.7871	27.72/0.7419	26.61/0.8028	31.00/0.9151
USRNet (Zhang et al. 2020)	$\times 4$	32.42/0.8978	28.83/0.7871	27.69/0.7404	26.44/0.7976	31.11/0.9154
SRGAT (Yan et al. 2021)	$\times 4$	32.57/0.8997	28.86/0.7879	27.77/0.7421	26.76/0.8052	31.41/0.9181
RGCN (Schlichtkrull et al. 2018)	$\times 4$	32.65/0.9005	28.91/0.7892	27.79/0.7440	26.85/0.8089	31.24/0.9176
ESRT (Lu et al. 2022)	$\times 4$	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100
LBNet (Gao et al. 2022)	$\times 4$	32.29/0.8960	28.68/0.7832	27.62/0.7382	26.27/0.7906	30.76/0.9111
SwinIR (Liang et al. 2021)	$\times 4$	32.72/0.9021	28.94/0.7914	27.83/0.7459	27.07/0.8164	31.67/0.9226

9.7 Deep learning for video processing

Table 53 compares the performance of several video understanding models, including CoVGT (Xiao et al. 2023), HiTeA (Ye et al. 2023), InternVideo (Wang et al. 2022), ImageViT (Papalampidi et al. 2024), ShortViIT (Papalampidi et al. 2024), Flamingo (Alayrac et al. 2022), SeViLA Localizer + ShortViIT (Papalampidi et al. 2024), SeViLA, TimeS-L (Bertasius et al. 2021), VideoSwin-B (Liu et al. 2022), BEVT (Wang et al. 2022), SIFAR-B-14 (Fan et al. 2021), ORViT (Herzig et al. 2022), AIM ViT-B (Yang et al. 2023), AIM ViT-L (Yang et al. 2023), Bard variants (Papalampidi et al. 2024), GPT-4 Turbo, GPT-4V, and Gemini Ultra (Team et al. 2023). These models differ in parameter size, frame usage, and performance across tasks like long video question answering (EgoSchema, Next-QA) and fine-grained action classification (Diving48), each offering unique strengths for specific scenarios. For example, GPT-4V excels in EgoSchema with 63.5% on the subset and 55.6% on the full set, while SeViLA leads on Next-QA with 73.8%, though it underperforms on EgoSchema (25.7% on subset). In action classification, TimeS-L achieves the highest Top-1 accuracy on Diving48 at 91.0%, followed by AIM ViT-L at 90.6%, but both require additional spatial cropping (SC). LongViT, with 256 frames, performs well on EgoSchema (56.8% on subset), but its high computational demand may limit its applicability.

Fig. 8 Comparisons of super-resolution effects of different deep learning methods, i.e., **a** ACNet, **b** CARN, **c** DRCN and **d** VDSR on B100 dataset for scales of 2, 3 and 4

9.8 Deep learning for natural language processing

Table 54 evaluates the performance of various large language models (LLMs) across a diverse set of benchmarks, including AIME 2024, GPQA, SWE Bench, MATH 500, BFCL, and Alder Polyglot. The models compared—such as GPT-4o, Claude 3.5 Sonnet, OpenAI o1, DeepSeek-R1, Gemini 2.5 Pro, Grok 3 [Beta], and several Llama variants—exhibit differences in parameter scale, training approaches, and task-specific strengths, offering distinct advantages depending on the evaluation context. For instance, Gemini 2.5 Pro achieves an impressive 92% on AIME 2024, closely followed by Grok 3 [Beta] at 93.3%, showcasing their prowess in mathematical reasoning, while Llama 4 Behemoth dominates GPQA with a striking 95%. On the SWE Bench, a coding-focused task, OpenAI o1-mini leads with a remarkable 90%, highlighting its efficiency in software engineering scenarios, though it lags on MATH 500 at 52.2%.

In contrast, models like DeepSeek-R1 and OpenAI o3-mini excel on MATH 500, scoring 97.3 and 97.9% respectively, demonstrating superior capability in advanced mathematical problem-solving. However, their performance diverges elsewhere—DeepSeek-R1 scores

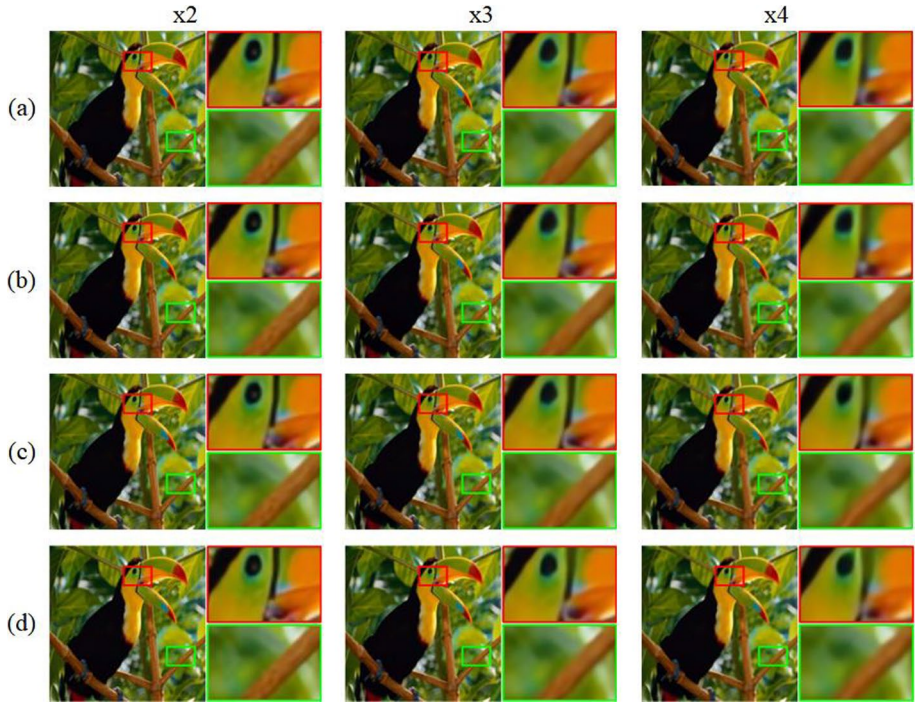


Fig. 9 Comparisons of super-resolution effects of different deep learning methods, i.e., **a** ACNet, **b** CARN, **c** DRCN and **d** VDSR on Set5 dataset for scales of 2, 3 and 4

Table 40 PSNR(dB)/SSIM of different methods on NTIRE 2018 Track 2 dataset

Methods	Bicubic	FSRCNN (Dong et al. 2016)	EDSR (Lim et al. 2017)	SRGAN (Ledig et al. 2017)	BM3D+EDSR (Lim et al. 2017)	CinCGAN (Yuan et al. 2018)
PSNR	22.85	22.67	25.77	24.33	22.88	24.33
SSIM	0.65	0.61	0.71	0.67	0.68	0.69

a modest 49.2% on SWE Bench, while OpenAI o3-mini reaches 61%. Claude 3.7 Sonnet [R], with a strong 96.2% on MATH 500 and 70.3% on SWE Bench, emerges as a versatile contender, though it requires additional resources (denoted by [R]). Meanwhile, lighter models like GPT-4o mini struggle, with a mere 3.6% on MATH 500, despite a respectable 64.1% on SWE Bench. In multilingual tasks like Alder Polyglot, Claude 3.7 Sonnet [R] and DeepSeek-R1 lead with 64.9 and 64%, respectively, while others, such as Gemma 3 27b, falter across multiple domains (e.g., 4.9% on MATH 500). These disparities underscore the trade-offs between computational complexity, task specialization, and generalization in modern LLMs.

Table 41 PSNR(dB)/SSIM of different methods for scales 2, 3 and 4

Methods	Scale	Set5	Set14	Urban100	BSD100
SRCNN (Liu et al. 2009)	×2	36.66 / 0.9542	32.45 / 0.9067	29.50 / 0.8946	31.36 / 0.8879
FSRCNN (Dong et al. 2016)	x2	37.05 / 0.9560	32.66 / 0.9090	29.88 / 0.9020	31.53 / 0.8920
VDSR (Patra and Kot 2002)	×2	37.53 / 0.9590	33.05 / 0.9130	30.77 / 0.9140	31.90 / 0.8960
RDN (Song et al. 2015)	×2	38.24 / 0.9614	34.01 / 0.9212	32.89 / 0.9353	32.34 / 0.9017
D-DBPN (Haris et al. 2018)	×2	38.09 / 0.9600	33.85 / 0.9190	32.55 / 0.9324	32.27 / 0.9000
EDSR (Srivastava 2013)	×2	38.11 / 0.9602	33.92 / 0.9195	32.93 / 0.9351	32.32 / 0.9013
SRCNN (Liu et al. 2009)	×4	30.48 / 0.8628	27.50 / 0.7513	24.52 / 0.7221	26.90 / 0.7101
FSRCNN (Dong et al. 2016)	×4	30.72 / 0.8660	27.61 / 0.7550	24.62 / 0.7280	26.98 / 0.7150
VDSR (Patra and Kot 2002)	×4	31.35 / 0.8830	28.02 / 0.7680	25.18 / 0.7540	27.29 / 0.7260
RDN (Song et al. 2015)	×4	32.47 / 0.8990	28.81 / 0.7871	26.61 / 0.8028	27.72 / 0.7419
D-DBPN (Haris et al. 2018)	×4	32.47 / 0.8980	28.82 / 0.7860	26.38 / 0.7946	27.72 / 0.7400
EDSR (Srivastava 2013)	×4	32.46 / 0.8968	28.80 / 0.7876	26.64 / 0.8033	27.71 / 0.7420
SRCNN (Liu et al. 2009)	×8	25.33 / 0.6900	23.76 / 0.5910	21.29 / 0.5440	24.13 / 0.5660
FSRCNN (Dong et al. 2016)	×8	20.13 / 0.5520	19.75 / 0.4820	21.32 / 0.5380	24.21 / 0.5680
VDSR (Patra and Kot 2002)	×8	25.93 / 0.7240	24.26 / 0.6140	21.70 / 0.5710	24.49 / 0.5830
RDN (Song et al. 2015)	×8	27.21 / 0.7840	25.13 / 0.6480	22.73 / 0.6312	24.88 / 0.6010
D-DBPN (Haris et al. 2018)	×8	26.96 / 0.7762	24.91 / 0.6420	22.51 / 0.6221	24.81 / 0.5985

9.9 Deep learning for 3D data processing

Table 55 compares the performance of various 3D point cloud classification algorithms evaluated on the ModelNet40 dataset (Wu et al. 2015), including PointGST (Liang et al. 2024), Mamba3D + Point-MAE (Han et al. 2024), OTMae3D (Wang et al. 2024), PointNeXt (Qian et al. 2022), MVTN (Hamdi et al. 2021), Feature Geometric Net (FG-Net) (Liu et al. 2020), DeepGCN (Li et al. 2019), PointNet++ + SageMix (Lee et al. 2022), DGCNN (Wang et al. 2019), RS-CNN (Liu et al. 2019), and PointNet++ (Qi et al. 2017b). These models differ in overall accuracy, mean accuracy, and the number of parameters, each tailored to specific use cases. PointGST achieves the highest overall accuracy at 95.3%, though it lacks data on mean accuracy and parameters. Mamba3D + Point-MAE follows closely with 95.1% accuracy and 16.9M parameters, while PointNeXt offers 94.0% overall accuracy and 91.1% mean accuracy with a leaner 4.5M parameters, making it efficient for resource-constrained

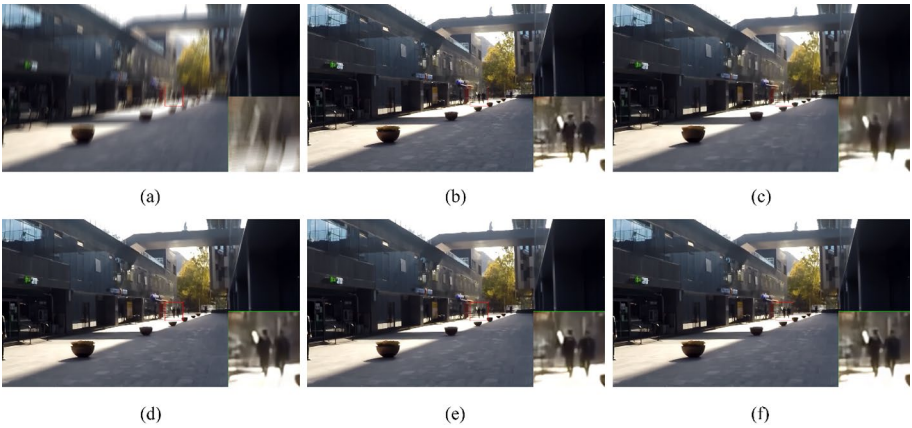


Fig. 10 Different methods, i.e., **a** blurry, **b** Clean image, **c** DeepDeblur (Nah et al. 2017), **d** MIMO-UNet (Cho et al. 2021), **e** MPRNet (Zamir et al. 2021) and **f** MSSNet (Kim et al. 2022) on GOPRO dataset for image deblurring

Table 42 Performance comparison of deep networks on the ISIC2018 dataset

Methods	Params (M)	Flops (G)	Acc%	F1%	Prec%	Recall%
ConvNeXt-B (Li et al. 2022)	88.59	15.36	76.52	50.94	66.84	50.52
VGG-19 (Mateen et al. 2018)	143.68	19.67	79.25	61.83	63.71	60.89
Mixer-L/16 (Tolstikhin et al. 2021)	208.2	44.57	78.92	59.88	61.36	59.16
T2T-ViT_t-24 (Zhao et al. 2022)	64	12.69	77.59	57.21	59.60	55.94
DeiT-base (Yu et al. 2022)	86.57	16.86	72.31	41.01	47.19	44.09
VIT-B/16 (Hong et al. 2024)	86.86	33.03	78.32	60.93	64.16	60.52
VIT-B/32 (Garcia-Martin and Sanchez-Reillo 2023)	88.3	8.56	77.92	57.52	58.74	56.90
HiFuse_Base (Huo et al. 2024)	127.8	10.97	84.12	75.32	76.52	74.74

Table 43 Performance comparison of deep networks on the Covid19 dataset

Methods	Acc%	F1%	Prec%	Recall%
ConvNeXt-B (Li et al. 2022)	55.38	54.68	54.95	54.81
VGG-19 (Mateen et al. 2018)	59.14	57.55	59.04	58.13
Mixer-L/16 (Tolstikhin et al. 2021)	70.43	70.12	70.38	70.06
T2T-ViT_t-24 (Zhao et al. 2022)	63.44	60.34	65.68	61.89
DeiT-base (Yu et al. 2022)	50.54	39.31	44.47	47.96
VIT-B/16 (Hong et al. 2024)	65.05	64.88	64.90	64.87
VIT-B/32 (Garcia-Martin and Sanchez-Reillo 2023)	61.83	60.59	61.89	60.94
HiFuse_Base (Huo et al. 2024)	76.34	76.17	76.30	76.11

Table 44 Deep learning for face image classification

Methods	Datasets	Accuracy (%)
RAN (Wang et al. 2020)	RAF-DB	86.9
ViT+SE (Aouayeb et al. 2021)		87.22
CNN+softlabel (Vo et al. 2020)		86.31
DACL (Farzaneh and Qi 2021)		87.78
AD-Corre (Fard and Mahoor 2022)	FER-2013	72.03
ResNet-18 (He et al. 2016)		72.3
AFAW (Xie et al. 2019)		72.67
CT-DBN (Liang et al. 2023)		72.81
ECNN (Chen et al. 2016)	JAFFE	94.3
Attention CNN (Minaee et al. 2021)		92.8
KECA and SSVM (Liu et al. 2019)		93.04
EFCN (Zhang et al. 2024)		91.05

Table 45 Performance of image classification algorithms on ImageNet

Methods	Top-1 %	Top-5 %	Parameters ($\times 10^6$)	FLOPs ($\times 10^6$)	Model size (MB)	Layers
LeNet-5 (Fırat et al. 2022)	95.5	98.5	0.06	–	–	5
Inception-v1 (Sam et al. 2019)	77.4	93.3	6.8	1550	53	22
Inception-v3 (Sam et al. 2019)	82.8	94.5	23.6	–	–	159
ResNet (He et al. 2016)	75.1	93.3	23.7	3800	–	50
Inception-v4 (Szegedy et al. 2017)	82.3	96.2	43	1650	83	144
Trimps-Soushen (Turay and Vladimirova 2022)	82.3	97.1	19	–	–	89
SqueezeNet (Hassanpour and Malek 2019)	57.5	80.3	1.2	352	0.5	19
Xception (Kassani et al. 2019)	79.0	94.5	22	1200	88	126
ResNetXt-50 (He et al. 2016)	77.7	93.8	22.2	4100	–	50
MobileNets (Howard 2017)	70.6	89.9	4.2	569	16	29
ShiftNet (Yan et al. 2018)	58.8	82.0	0.8	279	–	44
MobileNetV2 (Sandler et al. 2018)	72.0	90.1	3.4	300	300	114
ShuffleNet (Zhang et al. 2018)	73.7	89.7	143	140	75	50
SqueezeNext (Gholami et al. 2018)	60.3	83.5	0.9	310	–	112
ColorNet (Gowda and Yuan 2019)	84.6	96.1	19	250	26	–
AlexNet (Li et al. 2021)	57.2	80.3	62	720	240	8
ZFNet (Fu et al. 2018)	73.9	88.3	60	630	170	8
VGGNet (Wang et al. 2015)	68.5	88.5	138	15,500	575	19
NASNet Large (Zhang 2023)	82.7	96.2	89	2380	343	32
ResNeXt (Xie et al. 2017)	85.4	97.6	829	153,000	458	101
EfficientNet-L2 (Xie et al. 2020)	85.5	97.5	480	23,500	–	154
FixEfficientNetL2 (Touvron et al. 2019)	88.5	98.7	480	23,500	–	154
LambdaResNet200 (Bello et al. 2021)	84.3	96.3	42	34,000	–	200
MPL-EfficientNet-B6-Wide (Pham et al. 2021)	90.0	98.7	390	–	–	–
MPL-EfficientNet-L2 (Pham et al. 2021)	90.2	98.8	480	–	–	154

scenarios. DeepGCN and DGCNN provide solid performance at 93.6 and 92.9% overall accuracy, respectively, with fewer parameters (2.2M and 1.81M), whereas PointNet++ lags at 90.7%. Users should choose models based on their specific requirements, weighing accuracy against computational complexity.

Table 46 Performance comparisons of different surface detection methods on NEU-DET and DAGM datasets

Datasets	Methods	Accuracy	Precision	Recall	F1-Score	IoU	mAP
NEU-DET	KNN (Cover and Hart 1967)	64.13±0.22	63.08±0.5	63.07±0.01	63.07±0.03	–	60.41±0.02
	SVM (Suthaharan and Suthaharan 2016)	75.82±0.13	74.69±0.06	74.14±0.03	74.41±0.04	–	69.08±0.06
	BP (LeCun et al. 1988)	79.99±0.09	78.71±0.08	78.02±0.16	78.36±0.12	–	73.23±0.03
	SDNet (Zhang and Ma 2021)	67.21±0.02	50.04±0.01	58.04±0.09	53.74±0.04	50.26±0.01	52.17±0.02
	ViT_transformer (Dosovitskiy 2020)	74.56±0.09	62.35±0.05	61.19±0.02	61.77±0.03	56.59±0.03	58.05±0.01
	DeepLab (Chen et al. 2017)	81.65±0.14	86.12±0.01	83.22±0.12	84.64±0.07	74.08±0.03	76.61±0.02
	DenseNet (Huang et al. 2017)	63.11±0.13	47.62±0.02	34.77±0.07	40.19±0.06	45.32±0.09	39.55±0.07
	FCN (Long et al. 2015)	77.99±0.07	79.81±0.04	82.19±0.03	80.98±0.04	67.41±0.11	71.34±0.05
	PAN (Li et al. 2018)	88.85±0.06	80.70±0.12	85.71±0.03	83.13±0.08	81.40±0.04	73.23±0.03
	U-Net (Ronneberger et al. 2015)	80.05±0.10	83.45±0.06	84.18±0.04	83.81±0.07	78.87±0.05	75.09±0.02
	PGA-Net (Dong et al. 2019)	89.98±0.03	86.33±0.10	87.13±0.02	86.73±0.06	82.09±0.02	80.02±0.02
	CADN (Zhang et al. 2021)	93.48±0.08	89.48±0.07	88.24±0.03	88.86±0.05	83.47±0.10	82.04±0.01
	RetinaNet (Cheng and Yu 2020)	95.79±0.12	90.23±0.04	89.95±0.06	89.59±0.05	44.18±0.08	83.11±0.04

Table 46 (continued)

Datasets	Methods	Accuracy	Precision	Recall	F1-Score	IoU	mAP
	MF-GAN (Yang et al. 2022)	97.02±0.11	92.92±0.03	89.69±0.03	91.28±0.03	89.27±0.17	89.77±0.03
DAGM	KNN (Cover and Hart 1967)	72.44±0.05	70.01±0.10	69.24±0.04	69.62±0.07	–	66.29±0.04
	SVM (Suthaharan and Suthaharan 2016)	80.07±0.03	79.67±0.08	79.09±0.06	79.38±0.07	–	73.38±0.03
	BP (LeCun et al. 1988)	85.16±0.09	84.71±0.04	84.66±0.05	83.94±0.05	73.21±0.08	75.54±0.04
	SDNet (Zhang and Ma 2021)	69.45±0.09	51.28±0.02	59.52±0.04	55.09±0.07	33.21±0.01	52.08±0.02
	ViT_transformer (Dosovitskiy 2020)	54.86±0.11	50.83±0.01	49.78±0.03	50.30±0.02	41.22±0.02	48.46±0.06
	DeepLab (Chen et al. 2017)	87.33±0.03	84.03±0.14	82.55±0.05	83.28±0.09	74.61±0.02	79.35±0.01
	DenseNet (Huang et al. 2017)	51.04±0.04	48.49±0.03	40.17±0.06	43.94±0.05	32.84±0.08	39.61±0.02
	FCN (Long et al. 2015)	85.35±0.13	83.25±0.06	81.59±0.04	82.41±0.05	73.79±0.11	78.27±0.05
	PAN (Li et al. 2018)	87.22±0.03	84.02±0.01	82.49±0.02	83.25±0.02	74.47±0.05	79.65±0.02
	U-Net (Ronneberger et al. 2015)	86.96±0.05	83.66±0.05	82.19±0.04	82.92±0.04	70.09±0.14	78.92±0.03
	PGA-Net (Dong et al. 2019)	89.34±0.08	86.23±0.02	87.87±0.04	86.56±0.03	76.52±0.07	82.69±0.04
	CADN (Zhang et al. 2021)	90.04±0.09	85.04±0.13	83.03±0.01	84.02±0.07	87.37±0.07	90.15±0.02

Table 46 (continued)

Datasets	Methods	Accuracy	Precision	Recall	F1-Score	IoU	mAP
	RetinaNet (Cheng and Yu 2020)	90.45±0.02	86.55±0.03	84.85±0.07	85.69±0.05	75.05±0.01	82.69±0.04
	MF-GAN (Yang et al. 2022)	97.92±0.10	93.43±0.04	93.69±0.04	93.56±0.02	87.37±0.07	90.15±0.02

Table 47 Accuracies of segmentation models based deep learning on the PASCAL VOC

Method	Backbone	MIoU
FCN (Long et al. 2015)	VGG-16	62.2
DeepLab-CRF (Chen et al. 2017)	ResNet-101	77.5
GCN* (Kipf and Welling 2016)	ResNet-152	79.7
RefineNet (Lin et al. 2017)	ResNet-152	82.2
Wide ResNet (Zagoruyko and Komodakis 2016)	WideResNet-38	84.2
PSPNet (Zhao et al. 2017)	ResNet-101	84.9
DeeplabV3 (Chen et al. 2017)	ResNet-101	85.4
PSANet (Zhao et al. 2018)	ResNet-101	85.7
EncNet (Zhang et al. 2018a)	ResNet-101	85.7
Exfuse (Zhang et al. 2018b)	ResNet-101	86.2
DM-Net* (He et al. 2019a)	ResNet-101	86.8
APC-Net* (He et al. 2019b)	ResNet-101	87.1
EMANet (Li et al. 2019)	ResNet-101	87.7
DeeplabV3+ (Chen et al. 2018)	XNeXI-131	87.8
Exfuse (Zhang et al. 2018b)	ResNet-131	87.9
EMANet (Li et al. 2019)	ResNet-152	88.2

10 Discussion

The application of deep learning in multiple fields has shown extensive potential, and the interaction between different tasks provides important opportunities for technological progress. Low-level visual tasks, high-level visual tasks, video processing, natural language processing, and 3D data processing together form the core areas of deep learning applications. For example, denoising or super-resolution processed images can improve the accuracy of object detection, while advances in video analysis can support cross-modal NLP tasks such as video description generation, and 3D scene understanding can enhance robot navigation performance. This cross-domain interdependence highlights the importance of continued research and development in all deep learning applications to achieve overall technological progress. This section will provide an overview of potential research directions in each field and identify some of the unresolved challenges.

Low-level vision tasks aim to restore degraded images to higher quality, encompassing tasks i.e., image denoising, super-resolution, and deblurring. These foundational tasks are vital for enhancing image quality and are integral to more complex high-level vision tasks. Deep learning has significantly advanced low-level vision tasks, presenting new research

Table 48 Accuracies of segmentation models based deep learning on the Cityscapes dataset

Method	Backbone	MIoU
FCN (Long et al. 2015)	–	65.3
DeeplabV2 (Chen et al. 2017)	ResNet-101	70.4
RefineNet (Lin et al. 2017)	ResNet-101	73.6
FoveaNet (Li et al. 2017)	ResNet-101	74.1
GCN (Kipf and Welling 2016)	ResNet-101	76.9
DUC-HDC (Wang et al. 2018)	ResNet-101	77.6
Wide ResNet (Zagoruyko and Komodakis 2016)	WideResNet-38	78.4
PSPNet (Zhao et al. 2017)	ResNet-101	85.4
BiSeNet (Yu et al. 2018)	ResNet-101	78.9
PSANet (Li et al. 2020)	ResNet-101	80.1
DenseASPP (Hu et al. 2020)	DenseNet-161	80.6
DANet (Xue et al. 2019)	ResNet-101	81.5
CCNet (Huang et al. 2019)	ResNet-101	81.4
DeeplabV3 (Chen et al. 2017)	ResNet-101	81.3
ACNet (Hu et al. 2019)	ResNet-101	82.3

Table 49 Results of deep networks on COCO dataset for instance segmentation

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
MNC (Dai et al. 2016)	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS (Kobashi et al. 2013)	ResNet-101-C5-dilated	29.2	49.5	–	7.1	31.3	50
Mask R-CNN (He et al. 2017)	ResNeXt-101-FPN	37.1	60.0	39.4	16.9	39.9	53.5
Faster R-CNN (He et al. 2016)	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
One-stage methods							
ExtremeNet (Zhou et al. 2019)	Hourglass-104	18.9	44.5	13.7	10.4	20.4	28.3
TensorMask (Chen et al. 2019)	ResNet-101-FPN	37.1	59.3	39.4	17.1	39.1	51.6
YOLACT (Zeng et al. 2022)	ResNet-101-FPN	31.2	50.6	32.8	12.1	33.3	47.1
PolarMask (Xie et al. 2021)	ResNeXt-101-FPN	32.9	55.4	33.8	15.5	35.1	46.3
SOLO (Wang et al. 2020)	ResNet-50-FPN	36.8	58.6	39.0	15.9	39.5	52.1

opportunities alongside unique challenges. Potential research directions of several image restoration tasks summarized as follows.

1. Unify a framework for complex image restoration tasks. Due to varying scenes, designing a universal network can be suitable to different image restoration tasks, i.e., single image denoising, super-resolution and deblurring, hybrid image restoration tasks, i.e., noisy image super-resolution, noisy image deblurring, low-resolution image deblurring, according to properties of image restoration, which is suitable to mobile digital devices in the real world, i.e., smart phones and cameras.
2. Lightweight networks for image restoration tasks. Due to limited load capacity of mobile devices, devolved lightweight networks are necessary, according to deep learning theory, i.e., channel relations, optimizing loss function and relations of hierarchical information, etc.
3. Self-supervised learning for image restoration tasks. Because clean images (also regraded as reference images) are difficult to be obtained caused complex shooting

Table 50 Performance comparison using AP and MIoU metrics of different deep networks

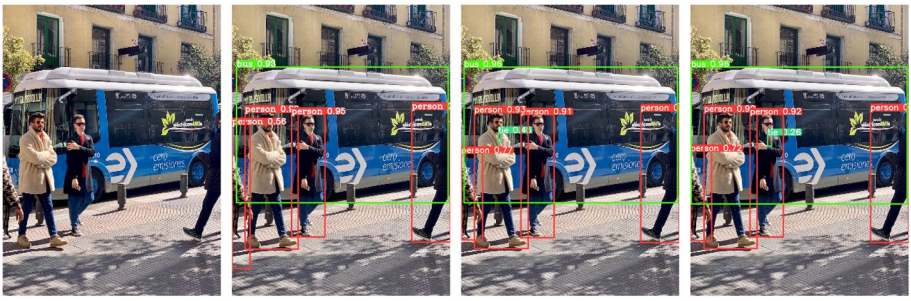
Datasets	Task/val	Methods	AP	MIOU
Cityscapes	val	EfficientPS Single-scale (Mohan and Valada 2021)	38.3	79.3
		EfficientPS Multi-scale (Mohan and Valada 2021)	43.8	82.1
		PanoNet (Chen et al. 2020)	23.1	–
		CASNet (Ji et al. 2020)	35.8	–
		Panoptic-deeplab (Cheng et al. 2020)	42.5	83.1
		UPSNet (Xiong et al. 2019)	39.0	79.2
		SDC-Depth (Wang et al. 2020)	–	64.8
		PanopticDeepLab+ (Cheng et al. 2020)	46.8	85.3
		SPINet (Hwang et al. 2022)	35.3	80.0
	test	Panoptic-deeplab (Cheng et al. 2020)	39.0	84.2
		PanopticDeepLab+ (Cheng et al. 2020)	42.2	84.1
COCO	val	SDC-Depth (Wang et al. 2020)	31.0	38.6
		UPSNet (Xiong et al. 2019)	34.3	55.8
		SPINet (Hwang et al. 2022)	33.2	43.2
Mapillary Vistas	val	Panoptic-deeplab (Cheng et al. 2020)	17.2	56.8
		EfficientPS Single-scale (Mohan and Valada 2021)	18.7	52.6
		EfficientPS Multi-scale (Mohan and Valada 2021)	20.8	54.1
		PanopticDeepLab+ (Cheng et al. 2020)	21.8	60.3
	test	Panoptic-deeplab (Cheng et al. 2020)	16.9	57.6
ADE20K	val	PanopticDeepLab+ (Cheng et al. 2020)	–	50.35
	test	PanopticDeepLab+ (Cheng et al. 2020)	–	40.47
KITTI	val	EfficientPS Single-scale (Mohan and Valada 2021)	27.1	55.3
		EfficientPS Multi-scale (Mohan and Valada 2021)	27.9	56.4
SemanticKITTI	val	EfficientLPS (Sirohi et al. 2021)	–	64.9
		Panoster (Gasperini et al. 2021)	–	61.1
	test	EfficientLPS (Sirohi et al. 2021)	–	61.4
		Panoster (Gasperini et al. 2021)	–	59.9

Table 51 Performance of two-stage object detection algorithms based deep learning techniques

Algorithms	Backbone	FPS	VOC 2007 (Everingham et al. 2010) (%)	VOC 2012 (Shetty 2016) (%)	MSCOCO (Lin et al. 2014) (%)
R-CNN (Girshick et al. 2014)	AlexNet	–	58.8(VOC07+VOC12)	–	–
Fast R-CNN (Girshick 2015)	VGG-16	3	70.0(VOC07+VOC12)	68.4(VOC07+VOC12)	–
Faster R-CNN (Ren et al. 2015)	VGG-16	7	73.2(VOC07+VOC12)	70.4(VOC07+VOC12)	21.2(MSCOCO)
Mask R-CNN (He et al. 2017)	ResNet-50-FPN	11	76.4(VOC07+VOC12)	73.8(VOC07+VOC12)	34.9(MSCOCO)
R-FCN (Dai et al. 2016)	VGG-16	–	76.5(VOC07+VOC12)	–	29.9(MSCOCO)
R-FCN (Dai et al. 2016)	ResNet-101	6	80.5(VOC07+VOC12)	77.6(VOC07+VOC12)	–

Table 52 Performance of one-stage object detection algorithms based deep learning techniques

One-stage detection	Backbone	Datasets	Max FPS	AP
YOLOv1-448 (Redmon 2016)	GoogleNet like	VOC2007+2012	45	63.4
YOLOv2-416 (Redmon and Farhadi 2017)	Darknet-19	VOC2007+2012	67	78.6
SSD-300 (Liu et al. 2016)	VGGNet-16	MSCOCO	16.4	31
DSSD-321 (Fu et al. 2017)	ResNet-101	MSCOCO	11.8	33.2
YOLOv3 (Redmon and Farhadi 2018)	Darknet-53	MSCOCO	34.5	33
YOLOv4 (Bochkovskiy et al. 2020)	CSPDarknet-53	MSCOCO	125	45.8
YOLOv5 (Bochkovskiy et al. 2020)	CSPDarknet-53 like	MSCOCO	200	45.5



(a) (b) (c) (d)

Fig. 11 Different methods based YOLO, i.e., **a** original image, **b** YOLOv6, **c** YOLOv7 and **d** YOLOv9 for object detecting on COCO dataset

environments, moving objects, unstable equipment, man-made factors, existing methods based deep learning techniques use paired images to train models of image restoration, where robustness of obtained models get poor for real world. Thus, using self-supervised learning to guide deep networks to achieve adaptive models for complex real scenes is important for low-level vision.

4. Muti-modal techniques for image restoration tasks. Because collected images are compressed by cameras, they may loss key information for shooting scenes. Multi-modal techniques can overcome this issue for image restoration.
5. Large model techniques for image restoration tasks. To enhance robustness of obtained image restorations, large model techniques are used to fully learn dependencies of different data to achieve better performance in low-level vision tasks.
6. Developing new metrics can better verify performance of deep learning methods for high-level vision.

Table 53 Performance of video processing models on multiple tasks

Method	Params	Frames	EgoSchema Subset	EgoSchema Full	Next-QA	Diving48 Top-1	Remarks
CoVGT (Xiao et al. 2023)	149 M	32	–	–	60.0	–	
HiTeA (Ye et al. 2023)	297 M	16	–	–	63.1	–	
InternVideo (Wang et al. 2022)	478 M	90	–	32.1	63.2	–	
ImageViT (Papalampidi et al. 2024)	1B	16	40.8	30.9	–	–	
ShortViIT (Papalampidi et al. 2024)	1B	16	47.9	31.0	–	–	
Flamingo (Alayrac et al. 2022)	3B	32	–	–	–	–	
SeViLA Localizer + ShortViIT (Papalampidi et al. 2024)	5B	32	49.6	31.3	–	–	
SeViLA (Yu et al. 2023)	4B	32	25.7	22.7	46.2	–	73.8 –
TimeS-L (Bertasius et al. 2021)	121 M	–	–	–	–	91.0	Needs SC
VideoSwin-B (Liu et al. 2022)	88 M	–	–	–	–	81.9	Needs SC
BEVT (Wang et al. 2022)	88 M	–	–	–	–	86.7	Needs SC
SIFAR-B-14 (Fan et al. 2021)	87 M	–	–	–	–	87.3	Needs SC
ORViT (Herzig et al. 2022)	160 M	–	–	–	–	88.0	Needs SC+BB
AIM ViT-B (Yang et al. 2023)	97 M	–	–	–	–	88.9	Needs SC
AIM ViT-L (Yang et al. 2023)	341 M	–	–	–	–	90.6	Needs SC
Bard only (blind) (Papalampidi et al. 2024)	–	–	27.0	33.2	–	–	Blind model
Bard + ImageViT (Papalampidi et al. 2024)	–	–	35.0	35.0	–	–	
Bard + ShortViIT (Papalampidi et al. 2024)	–	–	42.0	36.2	–	–	
Bard + PALI (Papalampidi et al. 2024)	–	–	44.8	39.2	–	–	
GPT-4 Turbo (blind)	–	–	31.0	30.8	–	–	Blind model
GPT-4V	–	–	63.5	55.6	–	–	
Gemini Ultra (Team et al. 2023)	–	–	–	–	–	–	

Classical high-level vision tasks including image classification, object detection, and image segmentation, which require understanding images rather than recovering images. Potential research can be concluded as follows.

1. Unifying a framework is used for high-level vision tasks, i.e., image classification, object detection and image segmentation, according to relations of three tasks.
2. Lightweight networks for high-level vision tasks. Some existing methods based deep learning have higher computational costs and complexities, thus refining networks are very necessary for trained GPU platforms and applied platforms, according to salient information of key layers and optimizing loss functions, etc.
3. Self-supervised learning methods can be guided deep networks in unsupervised ways rather than ground truth of manual labeling to improve applicability of several high-level vision tasks in the real world.
4. Muti-modal techniques can be used to extract complementary information to reduce data dependency and deep network dependency in high-level vision tasks.

Table 54 Performance of LLM models on multiple tasks

Models	AIME 2024	GPQA	SWE Bench	MATH 500	BFCL	Alder Polyglot
GPT-4o	13.4%	56.1%	31%	60.3%	72.08%	27.1%
Claude 3.5 Sonnet	16%	65%	49%	78%	56.46%	51.6%
Claude 3.7 Sonnet	23.3%	68%	62.3%	82.2%	58.3%	60.4%
GPT-4.5	36.7%	71.4%	38%	69.94%	44.9%	
DeepSeek V3 0324	59.4%	64.8%	38.8%	94%	58.55%	55.1%
Claude 3.7 Sonnet [R]	61.3%	78.2%	70.3%	96.2%	58.3%	64.9%
OpenAI o1-mini	63.6%	60%	90%	52.2%	32.9%	
OpenAI o1	79.2%	75.7%	48.9%	96.4%	67.87%	61.7%
DeepSeek-R1	79.8%	71.5%	49.2%	97.3%	57.53%	64%
OpenAI o3-mini	87.3%	79.7%	61%	97.9%	65.12%	60.4%
Gemini 2.5 Pro	92%	84%	63.8%	72.9%		
Grok 3 [Beta]	93.3%	84.6%				
Llama 4 Behemoth	73.7%	95%				
Llama 4 Scout	57.2%					
Llama 4 Maverick	69.8%	15.6%				
Gemma 3 27b	42.4%	10.2%	89%	4.9%		
Qwen2.5-VL-32B	46%	18.8%	82.2%	62.84%		
Gemini 2.0 Flash	62.1%	51.8%	89.7%	60.42%	22.2%	
Llama 3.3 70b	50.5%	77%	77.3%	51.43%		
Nova Pro	46.9%	76.6%	68.4%	61.38%		
Claude 3.5 Haiku	41.6%	40.6%	69.4%	54.31%	28%	
Llama 3.1 405b	49%	73.8%	81.1%	51.43%		
GPT-4o mini	40.2%	70.2%	64.1%	3.6%		

Table 55 Performance of 3D Point Cloud Classification on ModelNet40

Model	Overall accuracy	Mean accuracy	Number of params
PointGST (Liang et al. 2024)	95.3	–	–
Mamba3D + Point-MAE (Han et al. 2024)	95.1	–	16.9M
OTMae3D (Wang et al. 2024)	94.5	–	–
PointNeXt (Qian et al. 2022)	94.0	91.1	4.5M
MVTN (Hamdi et al. 2021)	93.8	92.2	–
Feature Geometric Net (FG-Net) (Liu et al. 2020)	93.8	91.1	–
DeepGCN (Li et al. 2019)	93.6	90.9	2.2M
PointNet++ + SageMix (Lee et al. 2022)	93.3	–	–
DGCNN (Wang et al. 2019)	92.9	90.2	1.81M
RS-CNN (Liu et al. 2019)	92.9	–	–
PointNet++ (Qi et al. 2017b)	90.7	–	1.74M

5. Large model techniques can further enhance data dependency to extract salient information for improving performance of high-level vision tasks.
6. Developing new metrics can better test performance of deep learning methods for high-level vision.

7. Real-time processing of high-level vision, improving processing abilities of hardware equipment and optimizing deep learning methods is very necessary.

Video processing tasks involve video analysis and understanding, video generation and editing, and video enhancement and restoration. These tasks require the processing of dynamic visual data and deal with the complexity introduced by the time dimension. The application of deep learning to video processing has brought new possibilities to multimedia technology. The following are potential research directions:

1. A unified framework for video processing tasks, such as video analysis, generation, and enhancement, designs general models based on the temporal and spatial characteristics of video data to adapt to different application scenarios (such as monitoring, entertainment, and education).
2. Lightweight networks for video processing tasks. The high-dimensional nature of video data leads to high computational costs, so lightweight networks need to be developed to optimise the extraction of temporal and spatial features and adapt to resource-constrained devices.
3. The application of self-supervised learning in video processing. Video data annotation is costly and complex. Self-supervised learning can take advantage of the temporal consistency of videos to reduce the dependence on labelled data and improve the robustness of models in real-world scenarios.
4. The application of multimodal technology in video processing. Combining multimodal data such as video, audio and text can improve the quality of video understanding and generation, for example in video subtitling or cross-modal video retrieval.
5. The application of large model technology in video processing. Using large model technology to learn long-term dependencies in video data improves the performance of video generation and enhancement tasks.
6. Developing new metrics to better evaluate the performance of deep learning methods in video processing tasks.
7. Optimisation of real-time performance in video processing. Optimisation of the collaboration between deep learning models and hardware devices to improve processing speed for real-time video analysis and enhancement.

NLP tasks include text representation and structured analysis, text generation and interactive applications, as well as cross-modal integration and advanced scenarios. These tasks require deep learning models to understand and generate natural language. The following are potential research directions:

1. A framework that unifies NLP tasks, such as text representation, generation, and cross-modal integration, and designs universal models based on the semantic and contextual characteristics of language tasks to improve model generalization.
2. Lightweight networks for NLP tasks. Develop lightweight language models for resource-constrained devices, optimising computational complexity and inference speed while maintaining the quality of language understanding and generation.
3. Application of self-supervised learning in NLP. Use large-scale unlabelled text data to improve the adaptability of models to complex language scenarios through

self-supervised learning, for example in low-resource languages or domain-specific tasks.
4. Applications of multimodal technologies in NLP. Combining multimodal data such as text, images and videos to improve the performance of cross-modal tasks such as visual question answering or image caption generation.
5. Applications of large model technologies in NLP. Using large model technologies such as pre-trained language models to enhance language understanding and generation capabilities, especially in interactive applications and advanced scenarios.
6. Developing new metrics to better evaluate the performance of deep learning methods in NLP tasks.
7. Optimisation of real-time performance for NLP tasks. For real-time interactive applications (e.g. chatbots), optimise model inference speed and hardware support to improve the user experience.

3D data processing tasks include 3D object recognition and classification, 3D scene understanding and segmentation, and 3D reconstruction and generation. These tasks require processing spatial data and coping with the complexity of three-dimensional structures. The following are potential research directions:

1. A unified framework for 3D data processing tasks, such as 3D object recognition, scene segmentation, and reconstruction, designs a general model based on the geometric and topological characteristics of 3D data to adapt to different application scenarios (such as robot navigation and virtual reality). Lightweight networks for 3D data processing. The high dimensionality of 3D data leads to high computational costs. Lightweight networks are developed to optimise the processing efficiency of point cloud or stereo microscope data and adapt to resource-constrained devices. The application of self-supervised learning in 3D data processing. The cost of annotating 3D data is high, and self-supervised learning can use geometric consistency or unsupervised pre-training to improve the adaptability of models to complex 3D scenes.
2. The application of multimodal technology in 3D data processing. Combining 3D data with multimodal information such as images and text improves the quality of 3D scene understanding and reconstruction, for example in augmented reality (AR) applications.
3. The application of large model technology in 3D data processing. Using large model technology to learn complex spatial dependencies in 3D data improves the performance of 3D object recognition and reconstruction tasks.
4. Development of new metrics to better evaluate the performance of deep learning methods for 3D data processing tasks.
5. Optimisation of real-time processing of 3D data. Optimisation of the collaboration between deep learning models and hardware devices to improve processing speed for real-time 3D scene understanding and reconstruction.

The challenges of low- and high-level vision tasks, video processing, NLP and 3D data processing tasks can be summarised as follows:

1. Models are less robust to different scenarios and applications. Whether it is images, videos, text or 3D data, the diversity of complex real-world scenarios challenges the generalisation ability of models.
2. Deep networks may lead to high computational costs and complexity, which is not conducive to practical applications, especially in video and 3D data processing, where the data dimension is higher and the computational requirements are greater.
3. The real labels of deep learning methods are usually obtained through manual operation or ideal conditions, which is challenging in practical applications. For example, the labeling of video and 3D data is costly, and the labeling of low-resource languages is scarce in NLP tasks.
4. Features extracted using only deep networks and raw data (e.g., images, videos, or text in a single modality) cannot fully express the overall information of the data, resulting in low performance or robustness in the real world for low-level and high-level vision tasks, video processing, NLP, and 3D data processing.
5. Evaluation metrics may affect the performance of models based on deep learning algorithms, especially in video processing, NLP, and 3D data processing, and existing metrics may not fully reflect the effectiveness of models. The load capacity and speed of the hardware device may affect the real-time processing of different tasks, such as real-time analysis of video processing, real-time interaction of NLP, and real-time reconstruction of 3D data.

11 Conclusion

In this paper, we provide a comprehensive overview of deep learning fundamentals, aiming to offer a foundational introduction for engineers and scholars. Specifically, we trace the development of deep learning techniques and highlight classical network architectures and key components to enhance readers' understanding of deep learning principles. To cater to diverse audiences, we systematically summarize the principles, differences, relationships, and applications of deep networks across various tasks, i.e., low-level vision tasks, such as image denoising, image super-resolution, and image deblurring, high-level vision tasks, such as image classification, image segmentation, and object detection, video processing, such as video analysis, generation, enhancement, natural language processing, such as text representation, generation, cross-modal integration and 3D data processing, such as 3D object recognition, scene understanding, reconstruction. Additionally, we evaluate their performance through both quantitative and qualitative analyses. Finally, we outline potential research directions and challenges in deep learning and summarise the contributions of this paper.

Author Contributions Chunwei Tian and Tongtong Cheng wrote the main manuscript text. Zhe Peng, Wangmeng Zuo, Yonglin Tian, Qingfu Zhang, Fei-Yue Wang and David Zhang provided revision suggestions for the manuscript. All authors reviewed the manuscript.

Funding This work was supported in part by National Natural Science Foundation of China under Grant 62201468, in part by the Shenzhen Science and Technology Program under Grant JCYJ20230807140412025.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467

Abdelhamed A, Lin S, Brown MS (20180 A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700

Aharon M, Elad M, Bruckstein A (2006) K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322

Aizawa A (2003) An information-theoretic perspective of tf-idf measures. Inf Process Manag 39(1):45–65

Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M et al (2022) Flamingo: a visual language model for few-shot learning. Adv Neural Inf Process Syst 35:23716–23736

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433

Anwar S, Huynh CP, Porikli F (2017) Chaining identity mapping modules for image denoising. arXiv preprint arXiv:1712.02933

Anwar S, Barnes N (2019) Real image denoising with feature attention In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3155–3164

Aouayeb M, Hamidouche W, Soladie C, Kpalma K, Seguier R (2021) Learning vision transformer with squeeze and excitation for facial expression recognition. arXiv preprint arXiv:2107.03107

Arjomand Bigdeli S, Zwicker M, Favaro P, Jin M (2017) Deep mean-shift priors for image restoration. In: Advances in Neural Information Processing Systems 30

Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, Savarese S (2016) 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534–1543

Ba JL, Kiros JR, Hinton GE (2018) Layer normalization. arXiv preprint arXiv:1607.06450

Bahdanau D, Cho K, Bengio Y (2016) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Bao S, He H, Wang F, Wu H, Wang H, Wu W, Guo Z, Liu Z, Xu X (2020) Plato-2: Towards building an open-domain chatbot via curriculum learning. arXiv preprint arXiv:2006.16779

Bar-Tal O, Chefer H, Tov O, Herrmann C, Paiss R, Zada S, Ephrat A, Hur J, Liu G, Raj A, et al. (2024) Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers, pp. 1–11

Bastien F, Lamblin P, Pascanu R, Bergstra J, Goodfellow I, Bergeron A, Bouchard N, Warde-Farley D, Bengio Y (2012) Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590

Beal J, Kim E, Tzeng E, Park DH, Kislyuk, D (2020) Toward transformer-based object detection. arXiv:2012.09958

Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, Gall J (2019) Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9297–9307

Bell RM, Koren Y (2007) Lessons from the Netflix prize challenge. ACM SIGKDD Explor Newsl 9(2):75–79
Bello I, Fedus W, Du X, Cubuk ED, Srinivas A, Lin T-Y, Shlens J, Zoph B (2021) Revisiting resnets: Improved training and scaling strategies. Adv Neural Inf Process Syst 34:22614–22627
Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166
Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4
Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding
Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934
Brock A, Lim T, Ritchie JM, Weston N (2016) Neural photo editing with introspective adversarial networks. arXiv preprint arXiv:1609.07093
Brown PF, Della Pietra VJ, Desouza PV, Lai JC, Mercer RL (1992) Class-based n-gram models of natural language. Comput Linguist 18(4):467–480
Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
Buades A, Coll B, Morel J-M (2005) A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 60–65 (2005). IEEE
Burger HC, Schuler CJ, Harmeling S (2012) Image denoising: Can plain neural networks compete with bm3d? In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2392–2399. IEEE
Byun J, Cha S, Moon T (2021) Fbi-denoiser: Fast blind image denoiser for poisson-gaussian noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5768–5777
Cai G, Cai Y (2020) Hierarchy spatial-temporal transformer for action recognition in short videos. In: FSDM, pp. 760–774
Cai Z, Vasconcelos N (2018) Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162
Cai C, Wang D, Wang Y (2021) Graph coarsening with neural networks. arXiv preprint arXiv:2102.01350
Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229. Springer
Carvalho P, Durupt A, Grandvalet Y (2022) A survey of machine learning approaches for visual inspection on the dagm dataset. Advances in manufacturing technology, vol XXXV. IOS Press, Amsterdam, pp 255–260
Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, et al. (2015) Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012
Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H (2015) Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012
Chan C, Ginosar S, Zhou T, Efros AA (2019) Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5933–5942
Chen Y-J, Tsai C-Y, Xu X, Shi Y, Ho T-Y, Huang M, Yuan H, Zhuang J (2021) Ct image denoising with encoder-decoder based graph convolutional networks. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 400–404. IEEE
Chen X, Girshick R, He K, Dollár P (2019) Tensormask: A foundation for dense object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2061–2069
Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587
Chen X, Wang J, Hebert M (2020) Panonet: Real-time panoptic segmentation through position-sensitive feature embedding. arXiv preprint arXiv:2008.00192
Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818
Chen Y (2015) Convolutional neural network for sentence classification. Master's thesis, University of Waterloo
Chen H, Wang Y Guo T Xu C Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2021) Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12299–12310
Chen H, Bhanu B (2007) 3d free-form object recognition in range images using local surface patches. Pattern Recogn Lett 28(10):1252–1262
Chen Y, Pock T (2016) Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. IEEE Trans Pattern Anal Mach Intell 39(6):1256–1272

Chen J, Chen Z, Chi Z, Fu H (2016) Facial expression recognition in video with multiple feature fusion. IEEE Trans Affect Comput 9(1):38–50
Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
Chen J, Chen J, Chao H, Yang M (2018) Image blind denoising with generative adversarial network based noise modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3155–3164
Chen J, Chen J, Chao H, Yang M (2018) Image blind denoising with generative adversarial network based noise modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3155–3164
Cheng B, Collins MD, Zhu Y, Liu T, Huang TS, Adam H, Chen L-C (2020) Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12475–12485
Cheng X, Yu J (2020) Retinanet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. IEEE Trans Instrum Meas 70:1–11
Chen S, Sun P, Song Y, Luo P (2023) Diffusiondet: Diffusion model for object detection In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19830–19843
Chen Z, Sun K, Zhou Z, Lin X, Sun X, Cao L, Ji R (2024) Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. arXiv preprint arXiv:2403.18471
Chetlur S, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, Shelhamer E (2014) cudnn: Efficient primitives for deep learning arXiv preprint arXiv:1410.0759
Cho S-J, Ji S-W, Hong J-P, Jung S-W, Ko S-J (2021) Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4641–4650
Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Computer vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part VIII 14, pp. 628–644. Springer
Ciresan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J (2011) Flexible, high performance convolutional neural networks for image classification. In: Twenty-second International Joint Conference on Artificial Intelligence . Citeseer
Clauset A (2011) A brief primer on probability distributions. In: Santa Fe Institute
Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223
Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
Croitoru F-A, Hondru V, Ionescu RT, Shah M (2023) Diffusion models in vision: A survey. IEEE Trans Pattern Anal Mach Intell 45(9):10850–10869
Cui L, Jiang X, Xu M, Li W, Lv P, Zhou B (2021) Sddnet: A fast and accurate network for surface defect detection. IEEE Trans Instrum Meas 70:1–13
Dabov K, Foi A, Katkovnik V, Egiazarian K (2007) Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Trans Image Process 16(8):2080–2095
Dai T, Cai J, Zhang Y, Xia S-T, Zhang L (2019) Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11065–11074
Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M (2047) Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5828–5839
Dai J, He K, Sun J (2016) Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158
Dai J, Li Y, He K, Sun J (2016) R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems 29
Danielyan A, Katkovnik V, Egiazarian K (2011) Bm3d frames and variational image deblurring. IEEE Trans Image Process 21(4):1715–1728
DasGupta B, Schnitger G (1992) The power of approximating: a comparison of activation functions. In: Advances in neural information processing systems 5
De Boer P-T, Kroese DP, Mannor S, Rubinstein RY (2005) A tutorial on the cross-entropy method. Ann Oper Res 134:19–67

Denton EL, Chintala S, Fergus R (2015) Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in neural information processing systems 28
Devereux B, Amable G, Posada CC (2004) An efficient image segmentation algorithm for landscape analysis. Int J Appl Earth Obs Geoinf 6(1):47–61
Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186
Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14, pp. 391–407. Springer
Dong W, Zhang L, Shi G, Li X (2012) Nonlocally centralized sparse representation for image restoration. IEEE Trans Image Process 22(4):1620–1630
Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell 38(2):295–307
Dong H, Song K, He Y, Xu J, Yan Y, Meng Q (2019) Pga-net: Pyramid feature fusion and global context attention network for automated surface defect detection. IEEE Trans Ind Inf 16(12):7448–7458
Dong J, Roth S, Schiele B (2021) Dwdn: Deep wiener deconvolution network for non-blind image deblurring. IEEE Trans Pattern Anal Mach Intell 44(12):9960–9976
Dong C, Deng Y, Loy CC, Tang X (2015) Compression artifacts reduction by a deep convolutional network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 576–584
Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14, pp. 391–407. Springer
Dosovitskiy A (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
Dunnmon JA, Ratner AJ, Saab K, Khandwala N, Markert M, Sagreiya H, Goldman R, Lee-Messer C, Lungren MP, Rubin DL et al (2020) (2020) Cross-modal data programming enables rapid medical machine learning. Patterns 1(2):100019
Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6(3):361–365
Esser P, Sutter E, Ommer B (2018) A variational u-net for conditional appearance and shape generation In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8857–8866
Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88:303–338
Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88:303–338
Faltings F, Galley M, Peng B, Brantley K, Cai W, Zhang Y, Gao J, Dolan B (2023) Interactive text generation. arXiv preprint arXiv:2303.00908
Fan Q, Panda R, et al. (2021) Can an image classifier suffice for action recognition? arXiv preprint arXiv:2106.14104
Fang H, Xia M, Zhou G, Chang Y, Yan L (2021) Infrared small uav target detection based on residual image prediction via global and local dilated residual networks. IEEE Geosci Remote Sens Lett 19:1–5
Fang K, Lu W, Zhou X, Xu J, Mao K (2022) A multitarget interested region extraction method for wrist x-ray images based on optimized alexnet and two-class combined model. IEEE Trans Comput Soc Syst 9(6):1624–1634. https://doi.org/10.1109/TCSS.2021.3132040
Fang H, Han B, Zhang S, Zhou S, Hu C, Ye W-M (2024) Data augmentation for object detection via controllable diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1257–1266
Fard AP, Mahoor MH (2022) Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access 10:26756–26768
Farzaneh AH, Qi X (2021) Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2402–2411
Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941
Fırat H, Asker ME, Bayindir Mİ, Hanbay D (2022) Spatial-spectral classification of hyperspectral remote sensing images using 3d cnn based lenet-5 architecture. Infrared Phys Technol 127:104470
Franzen, R.: Kodak lossless true color image suite. source: http://r0k.us/graphics/kodak 4(2), 9 (1999)
Freeman I, Roese-Koerner L, Kummert A (2018) Effnet: An efficient structure for convolutional neural networks. In: 2018 25th IEEE International Conference on Image Processing (icip), pp. 6–10. IEEE

Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems 26
Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2017) Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
Fu L, Feng Y, Majeed Y, Zhang X, Zhang J, Karkee M, Zhang Q (2018) Kiwifruit detection in field images using faster r-cnn with zfnet. IFAC-PapersOnLine 51(17):45–50
Gallagher AC, Chen T (2009) Understanding images of groups of people In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 256–263 . IEEE
Gao G, Wang Z, Li J, Li W, Yu Y, Zeng T (2022) Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer. arXiv preprint arXiv:2204.13286 (2022)
Gao L, Zhang J, Yang C, Zhou Y (2022) Cas-vswin transformer: A variant swin transformer for surface-defect detection. Comput Ind 140:103689
Garcia-Martin R, Sanchez-Reillo R (2023) Vision transformers for vein biometric recognition. IEEE Access 11:22060–22080
Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. Atmos Environ 32(14–15):2627–2636
Garibi D, Patashnik O, Voynov A, Averbuch-Elor H, Cohen-Or D (2024) Renoise: Real image inversion through iterative noising. arXiv preprint arXiv:2403.14602
Gasperini S, Mahani M-AN, Marcos-Ramiro A, Navab N, Tombari F (2021) Panoster: End-to-end panoptic segmentation of lidar point clouds. IEEE Robot Autom Lett 6(2):3216–3223
Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The kitti dataset. Int J Robot Res 32(11):1231–1237
Ge Z, Liu S, Wang F, Li Z, Sun J (2021) Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430
Gholami, A., Kwon, K., Wu, B., Tai, Z., Yue, X., Jin, P., Zhao, S., Keutzer, K.: Squeezenext: Hardware-aware neural network design. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1638–1647 (2018)
Giannopoulos P, Perikos I, Hatzilygeroudis I (2018) Deep learning approaches for facial emotion recognition: A case study on fer-2013. Advances in hybridization of intelligent methods: Models, systems and applications. Springer, Cham, pp 1–16
Girshick R (2015) Fast r-cnn. arXiv preprint arXiv:1504.08083
Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587
Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587
Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587
Goodfellow I (2016) Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160
Gould S, Fulton R, Koller D (2019) Decomposing a scene into geometric and semantically consistent regions. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1–8. IEEE
Gow RD, Renshaw D, Findlater K, Grant L, McLeod SJ, Hart J, Nicol RL (2007) A comprehensive tool for modeling cmos image-sensor-noise performance. IEEE Trans Electron Devices 54(6):1321–1329
Gowda SN, Yuan C (2019) Colornet: Investigating the importance of color spaces for image classification. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14, pp. 581–596. Springer
Graham B (2014) Fractional max-pooling. arXiv preprint arXiv:1412.6071
Graves A, Mohamed A-r, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE
Gu S, Zhang L, Zuo W, Feng X (2014) Weighted nuclear norm minimization with application to image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2862–2869
Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. In: Advances in neural information processing systems 30
Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402–2410
Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, Cong R (2020) Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1780–1789

Guo S, Yan Z, Zhang K, Zuo W, Zhang L (2019) Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1712–1722

Gurrola-Ramos J, Dalmau O, Alarcón TE (2021) A residual dense u-net neural network for image denoising. IEEE Access 9:31742–31754

Hackel T, Savinov N, Ladicky L, Wegner JD, Schindler K, Pollefeys M (2017) Semantic3d. net: A new large-scale point cloud classification benchmark. arXiv preprint arXiv:1704.03847

Hadji I, Derpanis KG, Jepson AD (2021) Representation learning via global temporal alignment and cycle-consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11068–11077

Hafiz AM, Bhat GM (2020) A survey on instance segmentation: state of the art. Int J Multimed Inf Retr 9(3):171–189

Hamdi A, Giancola S, Ghanem B (2021) Mvtn: Multi-view transformation network for 3d shape recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11

Han X, Tang Y, Wang Z, Li X (2024) Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 4995–5004

Han B, Roy K (2018) Deltaframe-bp: An algorithm using frame difference for deep convolutional neural networks training and inference on video data. IEEE Trans Multi-Scale Comput Syst 4(4):624–634

Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1664–1673

Hassanpour M, Malek H (2019) Document image classification using squeezenet convolutional neural network. In: 2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pp. 1–4. IEEE

He J, Deng Z, Qiao Y (2019a) Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3562–3572

He J, Deng Z, Zhou L, Wang Y, Qiao Y (2019b) Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7519–7528

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778

He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916

He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916

He Y, Song K, Meng Q, Yan Y (2019) An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas 69(4):1493–1504

He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y (2022) Swin transformer embedding unet for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens 60:1–15

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969

Herzig R, Ben-Avraham E, Mangalam K, Bar A, Chechik G, Rohrbach A, Darrell T, Globerson A (2022) Object-region video transformers. In: Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition, pp. 3148–3159

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778

He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp. 630–645 . Springer

Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580

Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554

Hirose Y, Yamashita K, Hijiya S (1991) Back-propagation algorithm which varies the number of hidden units. Neural Netw 4(1):61–66

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851
Hong S, Wu J, Zhu L, Chen W (2024) Brain tumor classification in vit-b/16 based on relative position encoding and residual mlp. PLoS ONE 19(7):0298102
Hoppe H, DeRose T, Duchamp T, McDonald J, Stuetzle W (1992) Surface reconstruction from unorganized points. In: Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques, pp. 71–78
Hore A, Ziou D (2010) Image quality metrics: Psnr vs. ssim. In: 2010 20th International Conference on Pattern Recognition, pp. 2366–2369. IEEE
Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
Howard AG (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
Hu X, Yang K, Fei L, Wang K (2019) Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1440–1444. IEEE
Hu Q, Yang B, Xie L, Rosa S, Guo Y, Wang Z, Trigoni N, Markham A (2020) Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11108–11117
Hu L (2024) Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8153–8163
Hu P, Li X, Tian Y, Tang T, Zhou T, Bai X, Zhu S, Liang T, Li J (2020) Automatic pancreas segmentation in ct images with distance-based saliency-aware denseaspp network. IEEE J Biomed Health Inform 25(5):1601–1611
Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708
Huang K, Shi B, Li X, Li X, Huang S, Li Y (2022) Multi-modal sensor fusion for auto driving perception: A survey. arXiv preprint arXiv:2202.02703 (2022)
Huang J-B, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206
Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 603–612
Huang Z, Xu W, Yu K (2015) Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991
Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708
Hu J, Cao L, Lu Y, Zhang S, Wang Y, Li K, Huang F, Shao L, Ji R (2021) Istr: End-to-end instance segmentation with transformers arXiv preprint arXiv:2105.00637
Huo X, Sun G, Tian S, Wang Y, Yu L, Long J, Zhang W, Li A (2024) Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomed Signal Process Control 87:105534
Hwang S, Oh SW, Kim SJ (2022) Single-shot path integrated panoptic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3328–3337
Ioffe S (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167
Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134
Janocha K, Czarnecki WM (2017) On loss functions for deep neural networks in classification. arXiv preprint arXiv:1702.05659
Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y (2009) What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153 . IEEE
Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y (2009) What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153. IEEE
Ji Y, Zhang H, Jie Z, Ma L, Wu QJ (2020) Casnet: A cross-attention siamese network for video salient object detection. IEEE Trans Neural Netw Learn Syst 32(6):2676–2690
Jian M, Yu X, Zhang H, Yang C (2024) Swinct: feature enhancement based low-dose ct images denoising with swin transformer. Multimed Syst 30(1):1
Jiang B, Lu Y, Zhang B, Lu G (2023) Agp-net: Adaptive graph prior network for image denoising. IEEE Trans Ind Inform 20:4753–4764

Jiang P, Ergu D, Liu F, Cai Y, Ma B (2022) A review of yolo algorithm developments. Procedia Comput Sci 199:1066–1073
Jiang B, Lu Y, Chen X, Lu X, Lu G (2023) Graph attention in attention network for image denoising. IEEE Trans Syst Man Cybern Syst 53:7077–7088
Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678
Jin, M, Roth, S, Favaro, P (2017) Noise-blind image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3510–3518
Jobson DJ, Rahman Z-U, Woodell GA (1997) A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans Image Process 6(7):965–976
Johnson AE, Hebert M (1999) Using spin images for efficient object recognition in cluttered 3d scenes. IEEE Trans Pattern Anal Mach Intell 21(5):433–449
Joshi N, Szeliski R, Kriegman DJ (2008) Psf estimation using sharp edge prediction. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE
Kadimesetty VS, Gutta S, Ganapathy S, Yalavarthy PK (2018) Convolutional neural network-based robust denoising of low-dose computed tomography perfusion maps. IEEE Trans Radiat Plasma Med Sci 3(2):137–152
Kang K, Park S, Park H, Kang D, Paik J (2023) Action recognition using multi-stream 2d cnn with deep learning-based temporal modality. In: 2023 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–3. IEEE
Kappeler A, Yoo S, Dai Q, Katsaggelos AK (2016) Video super-resolution with convolutional neural networks. IEEE Trans Comput Imaging 2(2):109–122
Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
Karras T (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
Kassani SH, Kassani PH, Khazaeinezhad R, Wesolowski MJ, Schneider KA, Deters R (2019) Diabetic retinopathy classification using a modified xception architecture. In: 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 1–6. IEEE
Kazhdan M, Bolitho M, Hoppe H (20060 Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing, vol. 7
Ketkar N, Ketkar N (2017) Introduction to keras. Deep learning with python: a hands-on introduction. Apress, Berkeley, pp 97–111
Khmag A (2023) Additive gaussian noise removal based on generative adversarial network model and semi-soft thresholding approach. Multimed Tools Appl 82(5):7757–7777
Kim K, Lee S, Cho S (2022) Mssnet: Multi-scale-stage network for single image deblurring. In: European Conference on Computer Vision, pp. 524–539. Springer
Kim D, Woo S, Lee J-Y, Kweon IS (2019) Deep video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5792–5801
Kim H, Garrido P, Tewari A, Xu W, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. ACM Trans Graph (TOG) 37(4):1–14
Kim J, Lee JK, Lee KM (2016a) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1645
Kim J, Lee JK, Lee KM (2016b) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654
Kim K, Lee S, Cho S (2022) Mssnet: Multi-scale-stage network for single image deblurring. In: European Conference on Computer Vision, pp. 524–539. Springer
Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
Kirillov A, Girshick R, He K, Dollár P (2019) Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6399–6408
Kirillov A, He K, Girshick R, Rother C, Dollár P (2019) Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9404–9413
Kitaev N, Klein D (2018) Constituency parsing with a self-attentive encoder. arXiv preprint arXiv:1805.01052
Kobashi S, Kuramoto K, Hata Y (2013) Interactive fuzzy connectedness image segmentation for neonatal brain mr image segmentation. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1799–1804. IEEE

Kong L, Dong J, Ge J, Li M, Pan J (2023) Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5886–5895
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25
Kruse J, Rother C, Schmidt U (2017) Learning to push the limits of efficient fft-based image deconvolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4586–4594
Kupyn O, Budzan V, Mykhailych M, Mishkin D, Matas J (2018) Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8183–8192
Kupyn O, Martyniuk T, Wu J, Wang Z (2019) Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8878–8887
Lai W-S, Huang J-B, Ahuja N, Yang M-H (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632
Lai W-S, Huang J-B, Ahuja N, Yang M-H (2018) Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Trans Pattern Anal Mach Intell 41(11):2599–2613
Lai Z, Fu Y, Zhang J (2024) Hyperspectral image super resolution with real unaligned rgb guidance. IEEE Trans Neural Netw Learn Syst 36:2999–3011
Land EH, McCann JJ (1971) Lightness and retinex theory. J Opt Soc Am 61(1):1–11
Lan X, Roth S, Huttenlocher D, Black MJ (2006) Efficient belief propagation with learned higher-order markov random fields. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part II 9, pp. 269–282. Springer
Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750
Lawrence S, Giles CL, Tsoi AC, Back AD (1997) Face recognition: A convolutional neural-network approach. IEEE Trans Neural Netw 8(1):98–113
Lebrun M, Colom M, Morel J-M (2015) The noise clinic: a blind image denoising algorithm. Image Process On Line 5:1–54
LeCun Y, Touresky D, Hinton G, Sejnowski T (1988) A theoretical framework for back-propagation. In: Proceedings of the 1988 Connectionist Models Summer School, vol. 1, pp. 21–28
LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551
LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690
Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690
Lee C-Y, Gallagher PW, Tu Z (2016) Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: Artificial Intelligence and Statistics, pp. 464–472 . PMLR
Lee D, Yannakakis M (1996) Principles and methods of testing finite state machines-a survey. Proc IEEE 84(8):1090–1123
Lee S, Jeon M, Kim I, Xiong Y, Kim HJ (2022) Sagemix: Saliency-guided mixup for point clouds. Adv Neural Inf Process Syst 35:23580–23592
Lee J, Lee I, Kang J (2019) Self-attention graph pooling. In: International Conference on Machine Learning, pp. 3734–3743. pmlr
Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, Aittala M, Aila T (2018) Noise2noise: Learning image restoration without clean data. arXiv preprint arXiv:1803.04189
Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861
Li X, Jie Z, Wang W, Liu C, Yang J, Shen X, Lin Z, Chen Q, Yan S, Feng J (2017) Foveanet: Perspective-aware urban scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 784–792
Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D (2016) Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541
Li G, Muller M, Thabet A, Ghanem B (2019) Deepgcns: Can gcns go as deep as cnns? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9267–9276

Li S, Wang L, Li J, Yao Y (2021) Image classification algorithm based on improved alexnet. J Phys: Conf Ser 1813:012051
Li H, Xiong P, An J, Wang L (2018) Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180
Li X, Zhong Z, Wu J, Yang Y, Lin Z, Liu H (2019) Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9167–9176
Li, J, Raventos A, Bhargava A, Tagawa T, Gaidon A (2018) Learning to fuse things and stuff. arXiv preprint arXiv:1812.01192
Li F, Ye Y, Tian Z, Zhang X (2019) Cpu versus gpu: which can perform matrix computation faster–performance comparison for basic linear algebra subprograms. Neural Comput Appl 31:4353–4365
Li F, Jin W, Fan C, Zou L, Chen Q, Li X, Jiang H, Liu Y (2020) Psanet: Pyramid splitting and aggregation network for 3d object detection in point cloud. Sensors 21(1):136
Li Z, Liu F, Yang W, Peng S, Zhou J (2021) A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst 33(12):6999–7019
Li Z, Gu T, Li B, Xu W, He X, Hui X (2022) Convnext-based fine-grained image classification and bilinear attention mechanism model. Appl Sci 12(18):9016
Li Y, Fan Q, Huang H, Han Z, Gu Q (2023) A modified yolov8 detection network for uav aerial image recognition. Drones 7(5):304
Li M, Fu Y, Zhang T, Wen G (2024) Supervise-assisted self-supervised deep-learning method for hyperspectral image restoration. IEEE Trans Neural Netw Learn Syst 36:7331–7344
Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844
Liang D, Feng T, Zhou X, Zhang Y, Zou Z, Bai X (2024) Parameter-efficient fine-tuning in spectral domain for point cloud learning. arXiv preprint arXiv:2410.08114
Liang X, Xu L, Zhang W, Zhang Y, Liu J, Liu Z (2023) A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. Vis Comput 39(6):2277–2290
Liang P, Jiang J, Liu X, Ma J (2024) Image deblurring by exploring in-depth properties of transformer. IEEE Trans Neural Netw Learn Syst 36:4652–4663
Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844
Li Y, Chen X, Zhu Z, Xie L, Huang G, Du D, Wang X (2019) Attention-guided unified network for panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7026–7035
Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5457–5466
Li C, Li L, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W (2022) Yolov6: A single-stage object detection framework for industrial applications arXiv preprint arXiv:2209.02976
Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144
Lin M (2013) Network in network. arXiv preprint arXiv:1312.4400
Lin T (2017) Focal loss for dense object detection. arXiv preprint arXiv:1708.02002
Lin B, Ge Y, Cheng X, Li Z, Zhu B, Wang S, He X, Ye Y, Yuan S, Chen L, et al (2024) Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131
Lin K, Li TH, Liu S, Li G (2019) Real photographs denoising with noise domain adaptation and attentive generative adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops
Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pp. 740–755. Springer
Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pp. 740–755. Springer
Lin G, Milan A, Shen C, Reid I (2017) Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934
Lin C-T, Huang S-W, Wu Y-Y, Lai S-H (2020) Gan-based day-to-night image style transfer for nighttime vehicle detection. IEEE Trans Intell Transp Syst 22(2):951–963
Lin T, Wang Y, Liu X, Qiu X (2022) A survey of transformers. AI Open 3:111–132

Lin M, Chen Q, Yan S (2013) Network in network. arXiv preprint arXiv:1312.4400
Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
Li Y, Tofighi M, Geng J, Monga V, Eldar YC (2019) Deep algorithm unrolling for blind image deblurring. arXiv preprint arXiv:1902.03493
Liu Y, Fan B, Xiang S, Pan C (2019) Relation-shape convolutional neural network for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8895–8904
Liu K, Gao Z, Lin F, Chen BM (2020) Fg-net: Fast large-scale lidar point clouds understanding network leveraging correlated feature mining and geometric-aware modelling. arXiv preprint arXiv:2012.09439
Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022) Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211
Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis, M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
Liu D, Wen B, Fan Y, Loy CC, Huang TS (2017) Non-local recurrent network for image restoration. In: Advances in neural information processing systems 31 (2018)
Liu L, Yang L, Chen Y, Zhang X, Hu L, Deng F (2019) Facial expression recognition based on ssvm algorithm and multi-source texture feature fusion using keca. In: Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2017, pp. 659–666. Springer
Liu C, Yuen J, Torralba A (2009) Nonparametric scene parsing: Label transfer via dense scene alignment. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1972–1979. IEEE
Liu P, Zhang H, Zhang K, Lin L, Zuo W (2018) Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 773–782
Liu P, Zhang H, Zhang K, Lin L, Zuo W (2018) Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 773–782
Liu J, Sun Y, Xu X, Kamilov US (2019) Image restoration using total variation regularized deep image prior. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7715–7719. IEEE
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37 . Springer
Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo, B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022
Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768
Liu H, Wan Z, Huang W, Song Y, Han X, Liao J (2021) Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9371–9381
Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 212–220
Liu W, Wen Y, Yu Z, Yang M (2016) Large-margin softmax loss for convolutional neural networks. arXiv preprint arXiv:1612.02295
Liu J, Zhang W, Tang Y, Tang J, Wu G (2020) Residual feature aggregation network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2359–2368
Li K, Wang Y, Gao P, Song G, Liu Y, Li H, Qiao Y (2022) Uniformer: Unified transformer for efficient spatiotemporal representation learning. arXiv preprint arXiv:2201.04676
Li Z, Wang W, Xie E, Yu Z, Anandkuma, A, Alvarez JM, Luo P, Lu T (2022) Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1280–1289
Li Z, Yang J, Liu Z, Yang X, Jeon G, Wu W (2019) Feedback network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3867–3876
Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440
Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440
Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440

Longuet-Higgins HC (1981) A computer algorithm for reconstructing a scene from two projections. Nature 293(5828):133–135
Lu H, Fei N, Huo Y, Gao Y, Lu Z, Wen J-R (2022) Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15692–15701
Lu Z, Li J, Liu H, Huang C, Zhang L, Zeng T (2022) Transformer for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 457–466
Lu T, Wang Y, Zhang Y, Jiang J, Wang Z, Xiong Z (2022) Rethinking prior-guided face super-resolution: A new paradigm with facial component prior. IEEE Trans Neural Netw Learn Syst 35(3):3938–3952
Luccioni AS, Viviano JD (2021) What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. arXiv preprint arXiv:2105.02732
Luc P, Neverova N, Couprie C, Verbeek J, LeCun Y (2017) Predicting deeper into the future of semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 648–657
Lyu Q, Guo M, Pei Z (2020) Degan: Mixed noise removal via generative adversarial networks. Appl Soft Comput 95:106478
Ma J, Peng C, Tian X, Jiang J (2021) Dbdnet: A deep boosting strategy for image denoising. IEEE Trans Multimed 24:3157–3168
Ma J, Xiong G, Xu J, Chen X (2023) Cvtnet: A cross-view transformer network for lidar-based place recognition in autonomous driving environments. IEEE Trans Ind Inform 20:4039–4048
Ma J, He Y, Li F, Han L, You C, Wang B (2024) Segment anything in medical images. Nat Commun 15(1):654
Mansimov E, Parisotto E, Ba JL, Salakhutdinov R (2015) Generating images from captions with attention. arXiv preprint arXiv:1511.02793
Mansour Y, Heckel R (2023) Zero-shot noise2noise: Efficient image denoising without any data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14018–14027
Mao X, Shen C, Yang Y-B (2016) Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Advances in neural information processing systems 29
Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 2, pp. 416–423. IEEE
Mateen M, Wen J, Nasrullah, Song S, Huang Z (2018) Fundus image classification using vgg-19 architecture with pca and svd. Symmetry 11(1):1
Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K (2017) Sketch-based manga retrieval using manga109 dataset. Multimed Tools Appl 76:21811–21838
Mei Y, Fan Y, Zhang Y, Yu J, Zhou Y, Liu D, Fu Y, Huang TS, Shi H (2023) Pyramid attention network for image restoration. Int J Comput Vis 131(12):3207–3225
Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4460–4470
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2021) Nerf: Representing scenes as neural radiance fields for view synthesis. Commun ACM 65(1):99–106
Milton MAA (2019) Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. arXiv preprint arXiv:1901.10802
Minaee S, Minaei M, Abdolrashidi A (2021) Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors 21(9):3046
Miranda E, Aryuni M, Irwansyah E (2016) A survey of medical image classification techniques. In: 2016 International Conference on Information Management and Technology (ICIMTech), pp. 56–61. IEEE
Mnih V, Heess N, Graves A (2014) Recurrent models of visual attention. In: Advances in neural information processing systems 27
Mohammadshahi A, Henderson J (2019) Graph-to-graph transformer for transition-based dependency parsing. arXiv preprint arXiv:1911.03561
Mohan R, Valada A (2021) Efficientps: Efficient panoptic segmentation. Int J Comput Vis 129(5):1551–1579
Mottaghi R, Chen X, Liu X, Cho N-G, Lee S-W, Fidler S, Urtasun R, Yuille A (2014) The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 891–898
Mou C, Zhang J, Wu Z (2021) Dynamic attentive graph learning for image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4328–4337
Nah S, Hyun Kim T, Mu Lee K (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3883–3891

Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814
Nam S, Hwang Y, Matsushita Y, Kim SJ (2016) A holistic approach to cross-channel image noise modeling and its application to image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1683–1691
Neuhold G, Ollmann T, Rota Bulo S, Kontschieder P (2017) The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4990–4999 (2017)
Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR
Nielsen MA (2015) Neural networks and deep learning, vol 25. Determination Press, San Francisco
Oei K, Gomaa A, Feit AM, Belo J (2024) Self-supervised contrastive learning for videos using differentiable local alignment. arXiv preprint arXiv:2409.04607
Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B (2018) Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999
Pan X, Shi J, Luo P, Wang X, Tang X (2018) Spatial as deep: Spatial cnn for traffic scene understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
Pan J, Ferrer CC, McGuinness K, O'Connor NE, Torres J, Sayrol E, Giro-i-Nieto X (2017) Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081
Papalampidi, P, Koppula S, Pathak S, Chiu J, Heyward J, Patraucean V, Shen J, Miech A, Zisserman A, Nematzdeh A (2024) A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14386–14397
Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174
Parkhi O, Vedaldi A, Zisserman A (2015) Deep face recognition In: BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association
Paszke A, Gross S, Chintala S, Chanan G (2017) Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration 6(3), 67
Patra JC, Kot AC (2002) Nonlinear dynamic system identification using chebyshev functional link artificial neural networks. IEEE Trans Syst Man Cybern B 32(4):505–511
Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543
Pham H, Dai Z, Xie Q, Le QV (2021) Meta pseudo labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11557–11568
Plotz T, Roth S (2017) Benchmarking denoising algorithms with real photographs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1586–1595
Qi CR, Su H, Mo K, Guibas LJ (2017a) Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660
Qi CR, Yi L, Su H, Guibas LJ (2017b) Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems 30
Qian G, Li Y, Peng H, Mai J, Hammoud H, Elhoseiny M, Ghanem B (2022) Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Adv Neural Inf Process Syst 35:23192–23204
Qing Z, Zhang S, Wang J, Wang X, Wei Y, Zhang Y, Gao C, Sang N (2024) Hierarchical spatio-temporal decoupling for text-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6635–6645
Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5533–5541
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PmLR
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K (2017) Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225
Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. Pmlr

Raza A, Ayub H, Khan JA, Ahmad I, Salama AS, Daradkeh YI, Javeed D, Ur Rehman A, Hamam H (2022) A hybrid deep learning-based approach for brain tumor classification. Electronics 11(7):1146
Redmon J (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271
Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069. PMLR
Ren R, Hung T, Tan KC (2017) A generic deep-learning-based approach for automated surface inspection. IEEE Trans Cybern 48(3):929–940
Ren W, Pan J, Zhang H, Cao X, Yang M-H (2020) Single image dehazing via multi-scale convolutional neural networks with holistic edges. Int J Comput Vis 128:240–259
Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems 28
Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems 28
Ren W, Zhang J, Ma L, Pan J, Cao X, Zuo W, Liu W, Yang M-H (2018) Deep non-blind deconvolution via generalized low-rank approximation. In: Advances in neural information processing systems 31
Reza AM (2004) Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. J VLSI Signal Process Syst Signal Image Video Technol 38:35–44
Rodríguez P, Bautista MA, Gonzalez J, Escalera S (2018) Beyond one-hot encoding: Lower dimensional target embedding. Image Vis Comput 75:21–31
Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695
Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241. Springer
Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241 . Springer
Roth S, Black MJ (2005) Fields of experts: A framework for learning image priors. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 860–867. IEEE
Rozenberszki D, Litany O, Dai A (2024) Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19957–19967
Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T et al (2022a) Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inf Process Syst 35:36479–36494
Saharia C, Ho J, Chan W, Salimans T, Fleet DJ, Norouzi M (2022b) Image super-resolution via iterative refinement. IEEE Trans Pattern Anal Mach Intell 45(4):4713–4726
Sam SM, Kamardin K, Sjarif NNA, Mohamed N et al (2019) Offline signature verification using deep learning convolutional neural network (cnn) architectures googlenet inception-v1 and inception-v3. Procedia Comput Sci 161:475–483
Sanchez-Lengeling B, Reif E, Pearce A, Wiltschko AB (2021) A gentle introduction to graph neural networks. Distill 6(9):33
Sanders J, Kandrot E (2010) CUDA by example: an introduction to general-purpose GPU programming. Addison-Wesley Professional, Boston
Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520
Sara U, Akter M, Uddin MS (2019) Image quality assessment through fsim, ssim, mse and psnr–a comparative study. J Comput Commun 7(3):8–18
Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2008) The graph neural network model. IEEE Trans Neural Netw 20(1):61–80
Schalkoff RJ (1997) Artificial neural networks. McGraw-Hill Higher Education, New York
Schlichtkrull M, Kipf TN, Bloem P, Van Den Berg R, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, pp. 593–607. Springer

Schmidt U, Roth S (2014) Shrinkage fields for effective image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2774–2781
Schmidt U, Roth S (2014) Shrinkage fields for effective image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2774–2781
Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823
Schuler CJ, Christopher Burger H, Harmeling S, Scholkopf B (2013) A machine learning approach for non-blind image deconvolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1067–1074
Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, pp. 519–528. IEEE
Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229
Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229
Setiadi DRIM (2021) Psnr vs ssim: imperceptibility quality assessment for image steganography. Multimed Tools Appl 80(6):8423–8444
Shang H, Sun C, Liu J, Chen X, Yan R (2023) Defect-aware transformer network for intelligent visual surface defect detection. Adv Eng Inform 55:101882
Shetty S (2016) Application of convolutional neural network for image classification on pascal voc challenge 2012 dataset. arXiv preprint arXiv:1607.03785
Shih FY, Chuang C-F, Wang PS (2008) Performance comparisons of facial expression recognition in jaffe database. Int J Pattern Recognit Artif Intell 22(03):445–459
Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N (2019) First order motion model for image animation. In: Advances in neural information processing systems 32
Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12, pp. 746–760. Springer
Simonovsky M, Komodakis N (2017) Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3693–3702
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
Singh A, Kushwaha S, Alarfaj M, Singh M (2022) Comprehensive overview of backpropagation algorithm for digital image denoising. Electronics 11(10):1590
Sirohi K, Mohan R, Büscher D, Burgard W, Valada A (2021) Efficientlps: Efficient lidar panoptic segmentation. IEEE Trans Rob 38(3):1894–1914
Soh JW, Cho NI (2021) Deep universal blind image denoising. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 747–754 . IEEE
Soltanayev S, Chun SY (2018) Training deep learning based denoisers without ground truth data. In: Advances in neural information processing systems 31
Song S, Lichtenberg SP, Xiao J (2015) Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 567–576
Song Y, Zhu Y, Du X (2019) Dynamic residual dense network for image denoising. Sensors 19(17):3809
Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B (2021) Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations
Soni A, Dutta T, Nigam N, Verma D, Gupta HP (2023) Supervised attention network for arbitrary-shaped text detection in edge-fainted noisy scene images. IEEE Trans Comput Soc Syst 10(3):1179–1188. https://doi.org/10.1109/TCSS.2022.3153557
Srivastava N (2013) Improving neural networks with dropout. Thesis, University of Toronto
Stewart S, Layton M, Williams M, Ingram D, Maily W (2001) Response of cotton to prebloom square loss. J Econ Entomol 94(2):388–396
Stone JE, Gohara D, Shi G (2010) Opencl: A parallel programming standard for heterogeneous computing systems. Comput Sci Eng 12(3):66
Strudel R, Garcia R, Laptev I, Schmid C (2021) Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272
Su W-t, Cheung G, Wildes R, Lin C-W (2020) Graph neural net using analytical graph filters and topology optimization for image denoising. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8464–8468. IEEE

Su S, Delbracio M, Wang J, Sapiro G, Heidrich W, Wang O (2017) Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1279–1288
Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems 27
Suthaharan S, Suthaharan S (2016) Support vector machine. Machine learning models and algorithms for big data classification, vol 36. Integrated series in information system. Springer, Boston
Sutton C, McCallum A et al (2012) An introduction to conditional random fields. Found Trends Mach Learn 4(4):267–373
Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9
Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826
Tabernik D, Šela S, Skvarč J, Skočaj D (2020) Segmentation-based deep-learning approach for surface-defect detection. J Intell Manuf 31(3):759–776
Tai Y, Yang J, Liu X, Xu C (2017) Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4539–4547
Tai Y, Yang J, Liu X, Xu C (2017) Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4539–4547
Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708
Tan Z, Wang M, Xie J, Chen Y, Shi X (2018) Deep semantic role labeling with self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR
Tao X, Zhang D, Ma W, Liu X, Xu D (2018) Automatic metallic surface defect detection and recognition with convolutional neural networks. Appl Sci 8(9):1575
Tao X, Gao H, Shen X, Wang J, Jia J (2018) Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8174–8182
Tchapmi L, Choy C, Armeni I, Gwak J, Savarese S (2017) Segcloud: Semantic segmentation of 3d point clouds. In: 2017 International Conference on 3D Vision (3DV), pp. 537–547. IEEE
Team G, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K, et al (2023) Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805
Thomas H, Qi CR, Deschaud J-E, Marcotegui B, Goulette F, Guibas LJ (2019) Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6411–6420
Tian C, Zheng M, Zuo W, Zhang S, Zhang Y, Lin C-W (2024) A cross transformer for image denoising. Inf Fusion 102:102043. https://doi.org/10.1016/j.inffus.2023.102043
Tian C, Xu Y, Fei L, Wang J, Wen J, Luo N (2019) Enhanced cnn for image denoising. CAAI Trans Intell Technol 4(1):17–23
Tian C, Xu Y, Zuo W (2020) Image denoising using deep cnn with batch renormalization. Neural Netw 121:461–473
Tian C, Xu Y, Zuo W (2020) Image denoising using deep cnn with batch renormalization. Neural Netw 121:461–473
Tian C, Fei L, Zheng W, Xu Y, Zuo W, Lin C-W (2020) Deep learning on image denoising: An overview. Neural Netw 131:251–275
Tian C, Xu Y, Li Z, Zuo W, Fei L, Liu H (2020) Attention-guided cnn for image denoising. Neural Netw 124:117–129
Tian C, Xu Y, Zuo W, Du B, Lin C-W, Zhang D (2021) Designing and training of a dual cnn for image denoising. Knowl-Based Syst 226:106949
Tian Z, Shen C, Chen H, He T (2019) Fcos: Fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355
Timofte R, Agustsson E, Van Gool L, Yang M-H, Zhang L (2017) Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 114–125

Timofte, R, Agustsson, E, Van Gool, L, Yang, M-H, Zhang, L (2017) Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 114–125
Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J et al (2021) Mlp-mixer: An all-mlp architecture for vision. Adv Neural Inf Process Syst 34:24261–24272
Tong T, Li G, Liu X, Gao Q (2017) Image super-resolution using dense skip connections In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4799–4807
Touvron H, Vedaldi A, Douze M, Jégou H (2019) Fixing the train-test resolution discrepancy. In: Advances in neural information processing systems 32
Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497
Tsai F-J, Peng Y-T, Lin Y-Y, Tsai C-C, Lin C-W (2022) Stripformer: Strip transformer for fast image deblurring. In: European Conference on Computer Vision, pp. 146–162. Springer
Tu Z, Talebi H, Zhang H, Yang F, Milanfar P, Bovik A, Li Y (2022) Maxim: Multi-axis mlp for image processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5769–5780
Tu J, Ji W, Zhao H, Zhang C, Zimmermann R, Qian H (2024) Driveditfit: Fine-tuning diffusion transformers for autonomous driving. arXiv preprint arXiv:2407.15661
Tulyakov S, Liu M-Y, Yang X, Kautz J (2018) Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1535
Turay T, Vladimirova T (2022) Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey. IEEE Access 10:14076–14119
Ulyanov D, Vedaldi A, Lempitsky V (2018) Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454
Valsesia D, Fracastoro G, Magli E (2020) Deep graph-convolutional image denoising. IEEE Trans Image Process 29:8226–8237
Van Gansbeke W, De Brabandere B (2025) A simple latent diffusion approach for panoptic segmentation and mask inpainting. In: European Conference on Computer Vision, pp. 78–97. Springer
Vaswani A (2017) Attention is all you need. In: Advances in Neural Information Processing Systems
Vedaldi A, Lenc K (2015) Matconvnet: Convolutional neural networks for matlab In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 689–692
Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903
Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164
Vo T-H, Lee G-S, Yang H-J, Kim S-H (2020) Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access 8:131988–132001
Walmer M, Kanjirathinkal R, Tai KS, Muzumdar K, Tian T, Shrivastava A (2023) Multi-entity video transformers for fine-grained video representation learning. arXiv preprint arXiv:2311.10873
Wang C-Y, Bochkovskiy A, Liao H-YM (2023) Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464–7475
Wang C-Y, Yeh I-H, Liao H-YM (2024) Yolov9: Learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616
Wang T-C, Liu M-Y, Zhu J-Y, Liu G, Tao A, Kautz J, Catanzaro B (2018) Video-to-video synthesis. arXiv preprint arXiv:1808.06601
Wang R, Chen D, Wu Z, Chen Y, Dai X, Liu M, Jiang Y-G, Zhou L, Yuan L (2022) Bevt: Bert pretraining of video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14733–14743
Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G (2018) Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1451–1460. IEEE
Wang L, Guo S, Huang W, Qiao Y (2015) Places205-vggnet models for scene recognition. arXiv preprint arXiv:1508.01667
Wang Y, Li K, Li Y, He Y, Huang B, Zhao Z, Zhang H, Xu J, Liu Y, Wang Z, et al (2022) Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191
Wang T, Sun M, Hu K (2017) Dilated deep residual network for image denoising. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1272–1279. IEEE

Wang T, Sun M, Hu K (2017) Dilated deep residual network for image denoising. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1272–1279. IEEE
Wang C, Zha Y, He J, Yang W, Zhang T (2024) Rethinking masked representation learning for 3d point cloud understanding. IEEE Trans Image Process 34:247–262
Wang L, Zhang J, Wang O, Lin Z, Lu H (2020) Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 541–550
Wang Q, Hu X, Wang H, Men A, Jiang Z (2021) Multi-dip: A general framework for unsupervised multi-degraded image restoration. In: International Conference on Neural Information Processing, pp. 378–389 Springer
Wang Y, Perazzi F, McWilliams B, Sorkine-Hornung A, Sorkine-Hornung O, Schroers C (2018) A fully progressive approach to single-image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 864–873
Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM (2019) Dynamic graph cnn for learning on point clouds. ACM Trans Graph 38(5):1–12
Wang Z, Chen J, Hoi SC (2020) Deep learning for image super-resolution: A survey. IEEE Trans Pattern Anal Mach Intell 43(10):3365–3387
Wang K, Peng X, Yang J, Meng D, Qiao Y (2020) Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans Image Process 29:4057–4069
Wang X, Zhang R, Kong T, Li L, Shen C (2020) Solov2: Dynamic and fast instance segmentation. Adv Neural Inf Process Syst 33:17721–17732
Wang Y, Lu T, Zhang Y, Wang Z, Jiang J, Xiong Z (2022) Faceformer: Aggregating global and local representation for face hallucination. IEEE Trans Circuits Syst Video Technol 33(6):2533–2545
Wang Y, Lu T, Yao Y, Zhang Y, Xiong Z (2023) Learning to hallucinate face in the dark. IEEE Trans Multimed 26:2314–2326
Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H (2022) Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17683–17693
Wang Z, Fu Y, Liu J, Zhang Y (2023) Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18156–18165
Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36. Springer
Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 157–166
Wei T, Guo L-Z, Li Y-F, Gao W (2018) Learning safe multi-label prediction for weakly labeled data. Mach Learn 107:703–725
Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Computer vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part VII 14, pp. 499–515. Springer
Williams T, Li R (2018) Wavelet pooling for convolutional neural networks. In: International Conference on Learning Representations
Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920
Wu J, Zhang C, Xue T, Freeman B, Tenenbaum J (2016) Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in neural information processing systems 29
Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24
Wu W, Liu H, Li L, Long Y, Wang X, Wang Z, Li J, Chang Y (2021) Application of local fully convolutional neural network combined with yolo v5 algorithm in small target detection of remote sensing image. PLoS ONE 16(10):0259283
Wu Y, He K (2018) Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19
Wu X, Liu M, Cao Y, Ren D, Zuo W (2020) Unpaired learning of deep image denoising. In: European Conference on Computer Vision, pp. 352–368 . Springer
Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L (2021) Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31

Xiang Y, Kim W, Chen W, Ji J, Choy C, Su H, Mottaghi R, Guibas L, Savarese S (2016) Objectnet3d: A large scale database for 3d object recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14, pp. 160–176. Springer
Xiao J, Owens A, Torralba A (2013) Sun3d: A database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1625–1632
Xiao J, Zhou P, Yao A, Li Y, Hong R, Yan S, Chua T-S (2023) Contrastive video question answering via video graph transformer. IEEE Trans Pattern Anal Mach Intell 45(11):13265–13280
Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500
Xie Q, Luong M-T, Hovy E, Le QV (2020) Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10687–10698
Xie W, Shen L, Duan J (2019) Adaptive weighting of handcrafted feature losses for facial expression recognition. IEEE Trans Cybern 51(5):2787–2800
Xie E, Wang W, Ding M, Zhang R, Luo P (2021) Polarmask++: Enhanced polar representation for single-shot instance segmentation and beyond. IEEE Trans Pattern Anal Mach Intell 44(9):5385–5400
Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500
Xie J, Xu L, Chen E (2012) Image denoising and inpainting with deep neural networks. In: Advances in neural information processing systems 25
Xiong Y, Liao R, Zhao H, Hu R, Bai M, Yumer E, Urtasun R (2019) Upsnet: A unified panoptic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8818–8826
Xiong Y, Liao R, Zhao H, Hu R, Bai M, Yumer E, Urtasun R (2019) Upsnet: A unified panoptic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8818–8826
Xu J, Zhang L, Zhang D (2018) A trilateral weighted sparse coding scheme for real-world image denoising. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 20–36
Xu J, Zhang L, Zhang D (2018) External prior guided internal prior learning for real-world noisy image denoising. IEEE Trans Image Process 27(6):2996–3010
Xu B, Shi X, Zhao Z, Zheng W (2018) Leveraging biomedical resources in bi-lstm for drug-drug interaction extraction. IEEE Access 6:33432–33439
Xue H, Liu C, Wan F, Jiao J, Ji X, Ye Q (2019) Danet: Divergent activation for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6589–6598
Xu L, Ren JS, Liu C, Jia J (2014) Deep convolutional neural network for image deconvolution. In: Advances in neural information processing systems 27
Yan Z, Li X, Li M, Zuo W, Shan S (2018) Shift-net: Image inpainting via deep feature rearrangement. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–17
Yan Y, Ren W, Hu X, Li K, Shen H, Cao X (2021) Srgat: Single image super-resolution with graph attention network. IEEE Trans Image Process 30:4905–4918
Yang T-J, Collins MD, Zhu Y, Hwang J-J, Liu T, Zhang X, Sze V, Papandreou G, Chen L-C (2019) Deeperlab: Single-shot image parser. arXiv preprint arXiv:1902.05093
Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P (2020) Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865
Yang Y, Ni X, Hao Y, Liu C, Wang W, Liu Y, Xie H (2022) Mf-gan: Multi-conditional fusion generative adversarial network for text-to-image synthesis. In: International Conference on Multimedia Modeling, pp. 41–53. Springer
Yang T, Zhu Y, Xie Y, Zhang A, Chen C, Li M (2023) Aim: Adapting image models for efficient video action recognition. arXiv preprint arXiv:2302.03024
Yang D, Sun J (2017) Bm3d-net: A convolutional neural network for transform-domain collaborative filtering. IEEE Signal Process Lett 25(1):55–59
Yang H, Yuan C, Zhang L, Sun Y, Hu W, Maybank SJ (2020) Sta-cnn: Convolutional spatial-temporal attention learning for action recognition. IEEE Trans Image Process 29:5783–5793
Yang Y, Fu H, Aviles-Rivero AI, Schönlieb C-B, Zhu L (2023) Diffmic: Dual-guidance diffusion network for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 95–105. Springer
Yang T, Ren P, Xie X, Zhang L (2021) Gan prior embedded network for blind face restoration in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 672–681

Ye Q, Xu G, Yan M, Xu H, Qian Q, Zhang J, Huang F (2023) Hitea: Hierarchical temporal-aware video-language pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15405–15416
Ye W, Zhang W, Lei W, Zhang W, Chen X, Wang Y (2023) Remote sensing image instance segmentation network with transformer and multi-scale feature representation. Expert Syst Appl 234:121007
Yi X, Babyn P (2018) Sharpness-aware low-dose ct denoising using conditional generative adversarial network. J Digit Imaging 31:655–669
Yi L, Li G, Jiang M (2017) An end-to-end steel strip surface defects recognition system based on convolutional neural networks. Steel Res Int 88(2):1600068
Yu F, Huang K, Wang M, Cheng Y, Chu W, Cui L (2022) Width & depth pruning for vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. **36**, pp. 3143–3151
Yu X, Tang L, Rao Y, Huang T, Zhou J, Lu J (2022) Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19313–19322
Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341
Yu S, Cho J, Yadav P, Bansal M (2023) Self-chained image-language model for video localization and question answering. Adv Neural Inf Process Syst 36:76749–76771
Yu Y, Yuan J, Liao L, Li X, Zhong X, Wu J (2024) Ensemble cross unet transformers for augmentation of atomic electron tomography. IEEE Trans Instrum Meas 73:5021714
Yuan Y, Liu S, Zhang J, Zhang Y, Dong C, Lin L (2018) Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 701–710
Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z-H, Tay FE, Feng J, Yan S (2021) Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 558–567
Yue Z, Yong H, Zhao Q, Meng D, Zhang L (2019) Variational denoising network: Toward blind noise modeling and removal. In: Advances in neural information processing systems 32
Yu Z, Li A, Au OC, Xu C (2012) Bag of textons for image segmentation via soft clustering and convex shift. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 781–788. IEEE
Yu D, Wang H, Chen P, Wei Z (2014) Mixed pooling for convolutional neural networks. In: Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24–26, 2014, Proceedings 9, pp. 364–375 . Springer
Zagoruyko S, Komodakis N (2016) Wide residual networks. arXiv preprint arXiv:1605.07146
Zagoruyko S, Komodakis N (2017) Diracnets: Training very deep neural networks without skip-connections. arXiv preprint arXiv:1706.00388
Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H (2022) Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728–5739
Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, Shao L (2021) Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14821–14831
Zeiler M (2014) Visualizing and understanding convolutional networks. In: European Conference on Computer vision/arXiv, vol. 1311
Zeiler MD, Krishnan D, Taylor GW, Fergus R (2010) Deconvolutional networks In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2528–2535. IEEE
Zeng Y, Wei G, Zheng J, Zou J, Wei Y, Zhang Y, Li H (2024) Make pixels dance: High-dynamic video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8850–8860
Zeng J, Ouyang H, Liu M, Leng L, Fu X (2022) Multi-scale yolact for instance segmentation. J King Saud Univ Comput Inf Sci 34(10):9419–9427
Zeyde R, Elad M, Protter M (2012) On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24–30, 2010, Revised Selected Papers 7, pp. 711–730. Springer
Zhang R (2019) Making convolutional networks shift-invariant again. In: International Conference on Machine Learning, pp. 7324–7334. PMLR
Zhang H, Dana K, Shi J, Zhang Z, Wang X, Tyagi A, Agrawal A (2018a) Context encoding for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7151–7160

Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J (2021) Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588

Zhang Y, Sun S, Galley M, Chen Y-C, Brockett C, Gao X, Gao J, Liu J, Dolan B (2019) Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536

Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481

Zhang Z, Zhang X, Peng C, Xue X, Sun J (2018b) Exfuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–284 (2018)

Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)

Zhang Y (2023) Lung segmentation with nasnet-large-decoder net. arXiv preprint arXiv:2303.10315

Zhang H, Ma J (2021) Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. Int J Comput Vis 129(10):2761–2785

Zhang D, Zhou F (2023) Self-supervised image denoising for real-world images with context-aware transformer. IEEE Access 11:14340–14349

Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. Int J Forecast 14(1):35–62

Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 1:43–52

Zhang L, Wu X, Buades A, Li X (2011) Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. J Electron Imaging 20(2):023016–02301616

Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans Image Process 26(7):3142–3155

Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans Image Process 26(7):3142–3155

Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans Image Process 26(7):3142–3155

Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans Image Process 26(7):3142–3155

Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans Image Process 26(7):3142–3155

Zhang K, Zuo W, Zhang L (2018) Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Trans Image Process 27(9):4608–4622

Zhang K, Zuo W, Zhang L (2018) Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Trans Image Process 27(9):4608–4622

Zhang J, Zhu Y, Li W, Fu W, Cao L (2021) Drnet: A deep neural network with multi-layer residual blocks improves image denoising. IEEE Access 9:79936–79946

Zhang J, Su H, Zou W, Gong X, Zhang Z, Shen F (2021) Cadn: A weakly supervised learning-based category-aware object detection network for surface defect detection. Pattern Recogn 109:107571

Zhang K, Li Y, Liang J, Cao J, Zhang Y, Tang H, Fan D-P, Timofte R, Gool LV (2023) Practical blind image denoising via swin-conv-unet and data synthesis. Mach Intell Res 20(6):822–836

Zhang X, Fu X, Qi G, Zhang N (2024) A multi-scale feature fusion convolutional neural network for facial expression recognition. Expert Syst 41(4):13517

Zhang Z, Bu J, Ester M, Zhang J, Yao C, Yu Z, Wang C (2019) Hierarchical graph pooling with structure learning. arXiv preprint arXiv:1911.05954

Zhang K, Gool LV, Timofte R (2020) Deep unfolding network for image super-resolution In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3217–3226

Zhang Y, Li D, Law KL, Wang X, Qin H, Li H (2022) Idr: Self-supervised image denoising via iterative data refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2098–2107

Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301

Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481

Zhang L, Xiang T, Gong S (2017) Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2021–2030

Zhang H, Zhang L, Qi X, Li H, Torr PH, Koniusz P (2020) Few-shot action recognition via improved attention with self-supervision. arXiv preprint arXiv:2001.03905
Zhang K, Zuo W, Gu S, Zhang L (2017) Learning deep cnn denoiser prior for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3929–3938
Zhao H, Jiang L, Jia J, Torr PH, Koltun V (2021) Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16259–16268
Zhao H, Shao W, Bao B, Li H (2019) A simple and robust deep convolutional approach to blind image denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops
Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890
Zhao H, Zhang Y, Liu S, Shi J, Loy CC, Lin D, Jia J (2018) Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 267–283
Zhao C, Shuai R, Ma L, Liu W, Wu M (2022) Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. Multimed Tools Appl 81(17):24265–24300
Zhao Y, Jiang Z, Men A, Ju G (2019) Pyramid real image denoising network. In: 2019 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4 . IEEE
Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890
Zhong X, Tu S, Ma X, Jiang K, Huang W, Wang Z (2022) Rainy wcity: A real rainfall dataset with diverse conditions for semantic driving scene understanding. In: IJCAI, pp. 1743–1749
Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A (2017) Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641
Zhou X, Zhuo J, Krahenbuhl P (2019) Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 850–859
Zhou B, Zhao H, Puig X, Xiao T, Fidler S, Barriuso A, Torralba A (2019) Semantic understanding of scenes through the ade20k dataset. Int J Comput Vis 127:302–321
Zhou X, Wang D, Krähenbühl P (2019) Objects as points. arXiv preprint arXiv:1904.07850
Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232
Zhu W, Zhang H, Zhang C, Zhu X, Guan Z, Jia J (2023) Surface defect detection and classification of steel using an efficient swin transformer. Adv Eng Inform 57:102061
Zoran D, Weiss Y (2011) From learning models of natural image patches to whole image restoration. In: 2011 International Conference on Computer Vision, pp. 479–486. IEEE
Zou Y, Yan C, Fu Y (2023) Iterative denoiser and noise estimator for self-supervised image denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13265–13274

Authors and Affiliations

Chunwei Tian[1] · Tongtong Cheng[2] · Zhe Peng[3] · Wangmeng Zuo[1] · Yonglin Tian[4] · Qingfu Zhang[5] · Fei-Yue Wang[4] · David Zhang[2,6]

✉ Zhe Peng
jeffrey-zhe.peng@polyu.edu.hk

Chunwei Tian
chunweitian@163.com

Tongtong Cheng
tongtongcheng@link.cuhk.edu.cn

Wangmeng Zuo
wmzuo@hit.edu.cn

Yonglin Tian
tyldyx@mail.ustc.edu.cn

Qingfu Zhang
qingfu.zhang@cityu.edu.hk

Fei-Yue Wang
feiyue.wang@ia.ac.cn

David Zhang
davidzhang@cuhk.edu.cn

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

[2] School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China

[3] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[4] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[5] Department of Computer Science, City University of Hong Kong, Hong Kong, China

[6] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China