Leveraging Pretrained Diffusion Model for Semantic 3D Reconstruction from Monocular Remote Sensing Image

Xin Xu, Ruizhe Deng, Qinglong Cao, Zhiling Guo*, Yuntian Chen*, *Member, IEEE,* and Jinyue Yan*

***Abstract*—Semantic 3D reconstruction from monocular imagery serves as a cost-effective tool for many urban applications, such as energy system modeling, resilience analysis, and urban planning. However, the generalization of task-specific models for semantic 3D reconstruction remains limited by the available data scale and diversity. In contrast, visual foundation models (VFMs) are trained on large-scale, diverse datasets, enabling stronger adaptability and richer visual knowledge across different tasks. Unlike most VFMs that focus on discrimination or feature extraction, pretrained diffusion models (PDMs) are generative, combining high-level semantic understanding with the ability to produce high-fidelity details and textures. Building upon these advantages, this study proposes a novel task-adaptive framework that harnesses PDMs for semantic 3D reconstruction from monocular remote sensing images. Our framework employs low-rank adaptation to efficiently fine-tune the denoising network, effectively modeling the high-dimensional features required for semantic 3D reconstruction while only training a minimal fraction of parameters. We further design a lightweight, task-specific decoder to map these features into target elevation and semantic maps. In addition, we introduce an evidential height regression method, which incorporates uncertainty awareness into height estimation without introducing additional computational overhead. Experiments on the public US3D JAX and Open Data DC datasets demonstrate that our framework significantly outperforms other existing methods in both subtasks of height estimation and semantic segmentation, achieving high-fidelity semantic 3D reconstruction of remote sensing scenes. This technology holds significant potential for advancing urban modeling, enabling more accurate and efficient large-scale geographic analysis.**

***Index Terms*—Semantic 3D reconstruction, visual foundation models, pretrained diffusion model, task adaptation, low-rank adaptation.**

Manuscript received XXX XXX, 20XX; revised XXX XXX, 20XX. This work was supported in part by following projects: P0047700 - International Centre of Urban Energy Nexus; P0052733 - RISUD: Cutting-edge Solar Synergies Integrated with 3D Urban Environments towards a Carbon-Neutral City; P0052743 - MOST National Key R&D Program: Urban Photovoltaic System Planning Method Considering Carbon Footprint and Environmental Benefits; P0056532 - RICRI: IPT4U: Intelligent Platform & Toolbox for Urban Infrastructure Resilience.

X. Xu and R. Deng are with the Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China; Centre of Urban Energy Nexus, The Hong Kong Polytechnic University, Hong Kong SAR, China; The Zhejiang Key Laboratory of Industrial Intelligence and Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, P.R. China (e-mail: {xin666.xu, ruizhe.deng}@connect.polyu.hk).

Q. Cao is with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China, and also with the Zhejiang Key Laboratory of Industrial Intelligence and Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, P.R. China (e-mail: caoql2022@sjtu.edu.cn).

Z. Guo and J Yan are with the State Key Laboratory of Climate Resilience for Coastal Cities, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China; International Centre of Urban Energy Nexus, The Hong Kong Polytechnic University, Hong Kong SAR, China; Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: zhiling.guo, j-jerry.yan@polyu.edu.hk),

Y. Chen is with the Zhejiang Key Laboratory of Industrial Intelligence and Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, P.R. China (e-mail: ychen@eitech.edu.cn).

Corresponding author: Zhiling Guo, Yuntian Chen, Jinyue Yan

I. Introduction

SEMANTIC 3D reconstruction from a remote sensing perspective is an efficient and wide-reaching approach for urban geographic analysis, and has been widely applied in geographic information systems, environmental and energy monitoring, and digital twin technologies [1], [2], [3]. In this process, ground objects' height and semantic category information are the core elements for achieving a semantic 3D model. In particular, by estimating height information and performing semantic recognition, a detailed scene representation model can be constructed to support a wide range of downstream applications. Therefore, achieving semantic 3D reconstruction requires addressing two core sub-tasks: pixel-wise height estimation and semantic segmentation [4]. The goal of height estimation is to map the texture and structural features of a 2D remote sensing image to a height value for each pixel and, in combination with geospatial coordinates, reconstruct the complete 3D geometry. Semantic segmentation, on the other hand, is a pixel-level classification task that assigns each pixel a semantic label, such as building, road, or vegetation, by identifying the object categories present in the remote sensing image [5].

In remote sensing, data representing surface feature heights is typically stored in the form of digital surface models (DSMs). Traditional methods for acquiring DSMs mainly include airborne LiDAR and multi-view image reconstruction [6]. The former relies on expensive sensors mounted on specialized aerial platforms and requires complex post-processing to obtain high-precision results, with operational coverage generally limited to local areas. The latter achieves large-scale modeling by fusing images from different viewpoints and incorporating camera parameters, but it requires strict conditions for multi-view image pairing, including consistent illumination, precise camera calibration, and controlled acquisition timing. These constraints greatly limit the large-scale, cost-effective application of semantic 3D reconstruction [7]. Monocular vision-based methods have recently emerged as a promising alternative. By leveraging knowledge-driven neural networks to estimate height information from a single image, these methods eliminate the need for strict multi-view correspondence and reduce data acquisition costs [8]. As a

result, monocular approaches offer significant advantages for large-scale semantic 3D reconstruction and have become a major focus of current research.

With the advancement of deep learning, semantic segmentation—a technique for extracting semantic attributes of ground objects from remote sensing imagery—has reached a relatively high level of maturity [9]. In the context of semantic 3D reconstruction, joint modeling of height estimation and semantic segmentation generally follows two main approaches. The first employs separate networks for height estimation and semantic segmentation, enabling targeted optimization for each task but leading to redundant computation and inefficient use of data [10]. The second uses a unified network to perform both tasks simultaneously, typically by sharing a backbone to extract common features and adding task-specific output heads [11]. This shared architecture not only preserves the independence of each task but also facilitates feature-level information exchange and mutual reinforcement. As a result, height estimation benefits from the precise localization of semantic boundaries, while semantic classification takes advantage of height distribution patterns to improve category discrimination.

Previous CNN- and transformer-based studies have advanced semantic 3D reconstruction through targeted designs for height estimation and segmentation, yet their reliance on limited datasets constrains performance on complex scenarios. Recently, vision foundation models (VFMs) have emerged as a promising solution to this bottleneck. Such models are typically pretrained through self-supervised or unsupervised fundamental vision tasks, such as text–image feature alignment, image reconstruction, and generation, enabling them to acquire broader and more general prior knowledge by leveraging large-scale datasets, thereby providing richer feature representations for downstream tasks [12]. Applying VFMs to solve domain-specific problems has become a current research hotspot [13]. Although most VFMs were originally developed for general visual scenes, an increasing number of studies have demonstrated their great potential in core remote sensing tasks such as height estimation and semantic segmentation, laying a solid foundation for achieving high-precision, large-scale semantic 3D reconstruction.

Pretrained diffusion models (PDMs), represented by Stable Diffusion, have recently demonstrated remarkable ability to produce diverse, high-fidelity images with both coherent semantics and fine-grained textures [14]. Their core mechanism lies in a progressive denoising process: starting from random noise, a neural network iteratively refines the image under the guidance of conditional information, gradually recovering both global structures and local details [15], [16]. This generative process requires the model to capture high-level semantic concepts as well as low-level visual patterns. Given the ability to simultaneously capture high-level and low-level visual features, PDMs are particularly attractive for remote sensing semantic 3D reconstruction, where accurate pixel-wise height estimation and semantic segmentation require both global scene understanding and fine-grained spatial details [17]. However, DMs are originally trained for generative tasks on natural image datasets, and direct application to remote sensing image may lead to suboptimal feature extraction due to domain differences in scale, texture, and spectral characteristics [18]. This motivates the development of a dedicated adaptation framework to fully harness the prior knowledge of PDM for remote sensing sub-tasks. In this study, we take a pioneering step in exploring the potential of PDM for semantic 3D reconstruction, demonstrating how their rich generative priors can be effectively transferred to dense prediction tasks.

As illustrated in Fig. 1, the proposed framework is built upon the representative pretrained latent diffusion model architecture. It first encodes remote sensing images into latent representations, which are then adapted for dense prediction tasks through the low-rank adaptation strategy. This efficient fine-tuning approach enables the denoising network to shift from generating image priors to producing features suited for height estimation and semantic segmentation. The adapted network models the latent features into high-dimensional representations, which are mapped by the task-specific decoder into height and semantic spaces. In addition, shortcut connections between the latent encoder and the decoders preserve global contextual information while capturing fine-grained details, enabling accurate multi-task predictions. Together, these components support precise semantic 3D reconstruction from monocular remote sensing images. Beyond that, to improve the robustness of height estimation in complex scenarios, we introduce evidential height regression (EHR), which simultaneously predicts both the height values and their associated uncertainties. This approach adopts a distribution-constrained supervision strategy, enabling the model to estimate the expected height while maintaining prediction accuracy, and to output corresponding evidential parameters as quantitative indicators of prediction confidence. Overall, this study makes the following contributions.

- We propose a novel adaptation framework that leverages PDM for semantic 3D reconstruction from monocular remote sensing images. By incorporating low-rank adaptation and a specially designed task-specific decoder, the denoising network is efficiently fine-tuned to transform generative priors into high-quality features for dense prediction, thereby enabling accurate and efficient multi-task learning for both height estimation and semantic segmentation.
- We introduce an uncertainty-aware regression method, i.e., EHR, that jointly estimates height values and their associated uncertainties. EHR preserves estimation accuracy while providing quantitative confidence measures, thereby enhancing robustness in challenging remote sensing scenarios.
- We validate the effectiveness of the proposed framework through extensive experiments on two publicly available datasets, US3D JAX and Open Data DC, achieving state-of-the-art performance in both height estimation and semantic segmentation. This demonstrates that our framework can generate accurate semantic 3D reconstructions of scenes from monocular remote sensing images.

The remainder of this paper is organized as follows: Section II provides a review of related work. Section III details the pro-

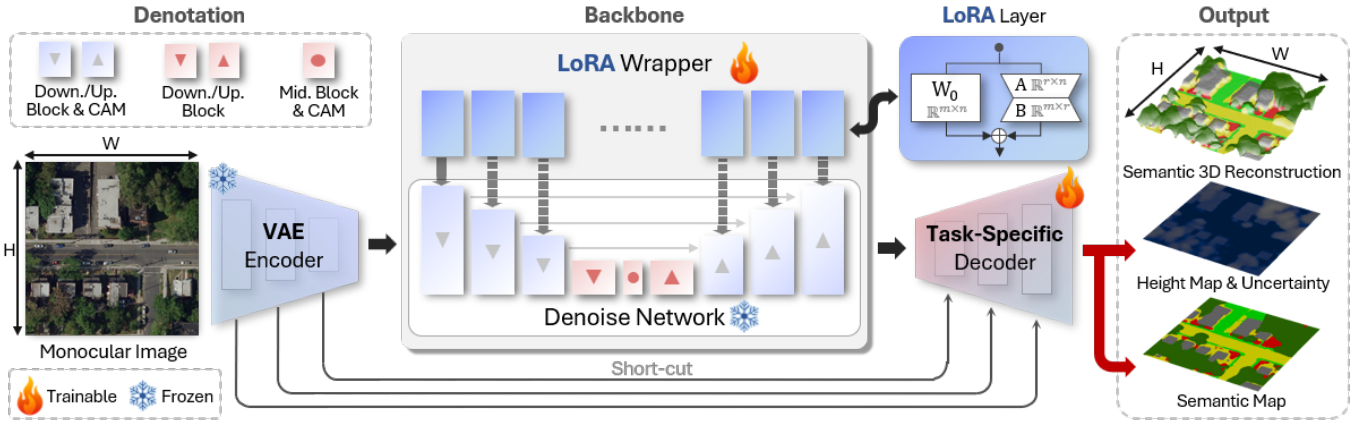

Fig. 1. Illustration of the proposed PDM adaptation framework. To leverage the prior knowledge embedded in the denoising network, we employ LoRA-based fine-tuning to facilitate task adaptation, thereby transforming the network's capability to effective feature modeling. Subsequently, the task-specific decoder is utilized to map the modeled features to precise semantic and height maps, enabling semantic 3D reconstruction.

posed framework. Section IV presents the experimental setup and implementation details. Section V reports and analyzes the experimental results. Finally, the conclusion is given in Section VI.

II. Related Work

A. Semantic 3D Reconstruction

Semantic 3D reconstruction is a joint modeling task that integrates 3D structure estimation with semantic recognition, aiming to accurately align semantic labels with the reconstructed geometry. Fundamentally, both image segmentation and dense 3D modeling from imagery are intrinsically ill-posed problems, lacking a unique and stable solution [19]. From an implementation perspective, semantic 3D reconstruction can be realized through either two separate modules or a coupled design for joint optimization. In aerial and satellite remote sensing scenarios, the horizontal coordinates of ground objects remain largely constant, making height estimation the primary challenge in reconstruction. Consequently, the task can be formulated as per-pixel height regression. While multi-view reconstruction imposes strict requirements on data acquisition, recent advances in monocular depth estimation (MDE) from the computer vision community have spurred the development of monocular height estimation (MHE) in remote sensing, providing an efficient end-to-end alternative [1].

Previous research has introduced numerous deep learning-driven methods for MHE, most of which are inspired by MDE. Mou et al. [20] proposed IM2HEIGHT, one of the first CNN models for MHE, which is based on an encoder-decoder architecture. Inspired by a module known as the progressive refinement module, originally used in MDE, Xing et al. [21] further developed a progressive learning network to gradually refine the coarse-scale height map by integrating multiscale features. To address the bias caused by the long-tailed distribution of height values between foreground and background pixels in MHE, Chen et al. [8] employed a classification-regression paradigm. They introduced head-tail cut and distribution-based constraints to mitigate the issue of distribution imbalance. Chen et al. [22] explored the application of vision transformer architectures to MHE and proposed the HeightFormer architecture that employs a multilevel interaction backbone alongside an adaptive height generator, substantially enhancing instance-level height estimation quality, particularly in preserving sharp object boundaries.

Semantic segmentation and MHE are both dense perception tasks, and previous studies have demonstrated that their joint modeling can facilitate mutual benefits, thereby improving overall performance [23]. Rao et al. [24] proposed the bidirectional guided attention network to promote effective information exchange and sharing between the two tasks. This network employs a shared backbone to extract unified features and introduces the Bidirectional guided attention module to enable cross-task feature interaction and fusion. Zhang et al. [23] designed an end-to-end multi-task network that incorporates a context-aggregation skip connection module to alleviate the semantic gap between the encoder and decoder. They further proposed a task-specific distillation module and a cross-task propagation module, which model task relationships using graph structures. The cross-task propagation module explicitly constructs local pattern graphs, also referred to as graphlets, and propagates high-order features across tasks. Gao et al. [11] were the first to introduce contrastive learning into the joint modeling of semantic segmentation and MHE, proposing cross-task contrastive loss and cross-pixel contrastive loss to maximize mutual information between tasks while enhancing intra-class feature consistency.

B. Vision Foundation Models

Although previous models designed specifically for semantic 3D reconstruction can achieve a promising performance on targeted tasks, they often rely on limited knowledge, struggling to handle complex scenes, resulting in restricted generalization capabilities. To address these limitations, VFMs have emerged in recent years. By pretraining on large-scale and diverse visual datasets, VFMs capture broader visual knowledge and general features, enabling robust adaptability and transferability across a wide range of downstream tasks. Notable examples include CLIP [25], DINO[26], and Stable Diffusion [15]. PDMs, exemplified by Stable Diffusion, represent a class of

generative VFMs built on the denoising diffusion probabilistic framework. Through a progressive denoising process, these models can generate diverse, high-fidelity images while simultaneously capturing both high-level semantic information and low-level spatial details.

Denoising Diffusion Probabilistic Models (DDPMs) provide the foundational framework for diffusion-based VFMs [27]. They are designed to learn the reverse of a diffusion process that incrementally degrades images with Gaussian noise. During inference, samples are generated by executing the reverse diffusion process, which incrementally denoises the data points. The denoising process of DDPMs is a Markov process, where each state is closely related to the previous state. This dependency necessitates multiple non-skippable iterations to complete the denoising. Based on DDPM, Song et al. [28] introduced a method called Denoising Diffusion Implicit Models (DDIMs), which accelerates the denoising process by employing a non-Markovian forward process. Conditional Diffusion Models are an extension of diffusion models that incorporate additional information into the diffusion process [29]. This allows the reverse process of the diffusion model to generate samples that conform to specific conditions, such as text prompts, semantic masks, or bounding boxes. Latent diffusion models [15] perform forward and reverse diffusion processes in the data latent space, resulting in more efficient computation.

In the realm of dense estimation, the diffusion model can also achieve state-of-the-art performance. Duan et al. designed a visual guidance conditional diffusion model for MDE, and achieved excellent results on both offline and online MDE evaluation with affordable inference costs [30]. Song et al. [31] proposed a novel single-step deterministic inference framework for MDE that incorporates a feature alignment module to balance generative and discriminative features, thereby mitigating overfitting to textural details during model training. On the other hand, Kolbeinsson et al. [32] further extended the application of diffusion to multi-class segmentation tasks. While VFMs provide strong general visual priors, effectively adapting them to specific remote sensing sub-tasks requires parameter-efficient techniques.

C. Task Adaptation

VFMs are primarily designed for large-scale pretraining on diverse datasets, typically using self-supervised or unsupervised approaches. While this enables broad visual knowledge acquisition, it also creates a gap when applying VFMs to complex downstream tasks. Addressing this gap has made task-specific adaptation a central focus in recent research. Efficient parameter update strategies can enhance the task adaptability of VFMs while avoiding redundant updates across all model parameters. For example, Hu et al. [33] introduced the low-rank adaptation (LoRA) method, representing a significant advancement in parameter-efficient fine-tuning of VFMs. Another line of research focuses on uncovering the latent capabilities of VFMs to maximize transfer performance. Rao et al. [34] proposed DenseCLIP, which reformulates text-guided image-level matching as a pixel-level matching problem. This approach effectively transfers the contrastive learning priors of CLIP, trained on image-text pairs, to dense visual estimation tasks. Although most VFMs are developed for general visual domains, recent studies have demonstrated their applicability to remote sensing tasks. Liu et al. [35] leveraged CLIP's text encoding capability to learn prototype representations from textual descriptions, enabling few-shot learning for remote sensing object detection. Cao et al. [36] proposed a domain-controlled prompt learning framework, where a lightweight model transfers essential domain-specific knowledge into domain biases. This mechanism jointly regulates both the visual and language branches to generate domain-adaptive prompts for CLIP.

As previously discussed, PDMs not only demonstrate powerful image generation capabilities but also show significant potential in downstream tasks. Tian et al. [37] analyzed the intermediate feature representations of PDMs and found that their unconditioned self-attention layers encode specific intra- and inter-attention similarities, which can be leveraged for unsupervised zero-shot segmentation. This indicates that, although PDMs are trained to estimate noise distributions, they can also capture general visual knowledge that can be transferred to other tasks. In the VDP method proposed by Zhao et al. [17], the denoising network of the PDMs is employed as a robust feature extractor, further demonstrating its applicability to dense visual perception tasks. More recently, Ke et al. [38] introduced Marigold, a method that applies minimal low-cost fine-tuning to successfully adapt PDMs for MDE in real-world scenarios. Although existing studies demonstrate the strong potential of PDMs in related vision tasks, their application to remote sensing sub-tasks has not yet been thoroughly explored.

III. Methodology

A. Preliminaries of Pretrained Diffusion Models

Intuitively, the diffusion model mechanism can be regarded as a Markovian diffusion process that progressively refines an initial random noise x_T into a noise-free image x_0. For any intermediate state x_t at timestep t (where T ranges from T to 0), its relationship with the preceding state can be expressed as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{a_t}x_{t-1}, \sqrt{1-a_t}I\right), \tag{1}$$

where a_t is a noise schedule that controls the amount of noise added at each timestep. By aggregating the process over t steps and leveraging the additive property of the noise distribution, we can derive the representation of any intermediate state at timestep t with respect to x_0:

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{a}_t}z_0, (1-\bar{a}_t)I\right). \tag{2}$$

By expanding the above distribution, we obtain a closed-form expression for the transition from x_0 to x_t:

$$x_t = \sqrt{\bar{a}_t}x_0 + (1-\bar{a}_t)\epsilon, \tag{3}$$

where ϵ denotes Gaussian noise that is randomly sampled, and $\bar{a}_t = \Pi_{s=1}^{t} a_t$. In the research community, the transition from $x_0 \rightarrow x_t$ is referred to as the forward process, while the transition from $x_t \rightarrow x_0$ is known as the reverse process. The

reverse process, $q(x_{t-1}|x_t)$, serves as the denoising procedure for progressively generating images. This process can be formulated as the following parameterized equation:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right). \quad (4)$$

The sampling process of diffusion models can be computed using discrete formulations of either diffusion stochastic differential equations (SDEs) or ordinary differential equations (ODEs). In the reverse process, x_{t-1} can be approximated using the following equation:

$$x_{t-1} = \frac{1}{\sqrt{a_t}}\left(x_t - \frac{1-a_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t \mathcal{N}(0, I), \quad (5)$$

where $\epsilon_\theta(x_t, t)$ is predicted by a parameterized denoising network, which takes as input the timestep and its corresponding noisy image.

Stable Diffusion (SD) is a well-known PDM that demonstrates state-of-the-art capabilities in both text-to-image and image-to-image generation [15]. As a two-stage diffusion model, SD first uses a variational auto-encoder (VAE) to map the input image to a latent space, and then uses a denoising network to iteratively refine the latent code. Given an image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ in the 3-channel image space, the encoder and decoder of the VAE are denoted by $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$, respectively. The latent representation z is obtained by the encoder for perceptual image compression, mapping the image to a lower-resolution feature space. The decoder is accountable for reconstructing the latent representation z back into an image in the RGB space as $\tilde{\mathbf{X}}$. The following equations are responsible for encoding and decoding, respectively:

$$z = \mathcal{E}(\mathbf{X}),\ z \in \mathbb{R}^{h\times w\times c}. \quad (6)$$

$$\tilde{\mathbf{X}} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(\mathbf{X})),\ \tilde{\mathbf{X}} \in \mathbb{R}^{H\times W\times 3}. \quad (7)$$

Specifically, $h = H/f$ and $w = W/f$, where f denotes the downsampling factor. The pretraining process of the VAE was supervised by a perceptual loss, a patch-based adversarial loss, a pixel-space loss, and a distribution regularization term, enabling it to preserve local realism and enhance reconstruction fidelity. SD leverages the latent space extracted by the VAE, which is perceptually equivalent to the image space, thereby significantly reducing the image resolution within the scope of the diffusion model. This approach substantially lowers computational complexity and training costs. In Fig. 2, we evaluate the performance of the VAE in encoding and subsequently decoding aerial images and height maps. It can be observed that although the reconstructed images generated by the VAE can preserve most of the original details for both modalities, there remain inherent errors in the VAE reconstructions of height maps.

The conditional U-Net is adopted as the denoising network in SD and uses the conditioning mechanism to guide the synthesis of specific images along designated pathways. This U-Net employs a cross-attention mechanism (CAM) for conditional interaction, enabling it to accept either images or text as input. The predicted noise of conditional denoising network can be represented as $\epsilon_\theta(x_t, t, \mathbf{C})$, where the conditional input $\mathbf{C}$ is transformed into an intermediate representation

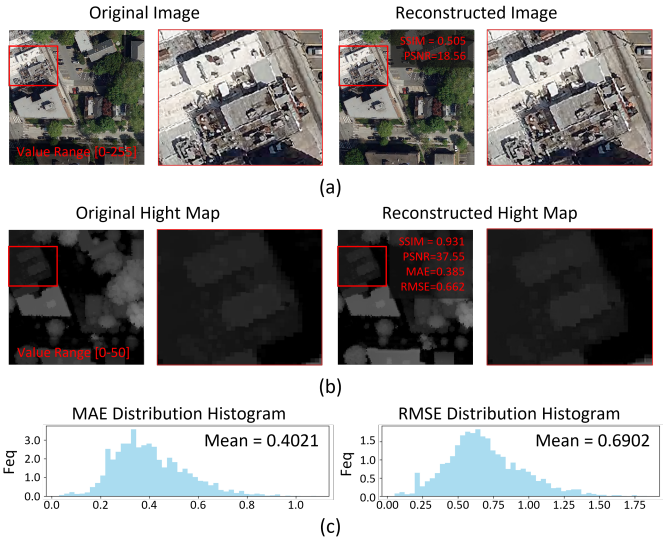

Fig. 2. Perceptual compression evaluation of pretrained VAE. (a) and (b) present example samples of aerial images and height maps, respectively, that have been compressed and reconstructed using the pretrained VAE; (c) illustrates the distribution of reconstruction accuracy loss, i.e mean absolute error (MAE) and root mean square error (RMSE), for height maps

$\tau_\theta(\mathbf{C}) \in \mathbb{R}^{M\times d_\tau}$ via a domain encoder τ_θ. The CAM is densely embedded at every stage of the denoising conditional U-Net, ensuring that the features at each stage are generated with the involvement of the conditional information. This design guarantees that the final generated results meet the specified requirements.

B. Overview of The Proposed Framework

To enhance the robustness of generation, SD has been extensively trained on the large-scale text-image pair dataset Laion-5B [39], enabling it to implicitly learn valuable and rich high-level and low-level visual representations from massive image-text pairs. Numerous studies have demonstrated that the general visual-linguistic knowledge acquired from such large-scale data can facilitate various sub-tasks in the field of remote sensing. It is important to emphasize, however, that unlike other VFMs, such as CLIP, SAM, or DINO, the SD learns the marginal distribution $\nabla_{z_t} \log p(z_t|y)$ of data in the latent feature space, rather than high-dimensional vector representations of images. To leverage the pretrained SD model for specific remote sensing sub-tasks, two key challenges must be addressed. The first is how to fully exploit the prior knowledge of the well-trained SD model, given the differences between the generative paradigm in general vision scenes of the diffusion processing and the dense regression (e.g., height estimation) and discrete classification (e.g., semantic segmentation) in the remote sensing domain. The second challenge is how to minimize training costs to ensure the feasibility of transferring the prior knowledge of the pretrained model to sub-tasks. To address these challenges, we propose a novel adaptation framework, as illustrated in Fig. 1, to exploit abundant prior knowledge of the pretrained SD model for tackling the semantic 3D reconstruction task.

The original random denoising process can be regarded as a form of stochastic generation. We remove the additional

noise input and instead utilize the latent features extracted from the initial image via the VAE encoder as the input to the denoising network. This approach is intended to ensure consistency in the distribution of latent features. As noise input is no longer necessary, the denoising process operates in a single-pass mode. Additionally, we redesigned a lightweight task decoder, on this basis, proposed a task-specific decoder (TSD) $\mathcal{D}(\cdot)$, as shown in Fig.3, to enhance domain-specific supervision and reduce the potential impact of the VAE's inherent errors on overall performance. Thus, for given an image $\mathbf{X}$, the forward process can be basicly represented as:

$$\tilde{y} = \mathcal{D}(\mathcal{F}(\mathcal{E}(\mathbf{X}), t^*, \mathbf{C}')), \tag{12}$$

where $\mathcal{F}(\cdot)$ denotes the denoising network, t^* represents the default fixed timestep, and $\mathbf{C}'$ refers to the condition information for guidance.

C. Parameter-Efficient Adaptation with Low-Rank Adaptation

We aim for the denoising network to directly sample a distribution from the samples for task prediction within the decoder, rather than merely refining the input. Therefore, task-adaptive fine-tuning is required. To this end, we employ LoRA for fine-tuning, which significantly reduces the number of trainable parameters. LoRA is an efficient fine-tuning technique designed to adapt large-scale language and vision models to downstream tasks [33]. The core premise of LoRA is that the weight updates (ΔW) applied to the pretrained model weights ($W_0 \in \mathbb{R}^{m\times n}$) exhibit a low intrinsic rank. Consequently, ΔW can be factorized into two low-rank matrices, $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$, such that $\Delta W = BA$, as shown in the top-right subfigure of Fig. 1. Here, the rank (r) is significantly smaller than the minimum dimension of W_0 (i.e., $r \ll \min(m, n)$). During fine-tuning, only the parameters of A and B are optimized to approximate ΔW, while W_0 remains frozen, thereby reducing computational overhead. During the forward pass, the computation of each hidden state, originally defined as $h = W_0 x$, is modified as follows:

$$h = W_0 x + \Delta W x = W_0 x + BAx. \tag{13}$$

Previous studies have demonstrated that LoRA requires fine-tuning only a small fraction of parameters to effectively adapt the base model to downstream tasks [40]. Moreover, LoRA does not introduce additional computational overhead during inference, as the LoRA weights can be merged with the pretrained weights at deployment by computing the new combined weights:

$$W' = W_0 + \lambda/rAB, \tag{14}$$

where λ is a hyperparameter that controls the contribution of the LoRA weights in the combined weights. In this work, we encapsulate the denoising network with a LoRA wrapper and fine-tune its convolutional and attention layers accordingly.

D. Task-specific Decoder

Due to the discrepancy between the supervision method used by the pretrained VAE decoder reconstruction loss and the task objective, the TSD is designed to reconstruct the latent features and transform them into specific prediction targets, allowing the incorporation of task-specific loss functions for optimization. The structure of the TSD is shown in Fig. 3 (c). Its basic body consists of two upsampling modules using deconvolution units. To meet the requirements of semantic 3D reconstruction, the TSD can simultaneously include a height estimation head and a semantic segmentation head. It is worth noting that, in the same image, there is an explicit relationship between the category and the height value of the target. Height information can serve as a basis for distinguishing target categories, such as differentiating buildings from the ground; conversely, the target category can also serve as a basis for height estimation. In addition, since both height regression and semantic segmentation are dense estimation tasks and utilize similar features, deep decoupling is not required. Therefore, in our design, the height estimation and semantic segmentation heads share the features of the TSD. The pyramid scene parsing (PSP) module is used to enhance the model's ability to utilize global contextual information. Incorporating both global and sub-region context helps distinguish between different categories, which is particularly beneficial for cross-scale perception, as it enables the model to accurately identify objects of varying scales by leveraging a larger receptive field. Given a feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, the process can be described as follows:

$$\mathcal{P}_s = \text{Upsampling}(\mathbf{Conv}_{1\times1}(\text{AVEPooling}(\mathbf{F}, s))), \tag{15}$$

$$\mathbf{F}' = \text{Concat}\left([\mathcal{P}_{s_1}(\mathbf{F}), \mathcal{P}_{s_2}(\mathbf{F}), \mathcal{P}_{s_3}(\mathbf{F}), \mathcal{P}_{s_4}(\mathbf{F}), \mathbf{F}]\right), \tag{16}$$

where Upsampling$(\cdot)$, AVEPooling$(\cdot, s)$, and Concat$(\cdot)$ denote the upsampling, average pooling, and concatenation opera-

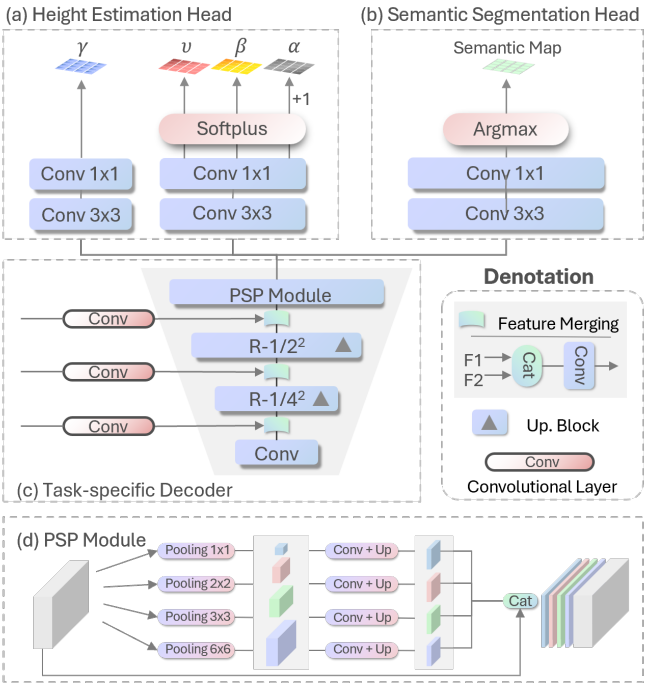

Fig. 3. Diagrams of the TSD and relative modules designing. (a) and (b) show the details of the height estimation head and the semantic segmentation head, respectively; (c) shows the structure of the TSD, where R-$1/N^2$ represents the resolution scale of the feature map; (d) is an illustration of the PSP Module.

tions, respectively, where s represents the scale parameter of the pooling layer.

1) Height Estimation Head: For the height estimation, we employ evidential height regression to estimate both the expectation and uncertainty of pixel height values. We assume that the target height value is independently and identically distributed according to a Gaussian distribution $q(\mu, \sigma^2)$, with the following Gaussian prior and Inverse-Gamma prior imposed on its unknown mean μ and variance σ^2:

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1}), \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \tag{17}$$

where $\gamma \in \mathbb{R}$, $\alpha > 1$, $\upsilon > 0$ and $\beta > 0$. The posterior distribution of the mean and variance can be factorized as $q(\mu)q(\sigma^2)$, its approximate form can then be derived as a Normal Inverse-Gamma (NIG) distribution:

$$p(\theta|\mathbf{m}) = \frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(a)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\gamma^2}\right\}, \tag{18}$$

where $\theta = (\mu, \sigma^2)$ and $\mathbf{m} = (\gamma, \upsilon, \alpha, \beta)$.

The target distribution can be estimated by sampling from the NIG distribution parameterized by $\mathbf{m}$. Consequently, $\mathbf{m}$ not only characterizes the expected value of the target, but also quantifies its dispersion, i.e., the associated uncertainty. The NIG distribution can be interpreted as a higher-order evidential distribution over a Gaussian distribution. Furthermore, given a specific NIG distribution, the prediction of height can be represented by $\mathbb{E}(\mu)$, while $\mathbb{E}(\sigma^2)$ and $Var(\sigma^2)$ correspond to the aleatoric and epistemic uncertainties, respectively:

$$\mathbb{E}(\mu) = \gamma, \quad \mathbb{E}(\sigma^2) = \frac{\beta}{\alpha - 1}, \quad Var(\sigma^2) = \frac{\beta}{\upsilon(\alpha-1)}. \tag{19}$$

According to the Bayesian probability theory, the likelihood estimation for the predicted h corresponds to the marginal distribution of θ under the evidential distribution $\mathbf{m}$. As derived in [41], this marginal distribution follows a Student-t distribution:

$$\begin{aligned} p(h|\mathbf{m}) &= \frac{p(h|\theta,\mathbf{m})p(\theta|\mathbf{m})}{p(\theta|h,\mathbf{m})} \\ &= \int_{\sigma^2=0}^{\infty}\int_{\mu=-\infty}^{\infty} p(h|\mu,\sigma^2)p(\mu,\sigma^2|\mathbf{m})\, d\mu\, d\sigma^2 \\ &= St\left(h; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\alpha}, 2\alpha\right). \end{aligned} \tag{20}$$

Subsequently, we can train the model by minimizing the negative log-likelihood (NLL) loss $\mathcal{L}_{NLL}$ of this distribution, thereby encouraging the model to fit the maximum likelihood estimation of the prediction by maximizing the model evidence, which is:

$$\begin{aligned} \mathcal{L}_{NLL} = &\frac{1}{2}\log\left(\frac{\pi}{\upsilon}\right) - \alpha\log(\Omega) \quad \backslash \\ &+ (\alpha + \frac{1}{2})\log((h-\gamma)^2\upsilon + \Omega) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}\right), \end{aligned} \tag{21}$$

where $\Omega = 2\beta(1+\upsilon)$. We employ the regularization loss $\mathcal{L}_R$ to ensure the minimization of erroneous predictions while simultaneously preventing gradient vanishing in cases of high-uncertainty predictions:

$$\mathcal{L}_R = |h - \gamma|(2\epsilon + \alpha), \tag{22}$$

where the first term represents the error scale, while the second term corresponds to the evidence scale, which is always greater than one. To prevent evidence inflation that may occur when optimizing $\mathcal{L}_{NLL}$ alone, the role of $\mathcal{L}_R$ is to penalize overconfidence—i.e., excessively high evidence values—when the prediction error increases. It should be emphasized that, in our work, the height expectation γ is not output in conjunction with other evidential parameters, but is instead predicted through an independent branch and supervised using the mean squared error (MSE) loss:

$$\mathcal{L}_{MSE} = (h - \gamma)^2. \tag{23}$$

Although γ participates in the calculation of the NLL loss and its regularization loss, gradients are not computed with respect to γ. The purpose of this approach is to reduce interference from other parameters during the convergence process of height prediction. In summary, the total loss for evidential height regression is denoted as:

$$L_{HR} = \mathcal{L}_{MSE} + \mathcal{L}_{NLL} + \lambda\mathcal{L}_R, \tag{24}$$

where λ is the weighting coefficient for the regularization loss.

2) Semantic Segmentation Head: For the semantic estimation, we employ a convolutional layer to produce per-class activation maps for pixel classification. The cross-entropy (CE) loss is used to supervise the semantic segmentation task:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right), \tag{25}$$

where y_i and $\hat{y}_i$ are the ground truth and predicted semantic probabilities, respectively.

IV. Experiments

A. Datasets

To evaluate the performance of the proposed framework for semantic 3D reconstruction, we conducted experiments on two public datasets: US3D JAX and Open Data DC.

- **US3D JAX** is a subset of the US3D dataset, which includes aerial images of the United States cities of Jacksonville [42]. The image's original size is uniformly 2048×2048. It also provides semantic maps with five categories and nDSM data. The dataset was cropped into non-overlapping 512×512 patches, yielding a total of 17,568 sample pairs for the training set and 1,920 sample pairs for the test set.
- **Open Data DC** is derived from publicly available geospatial data of Washington DC of the United States encompassing nDSM data, orthorectified aerial images, as well as building footprints and urban tree canopy distribution maps, where the building footprints and urban tree canopy serve as semantic labels [43]. The dataset was cropped into non-overlapping 512×512 patches, yielding a total of 12,066 sample pairs for the training set and 3,015 sample pairs for the test set.

In Appendix. A, we provide a more detailed feature analysis of the two datasets.

B. Evaluation Metrics

We used the following numerical metrics to evaluate the quality of the height estimation: mean absolute error (MAE), root mean squared error (RMSE), and δ_n accuracy:

$$\text{MAE} = \sum_{i=1}^{M} |P_i - GT_i|, \tag{26}$$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(P_i - GT_i)^2}, \tag{27}$$

$$\delta_n = \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(\max(P_i/GT_i, GT_i/P_i) < 1.25^n), \tag{28}$$

For semantic segmentation, we use intersection over union (IoU), accuracy (Acc) to evaluate the performance of the model. Additionally, those metrics with notation of "*-A*" indicates the prediction results for the entire image, "*-F*" indicates the prediction results for the foreground part (i.e., buildings), and "*-B*" indicates the prediction results for the background part. Both MAE and RMSE are computed in the units of the original data, i.e. meters, whereas δ_n is expressed in decimal form of percentage.

C. Implementation Details

The pretrained base model of Stable Diffusion v2.0[1] (SD-v2.0) was employed to conduct most experiments on a single Nvidia A100 GPU with 80 GB of memory. The structure of the denoising network remained unchanged, except for the training encapsulation using the LoRA wrapper. The AdamW optimizer was adopted to update the gradients of trainable modules, with a learning rate of 3×10^{-4} and a weight decay rate of 0.1. The weight λ of the regularization loss in height estimation loss was set to 0.2. Training was conducted for a maximum of 40k iterations with a batch size of 4. The number of channels per layer in the TSD is configured as 128.

V. Results and Discussions

A. Comparion with Existing Methods

To validate the effectiveness of our method, we compared it with several previous approaches, including traditional network baselines (such as U-Net [44], HR-Net [45], Swin-Transformer (Swin-T) [46], Adabins [47], IM2HEIGHT [20], HTC-DC [8], and DenseCLIP [34]), as well as variants based on diffusion models (such as DDIM [16], DDP [48], Marigold [38], and VDP [17]). Among them, DenseCLIP is a variant based on the CLIP designed for dense estimation tasks, and DDP is a diffusion model pipeline specifically designed for dense perception. Marigold and VDP are methods that adapt pretrained SD models for downstream tasks, similar to the conception in this work. For each experiment, we provide results using LoRA ranks r of 64, 128, and 256. This choice is primarily motivated by considerations of the number of trainable parameters and fairness in training time comparison. When the LoRA rank is set to $r = 64$, the SD model has 68M trainable parameters and TSD has 1.2M trainable parameters, and the training times on the two datasets are about 9.6 h and 13.9 h, respectively, comparable to other approaches.

1) Results on US3D JAX: The experimental results for height estimation and semantic segmentation on the US3D JAX dataset are presented in Table I and Table III. As can be seen, our method achieves significant improvements over previous approaches in height estimation. At $r = 64$ (only 6.5% of SD-v2.0 parameters fine-tuned), we achieve MAE-A of 1.4775 and RMSE-A of 3.5807, improving over the best baseline VDP by 8.1% and 5.2%, respectively. The results also show that as the LoRA rank increases, the model performance improves significantly. When $r = 128$, the number of trainable parameters doubles, and MAE-A and RMSE-A decrease by 5.2% and 3.3%, respectively, while σ_1 increases substantially to 0.3620, representing a 9.1% improvement. When we further increase r to 256, the performance improvements diminish; although the best results among all experiments are achieved, the key error metrics MAE-A and RMSE-A decrease by less than 3% compared to $r = 128$. As shown in Table III, the differences in performance among various methods on the semantic segmentation task are relatively small, but our method still achieves the best results. Among the comparison methods, VDP and DenseCLIP, both of which use pretrained models, perform the best, with the VDP model even slightly outperforming our method when $r = 64$. However, when $r = 128$ and $r = 256$, our method achieves the best performance. In particular, for the mIoU metric, the result at $r = 256$ surpasses the best comparison method, VDP, by 0.043, representing a relative improvement of 5.8%.

2) Results on Open Data DC: The experimental results for height estimation and semantic segmentation on the Open Data DC dataset are presented in Table II and Table IV. Among the comparison methods, VDP again achieves the best performance, followed by DenseCLIP, which adopts a traditional network paradigm. Consistent with the results on US3D JAX, our method outperforms existing approaches in height estimation by a significant margin. When $r = 64$, our method reduces the MAE-A by 0.28 compared to VDP, representing a 13.1% decrease. The best result is achieved at $r = 128$, where the MAE-A reaches 1.7784, a 17.04% reduction compared to VDP. For the RMSE-A metric, it decreases by 0.1141 (a 3.6% reduction), and for the σ_1 metric, it increases by 0.0235 (a 4.7% improvement). Consistent with the observations on the US3D JAX dataset, increasing r from 64 to 128 leads to a significant improvement in model performance, while increasing r from 128 to 256 does not result in a noticeable change. Table IV presents the performance on the semantic segmentation task, where the differences among the compared methods are relatively small. However, even at $r = 64$, our method achieves performance comparable to the state-of-the-art, and the best results obtained at $r = 256$, with Acc-A reaching 0.8547 and mIoU reaching 0.7481.

Fig. 4 shows the results of height estimation and semantic segmentation produced by our method on the US3D JAX and Open Data DC datasets, as well as comparisons with the ground truth. It can be observed that our method is able to simultaneously obtain accurate segmentation and height

[1]https://huggingface.co/stabilityai/stable-diffusion-2

TABLE I
THE HEIGHT ESTIMATION COMPARISON OF EXISTING METHODS IN **US3D JAX** DATASET. **BOLD** INDICATES THE BEST RESULTS, UNDERLINE INDICATES THE SECOND-BEST. THE SAME APPLIES HEREINAFTER.

Methods	MAE-A ↓	MAE-F ↓	MAE-B ↓	RMSE-A ↓	RMSE-F ↓	RMSE-B ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Conventional Network Baselines									
U-Net[44]	2.4243	4.9953	1.7949	3.8916	6.5081	2.6042	0.2835	0.4773	0.6019
HRNet[45]	2.4187	4.7447	1.8887	3.7740	6.1156	2.7115	0.2739	0.4641	0.5907
SwinT[46]	3.2563	6.0664	2.5194	4.8183	7.5893	3.4524	0.2082	0.3794	0.5102
Adabins[47]	3.2285	6.1091	2.5471	4.7320	7.6965	3.4268	0.2108	0.3827	0.5139
IM2HEIGHT[20]	2.5271	4.8070	1.8424	3.7778	6.6915	2.5261	0.2704	0.4521	0.5892
HTC-DC[8]	2.3188	4.5067	1.7022	3.5791	6.5818	2.3918	0.3043	0.4925	0.5859
DenseCLIP[34]	2.1234	4.3308	1.6061	3.4268	5.5831	2.3211	0.3113	0.4821	0.5971
Diffusion Model-based Methods									
DDIM[16]	4.3177	7.2010	3.0021	6.1081	9.1116	4.3744	0.1031	0.2892	0.3914
DDP[48]	2.0121	4.2981	1.3109	3.3382	5.7042	2.4228	0.3083	0.4991	0.5965
Marigold[38]	2.4257	4.8882	1.7816	4.0117	6.4118	2.8271	0.2811	0.4496	0.5565
VDP[17]	1.6077	4.0259	1.0889	3.1629	5.2365	2.2354	0.3256	0.5030	0.6093
Pretrain Diffusion Model Adaptation (Ours)									
SD-v2.0(r=64)	1.4775	3.5807	1.0503	2.9976	4.7182	2.2482	0.3319	0.4936	0.5992
SD-v2.0(r=128)	1.4105	3.4397	0.9955	2.8972	4.5432	**2.1515**	0.3620	0.5150	0.6097
SD-v2.0(r=256)	**1.3809**	**3.3393**	**0.9701**	**2.8112**	**4.4401**	2.1548	**0.3729**	**0.5177**	**0.6138**

TABLE II
THE HEIGHT ESTIMATION COMPARISON OF EXISTING METHODS IN **OPEN DATA DC** DATASET.

Methods	MAE-A ↓	MAE-F ↓	MAE-B ↓	RMSE-A ↓	RMSE-F ↓	RMSE-B ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Conventional Network Methods									
U-Net[44]	3.0308	3.1190	2.8420	3.9981	3.8829	3.7370	0.4008	0.6511	0.7438
HRNet[45]	2.7565	2.8291	2.6770	3.7889	3.6336	3.6783	0.4259	0.6516	0.7683
SwinT[46]	3.8170	3.9705	3.5468	4.8944	4.8286	4.5222	0.3222	0.5599	0.7054
Adabins[47]	3.4468	3.8266	3.2558	4.5359	4.7148	4.2523	0.3355	0.5632	0.7015
IM2HEIGHT[20]	2.8763	2.9061	2.6153	3.8533	3.7246	3.6521	0.4082	0.6431	0.7419
HTC-DC[8]	2.6188	2.7073	2.5921	3.8879	3.7911	3.4616	0.4115	0.6493	0.7681
DenseCLIP[34]	2.2183	2.1019	2.2037	3.1792	2.8913	3.3010	0.4864	0.6616	0.7791
Diffusion Model Methods									
DDIM[16]	4.1124	4.2376	4.0004	5.1023	5.3204	5.0241	0.2933	0.4423	0.6571
DDP[48]	2.4840	2.3567	2.6107	3.4854	3.2636	3.5901	0.4528	0.6518	0.7644
Marigold[38]	2.7142	2.8559	2.6347	3.8174	3.8002	3.4452	0.3998	0.6021	0.7348
VDP[17]	2.1439	2.1979	2.1027	3.1561	2.9262	3.2120	0.5002	0.6670	0.7808
Pretrain Diffusion Model Adaptation (Ours)									
SD-v2.0(r=64)	1.8634	2.0005	1.7465	3.0848	2.7021	3.0621	0.5096	0.6747	0.7799
SD-v2.0(r=128)	**1.7784**	1.9403	**1.7304**	**3.0419**	2.6437	**3.0279**	**0.5237**	**0.6886**	0.7857
SD-v2.0(r=256)	1.7880	**1.9277**	1.7562	3.0502	**2.6384**	3.0539	0.5193	0.6812	**0.7909**

TABLE III
THE SEMANTIC SEGMENTATION COMPARISON OF EXISTING METHODS IN **US3D JAX** DATASET.

Methods	Acc-A	Acc-F	Acc-B	IoU-F	mIoU
Unet[44]	0.8338	0.7752	0.8337	0.6311	0.6883
DeeplabV3+[49]	0.8411	0.8140	0.8463	0.6719	0.7112
BiseNetV2[50]	0.8292	0.7713	0.8302	0.6245	0.6831
OCRNet[51]	0.8448	0.8224	0.8411	0.6749	0.7087
SegFormer[52]	0.8341	0.7985	0.8313	0.6406	0.6913
MaskFormer[53]	0.7822	0.7447	0.7793	0.5181	0.6221
VDP[17]	0.8611	0.8635	0.8586	0.7109	0.7404
DenseCLIP[34]	0.8545	0.8292	0.8453	0.6906	0.7122
SD-v2.0(r=64)	0.8605	0.8589	0.8553	0.7035	0.7543
SD-v2.0(r=128)	0.8656	**0.8765**	0.8582	**0.7155**	0.7712
SD-v2.0(r=256)	**0.8704**	0.8727	**0.8621**	0.7126	**0.7834**

TABLE IV
THE SEMANTIC SEGMENTATION COMPARISON OF EXISTING METHODS IN **OPEN DATA DC** DATASET.

Methods	Acc-A	Acc-F	Acc-B	IoU-F	mIoU
Unet[44]	0.8047	0.8452	0.7842	0.7474	0.6872
DeeplabV3+[49]	0.8262	0.8798	0.8058	0.7935	0.7080
BiseNetV2[50]	0.8109	0.8530	0.7894	0.7580	0.6921
OCRNet[51]	0.8331	0.8667	0.8130	0.7973	0.7199
SegFormer[52]	0.8082	0.8532	0.7836	0.7414	0.6858
MaskFormer[53]	0.7772	0.8142	0.7573	0.7092	0.6533
VDP[17]	0.8542	0.8943	0.8317	0.8301	0.7323
DenseCLIP[34]	0.8392	0.8852	0.8163	0.7944	0.7228
SD-v2.0(r=64)	0.8482	0.8951	0.8267	0.8306	0.7293
SD-v2.0(r=128)	0.8503	**0.9005**	0.8284	0.8276	0.7300
SD-v2.0(r=256)	**0.8547**	0.8971	**0.8359**	**0.8312**	**0.7481**

estimation results using only a single RGB image. Compared to semantic segmentation, height estimation requires capturing more detailed information from the image. Our method can predict accurate height maps corresponding to the shapes of different objects. In Fig. 5, we present the height maps predicted by different methods for the same input on the US3D JAX and Open Data DC datasets. This figure allows for a qualitative comparison between our method and others. Although some methods also seem to achieve good results, the devil is in the details. As indicated by the red boxes in the figure, our method can more clearly present height variations caused by different objects within buildings. In the platform areas indicated by the yellow boxes, our method is also able to estimate smoother results.

B. Method Robustness Analysis

1) Qualitative Results on Non-building Categories: In Table.V, we present the results of height estimation and semantic segmentation for non-building categories. It can be observed that, in terms of height estimation performance, the two datasets exhibit a certain commonality: the models perform worst on the shared category *Tree*. In contrast, the performance

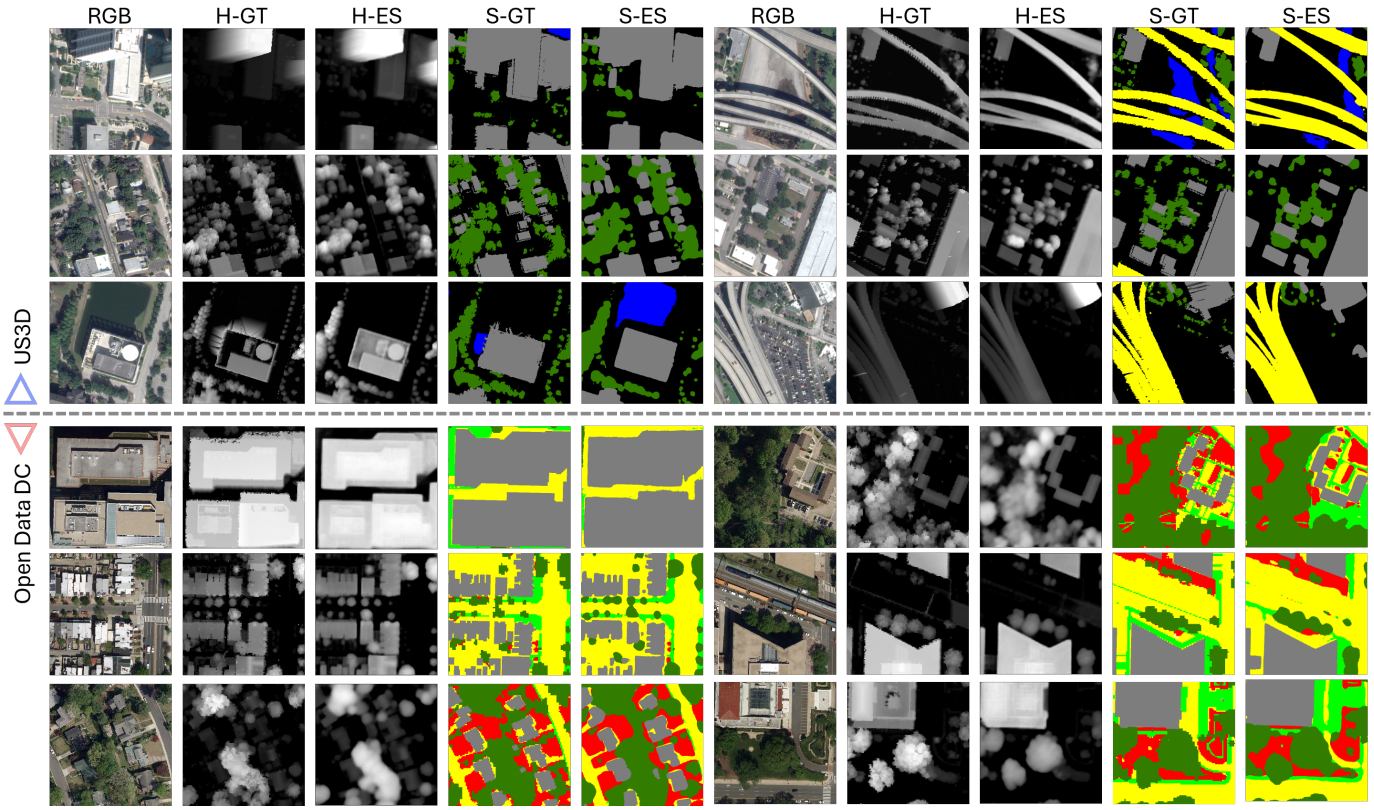

Fig. 4. Results on US3D JAX and Open Data DC. H-GT and H-ES denote the ground truth and estimated height map, respectively; S-GT and S-ES represent the ground truth and estimated segmentation map, respectively.

TABLE V
QUANTITATIVE RESULTS OF NON-BUILDING CATEGORIES ON **US3D JAX** AND **OPEN DATA DC** DATASETS. THE LORA RANK $r = 64$

Category	MAE ↓	RMSE ↓	Acc ↑	IoU ↑
US3D JAX				
Ground	0.5994	1.6163	0.9289	0.8101
Tree	2.5607	3.4714	0.5990	0.5109
Water	0.9109	1.3710	0.8097	0.7439
Bridge/Elevated/Road	2.6042	3.2211	0.9153	0.8154
Open Data DC				
Tree	2.6060	3.6699	0.8565	0.7418
Understory Vegetation	1.4546	2.5963	0.6030	0.5077
Impervious Surface	1.0820	2.2724	0.6203	0.4904
Road	1.2022	2.5376	0.7608	0.6264
Water	1.0369	1.7245	0.9732	0.9583

on the *Water* is much better, with the RMSE being the lowest in both datasets. This observation is intuitive, as the water surface exhibits the highest consistency in height, whereas *Tree* possess irregular vertical structures in both datasets. This further reveals that there are significant differences in the model's height estimation across different object categories, especially due to the influence of the heterogeneous height distributions among these categories. Regarding the semantic segmentation metrics, in US3D JAX, the Acc and IoU for trees are among the lowest, while in Open Data DC they are relatively good. This phenomenon may be attributed to the fact that the images in US3D JAX are primarily from urban scenes where trees are sparsely distributed, whereas Open Data DC contains larger contiguous forested areas.

2) Validation on Different Land Use Types: To evaluate the method's capability in height estimation across different land-use types, we conducted separate evaluations on the subsets of two datasets according to their parcel categories. Specifically, the test set of the Open Data DC dataset was divided into four subsets: high-rise districts (HD), residential zones (RZ), industrial zones (IZ), and vegetated zones (VZ). Similarly, the test set of the US3D JAX dataset was categorized into HD, RZ, IZ, and road networks (RN). In Appendix. A, detailed information on the subset categorization is provided. The quantitative results of each subset are reported in Table VI. For the US3D JAX subsets, the RZ group achieves the lowest mean absolute error (MAE) in height estimation, whereas the HD group yields the highest. A similar trend is observed in the Open Data DC subsets, where height estimation performance deteriorates in both the HD and VZ groups. In contrast, for semantic segmentation, the VZ subset demonstrates the best performance, while the HD subset performs the worst. This discrepancy suggests that the tasks of height estimation and semantic segmentation are decoupled, despite being jointly optimized during training.

3) Impact of Shadow on Height Estimation: To quantitatively assess the potential impact of shadows in remote sensing imagery on height estimation, we first employed an effective shadow segmentation algorithm [54] to detect shadow regions within the test images of both datasets. As illustrated in Fig. 6, shadow areas can be clearly delineated from the remote sensing imagery. Statistical analysis indicates that shadows constitute approximately 14.39% of the test set in the US3D JAX dataset and 13.77% in the Open Data DC dataset. As shown in Table VII, shadow regions indeed exert a measurable

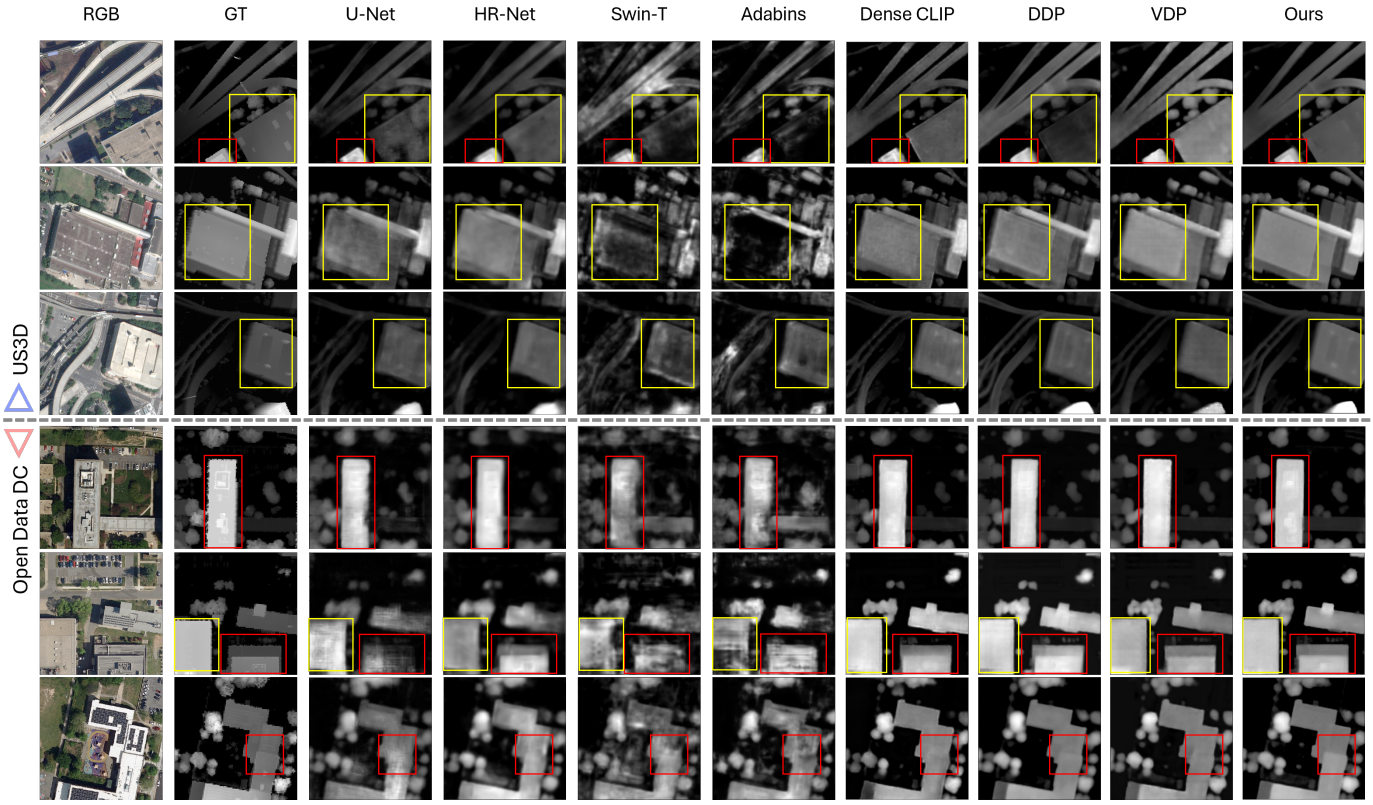

Fig. 5. Method comparison results on the US3D JAX and Open Data DC datasets. The red boxes indicate areas with height variations in buildings, while the yellow boxes indicate platform areas.

TABLE VI
QUANTITATIVE RESULTS OF SUBSETS OF **US3D JAX** AND **OPEN DATA DC** DATASETS. THE LORA RANK $r = 64$

Subsets	MAE-A ↓	RMSE-A ↓	Acc-A ↑	mIoU ↑
Subsets of US3D JAX				
HD	2.1960	4.8300	0.8514	0.7152
RZ	1.1109	2.2022	0.8681	0.8217
IZ	1.6175	3.2797	0.8592	0.7163
RN	1.4228	2.7396	0.8513	0.7325
Subsets of Open Data DC				
HD	2.1828	3.7209	0.8285	0.7029
RZ	1.6558	2.7535	0.8339	0.7301
IZ	1.0350	2.1213	0.8338	0.7183
VZ	2.8135	4.2487	0.9076	0.7526

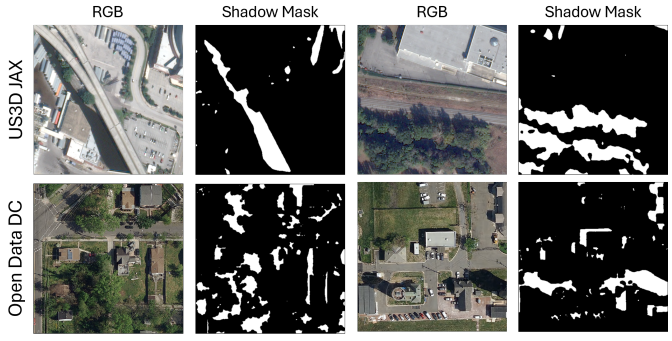

Fig. 6. Shadow detection samples of US3D JAX and Open Data DC datasets.

influence on model performance: height estimation results in shadowed areas are relatively worse, with the mean absolute error (MAE) increasing by 11.97% in the Open Data DC dataset and by 5.26% in the US3D JAX dataset.

TABLE VII
COMPARISON OF HEIGHT ESTIMATION PERFORMANCE FOR SHADOWED AND NON-SHADOWED AREAS. THE LORA RANK $r = 64$

Dataset	Shadow Area		Non-shadow Area	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Open Data DC	2.0295	3.3017	1.8125	3.0586
US3D JAX	1.5407	3.0841	1.4637	2.6010

C. Ablation Study

To thoroughly evaluate the effectiveness of the proposed approaches, we conducted a series of ablation experiments to investigate how different configurations affect the transfer performance. All experiments were carried out on the Open Data DC dataset. Below, we explain the different settings used in the ablation study and the numbering assigned to each:

- **S1**: Only fine-tuning the denoising unet, using the latent features of the target map as supervision, and then using the pretrained VAE decoder to restore the final output.
- **S2A**: Fine-tune the denoising unet while simultaneously training the height estimation and semantic segmentation TSD separately.
- **S2B**: Building upon S2A, the decoder employs the multi-task head and adopts multi-objective optimization during model training.
- **S3**: Establish a shortcut between the VAE encoder and the TSD.
- **S4**: Incorporate a pyramid scene parsing module into the TSD.

Table VIII and Table IX present the results of ablation experiments on height estimation and semantic segmentation,

TABLE VIII
ABLATION EXPERIMENT RESULTS FOR HEIGHT ESTIMATION EVALUATED ON **OPEN DATA DC** DATASET.

Settings	MAE-A	MAE-F	RMSE-A	RMSE-F	δ_1
S1	2.6209	2.7672	3.4721	3.3020	0.4281
S2A	2.0295	2.4139	3.1168	2.9090	0.4682
S2B	1.9826	2.3093	3.2192	2.9231	0.4624
S2B,S3	1.8084	2.0357	3.0335	2.6400	0.5091
S2B,S3,S4	**1.7784**	**1.9405**	**3.0419**	**2.6437**	**0.5237**

TABLE IX
ABLATION EXPERIMENT RESULTS FOR SEMANTIC SEGMENTATION EVALUATED ON **OPEN DATA DC** DATASET.

Setting	Acc-A	Acc-F	Acc-B	IoU-F	mIoU
S1	0.7110	0.6592	0.7015	0.6468	0.5630
S2A	0.8262	0.8671	0.8106	0.7797	0.7131
S2B	0.8207	0.8542	0.8112	0.7612	0.7107
S2B,S3	0.8419	0.8834	0.8294	0.8284	0.7203
S2B,S3,S4	**0.8503**	**0.9005**	**0.8278**	**0.8276**	**0.7300**

respectively.

1) Validity of Task-specific Decoder: The comparison between **S2A** and **S1** demonstrates that the TSD significantly enhances task adaptability. Whether for the individual height estimation or semantic segmentation, there are significant improvements across all metric dimensions. Specifically, MAE-A decreases from 2.6209 to 2.0295, and mIoU increases from 0.5630 to 0.7131. In addition to the inherent errors of the pretrained VAE decoder in height map restoration, as analyzed earlier, incomplete modeling in the latent space may also amplify errors in the decoder, making the TSD particularly important. It is worth mentioning that the redesigned decoder has less than one-tenth the number of parameters of the pretrained VAE decoder, making it more lightweight.

2) Impact of Multi-task Optimization: Compared to **S2A**, **S2B** adopts multi-task optimization, which not only reduces the computational load caused by separately inferring the two different tasks, but also shows no significant degradation in the performance of either task. In fact, **S2B** performs slightly better than **S2A** in the height estimation, while being slightly weaker in the semantic segmentation.

3) Validity of Micro-Design Implementation: In **S3**, establishing a shortcut between the VAE Encoder and TSD effectively improves performance in both tasks, with nearly all metrics showing significant enhancement. This demonstrates that pretrained VAE intermediate features are crucial for fine-grained height estimation and semantic segmentation. Adding the PSP module (**S4**) further enhances the decoder's cross-scale perception, yielding stable improvements across all metrics.

4) Impact of LoRA Rank and Training Convergence: Fig. 7 illustrates the impact of varying the LoRA rank r, which adjusts the number of trainable parameters in the SD model, on key performance metrics. The gray line represents the VDP baseline. It can be observed that, for the height estimation metrics, the performance at $r = 8$ already reaches a level comparable to the baseline. To investigate the convergence of the fine-tuning process, we present in Fig. 8 the variations of several evaluation metrics on the test set, including MAE-A, RMSE-A, δ_1, and Acc-A, measured at intervals of 4k iterations. As shown in the results, the model approaches

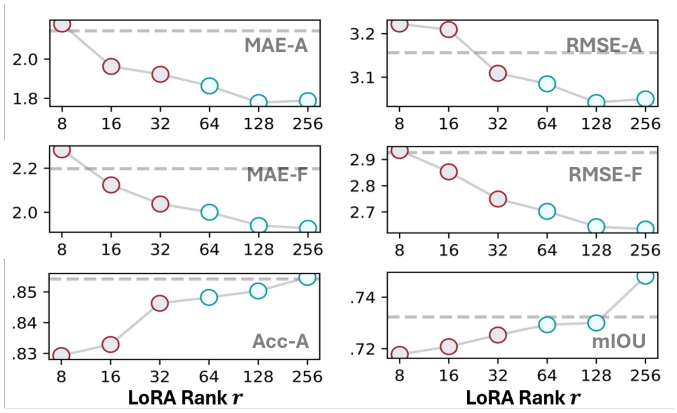

Fig. 7. Visualization of the performance metrics with different settings of LoRA rank r.

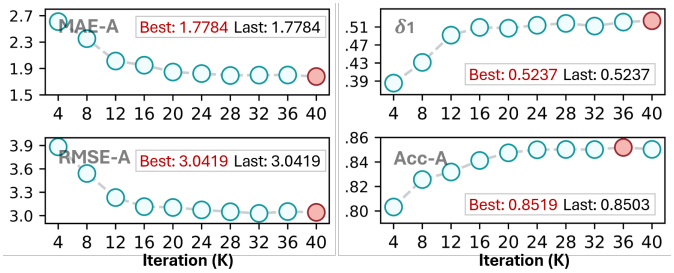

Fig. 8. Iterative changes in test set metrics during fine-tuning.

convergence around 24k iterations, beyond which the improvements across all metrics become marginal.

D. Generalization Analysis on Different Pretrained Diffusion Models

Our method is designed based on the SD pipeline and, therefore, can be adapted to different publicly available versions of SD models. To verify the generalizability of our method on different pretrained models, we conducted corresponding experiments on the Open Data DC dataset. We tested three different versions of SD: SD-v1.5[2], SD-v2.0, and SD-XL[3]. SD-v1.5 employs an 860M-parameter denoising network, trained from SD-v1.0 through pretraining on Laion-2B at 256×256 resolution, fine-tuning at 512×512, and additional 1.11M steps on Laion-Improved-Aesthetics and Laion-Aesthetics v2 5+. SD-v2.0 was trained from scratch for 550k steps on Laion-5B, then for 850k steps at 512×512 resolution. Its 865M-parameter network supports both 512×512 and 768×768 inputs and enables not only text-to-image generation but also inpainting, super-resolution, and depth-to-image tasks. SD-XL features a significantly larger 2.6B-parameter denoising network, with the increased capacity primarily from additional attention blocks and larger cross-attention context enabled by a second text encoder.

The results in Table X indicate that SD-v2.0 achieves marginally superior performance compared to SD-v1.5 across most metrics for both height estimation and semantic segmentation tasks. Although SD-XL achieves the best results, the improvement is not significant. Considering that SD-XL

[2]https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5
[3]https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

TABLE X
PERFORMANCE COMPARISON OF DIFFERENT BASE MODELS ON HEIGHT ESTIMATION AND SEMANTIC SEGMENTATION TASKS. #PARAMS INDICATES THE TRAINABLE PARAMETERS NUMBER WITH $r = 128$.

Base	*#Params*	*Height Estimation*					*Semantic Segmentation*				
		MAE-A	**MAE-F**	**RMSE-A**	**RMSE-F**	δ_1	**Acc-A**	**Acc-F**	**IoU-A**	**IoU-F**	**mIoU**
SD-v1.5	135M	1.7966	1.9798	3.0811	2.6764	0.5188	0.8456	0.8919	0.8297	**0.8326**	0.7297
SD-v2.0	135M	1.7784	1.9405	3.0419	2.6437	0.5237	0.8503	0.9005	0.8278	0.8276	0.7300
SD-XL	393M	**1.7621**	**1.9043**	**3.0112**	**2.5972**	**0.5331**	**0.8522**	**0.9040**	**0.8347**	0.8300	**0.7413**

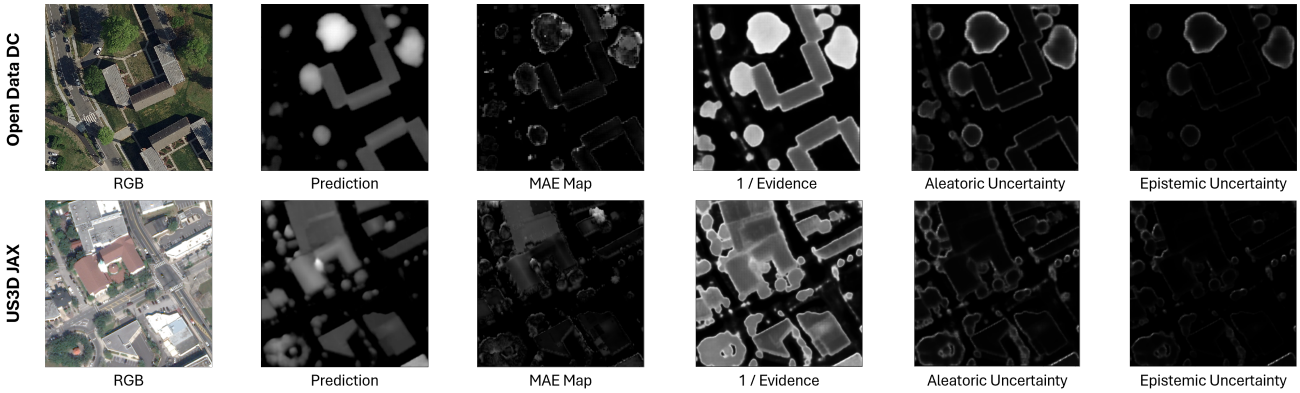

Fig. 9. Example results of uncertainty quantification. Brighter regions indicate lower evidence values, while darker regions correspond to higher evidence values. To enhance visualization, the values for aleatoric uncertainty and epistemic uncertainty were square-rooted, which preserves their monotonicity.

involves training approximately three times more parameters, the performance improvements appear insufficient to justify the substantially increased computational cost. In our experiments, the larger model scale also resulted in a training time nearly twice as long as that of the other two versions when using SD-XL as the base model. This is also why we chose SD-v2.0 as the primary base model for our experiments.

E. Uncertainty Perception for Height Estimation

Model uncertainty analysis serves as an important indicator of model reliability. In this study, we introduce evidential learning theory for the first time and propose EHR to assess the uncertainty of PDM in height estimation. To quantitatively evaluate uncertainty performance, we employ the correlation coefficient between uncertainty and error as the evaluation metric, which measures the relationship between the predicted height map uncertainty and the actual prediction error distribution. We refer to this metric as the uncertainty-error correlation coefficient (UECC):

$$\text{UECC} = \sum_{i=1}^{N} \frac{\sum(e_i - \hat{e}_i)(u_i - \hat{u}_i)}{\sqrt{\sum(e_i - \hat{e}_i)^2 \sum(u_i - \hat{u}_i)^2}}, \tag{29}$$

where e_i and u_i denote the pixel-wise error and uncertainty, respectively. Since UECC is a dimensionless metric, it can be used to compare the correlation between the uncertainty estimates provided by different methods and the actual error distribution. In height estimation, a higher correlation between uncertainty and error within an image indicates that the uncertainty estimation is more informative and reliable. We use the reciprocal of the evidence, i.e., $1/(2\epsilon + \alpha)$, as the uncertainty index for EHR. To investigate an appropriate value of λ in the context of EHR, we evaluated the performance under different λ settings in Table XI, where $\lambda = 0.2$ yielded comparatively better results.

TABLE XI
RESULTS OF ABLATION EXPERIMENTS UNDER DIFFERENT λ SETTINGS.

Metric	λ = 0.01	λ = 0.1	λ = 0.2	λ = 1
UECC ↑	0.4118	0.4672	**0.4834**	0.3805

TABLE XII
EXPERIMENTAL RESULTS ON THE **OPEN DATA DC** DATASET USING MSE, MC DROPOUT, AND EHR FOR SUPERVISION.

Settings	**MAE-A**	**MAE-F**	**RMSE-A**	**RMSE-F**	**UECC**
MSE	1.7816	**1.9348**	**3.0217**	2.6534	-
MC Dropout	2.1431	2.6561	3.3929	3.3466	0.4166
EHR	**1.7784**	1.9405	3.0419	**2.6437**	**0.4834**

Table XII presents a comparison among direct regression (i.e, MSE), MC Dropout, and EHR. The standard deviation from MC Dropout variational inference as its uncertainty index. Our method achieves an improvement of 0.0668 (a relative increase of 16%) over MC Dropout on the UECC metric, indicating that EHR provides more reliable uncertainty estimation. EHR does not require multiple sampling and can estimate uncertainty with a single forward pass, significantly reducing computational cost (whereas MC Dropout requires multiple inferences). Fig. 9 presents two uncertainty visualizations from US3D JAX and Open Data DC produced by EHR. The predicted height maps show that uncertainty estimates closely mirror the actual error distribution, with sharp increases along object boundaries consistent with geometric priors in 3D scenes. EHR effectively disentangles epistemic and aleatoric uncertainty: epistemic maps highlight pronounced edge responses around trees and buildings, while aleatoric maps suppress building-edge uncertainty but retain high values in vegetated areas.

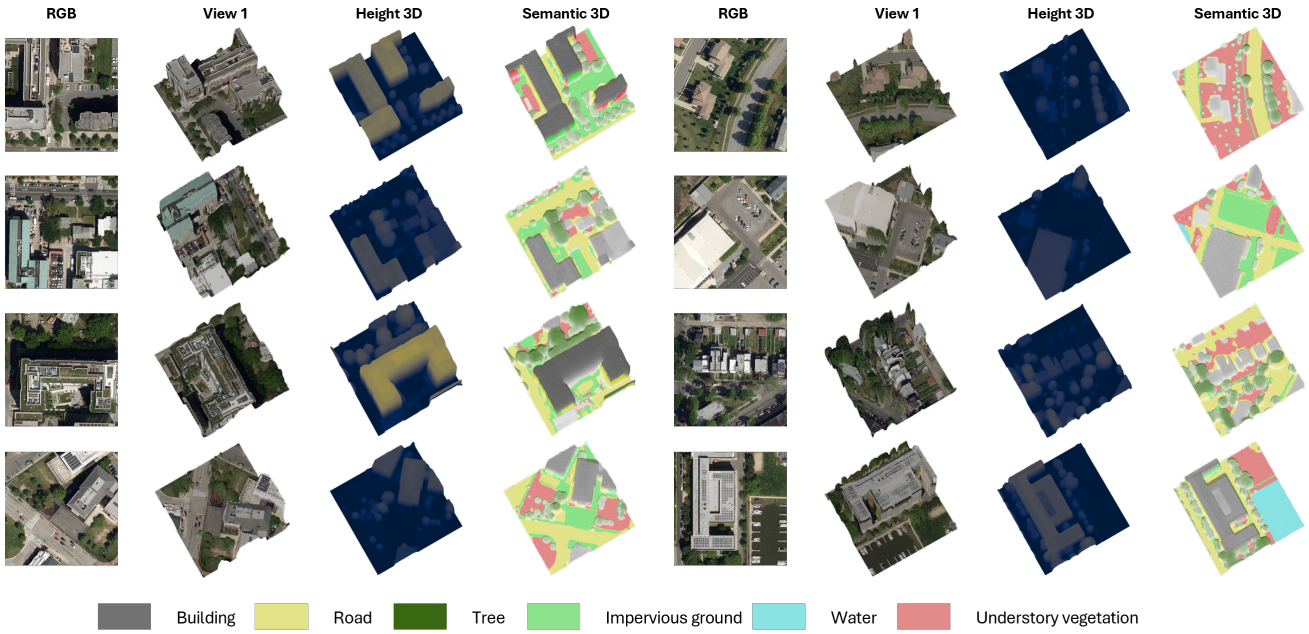

Fig. 10. The samples of 3D scene reconstruction, showing visualization results from two perspectives, as well as point clouds rendered according to height value and semantic 3D reconstruction.

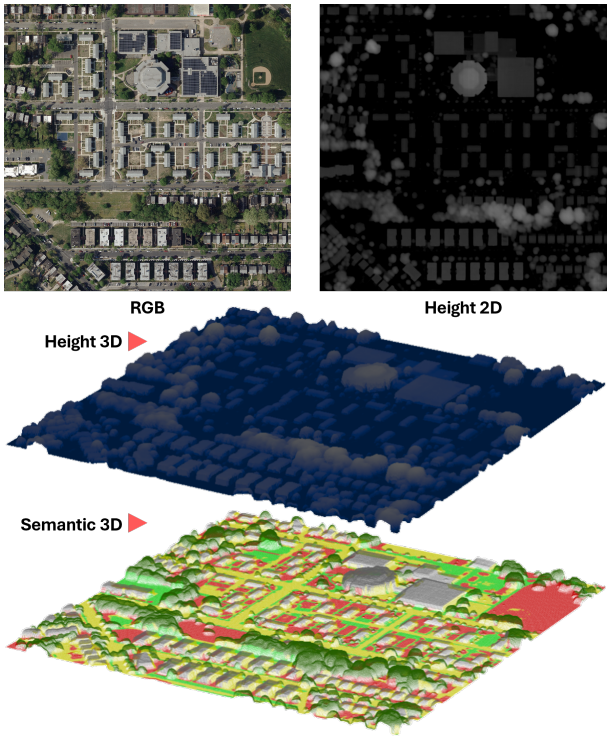

Fig. 11. The semantic 3D reconstruction visualization in a large region.

F. Evaluation of Semantic 3D Reconstruction

To better demonstrate the performance of our method in 3D reconstruction, we conducted an evaluation on the Open Data DC dataset using pseudo point cloud generation, with Chamfer distance and F1-score as evaluation metrics. Specifically, Chamfer distance measures the point-to-point distance between the reconstructed model and the reference model, while F1-score considers the balance between recall and precision [55]. The comparative results are presented in Table XIII, which demonstrate that our method achieves a 7.9% reduction in Chamfer distance compared to VDP, and more notably, a significant improvement of approximately 12.0% in F1@1.0.

TABLE XIII
QUANTITATIVE EVALUATION OF MODEL PERFORMANCE UNDER 3D RECONSTRUCTION METRICS. CHAMFER DISTANCE IS DENOTED BY d_{CD} WITH UNITS IN METERS, WHILE F1@δ REPRESENTS THE F1-SCORE AT A THRESHOLD OF δ.

Methods	$d_{\text{CD}} \downarrow$	F1@1.0 ↑	F1@0.5 ↑	F1@0.2 ↑
DenseCLIP [34]	6.4332	0.5322	0.4219	0.2928
VDP [17]	6.1955	0.5728	0.4627	0.3301
Ours *(r=64)*	**5.7037**	**0.6512**	**0.5199**	**0.3928**

In Fig. 10, we present the semantic 3D reconstruction results for several samples from the Open Data DC dataset. The figure includes visualizations from two different views, as well as 3D renderings generated using both height values and semantic masks. In Fig.11, we present the 3D visualization of a large-scale scene, where the image resolution is 2048×2048, which is 16 times larger than the model's input size. To ensure global consistency, we adopt a sliding-window processing method with a stride of 32 for overlapping sampling, and the multiple outputs are subsequently fused using a weighted averaging strategy. The qualitative results demonstrate that the proposed method is capable of reconstructing the 3D structure of buildings from a monocular remote sensing image while accurately identifying the attributes of various objects. Moreover, the proposed method can be further leveraged to generate high-quality 3D point-cloud data, providing a foundation for diverse downstream applications.

VI. CONCLUSIONS

In this work, we propose a novel framework that adapts the PDMs to jointly address the subtasks of height estimation and semantic segmentation for semantic 3D reconstruction. We employ efficient LoRA fine-tuning together with a task-

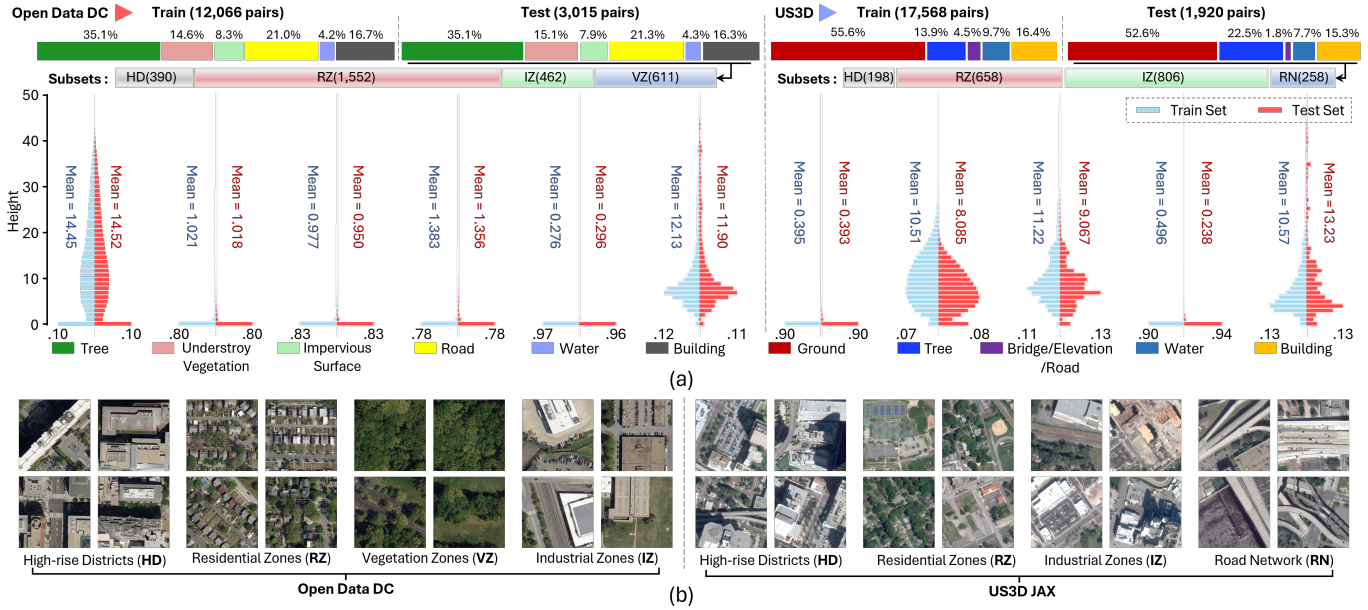

Fig. 12. Characteristic analysis of the Open Data DC and US3D JAX datasets. (a) shows the proportion of categories within the datasets, the height distributions in the training and test sets, and the number of subsets defined in the test set. (b) presents sample images of all the subsets.

specific decoder to achieve effective cross-domain adaptation of PDMs. Furthermore, we proposed evidential height regression, which not only computes accurate expected height values but also provides uncertainty estimates. Evaluations on two datasets demonstrate that our framework offers significant advantages over existing approaches. This pipeline enables high-fidelity semantic 3D reconstruction from remote sensing imagery, providing a practical tool for downstream applications.

Nevertheless, considering real-world deployment challenges, our work still has several limitations that warrant further investigation. For instance, due to data constraints, we are unable to assess the impact of viewpoint-induced roof–base displacement on building 3D reconstruction, nor can we conduct fine-grained evaluations using real 3D point cloud data to thoroughly measure the performance of converting the reconstructed 3D scene into point clouds. Methodologically, future work can investigate whether the iterative diffusion process can further refine height estimation, thereby mitigating inherent errors introduced by VAE decoding when the height map is treated as the final optimization target. Moreover, the proposed PDMs task-adaptive framework can be extended to a broader range of remote sensing tasks, including object detection, instance segmentation, and change detection.

Appendix A: Datasets Characteristic Analysis

In Fig.12 (a), we present an in-depth analysis of the dataset characteristics, including the proportion of semantic categories within the dataset, the height distributions of each category in both the training and test sets, as well as the number of subsets defined in the test set. In Fig.12 (b), we provide sample images of those subsets.

Acknowledgement

This work was supported by the High Performance Computing Centers at Eastern Institute of Technology, Ningbo, and Ningbo Institute of Digital Twin.

References

[1] L. Zhao, H. Wang, Y. Zhu, and M. Song, "A review of 3d reconstruction from high-resolution urban satellite images," International Journal of Remote Sensing, vol. 44, no. 2, pp. 713–748, 2023.

[2] Q. Yu, K. Dong, Z. Guo, J. Xu, J. Li, H. Tan, Y. Jin, J. Yuan, H. Zhang, J. Liu et al., "Global estimation of building-integrated facade and rooftop photovoltaic potential by integrating 3d building footprint and spatio-temporal datasets," Nexus, vol. 2, no. 2, 2025.

[3] K. Dong, Q. Yu, Z. Guo, J. Xu, H. Tan, H. Zhang, and J. Yan, "Advancing building facade solar potential assessment through aiot, gis, and meteorology synergy," Advances in Applied Energy, vol. 17, p. 100212, 2025.

[4] P. Wang, L. Shi, B. Chen, Z. Hu, J. Qiao, and Q. Dong, "Pursuing 3-d scene structures with optical satellite images from affine reconstruction to euclidean reconstruction," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022.

[5] Q. Cao, Y. Chen, C. Ma, and X. Yang, "Few-shot rotation-invariant aerial image semantic segmentation," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–13, 2023.

[6] W. Li, L. Meng, J. Wang, C. He, G.-S. Xia, and D. Lin, "3d building reconstruction from monocular remote sensing images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12 548–12 557.

[7] Q. Chen, W. Gan, P. Tao, P. Zhang, R. Huang, and L. Wang, "End-to-end multiview fusion for building mapping from aerial images," Information Fusion, vol. 111, p. 102498, 2024.

[8] S. Chen, Y. Shi, Z. Xiong, and X. X. Zhu, "Htc-dc net: Monocular height estimation from single remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–18, 2023.

[9] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," Expert Systems with Applications, vol. 169, p. 114417, 2021.

[10] Y. Mao, K. Chen, L. Zhao, W. Chen, D. Tang, W. Liu, Z. Wang, W. Diao, X. Sun, and K. Fu, "Elevation estimation-driven building 3-d reconstruction from single-view remote sensing imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–18, 2023.

[11] Z. Gao, W. Sun, Y. Lu, Y. Zhang, W. Song, Y. Zhang, and R. Zhai, "Joint learning of semantic segmentation and height estimation for remote sensing image leveraging contrastive learning," IEEE Transactions on Geoscience and Remote Sensing, 2023.

[12] L. P. Osco, Q. Wu, E. L. De Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. M. Junior, "The segment anything model (sam) for remote sensing applications: From zero to one shot," International Journal of

Applied Earth Observation and Geoinformation, vol. 124, p. 103540, 2023.
[13] Q. Cao, Y. Chen, L. Lu, H. Sun, Z. Zeng, X. Yang, and D. Zhang, "Generalized domain prompt learning for accessible scientific vision-language models," Nexus, vol. 2, no. 2, 2025.
[14] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," ACM computing surveys, vol. 56, no. 4, pp. 1–39, 2023.
[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695.
[16] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," arXiv preprint arXiv:2010.02502, 2020.
[17] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5729–5739.
[18] Y. Benigmim, S. Roy, S. Essid, V. Kalogeiton, and S. Lathuilière, "One-shot unsupervised domain adaptation with personalized diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 698–708.
[19] C. Häne, C. Zach, A. Cohen, and M. Pollefeys, "Dense semantic 3d reconstruction," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 9, pp. 1730–1743, 2016.
[20] L. Mou and X. X. Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," arXiv preprint arXiv:1802.10249, 2018.
[21] S. Xing, Q. Dong, and Z. Hu, "Gated feature aggregation for height estimation from single aerial images," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2021.
[22] Z. Chen, Y. Zhang, X. Qi, Y. Mao, X. Zhou, L. Wang, and Y. Ge, "Heightformer: A multilevel interaction and image-adaptive classification–regression network for monocular height estimation with aerial images," Remote Sensing, vol. 16, no. 2, p. 295, 2024.
[23] B. Zhang, Y. Wan, Y. Zhang, and Y. Li, "Jsh-net: joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery," International Journal of Remote Sensing, vol. 43, no. 17, pp. 6307–6332, 2022.
[24] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "Bidirectional guided attention network for 3-d semantic detection of remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 7, pp. 6138–6153, 2020.
[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PmLR, 2021, pp. 8748–8763.
[26] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2021.
[28] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in International Conference on Learning Representations, 2020.
[29] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.
[30] Y. Duan, X. Guo, and Z. Zhu, "Diffusiondepth: Diffusion denoising approach for monocular depth estimation," in European Conference on Computer Vision. Springer, 2025, pp. 432–449.
[31] Z. Song, Z. Wang, B. Li, H. Zhang, R. Zhu, L. Liu, P.-T. Jiang, and T. Zhang, "Depthmaster: Taming diffusion models for monocular depth estimation," arXiv preprint arXiv:2501.02576, 2025.
[32] B. Kolbeinsson and K. Mikolajczyk, "Multi-class segmentation from aerial views using recursive noise diffusion," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 8439–8449.
[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "Lora: Low-rank adaptation of large language models." ICLR, vol. 1, no. 2, p. 3, 2022.
[34] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18 082–18 091.
[35] T. Liu, S. Zhou, W. Li, Y. Zhang, and J. Guan, "Semantic prototyping with clip for few-shot object detection in remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, 2025.
[36] Q. Cao, Z. Xu, Y. Chen, C. Ma, and X. Yang, "Domain-controlled prompt learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 2, 2024, pp. 936–944.
[37] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco, "Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion," CVPR, 2024.
[38] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9492–9502.
[39] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman et al., "Laion-5b: An open large-scale dataset for training next generation image-text models," Advances in neural information processing systems, vol. 35, pp. 25 278–25 294, 2022.
[40] M. Yang, J. Chen, Y. Zhang, J. Liu, J. Zhang, Q. Ma, H. Verma, Q. Zhang, M. Zhou, I. King et al., "Low-rank adaptation for foundation models: A comprehensive review," arXiv preprint arXiv:2501.00365, 2024.
[41] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," Advances in neural information processing systems, vol. 33, pp. 14 927–14 937, 2020.
[42] K. Foster, G. Christie, and M. Brown, "Urban semantic 3d dataset," 2020. [Online]. Available: https://dx.doi.org/10.21227/9frn-7208
[43] "Open data dc," https://opendata.dc.gov/, accessed: 2025-01-17.
[44] O. Ronneberger, P. Fischer, and T. Brox, "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. Springer, 2015, pp. 234–241.
[45] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang et al., "Deep high-resolution representation learning for visual recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 10, pp. 3349–3364, 2020.
[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10 012–10 022.
[47] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4009–4018.
[48] Y. Ji, Z. Chen, E. Xie, L. Hong, X. Liu, Z. Liu, T. Lu, Z. Li, and P. Luo, "Ddp: Diffusion model for dense visual prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21 741–21 752.
[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[50] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," International Journal of Computer Vision, pp. 1–18, 2021.
[51] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer, 2020, pp. 173–190.
[52] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," Advances in neural information processing systems, vol. 34, pp. 12 077–12 090, 2021.
[53] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," Advances in neural information processing systems, vol. 34, pp. 17 864–17 875, 2021.
[54] G. F. Silva, G. B. Carneiro, R. Doth, L. A. Amaral, and D. F. de Azevedo, "Near real-time shadow detection and removal in aerial motion imagery application," ISPRS Journal of photogrammetry and remote sensing, vol. 140, pp. 104–121, 2018.
[55] C. Wen, Y. Zhang, Z. Li, and Y. Fu, "Pixel2mesh++: Multi-view 3d mesh generation via deformation," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1042–1051.