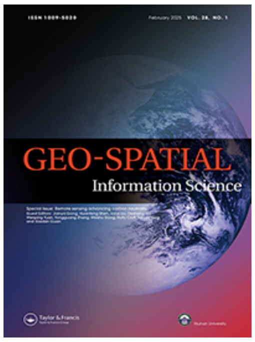

Taylor & Francis
Taylor & Francis Group

Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgsi20

Multi-level representation learning via ConvNeXt-based network for unaligned cross-view matching

Fangli Guan, Nan Zhao, Zhixiang Fang, Ling Jiang, Jianhui Zhang, Yue Yu & Haosheng Huang

To cite this article: Fangli Guan, Nan Zhao, Zhixiang Fang, Ling Jiang, Jianhui Zhang, Yue Yu & Haosheng Huang (17 Jan 2025): Multi-level representation learning via ConvNeXt-based network for unaligned cross-view matching, Geo-spatial Information Science, DOI: 10.1080/10095020.2024.2439385

To link to this article: https://doi.org/10.1080/10095020.2024.2439385

View supplementary material

Published online: 17 Jan 2025.

Submit your article to this journal

Article views: 529

View related articles

View Crossmark data

Citing articles: 1 View citing articles

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2024.2439385

Taylor & Francis
Taylor & Francis Group

Multi-level representation learning via ConvNeXt-based network for unaligned cross-view matching

Fangli Guan[a], Nan Zhao[a], Zhixiang Fang[b], Ling Jiang[c], Jianhui Zhang[a], Yue Yu[d] and Haosheng Huang[e]

[a]School of Computer Science, Hangzhou Dianzi University, Hangzhou, China; [b]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; [c]Anhui Province Key Laboratory of Physical Geographic Environment, Chuzhou University, Chuzhou, China; [d]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China; [e]Department of Geography, Ghent University, Ghent, Belgium

ABSTRACT

Cross-view matching refers to the use of images from different platforms (e.g. drone and satellite views) to retrieve the most relevant images, where the key is that the viewpoints and spatial resolution. However, most of the existing methods focus on extracting fine-grained features and ignore the connection of contextual information in the image. Therefore, we propose a novel ConvNeXt-based multi-level representation learning model for the solution of this task. First, we extract global features through the ConvNeXt model. In order to obtain a joint part-based representation learning from the global features, we then replicated the obtained global features, operating one copy with spatial attention and the other copy using a standard convolutional operation. In addition, the features of different branches are aggregated through the multilevel feature fusion module to prepare for cross-view matching. Finally, we created a new hybrid loss function to better limit these features and assist in mining crucial data regarding global features. The experimental results indicate that we have achieved advanced performance on two common datasets, University-1652 and SUES-200 at 89.79% and 95.75% in drone target matching and 94.87% and 98.80 in drone navigation.

ARTICLE HISTORY
Received 7 June 2024
Accepted 3 December 2024

KEYWORDS
Cross-view matching; ConvNeXt; satellite view; drone view; multilevel feature

1. Introduction

Cross-view matching aims to match an image without location information with another viewpoint‘s image or slice that contains GPS information. The application of cross-view matching covers various earth observation tasks (Chen et al. 2022; Li et al. 2023), such as drone precise navigation (Hodge, Hawkins, and Alexander 2021; Wei and Wang 2018; Zhu, Yang, and Chen 2021), abnormal change detection (Bai et al. 2022; Han et al. 2023; Zhang et al. 2019) and person re-identification (Zheng, Zheng, and Yang 2017). It has become one of the most important tools for increasing efficiency, reducing costs, and improving service quality.

Previous work has mainly focused on matching ground-view panoramic images and satellite images (Arandjelovic et al. 2016; Lin et al. 2015). As drone technology evolves, the addition of drone views expands the application of cross-view matching (Ji, Wei, and Lu 2018; Tian et al. 2021). Drones significantly differ in perspective, spatial resolution, and effectiveness compared to traditional ground panoramic images. The cross-view matching challenge was then extended to include the subject of drone-view and satellite-view picture matching. It can encourage two cutting-edge uses: (1) Drone navigation: matching and retrieving the satellite images or slices by inputting the drone images and obtaining the location range information of the remote sensing images. (2) Drone target matching: retrieving drone images of the area from the input satellite images or slices. Therefore, it is a great challenge to match satellite and UAV remote sensing images from different viewpoints to fully mine, fuse, and utilize the effective information in the above image sets.

Recently, significant progress has been made in deep learning-based drone matching techniques. Current cross-view matching methods focus on mining feature masks (Wang et al. 2022; Zhuang et al. 2021) and cross-view image conversion (Shi et al. 2019; Toker et al. 2021). However, the existing methods have the following problems. (1) Without capturing deep semantics, the majority of current approaches seek to segment feature maps or learn feature representations utilizing the complete image information, which leads to information loss. (2) Methods based on the Vision Transformer (ViT) (Dosovitskiy et al. 2020; Wang et al. 2023) have some

CONTACT Yue Yu michael-yue.yu@polyu.edu.hk

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10095020.2024.2439385

limitations regarding training data size and memory consumption, which might not be the best approach for drones that are available now (3) The CNN-based method (Bui, Kubo, and Sato 2022) can only focus on a smaller discrimination area because of the beneficial reception field's Gaussian distribution.

Inspired by previous research (Liu and Li 2019; Shen et al. 2023; Zhang and Zhu 2024), we put forward an innovative framework that achieved joint part-based learning of representations to solve this problem. We integrate spatial attention with the standard convolutional operation, which aims to make each set of part-based representation learning features spatially robust and well distributed. Meanwhile, the features of different branches are aggregated through the multi-level feature fusion module to prepare for cross-view matching. To further constrain these features, we designed a new joint hybrid function to help mine important information about global features. As shown in Figure 1, our approach addresses the shortcomings of VIT and CNN by utilizing the smallest possible memory and training data size to be able to focus on the context of the entire image. Our framework achieves advanced performance at University-1652 and SUES-200 datasets, with much lower computing costs (GFLOPs).

To summarize, this paper's primary contributions are as follows:

(1) We use spatial attention and standard convolutional operation, which achieve joint part-based learning of representations for drone-based matching.

(2) We present a hybrid function of cross-entropy loss and triplet loss metrics in performing inter-image similarity metrics, which improves the effective information utilization between image pairs.

(3) Our model achieves advanced performance on two publicly available datasets (University-1652 and SUES-200) at 89.79% and 95.75% in drone target matching and 94.87% and 98.80 in drone navigation.

2. Related work

2.1. Deep learning drone-based matching

In recent times, there has been notable advancement in cross-view matching through deep learning (Li et al. 2019; Liu et al. 2023) advances. There are two primary methods: the first one compares satellite and ground images, while the second one compares drone and satellite images. Image pairs for panoramic ground and satellite views are provided by CVUSA (Zhai et al. 2017) and CVACT (Liu and Li 2019). Moreover, a brand-new large-scale benchmark known as VIGOR (Zhu, Yang, and Chen 2021) has been suggested to close the gap between current datasets and real-world scenarios. A regular polar coordinate transformation was used by Shi et al. (2019) to distort satellite photos, bringing their domain closer to the panorama of the ground. To match the polar transform images more realistically, Toker et al. (2021) further trained a generative model based on true ground images. Ren et al. (2024) propose

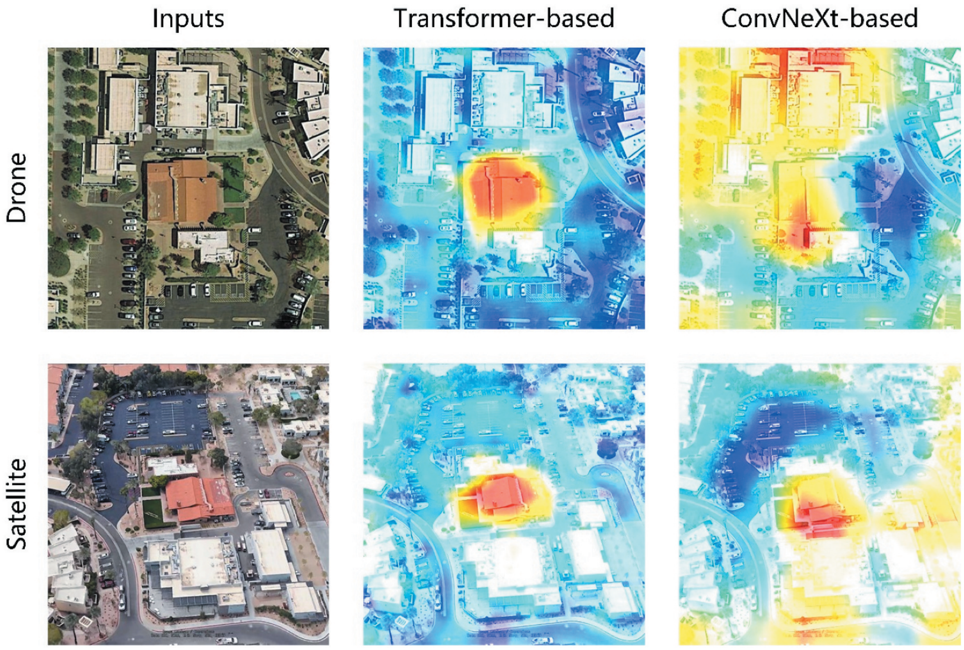

Figure 1. The Grad-cam heatmaps (Selvaraju et al. 2017). The input images from satellite and drone views are shown in the left column. The heatmap of the transform-based network is shown in the photos in the center column (Dai et al. 2022). The heatmap of our model, which is based on ConvNeXt, is shown in the figure in the right column.

a robust and accurate feature-matching method to weaken the radiation difference between data. Li, Qian, and Xia (2024) propose a new unsupervised framework that overcomes training limitations by utilizing initial pseudo labels and quicksort.

However, some research has started to look into using drone photos to make up for the difference in appearance between ground view and related satellite view photos. The most recent datasets are University-1652 (Zheng, Wei, and Yang 2020) and SUES-200 (Zhu et al. 2023), which include drone view photos of 1652 university buildings and 200 locations. Based on the drone view, two tasks are proposed: drone target matching and drone view navigation. To assess the quality of direction estimation, Hu et al. (2018) presented a set of assessment measures that use weighted soft margin sorting loss, which increases retrieval accuracy while simultaneously accelerating convergence speed. Ding et al. (2020) kept this network topology and extracted local features using context by applying fixed segmentation techniques, which made the features discriminative. Unit subtraction convolution was created by Lin et al. (2022) to automatically discover feature points without the need for extra annotations, enhance the model's ability to discriminate across regions using feature point information, and produce feature point attention masks. For drone-based matching, Dai et al. (2022) investigated a transformer-based feature separation and zone realignment technique. Ge et al. (2024) proposed a multi-branch joint representation learning network model based on an information fusion strategy.

2.2. Attention mechanism in drone-based matching

By distributing available resources according to weights, the attention mechanism can draw attention to key regions of an image. With new applications like person re-id (Chen et al. 2018), image recovery (Chen et al. 2022; Choi et al. 2020; Zhang et al. 2018), image classification (Roy et al. 2023; Wang et al. 2017), and geo-localization (Li et al. 2024; Shi et al. 2019), attention methods have also shown great promise in the field of computer vision in recent years. Altwaijry et al. (2016) introduced the spatial transformer module into the Siamese network to investigate a collection of potential matching patches to make use of the focus on objects of interest and patches. Bai et al. (2022) adopted a channel-based attention mechanism to achieve component-based representation learning through multi-level feature aggregation, as well as optional pooling strategies. Zhong and Wu (2024) propose a multi-branch spatial-spectral cross-attention module to fully mine and integrate detailed texture and semantic localization information. Gao et al. (2021) combining local and global attention mechanisms, encode information in image blocks to enable the model to learn substantive features in specific regions. To describe global dependencies, Yang, Lu, and Zhu (2021) developed cross-attention, which focuses more on identifying important fine-grained features from feature maps to provide a more discriminating visual representation. Wang et al. (2023) propose a cross-fusion framework to fuse the attention of satellite and drone views.

The above methods either rely on the polar transformer to narrow the domain gap or extract global features directly based on the whole image. Therefore, we propose a multi-level feature representation model for capturing high-level semantic features between images.

3. Methodology

3.1. Problem description and method summary

Given a dataset containing paired matching drone images and satellite images, we intend to train the model to obtain rich discriminative features and a hybrid function of cross-entropy loss and triplet loss is utilized to improve the effective information utilization between image pairs when performing the inter-image similarity metric. This model implements: 1) matching and retrieving the satellite images or slices by inputting the drone images and obtaining the location range information of the remote sensing images, and 2) retrieving drone images of the area from the input satellite images or slices.

3.1.1. Overview of method

As shown in Figure 2, the proposed method mainly consists of a feature extraction module and a multi-branch classifier module. The feature extraction module first extracts the feature maps of satellite images and drone images through the ConvNeXt network. Then it inputs the feature maps to the feature aggregation module to obtain the multi-level joint corresponding feature, and finally through the average pooling operation. The next is the multi-branch classifier module, where the extracted feature vectors are put into the classifier module to obtain the multi-level joint feature representation learning module.

3.2. ConvNext for geo-localization

We construct a ConvNeXt-based strong baseline for Drone-based matching, according to the general strong baseline for the University-1652 benchmark (Zheng, Wei, and Yang 2020). As in Figure 2 given the input images $\mathrm{I}_j \in R^{H \times W \times C}$. Here $H, W,$ and C denote the height, width, and channel numbers of I_j. The input is first processed by a depth-wise convolution to propagate information

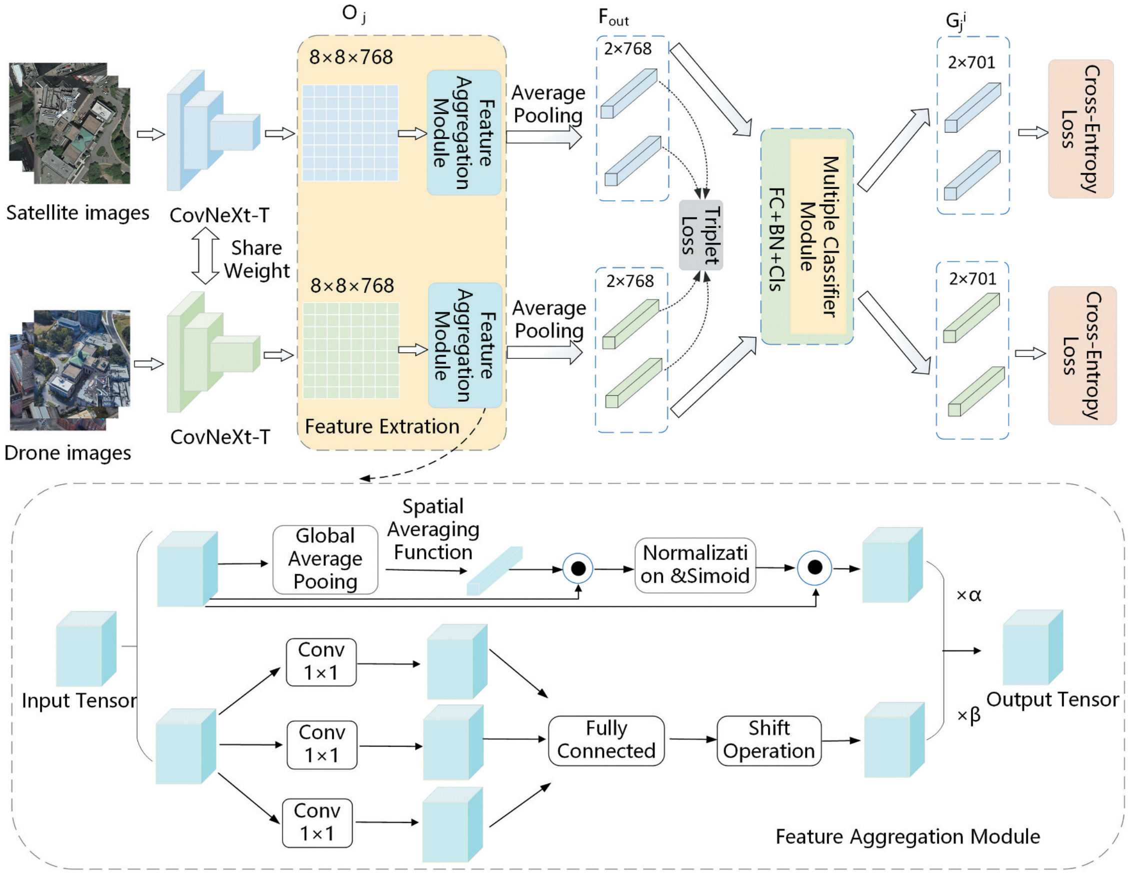

Figure 2. The pipeline of our proposed model. Given a sequence of input images, we first extract the feature maps from ConvNeXt and input them to the feature aggregation module, and then put the results to the average pooling layer to obtain the linear feature descriptions. Ultimately, the classifier module receives all of these feature descriptors and uses them to create a prediction vector that indicates how each segment will be geo-labeled. The details of the feature aggregation module are given in the lower left straight dashed box.

along spatial dimensions, producing features that are 1/4 the size of the input. Next, there are four stages: a ConvNeXt BlocksC_i and a Downsample layer, where $i = \{1, 2, 3, 4\}$.

ConvNeXt combines the advantages of ResNet and Inception models and adopts a split-transform -merge strategy, resulting in better performance than other networks. It utilizes skip connections, reduces the number of parameters, uses larger convolution kernels, and enhances training stability by substituting grouped convolutions and inverse bottleneck layers for regular convolutional layers. These strategies together build an efficient and effective network architecture.

3.3. Multi-level representation learning

Research on the impact of ConvNeXt in cross-view demonstrates that the robust baseline-based ConvNeXt performs well in drone-based matching. Nonetheless, the primary obstacles to be addressed continue to be the ambiguity surrounding position offset, separation, and scale. While it is crucial to extract robust global features with contextual relevance, numerous prior studies have also demonstrated that part-based features perform well for image retrieval.

Our approach acquires rich discriminative features through multilevel representation learning, allowing the model to learn the type to which each patch belongs. The process is described as follows.

3.3.1. Feature extraction

The following is an illustration of the ConvNeXt Layer $\mathcal{F}_{ConvNeXtLayer}$‘s output O_j:

$$O_j = \mathcal{F}_{ConvNeXtLayer}(I_j) \tag{1}$$

The final feature map O_j is handled by the feature aggregation module, which captures more robust and discriminative features, to provide rich part-based methods representations. The feature aggregation module is modified from spatial group-wise attention (SGE) (Li, Li, and Yang 2023) and convolution (Pan et al. 2022), which helps to achieve multi-level part-feature representation learning. To generate the matching feature map and distinct feature vectors, we replicate the original feature map into two branches, process one branch using spatially enhanced, and process the second branch using a mix of convolution and self-attention. Here is a representation of the process:

$$[F_{att}, F\,conv] = F(O_j) \tag{2}$$

$$F_{out} = \alpha F_{att} + \beta F_{conv} \tag{3}$$

where F_{att} and F_{conv} represent the results of spatial attention and convolution operations on the input tensor, respectively.

Then, the outputs of the two paths are added together, and the strength is controlled by two learnable scalars. Through this approach, the network can solve the problem of objects lacking unique features and enhance the effectiveness of feature encoders. Next, we perform a global average pooling operation on O_j and F_{out}, transforming the above feature maps into two $1 \times 1 \times C$-dim vectors of features. This is how the procedure is explained:

$$G_{I_j} = Averpool(F_{out}) \tag{4}$$

Among them, *Avepool* represents the global average pooling operation, and G_{I_j} represent the feature vectors corresponding to the input image I_j.

3.3.2. Multi-branch classifier module

Two feature vectors are created after feature extraction and fed into a dual-branch classifier to retrieve distinct characteristics. The layers that make up the classifier module are the Batch Normalization layer (BN), the Fully Connected layer (FC), and the Classification layer (Cl). By leveraging feature maps, it collects fine-grained features that improve the model's capacity to discriminate between instances. We employ cross-entropy loss to address this issue since it is not possible to use the features that are retrieved from several branches for matching directly.

3.4. Loss function

3.4.1. Cross-entropy loss based on cross-view

Since our dataset provides multiple images for each target position in cross-view tasks, we can train a classification model by considering each position as a class. Cross-entropy loss is frequently used in classification models to quantify the discrepancy between the model's anticipated probability and ground truth probability. We illustrate the cross-entropy loss in the following way:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}(\hat{q}(y|I_j)) \cdot \log(\hat{p}(y|I_j)) \tag{5}$$

where N is the quantity of classes, $\hat{q}(y|I_j)$ and $\hat{p}(y|I_j)$ is the ground truth probability and the anticipated probability that I_j is a member of geo-tag y.

3.4.2. Triplet loss based on cross-view

To make the model end-to-end, we will also use triplet loss when training the network. The objective is to maximize the distance from a preset boundary value between the anchored sample and the positive sample and minimize the distance between the anchored sample and the negative sample. The equation used for the triplet loss is as follows:

$$L_T = max(0, margin + d(a,p) - d(a,n)) \tag{6}$$

Where the Euclidean distance between samples x and y is represented by the symbol $d(x,y)$; the margin is a hyper-parameter used to specify the minimum interval between positive and negative samples.

Finally, outputs from both paths are added together and two learnable scalars control the strengths:

$$L_{total} = \alpha_1 L_{CE} + \alpha_2 L_T \tag{7}$$

4. Experiment and result analysis

4.1. Datasets and assessment criteria

Two large-scale cross-view geo-localization datasets, SUES-200 (Zhu et al. 2023) and University-1652 (Zheng, Wei, and Yang 2020), were used for the research. A selection of photos from the two datasets, each with a distinct angle, are displayed in Figure 3.

University-1652 (Zheng, Wei, and Yang 2020) is a multi-view multi-source dataset that includes images from Unnamed Aerial Vehicles, orbiting satellites, and ground cameras, covering 1652 constructions in 72 universities worldwide. There are two tasks supported by this dataset: drone target matching (Drone → Satellite) and drone navigation (Satellite → Drone), attempting to increase the precision of drone-based image matching and is widely used in geographic positioning research. The quantity of images from various viewpoints in the training and testing sets is displayed in Table 1. With 701 constructions from 33 institutions in the training set and 951 from 39 universities in the testing set, there is , generally, no overlap among the constructions in the dataset.

SUES-200 (Zhu et al. 2023) is a dataset with multiple heights and continuous scenes, which collected 200 positions from satellite and drone views. At every location, four distinct heights are captured by drone images: 150, 200, 250, and 300 meters. One satellite-view scene is equal to 50 drone-view photos at each altitude. A training set and a testing set comprising

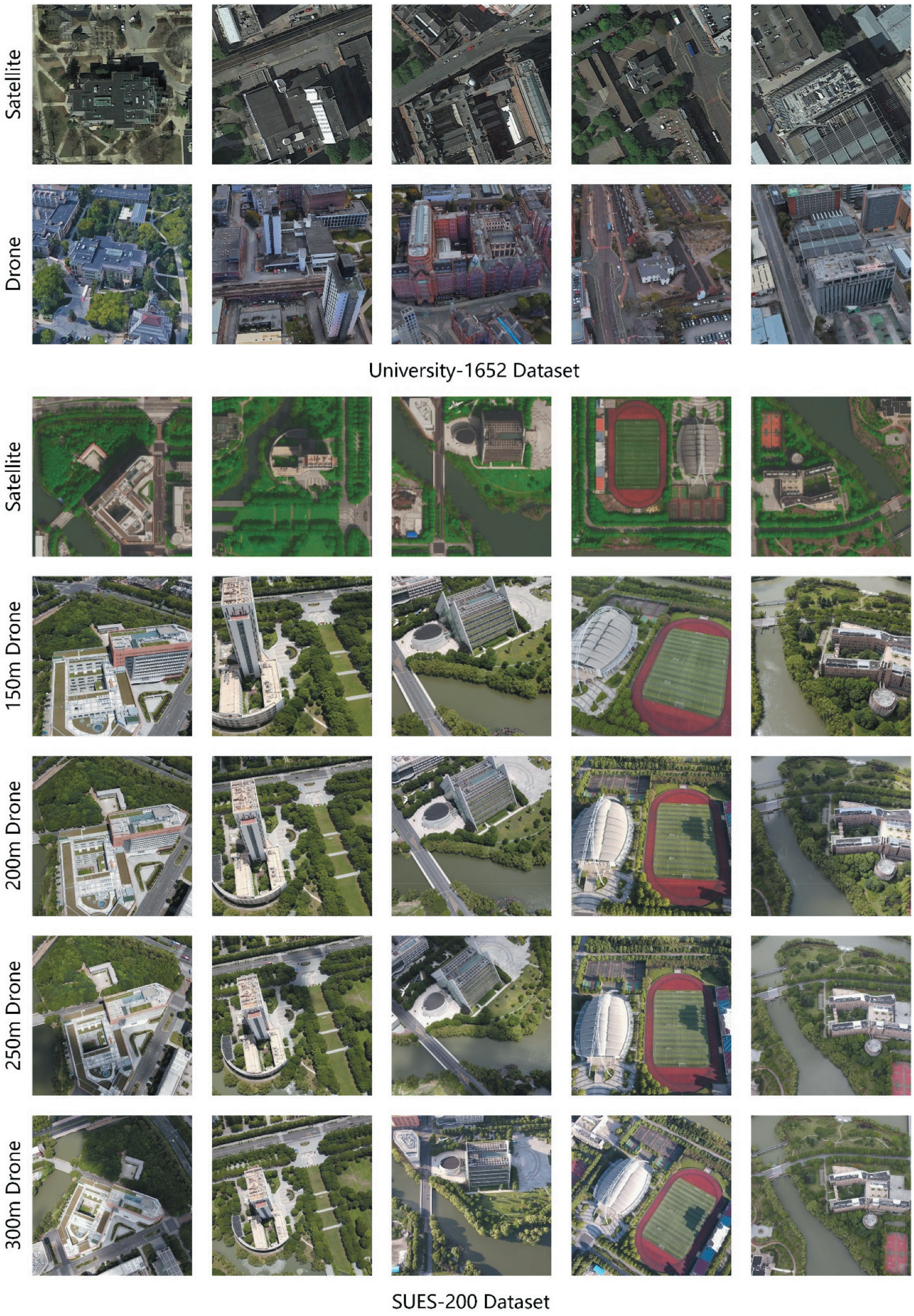

Figure 3. A few pictures of University-1652 and SUES-200 taken from various angles.

120 locations for training and 80 sites for testing were created.

4.1.1. Assessment criteria

We present the retrieval accuracy as top-k recall accuracy (R@K) and average precision (AP). R@K is the proportion of accurately identical images in the top-K of the rating list. A higher recall score indicates improved network performance. Furthermore, taking into account all ground-truth images in the gallery, AP indicates the region under the accuracy–recall curve. It also shows the average of the retrieval performance's accuracy and recall.

Table 1. The University-1652 dataset contains training and test sets of photos taken from various angles by ground-based and aerial vehicles, satellites, buildings, and university numbers.

Split	Views	Images	Classes	University
Train	Satellite	701	701	33
	Drone	37854	701	
	Ground	11640	701	
Test	Satellite Query	701	701	39
	Drone Query	37855	701	
	Ground Query	2579	701	
	Satellite Gallery	951	951	
	Drone Gallery	51355	51355	
	Ground Gallery	2921	793	

4.2. Experiment details

For obtaining visual features, we use the ConvNeXt-Tiny (Liu et al. 2022) on ImageNet (Deng et al. 2009) with pre-trained weights. To acquire a pair of fine-grained feature vectors, we next input the generated feature vectors into the feature aggregation module. We modify the Kaiming initialization in order to initialize the classifier block's parameters (He et al. 2015). During training and testing, we resized the input images to 256 × 256 while performing image enhancement. Because of the specificity of the task of cross-view matching, we also noticed that the number of images is usually different for different platforms due to the difficulty of capturing them. For example, we can easily acquire multiple drone images, while there is only one satellite image. Therefore, we give higher weight to the satellite image class and lower weight to the drone images, as a way to compensate for the imbalance between satellite and drone images.

We use Stochastic Gradient Descent (SGD) with a mini-batch of 8, momentum of 0.9, and weight decay of 0.0005 for our training. For the backbone layers and the remaining layers, the starting learning rate is 0.01. Our model is trained for 120 epochs, with a 0.1 decrease in the learning rate of all parameters between the 70 and 110 epochs. We compute the similarity between the query image and candidate images in the gallery using the Euclidean distance during testing. Our model is constructed upon the Pytorch framework, and three Nvidia GTX 3090Ti GPUs are used for every experiment.

4.3. Results analysis

4.3.1. University-1652 results

Table 2 illustrates that the proposed model attains 89.79% Recall@1 and 91.49% AP in the field of drone target matching (Drone → Satellite) and 94.87% Recall@1 and 89.71% AP in the field of drone navigation (Satellite → Drone). Drone and satellite images are the only training data used in any of our trials. The results have outperformed cutting-edge techniques like FSRA and MCCG. Recall@1 increases from 89.79% to 90.01% and AP increases from 91.49% to 91.78% in the task of (Drone → Satellite) when we scale the input photos to 384 × 384. In the task of (Satellite → Drone), the percentage of AP increases to 89.90% from 89.71%, and the result of Recall@1 climbs from 94.87% to 95.02%.

4.3.2. SUES-200 results

On the SUES-200, we also contrast our method with alternative approaches. Table 3 demonstrates that our solution obtains 83.05%, 88.65%, 94.05%, 95.07% Recall@1 and 86.00%, 90.81%, 95.02%, 96.30% AP at four altitudes in the job of (Drone → Satellite). Our model obtains 95.00%, 96.25%, 97.50%, 98.80% Recall@1 and 91.82%, 93.43%, 96.40%, and 97.06% AP at four altitudes in the challenge of (Satellite → Drone). The above observations show that our method has excellent and stable performance at different flight altitudes.

4.4. Ablation study

4.4.1. ConvNext vs. Transformer vs. ResNet

We compare the ConvNeXt network structure with Transformer- and Resnet-based networks. In terms of

Table 2. Compared to the state of the art on University-1652, it can be seen that the training set does not include the additional testing set that was gathered from Google images. S stands for the input image size. The best results are highlighted in bold. The drone target matching task is indicated by (Drone→satellite), and the drone navigation task is indicated by (Satellite→drone).

		Drone → Satellite		Satellite → Drone	
Backbone	Method	R@1	AP	R@1	AP
	Baseline(Zheng, Wei, and Yang 2020)	58.49	63.31	71.18	58.74
ResNet-50	LPN(Wang et al. 2022)	75.93	79.14	86.45	74.79
	MJRL(Ge et al. 2024)	86.06	88.08	91.44	85.73
	FSRA(Dai et al. 2022)	82.25	84.82	87.87	81.83
Vit-S	MSAM(Wang et al. 2023)	85.34	85.73	88.23	84.52
	TransFG(Zhao et al. 2024)	84.01	86.31	90.16	84.61
	MCCG(Shen et al. 2023)	89.28	91.01	94.29	89.29
ConvNeXt	Ours(s=256)	89.79	91.49	94.87	89.71
	Ours(s=384)	**90.01**	**91.78**	**95.02**	**89.90**

Table 3. Evaluating the performance of the SUES-200 at different altitudes.

	150m		200m		250m		300m	
Method	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP
Drone → Satellite								
SUES(Zhu et al. 2023)	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LPN(Wang et al. 2022)	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
FSRA(Dai et al. 2022)	68.25	73.45	83.00	85.99	90.68	92.27	91.95	93.46
MJRL(Ge et al. 2024)	77.57	81.30	85.50	89.73	92.58	93.21	92.96	94.21
Ours	83.05	86.00	89.65	91.81	94.05	95.62	95.75	96.30
Satellite → Drone								
SUES(Zhu et al. 2020)	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LPN(Wang et al. 2022)	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
FSRA(Dai et al. 2022)	83.75	76.67	90.00	85.34	93.75	90.17	95.00	92.03
MJRL(Ge et al. 2024)	93.75	79.49	94.03	90.52	95.35	92.43	97.50	94.62
Ours	95.00	91.82	96.25	93.43	97.50	96.40	98.80	97.06

Table 4. ResNet, transformer, and ConvNeXt inference time comparison. To ensure a fair comparison, every result is generated using the same apparatus. The best results are highlighted in bold.

		Drone → Satellite		Satellite → Drone	
Backbone	Inference time	R@1	AP	R@1	AP
ResNet-50	1x	68.93	72.31	82.61	67.69
ResNet-101	1.48x	72.33	75.32	85.44	70.43
Vit-S/16	1.21x	81.04	88.62	86.31	82.08
Vit-B/16	1.79x	83.50	89.35	88.93	84.26
ConvNeXt-T	**1.15x**	**89.79**	**91.49**	**94.87**	**89.71**
ConvNeXt-S	1.57x	89.90	91.62	95.42	90.37

R @ 1 and AP performance for two drone-based fields, Table 4 demonstrates the superior performance of the ConvNeXt network in comparison to both the ResNet and the vision transformer networks. The former is observed to achieve a performance gain of at least 20%, while the latter is found to exhibit a performance improvement of approximately 10%. Our network is faster in inference time compared to traditional Resnet and transformer methods. As in Figure 4 the multi-level representation learning allows the network to focus on fine-grained information, while the transformer-based approaches are limited in their ability to identify satellite elements in the context of the surrounding environment.

4.4.2. Effect of the SGE and convolution

We examine the efficacy of these two elements by taking one out of them, as indicated in Table 5. The proposed methodology yields a + 5% improvement in recall at 100% and a + 5% improvement in average precision on the field (Drone → Satellite) in comparison to the model without SGE and convolution. Additionally, it

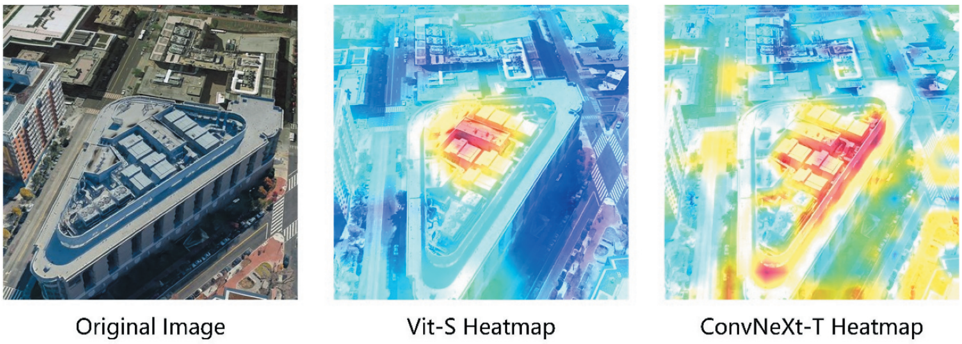

Figure 4. The ConvNeXt network and backbone network heatmaps. The initial picture is on the left, the heatmap of the final Vit-S layer is in the middle, and the heatmap of the final ConvNeXt-T layer is on the right.

Table 5. Effects of various modules on the dataset University-1652.

	Drone → Satellite		Satellite → Drone	
Method	R@1	AP	R@1	AP
w/o SGE and Convolution	83.45	85.97	91.44	83.83
w/o SGE	85.76	86.49	92.87	84.71
w/o Convolution	87.84	88.56	93.40	88.83
Ours	89.79	91.49	94.87	89.71

exhibits a + 3% improvement in recall and a + 5% improvement in average precision on the field (Satellite → Drone). We find that the performance is significantly decreased when any of these are removed, particularly the convolution. For a more discriminative feature representation, convolution enables the network to concentrate more on local feature extraction.

4.4.3. Robustness of position shifting

The goal of drone-based matching of the query image and the genuine matching image in the gallery typically shifts in real-world situations. As shown in Figure 5, we used FlipPad (FP) and BlackPad (BP) on the query image during testing to see if our model could handle the offset of the target matching in the actual matched image pairs. In particular, the query image is shifted from 20 pixels to the right, the black block of width 20 on the left side of the image is filled with BlackPad, and the piece of the image that is width 20 and filled is flipped to the right side. Table 6 displays the outcomes of the experiment.

4.4.4. Impact on the size of the input image

The input image size has a significant impact on the memory cost. While larger input images may be beneficial for network learning, they can also result in higher memory usage. Conversely, smaller input images may compress the detailed information of the original image, which may impact the discriminative representation learning process. Our objective is to alter the size of the input image while maintaining all other parameters constant in order to achieve a balance between memory usage and image size. When we increased the provided image resolution from 224 to 384, we saw a progressive improvement in performance in both tasks, as Table 7 illustrates. The more we increase the data input size to 512, the worse the performance becomes. With the help of this ablation experiment, we aim to determine the ideal input image size.

4.4.5. Effect of rotating images

The drone image's direction is random, but the University-1652 satellite image is directed northward. In this experiment, the gallery image is unaltered, and we merely rotate the question image. The query image was rotated in a range of 0° to 360°. As demonstrated in Table 8, where query and gallery images are rotated, the O° indicates that the image is not rotated, demonstrating how well our model adapts to rotation changes.

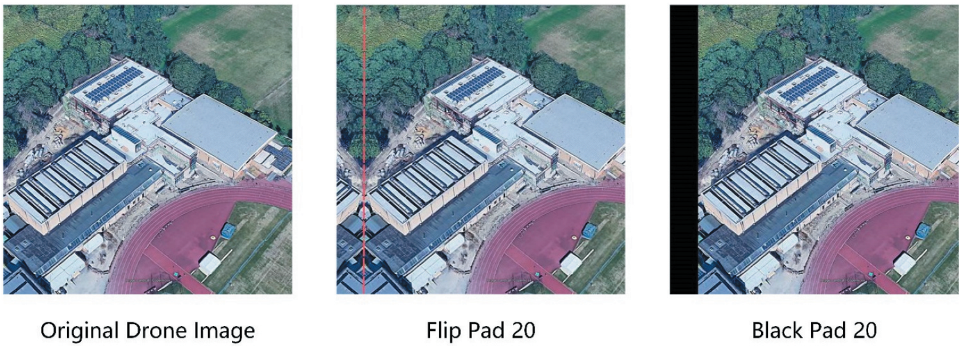

Figure 5. FlipPad (FP) and BlackPad (BP). FlipPad (FP) is a portion of the left side of the image with a width of 20 pixels that has been flipped after the image has been panned 20 pixels to the right. BlackPad (BP) is the left side of the image that has been expanded with black after the image has been panned 20 pixels to the right.

Table 6. For an ablation examination of moving query pictures during inference, the University-1652 dataset was utilized.

	Black Pad		Flip Pad	
	Drone → Satellite		Satellite → Drone	
Shifted Pixel	R@1	AP	R@1	AP
0	89.79	91.49	94.87	89.71
10	88.95	90.83	94.06	88.47
20	86.04	88.63	91.57	85.60

Table 7. The effects of various sizes of inputs on the dataset University-1652.

	Drone → Satellite		Satellite → Drone	
Image Size	R@1	AP	R@1	AP
224	85.65	89.53	93.52	86.51
256	89.79	91.49	94.87	89.71
384	90.01	91.78	95.02	89.90
512	89.50	91.32	93.67	89.71

Table 8. Ablation study of rotated images in the inference of the university-1652.

Rotation Angle		Drone → Satellite		Satellite → Drone	
Query	Gallery	R@1	AP	R@1	AP
0°	0°	89.79	91.49	94.87	89.71
65°	0°	86.52	88.62	92.84	86.63
90°	0°	83.45	85.97	91.44	83.83
180°	0°	86.87	87.65	91.87	84.25
63°	43°	88.94	91.05	93.48	88.70
267°	123°	84.72	86.06	91.67	84.46

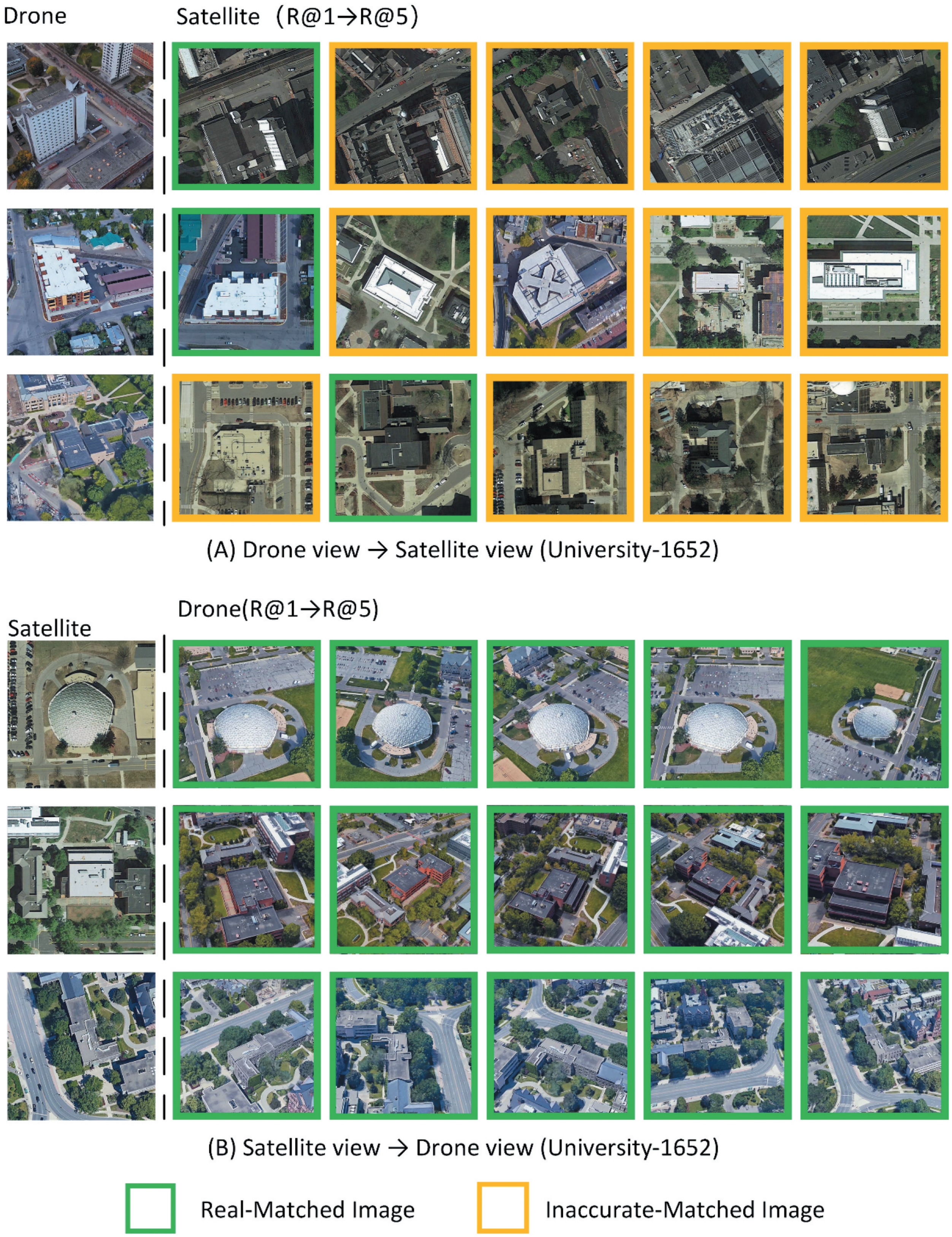
Drone
Satellite (R@1→R@5)
(A) Drone view → Satellite view (University-1652)
Satellite
Drone(R@1→R@5)
(B) Satellite view → Drone view (University-1652)
Real-Matched Image
Inaccurate-Matched Image

Figure 6. Qualitative outcomes of picture retrieval. (A) The top five retrieval outcomes for drone views' target matching on University-1652. (B) University-1652's top five retrieval outcomes for drone navigation. True matching images are marked by green boxes, and incorrect matching images are marked by yellow boxes.

4.5. Visualization of qualitative result

We display some retrieval outcomes in Figure 6 for the two fundamental fields on the University-1652 dataset. In both fields, acceptable images from the gallery sets can be retrieved using our model. Three drone view images are chosen at random from the test set for the drone view target localization job. The first five comparable satellite images from the gallery set are returned by the model for every drone image; the accurate findings are displayed in Figure 6(A). Additionally, a failure instance wherein one is unable to recall the matching image in top-1 is displayed in the third row of Figure 6(A). Three satellite images were selected at random from the test batch of images for the purpose of evaluating the drone navigation field. And then, the model delivers the first five drone images from the gallery collection that are similar to each satellite image. The accurate outcomes produced by our model are displayed in Figure 6(B).

5. Discussion and conclusions

This study developed a novel ConvNeXt-based multi-level representation learning model. Our model implements a comprehensive feature representation that extracts buildings, roads, and trees that contribute to matching, and the multi-branch classifier module enables the model to be more robust to positional offsets and scale changes. The experimental results reveal that our method achieves competitive matching accuracy results of 89.79% and 95.75% for drone target matching, and 94.87% and 98.80% for drone navigation. The results prove the effectiveness and adaptability of our model and highlight its potential in the geo-localization domain.

One limitation is the single-image training and query setting may bring some limitations for this task, which may lead to failure cases. In the future, we will use the generative model (Toker et al. 2021) or diffusion model (Xiao et al. 2023) to generate scarce satellite images to solve the failure cases and thus improve localization accuracy. Another limitation is that the complex network structure brings about an increase in the number of parameters and computations. Our future work will focus on developing a new end-to-end framework with a smaller number of parameters.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China [grant number 42401517], and was supported by the Open Fund of State Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University [grant number 23S01], and was supported by the Open Research Fund of Anhui Province Key Laboratory of Physical Geographic Environment, Chuzhou University [grant number 2023PGE001], and was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang [grant number GK249909299001-023], and was supported by the Zhiyuan Laboratory [grant number ZYL2024023]. And this work was also supported by Excellent Scientific Research and Innovation Team of Universities in Anhui Province [grant number 2023AH010071], Major Project on Natural Science Foundation of Universities in Anhui Province [grant number 2022AH040156], Excellent Young Scientists Project of Universities in Anhui province [grant number 2022AH030112], and Academic Foundation for Top Talents in Disciplines of Anhui Universities [grant number gxbjZD2022069].

Notes on contributors

Dr. Fangli Guan is currently a senior lecturer with the School of Computer Science, Hangzhou Dianzi University. His research interests include computer vision, AI model design and applications, embedded software application system.

Nan Zhao is currently pursuing his M.S. degree in College of Computer, Hangzhou Dianzi University. His research interests include deep learning, computer vision, and remote sensing.

Dr. Zhixiang Fang is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include crowd dynamics-oriented observation, human mobility, and intelligent navigation.

Dr. Ling Jiang is currently a Professor with the Anhui Province Key Laboratory of Physical Geographic Environment, Chuzhou University. His research interests include digital terrain modeling and analysis, urban scene geo-computing, and real-scene 3D modeling.

Dr. Jianhui Zhang is currently a Professor with the School of Computer Science, Hangzhou Dianzi University. His research interests include computer vision, AI model design and applications, embedded software application system.

Dr. Yue Yu is currently a postdoc in Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. His research interests include computer vision, Visual/LiDAR/IMU SLAM.

Dr. Haosheng Huang starts a tenure track professorship in Cartography and GIScience at Ghent University, Belgium, in February 2020. His research interests include location-based services (LBS), spatial cognition, computational mobility and activity analytics, and urban informatics. Please refer to https://users.ugent.be/~haohuang/ for more details.

ORCID

Fangli Guan http://orcid.org/0000-0001-7409-2129
Nan Zhao http://orcid.org/0009-0001-2951-4158

Jianhui Zhang http://orcid.org/0000-0002-0979-6514
Yue Yu http://orcid.org/0000-0003-3529-585X

Data availability statement

You can access the University-1652 dataset at https://github.com/layumi/University1652-Baseline/blob/master/Request.md. And the SUES-200 dataset is available at https://github.com/Reza-Zhu/SUES-200-Benchmark.

References

Altwaijry, H., E. Trulls, J. Hays, P. Fua, and S. Belongie. 2016. "Learning to Match Aerial Images with Deep Attentive Architectures." *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, United States, 3539–3547.

Arandjelovic, R., P. Gronat, A. Torii, T. Pajdla, and J. Sivic. 2016. "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition." *Paper presented at Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, United States, 5297–5307.

Bai, T., L. Wang, D. Yin, K. Sun, Y. Chen, W. Li, and D. Li. 2022. "Deep Learning for Change Detection in Remote Sensing: A Review." *Geo-Spatial Information Science* 26 (3): 262–288. https://doi.org/10.1080/10095020.2022.2085633.

Bui, D. V., M. Kubo, and H. Sato. 2022. "A Part-Aware Attention Neural Network for Cross-View Geo-Localization Between Uav and Satellite." *Journal of Robotics, Networking and Artificial Life* 9 (3): 275–284.

Chen, B., Q. Feng, B. Niu, F. Yan, B. Gao, J. Yang, J. Gong, and J. Liu. 2022. "Multi-Modal Fusion of Satellite and Street-View Images for Urban Village Classification Based on a Dual-Branch Deep Neural Network." *International Journal of Applied Earth Observation and Geoinformation* 109: 102794. https://doi.org/10.1016/j.jag.2022.102794.

Chen, D., S. Zhang, W. Ouyang, J. Yang, and Y. Tai. 2018. "Person Search via a Mask-Guided Two-Stream Cnn Model." *Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 734–750.

Choi, J., G. Sharma, M. Chandraker, and J. B. Huang. 2020. "Unsupervised and Semi-Supervised Domain Adaptation for Action Recognition from Drones." *Paper presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Snowmass, United States, 1717–1726.

Dai, M., J. Hu, J. Zhuang, and E. Zheng. 2022. "A Transformer-Based Feature Segmentation and Region Alignment Method for Uav-View Geo-Localization." *IEEE Transactions on Circuits and Systems for Video Technology* 32 (7): 4376–4389. https://doi.org/10.1109/TCSVT.2021.3135013.

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, United States, 248–255.

Ding, L., J. Zhou, L. Meng, and Z. Long. 2020. "A Practical Cross-View Image Matching Method Between UAV and Satellite for UAV-Based Geo-Localization." *Remote Sensing* 13 (1): 47. https://doi.org/10.3390/rs13010047.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al. 2020. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *Paper presented at the International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Gao, H., K. Xu, M. Cao, J. Xiao, Q. Xu, and Y. Yin. 2021. "The Deep Features and Attention Mechanism-Based Method to Dish Healthcare Under Social Iot Systems: An Empirical Study with a Hand-Deep Local–Global Net." *IEEE Transactions on Computational Social Systems* 9 (1): 336–347. https://doi.org/10.1109/TCSS.2021.3102591.

Ge, F., Y. Zhang, Y. Liu, G. Wang, S. Coleman, D. Kerr, and L. Wang. 2024. "Multibranch Joint Representation Learning Based on Information Fusion Strategy for Cross-View Geo-Localization." *IEEE Transactions on Geoscience & Remote Sensing* 62: 1–16. https://doi.org/10.1109/TGRS.2024.3378453.

Han, C., C. Wu, H. Guo, M. Hu, and H. Chen. 2023. "HANet: A Hierarchical Attention Network for Change Detection with Bitemporal Very-High-Resolution Remote Sensing Images." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 16: 3867–3878. https://doi.org/10.1109/JSTARS.2023.3264802.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification." *Paper presented at the Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 1026–1034.

Hodge, V. J., R. Hawkins, and R. Alexander. 2021. "Deep Reinforcement Learning for Drone Navigation Using Sensor Data." *Neural Computing & Applications* 33 (6): 2015–2033. https://doi.org/10.1007/s00521-020-05097-x.

Hu, S., M. Feng, R. M. Nguyen, and G. H. Lee. 2018. "Cvm-Net: Cross-View Matching Network for Image-Based Ground-To-Aerial Geo-Localization." *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, United States, 7258–7267.

Ji, S., S. Wei, and M. Lu. 2018. "Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set." *IEEE Transactions on Geoscience & Remote Sensing* 57 (1): 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

Li, D., F. Wang, F. Yang, and R. Dai. 2023. "Internet Intelligent Remote Sensing Scientific Experimental Satellite LuoJia3-01." *Geo-Spatial Information Science* 26 (3): 257–261. https://doi.org/10.1080/10095020.2023.2208472.

Li, G., M. Qian, and G. S. Xia. 2024. "Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, United States, 16719–16729.

Li, J., C. Yang, B. Qi, M. Zhu, and N. Wu. 2024. "4SCIG: A Four-Branch Framework to Reduce the Interference of Sky Area in Cross-View Image Geo-Localization." *IEEE Transactions on Geoscience & Remote Sensing* 62: 1–18. https://doi.org/10.1109/TGRS.2024.3504598.

Li, S., W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson. 2019. "Deep Learning for Hyperspectral Image Classification: An Overview." *IEEE Transactions on Geoscience & Remote Sensing* 57 (9): 6690–6709. https://doi.org/10.1109/TGRS.2019.2907932.

Li, Y., X. Li, and J. Yang. 2023. "Spatial Group-Wise Enhance: Enhancing Semantic Feature Learning in Cnn." *Paper presented at the Proceedings of the Asian Conference on Computer Vision*, Macau SAR, China, 687–702.

Lin, J., Z. Zheng, Z. Zhong, Z. Luo, S. Li, Y. Yang, and N. Sebe. 2022. "Joint Representation Learning and Keypoint Detection for Cross-View Geo-Localization." *IEEE Transactions on Image Processing*, 3780–3792.

Lin, T. Y., Y. Cui, S. Belongie, and J. Hays. 2015. "Learning Deep Representations for Ground-To-Aerial Geolocalization." *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, United States, 5007–5015.

Liu, J., J. Gao, S. Ji, C. Zeng, S. Zhang, and J. Gong. 2023. "Deep Learning Based Multi-View Stereo Matching and 3D Scene Reconstruction from Oblique Aerial Images." *Isprs Journal of Photogrammetry & Remote Sensing* 204:42–60. https://doi.org/10.1016/j.isprsjprs.2023.08.015.

Liu, L., and H. Li. 2019. "Lending Orientation to Neural Networks for Cross-View Geo-Localization." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, United States, 5624–5633.

Liu, Z., H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. 2022. "A Convnet for the 2020s." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, United States, 11976–11986.

Pan, X., C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang. 2022. "On the Integration of Self-Attention and Convolution." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, United States, 815–825.

Ren, K., W. Sun, X. Meng, G. Yang, J. Peng, B. Chen, and J. Li. 2024. "A Robust and Accurate Feature Matching Method for Multi-Modal Geographic Images Spatial Registration." *Geo-Spatial Information Science*: 1–20. https://doi.org/10.1080/10095020.2024.2354226.

Roy, S. K., A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot. 2023. "Multimodal Fusion Transformer for Remote Sensing Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 61:1–20.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization." *Paper presented at the Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 618–626.

Shen, T., Y. Wei, L. Kang, S. Wan, and Y. H. Yang. 2023. "MCCG: A ConvNext-Based Multiple-Classifier Method for Cross-View Geo-Localization." *IEEE Transactions on Circuits and Systems for Video Technology* 34 (3): 1456–1468.

Shi, Y., L. Liu, X. Yu, and H. Li. 2019. "Spatial-Aware Feature Aggregation for Image Based Cross-View Geo-Localization." *Advances in Neural Information Processing Systems* 32:10090–10100.

Tian, X., J. Shao, D. Ouyang, and H. T. Shen. 2021. "UAV-Satellite View Synthesis for Cross-View Geo-Localization." *IEEE Transactions on Circuits and Systems for Video Technology* 32 (7): 4804–4815. https://doi.org/10.1109/TCSVT.2021.3121987.

Toker, A., Q. Zhou, M. Maximov, and L. Leal-Taixé. 2021. "Coming Down to Earth: Satellite-To-Street View Synthesis for Geo-Localization." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* Nashville, United States, 6488–6497.

Wang, D., X. Chen, N. Guo, H. Yi, and Y. Li. 2023. "STCD: Efficient Siamese Transformers-Based Change Detection Method for Remote Sensing Images." *Geo-Spatial Information Science* 27 (4): 1192–1211. https://doi.org/10.1080/10095020.2022.2157762.

Wang, F., M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. 2017. "Residual Attention Network for Image Classification." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Honolulu, United States, 3156–3164.

Wang, T., Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang. 2022. "Each Part Matters: Local Patterns Facilitate Cross-View Geo-Localization." *IEEE Transactions on Circuits and Systems for Video Technology* 32 (2): 867–879. https://doi.org/10.1109/TCSVT.2021.3061265.

Wang, Y., Y. Xia, T. Lu, X. Zhang, and W. Yao. 2023. "An Efficient Method Based on Multi-View Semantic Alignment for Cross-View Geo-Localization." *Paper presented at the 2023 International Joint Conference on Neural Networks*, Gold Coast, Australia, 1–8.

Wei, H., and L. Wang. 2018. "Visual Navigation Using Projection of Spatial Right-Angle in Indoor Environment." *IEEE Transactions on Image Processing* 27 (7): 3164–3177. https://doi.org/10.1109/TIP.2018.2818931.

Xiao, Y., Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang. 2023. "EDiffSR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution." *IEEE Transactions on Geoscience & Remote Sensing* 62: 1–14. https://doi.org/10.1109/TGRS.2023.3341437.

Yang, H., X. Lu, and Y. Zhu. 2021. "Cross-View Geo-Localization with Layer-To-Layer Transformer." *Advances in Neural Information Processing Systems* 34: 29009–29020.

Zhai, M., Z. Bessinger, S. Workman, and N. Jacobs. 2017. "Predicting Ground-Level Scene Layout from Aerial Imagery." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Honolulu, United States, 867–875.

Zhang, H., G. Wang, Z. Lei, and J. N. Hwang. 2019. "Eye in the Sky: Drone-Based Object Tracking and 3d Localization." *Paper presented at the Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France, 899–907.

Zhang, Q., and Y. Zhu. 2024. "Aligning Geometric Spatial Layout in Cross-View Geo-Localization via Feature Recombination." *Paper Presented at the Proceedings of the AAAI Conference on Artificial Intelligence* 38 (7): 7251–7259. https://doi.org/10.1609/aaai.v38i7.28554.

Zhang, Y., K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. 2018. "Image Super-Resolution Using Very Deep Residual Channel Attention Networks." *Paper presenteds at the Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 286–301.

Zhao, H., K. Ren, T. Yue, C. Zhang, and S. Yuan. 2024. "TransFG: A Cross-View Geo-Localization of Satellite and UAVs Imagery Pipeline Using Transformer-Based Feature Aggregation and Gradient Guidance." *IEEE Transactions on Geoscience & Remote Sensing* 62: 1–12. https://doi.org/10.1109/TGRS.2024.3352418.

Zheng, Z., Y. Wei, and Y. Yang. 2020. "University-1652: A Multi-View Multi-Source Benchmark for Drone-Based Geo-Localization." *Paper presented at the Proceedings of the 28th ACM international conference on Multimedia*, Seattle, United States, 1395–1403.

Zheng, Z., L. Zheng, and Y. Yang. 2017. "A Discriminatively Learned Cnn Embedding for Person Reidentification." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14 (1): 1–20. https://doi.org/10.1145/3159171.

Zhong, H., and C. Wu. 2024. "T-UNet: Triplet UNet for Change Detection in High-Resolution Remote Sensing

Images." *Geo-Spatial Information Science*: 1–18. https://doi.org/10.1080/10095020.2024.2338224.

Zhu, P., J. Zheng, D. Du, L. Wen, Y. Sun, and Q. Hu. 2020. "Multi-Drone-Based Single Object Tracking with Agent Sharing Network." *IEEE Transactions on Circuits and Systems for Video Technology* 31 (10): 4058–4070. https://doi.org/10.1109/TCSVT.2020.3045747.

Zhu, R., L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu. 2023. "SUES-200: A Multi-Height Multi-Scene Cross-View Image Benchmark Across Drone and Satellite." *IEEE Transactions on Circuits and Systems for Video Technology* 33 (9): 4825–4839.

Zhu, S., T. Yang, and C. Chen. 2021. "Vigor: Cross-View Image Geo-Localization Beyond One-To-One Retrieval." *Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, United States, 3640–3649.

Zhuang, J., M. Dai, X. Chen, and E. Zheng. 2021. "A Faster and More Effective Cross-View Matching Method of Uav and Satellite Images for Uav Geolocalization." *Remote Sensing* 13 (19): 3979. https://doi.org/10.3390/rs13193979.