Simultaneous Super-resolution and Depth Estimation for Satellite Images Based on Diffusion Model

Yuwei Zhou and Yangming Lee

Abstract— **Satellite images provide an effective way to observe the earth surface on a large scale. 3D landscape models can provide critical structural information, such as forestry and crop growth. However, there has been very limited research to estimate the depth and the 3D models of the earth based on satellite images. LiDAR measurements on satellites are usually quite sparse. RGB images have higher resolution than LiDAR, but there has been little research on 3D surface measurements based on satellite RGB images. In comparison with in-situ sensing, satellite RGB images are usually low resolution. In this research, we explore the method that can enhance the satellite image resolution to generate super-resolution images and then conduct depth estimation and 3D reconstruction based on higher-resolution satellite images. Leveraging the strong generation capability of diffusion models, we developed a simultaneous diffusion model learning framework that can train diffusion models for both super-resolution images and depth estimation. With the super-resolution images and the corresponding depth maps, 3D surface reconstruction models with detailed landscape information can be generated. We evaluated the proposed methodology on multiple satellite datasets for both super-resolution and depth estimation tasks, which have demonstrated the effectiveness of our methodology.**

I. Introduction

In the current era of satellite Earth observation, a multitude of missions are operational, with their numbers continuing to rise [30]. Satellite remote sensing enables the rapid and efficient collection of global-scale geospatial data, establishing itself as a vital tool for accessing and understanding geographic information. Utilizing satellite imagery for large-scale 3D reconstruction of the Earth's surface provides precise digital models crucial for urban planning, ecological monitoring, disaster response, and other domains. This capability enhances spatial understanding and cognition of complex environments, highlighting its significant research and practical value.

However, satellite images are constrained by imaging conditions, storage, and transmission bandwidth, making it challenging to acquire high spatial resolution images. As remote sensing imagery finds increasingly diverse applications, the use of lower quality images significantly reduces the accuracy of key parameter estimates, severely limiting the research and application of satellite image data. Therefore, developing super-resolution methods for low-resolution (LR) image data to enhance spatial resolution is crucial for enabling more detailed analysis and applications of satellite imagery. In the actual satellite remote sensing image acquisition process, due to the long distance of the satellite orbit and the limitation of the volume and stability of the imaging system, the resolution of the remote sensing image data obtained after acquisition is often low. In order to obtain high-resolution remote sensing images, a direct way is to improve the imaging resolution from the hardware perspective, which can be very expensive and out of control for common users. Satellite images with higher resolutions can provide more details of the ground information.

The applications of 3D reconstruction technology have become widespread across various domains. It serves as a vital tool for modern geospatial analysis and urban planning, enabling the detailed reconstruction of natural landscapes and man-made structures. By integrating remote sensing data, aerial imagery, and ground-based surveys, detailed three-dimensional models of diverse landscape elements can now be created. 3D models based on satellite images can support the understanding of landscape, such as the growth of forestry and crops. However, 3D reconstruction techniques specifically tailored for satellite imagery remain scarce. Recently, diffusion models have garnered increasing attention for their powerful image generation capabilities across various computer vision tasks. In contrast to GANs, diffusion models can train to generate more diverse and complex images. Using the same training dataset, diffusion models mitigate the convergence issues often faced in GAN training. The algorithmic foundation of diffusion models involves training parameterized Markov chains through variational inference, demonstrating superior performance over other generative models like GANs in numerous tasks. As a conditional model dependent on priors, diffusion models can generate target data samples from noise sampled from a simple distribution. This involves both forward and inverse processes, where random noise is injected into data (forward process) and desired data samples are sampled from it (inverse process). In this paper, we develop diffusion methods specifically targeting satellite images, which can increase the resolution of the satellite images and build 3D models based on the enhanced satellite images. The significant contributions of this work are outlined as follows: 1) we have created a pipeline that can create the 3D models from the satellite images. 2) In dealing with the low-resolution issues, we have developed diffusion models that can create super-resolution images, which can leverage the low-resolution image to interpolate the pixels accurately. 3) We also have developed a diffusion model that can output the depth maps, which is one of the first methods targeting satellite images. The super-resolution and depth estimation tasks are learned simultaneously to further enhance each other. The entire

Yuwei Zhou and Yangming Lee are with Rochester Institute of Technology. `yz4891@rit.edu, yangming.lee@rit.edu`

framework is shown in Fig. 1.

II. Related Work

Satellites capture images of objects from great distances, resulting in low-resolution images from satellite remote sensing devices. Traditional methods to enhance resolution, such as nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation, rely solely on information provided by the low-resolution image itself. These techniques often struggle to accurately reconstruct high-resolution details, leading to mismatches when increasing resolution.

With the advancement of deep learning, Convolutional Neural Network (CNN)-based approaches have become prominent in the field of super-resolution. These strategies frequently employ techniques like residual learning [21], [20], [1], [22] or recursive learning [18]to develop network architectures, significantly improving super-resolution models. However, CNN-based methods may not effectively capture residual features and often fail to fully utilize hierarchical features in low-resolution images. Moreover, these networks have limited capacity for feature extraction within residual blocks, thereby restricting the learning capability of super-resolution networks.

To overcome these limitations, researchers have introduced Transformer-based networks. Networks such as Hrformer [28] and Restormer[29] leverage the Transformer's ability to model long-range dependencies and are pre-trained on large-scale datasets like ImageNet [4] and COCO[10] By employing the Vision Transformer [5], these approaches aim to achieve superior results in super-resolution tasks.

Diffusion models have recently gained significant attention in the field of super-resolution due to their robust generative capabilities and iterative refinement processes. Initial advancements, such as SRDiff (SISR diffusion probabilistic model) [9], demonstrated the effectiveness of using a forward process to progressively add noise to images and a reverse process to iteratively remove this noise, resulting in high-quality image reconstruction. This fundamental framework has been adapted and extended to tackle super-resolution challenges specifically.

Super-Resolution via Repeated Refinement (SR3) [16] exemplifies the application of diffusion models to enhance image resolution. SR3 employs a denoising diffusion process that iteratively refines low-resolution inputs into high-resolution outputs, achieving superior performance compared to traditional convolutional neural networks (CNNs). The SR3 framework demonstrates that diffusion models can effectively address the limitations of CNN-based methods, such as inadequate feature extraction and limited utilization of hierarchical information. Further advancements in the field have introduced Latent Diffusion Models (LDMs) [15], which operate within a lower-dimensional latent space. This approach significantly reduces computational costs while maintaining high fidelity in the reconstructed images. By focusing on the latent space, LDMs enable efficient processing of high-resolution image data, making them particularly suitable for real-time applications and large-scale deployments. In addition to these foundational works, recent research has explored the integration of cross-attention mechanisms and hybrid architectures that combine the strengths of diffusion models and transformers. For example, reference-based super-resolution (RefSR) [8] leverages cross-attention to incorporate contextual information from high-resolution reference images, further improving the quality and consistency of the super-resolved outputs. These hybrid approaches highlight the potential for combining diffusion models with other advanced machine learning techniques to push the boundaries of image enhancement. Overall, the integration of diffusion models into super-resolution frameworks represents a promising direction for future research and development. By leveraging their iterative refinement capabilities and ability to model long-range dependencies, diffusion models provide a powerful tool for overcoming the limitations of traditional and CNN-based methods in the quest for high-quality, high-resolution image reconstruction

3D reconstruction is pivotal in robotics and automation, leveraging both traditional and deep learning approaches. Traditional methods like Structure from Motion (SfM)[17] and Multi-View Stereo (MVS)[26] reconstruct 3D geometry from images, but these can be computationally intensive and sensitive to environmental conditions. Recent advancements in deep learning have transformed this field. Convolutional Neural Networks (CNNs) have shown efficacy in single-view depth estimation[27]and volumetric reconstruction[23]. For example, methods such as DeepMVS[2] combine deep learning with MVS for enhanced accuracy. Additionally, Neural Radiance Fields (NeRF)[13] and Transformer-based architectures[25] offer innovative solutions by modeling 3D scenes with photorealistic details and effectively capturing global context. Hybrid approaches that integrate geometric constraints with neural networks are also gaining traction. Techniques like differentiable rendering[11] enhance the accuracy and robustness of 3D reconstructions. These advancements indicate a promising future for 3D reconstruction in robotics and automation, focusing on scalability, accuracy, and real-time capabilities.

Diffusion models have recently emerged as a promising approach in the field of 3D reconstruction, leveraging their generative capabilities to produce high-quality 3D models through iterative refinement processes. Early works, such as Denoising Diffusion Probabilistic Models (DDPM) [6], demonstrated the effectiveness of diffusion processes in generating detailed structures by progressively refining noisy inputs.Recent advancements have extended diffusion models to 3D reconstruction tasks. The Diffusion Probabilistic Model for Point Cloud Generation [12] has shown how diffusion processes can be adapted to generate 3D point clouds from initial noisy distributions. This model iteratively refines the point cloud representation, resulting in high-fidelity reconstructions. Similarly, the application of diffusion models to voxel grids and mesh generation has been explored, offering new avenues for high-resolution 3D reconstructions with fine-grained details [12]. Hybrid approaches that integrate diffusion models with other deep learning techniques have also been investigated. For example, integrating diffusion

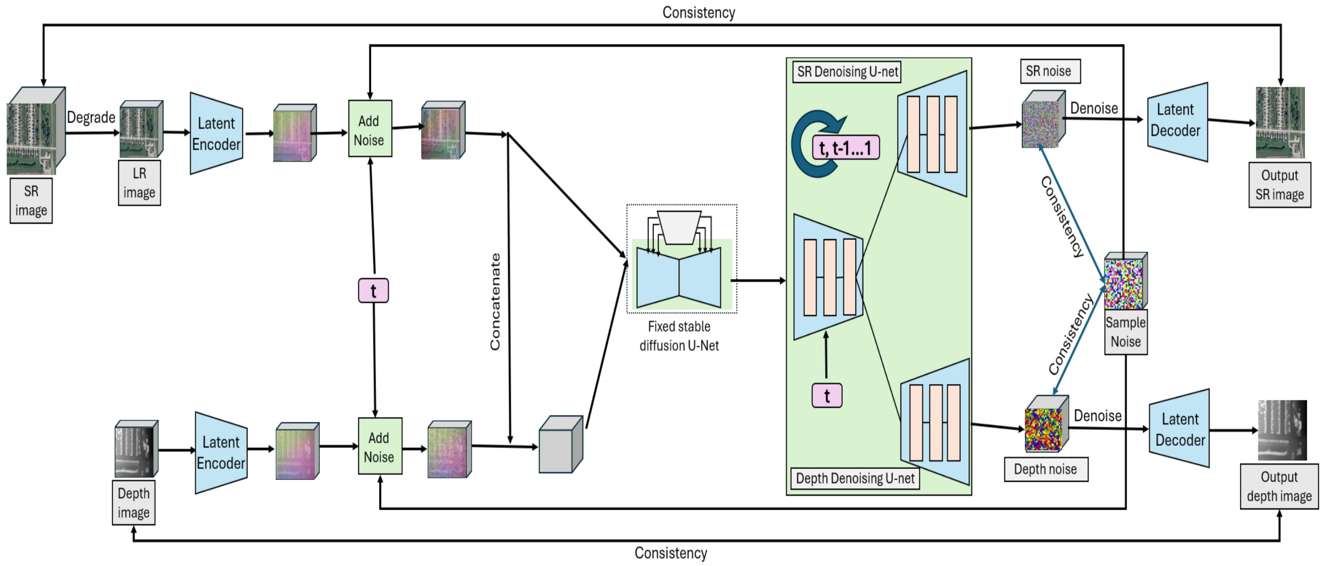

Fig. 1. The framework for our simultaneous super-resolution and depth estimation algorithm targeting satellite images. Gaussian noise is added to the low-resolution satellite images and the depth images. The noisy images are input to the latent encoder to learn the latent features. A fixed stable diffusion model followed by task-specific denoising U-net learns the noise distribution. As both tasks interpolate super-resolution images and depth maps based on the same image inputs and latent features, the super-resolution and depth estimation diffusion models share the same diffusion desnoising U-net encoder. With the shared encoded features, each task has its own diffusion desnoising decoder to output the noise, which we build the consistency constraint with the input noise. The denoised latent features are input to the latent decoder to output the estimated super-resolution images and depth maps, which are utilized to build consistency constraints with the original ground truth.

processes with convolutional neural networks (CNNs) and Transformer-based architectures has proven effective in capturing both local and global features essential for accurate 3D reconstructions [14]. These methods benefit from the iterative nature of diffusion models, which allows for progressive enhancement of 3D structures, resulting in more accurate and detailed reconstructions. Moreover, diffusion models have been applied to multi-view 3D reconstruction tasks, where they are used to integrate information from multiple 2D images to generate a coherent 3D model [19]. This approach leverages the ability of diffusion models to handle complex data distributions, enabling the reconstruction of 3D models with high precision from sparse and noisy input data.

III. Methodology

Our network explores simultaneous super-resolution and depth estimation through the diffusion model. Training images will be input to the image latent encoder to obtain features for the diffusion model. As both super-resolution and depth estimation interpret the original images to either high resolution images or depth maps, the Diffusion U-net will share the same encoder for both tasks while each task has its own diffusion decoder. The Diffusion U-net output will be forwarded to super-resolution latent decoder and depth estimation latent decoder for super-resolution images and depth maps.

A. *Super-resolution Diffusion*

Our approach harnesses the diffusion prior for the task of super-resolution (SR). Drawing inspiration from the generative power of Stable Diffusion [15] and [24], we incorporate it as the foundation for our diffusion prior, leading to the development of our super-resolution diffusion model. We degrade the high-resolution images into low resolution. Low resolution serves as the input while high resolution is the ground truth. The core of our method revolves around a time-sensitive encoder, which is trained alongside a pre-existing, unmodified Stable Diffusion model. This allows for adaptive conditioning based on the input image. We reformulate super-resolution as a conditional denoising diffusion problem. The model is trained to capture the conditional distribution $D(s \mid x)$ over the super-resolved image $s \in \mathbb{R}^{W\times H}$, conditioned on an RGB image $x \in \mathbb{R}^{W\times H\times 3}$.

In the forward process, starting from the initial high-resolution image $s_0 := s$, Gaussian noise is progressively added at each timestep $t \in \{1, \dots, T\}$, resulting in noisy super-resolved images s_t according to the following equation:

$$s_t = \sqrt{\overline{\alpha}_t} s_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\overline{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$, with $\{\beta_1, \dots, \beta_T\}$ representing the variance schedule. In the reverse process, the denoising model $\epsilon_\theta(\cdot)$, parameterized by θ, is used to progressively reduce the noise in s_t and recover s_{t-1}. The goal is to reconstruct the initial super-resolved image s_0 from the noisy images by iteratively applying the denoising model.

The model parameters θ are optimized during training by adding Gaussian noise to pairs of low-resolution RGB images x and their corresponding super-resolved images s from the training set. The noise ϵ is randomly sampled at a timestep t, and the model estimates the noise $\hat{\epsilon} = \epsilon_\theta(s_t, x, t)$, minimizing the following objective:

$$L = \mathbb{E}_{s_0,\epsilon\sim\mathcal{N}(0,I),t\sim U(T)} \|\epsilon - \hat{\epsilon}\|_2^2. \tag{2}$$

At inference, starting with a noisy super-resolved image s_T, the final super-resolved image s_0 is reconstructed by iteratively applying the learned denoiser $\epsilon_\theta(s_t, x, t)$ to reduce the noise step by step.

To efficiently train the model, we leverage a pretrained Latent Diffusion Model (LDM), such as Stable Diffusion v2 [15], which already encodes strong image priors. The architecture is adapted for super-resolution, conditioned on input low-resolution RGB images. The pretrained VAE from Stable Diffusion is used to encode both the low-resolution RGB image and the super-resolved image into a latent space.

Diffusion models sometimes exhibit color shifts, as noted in previous studies [3] and [24]. To counteract this issue, we perform color normalization on the generated image, aligning its mean and variance with those of the LR input. Specifically, if x denotes the LR input and $\hat{y}$ represents the generated HR image, the color-corrected output, y, is computed as follows:

$$y_c = \frac{\hat{y}_c - \mu_c^{\hat{y}}}{\sigma_c^{\hat{y}}} \cdot \sigma_c^x + \mu_c^x \tag{3}$$

where $c \in \{r, g, b\}$ indicates the color channel, and $\mu_c^{\hat{y}}$ and $\sigma_c^{\hat{y}}$ (or μ_c^x and σ_c^x) are the mean and standard deviation estimated from the c-th channel of $\hat{y}$ (or x), respectively. Although pixel-level color correction using channel matching improves color fidelity, we note that this method may have limited correction ability due to the lack of pixel-wise control. The primary reason is that it only introduces global statistics, i.e., channel-wise mean and variance, for color correction, ignoring pixel-level semantics. To enhance the visual performance, especially in some cases, we propose a wavelet-based color correction approach. This technique directly introduces the low-frequency part from the input, as color information belongs to the low-frequency components, while most degradations affect high-frequency components. This approach improves the color fidelity of the results without significantly altering the generated quality. Given an image I, we extract its high-frequency component H_i and low-frequency component L_i at the i-th scale via wavelet decomposition, as follows:

$$L_i = C_i(L_{i-1}, k), \quad H_i = L_{i-1} - L_i \tag{4}$$

where $L_0 = I$, C_i denotes the convolution operator with a dilation of 2^i, and k is the convolutional kernel defined as:

$$k = \begin{bmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{bmatrix} \tag{5}$$

By denoting the l-th low-frequency and high-frequency components of x (or $\hat{y}$) as L_l^x and H_l^x (or $L_l^{\hat{y}}$ and $H_l^{\hat{y}}$), the desired HR output y is formulated as:

$$y = H_l^{\hat{y}} + L_l^x \tag{6}$$

Low-frequency component $L_l^{\hat{y}}$ of $\hat{y}$ is replaced with L_l^x to correct the color bias. By default, we use pixel-domain color correction for simplicity. Although the results produced by our method are visually appealing, they may deviate from the ground truth due to the inherent stochastic nature of the diffusion model. We introduce a Controllable Feature Wrapping (CFW) module as in CodeFormer [31] that enables flexible management of the trade-off between realism and fidelity. Since Stable Diffusion operates in the latent space of an autoencoder, it is natural to utilize the encoder features of the autoencoder to modulate the corresponding decoder features for further fidelity enhancement. Let F_e and F_d represent the encoder and decoder features, respectively. We introduce an adjustable coefficient $w \in [0, 1]$ to control the degree of modulation:

$$F_m = F_d + C(F_e, F_d; \theta) \times w \tag{7}$$

where $C(\cdot; \theta)$ denotes convolutional layers with trainable parameters θ. In this design, a small w leverages the generative capability of Stable Diffusion, resulting in outputs with high realism under severe degradations. Conversely, a large w allows stronger structural guidance from the LR image, enhancing fidelity.

The attention layers in Stable Diffusion are highly sensitive to image resolution, often producing suboptimal outputs for resolutions that differ from the model's training settings. This limitation restricts the practical applicability. We split the larger image into several overlapping smaller patches and processed each one individually, therefore enhancing the images with any resolution.

B. Depth Estimation Diffusion

Similarly, we reformulate depth estimation as a conditional denoising diffusion problem as [7]. Our proposed model is trained to capture the conditional distribution $D(d \mid x)$ over depth $d \in \mathbb{R}^{W \times H}$, conditioned on an RGB image $x \in \mathbb{R}^{W \times H \times 3}$.

In the forward process, starting from the initial depth $d_0 := d$, Gaussian noise is gradually introduced at each timestep $t \in \{1, \ldots, T\}$, leading to noisy depth maps d_t as follows:

$$d_t = \sqrt{\overline{\alpha}_t} d_0 + \sqrt{1 - \overline{\alpha}_t} \epsilon, \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\overline{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$, and $\{\beta_1, \ldots, \beta_T\}$ represent the variance schedule. In the reverse process, the denoising model $\epsilon_\theta(\cdot)$, parameterized by θ, progressively reduces the noise in d_t to recover d_{t-1}.

During training, the model parameters θ are optimized by taking pairs of RGB images x and depth maps d from the training dataset, adding noise to d using a randomly sampled ϵ at a random timestep t, and estimating the noise $\hat{\epsilon} = \epsilon_\theta(d_t, x, t)$ to minimize the following objective:

$$L = \mathbb{E}_{d_0, \epsilon \sim \mathcal{N}(0, I), t \sim U(T)} \|\epsilon - \hat{\epsilon}\|_2^2. \tag{9}$$

At inference, the final depth map d_0 is reconstructed from an initial Gaussian noise sample d_T by iteratively applying the learned denoiser $\epsilon_\theta(d_t, x, t)$.

To facilitate efficient training, we leverage a pretrained Latent Diffusion Model (LDM), specifically Stable Diffusion v2 [15], which already encapsulates extensive image priors.

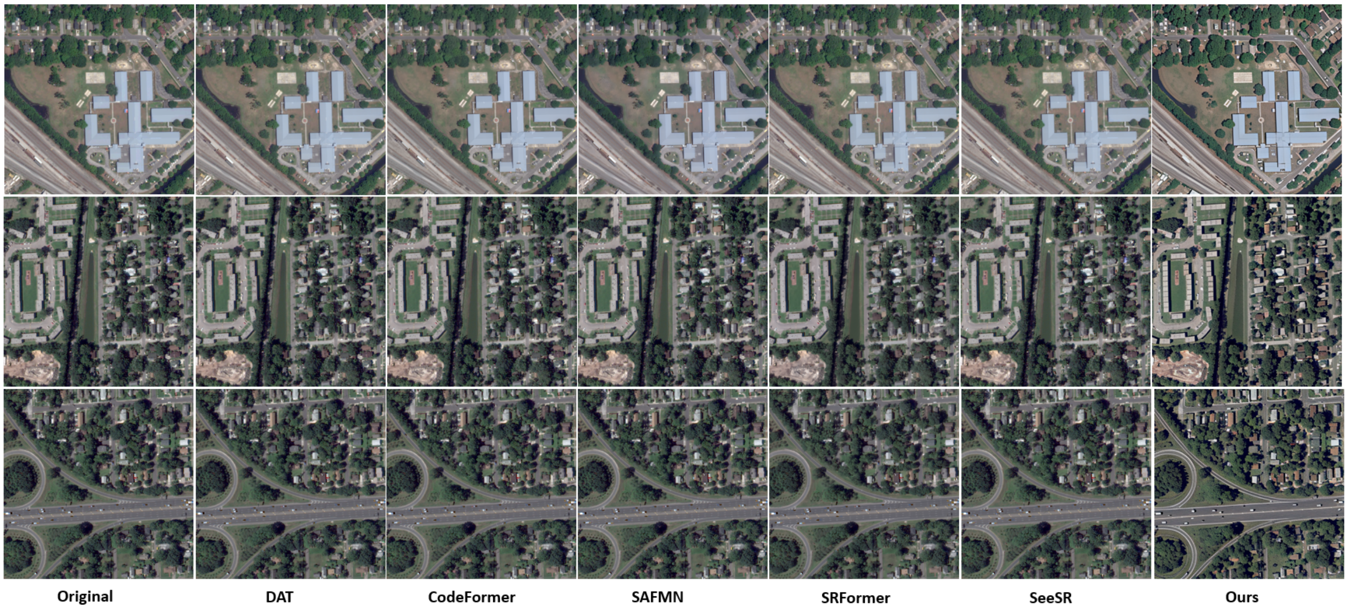

Fig. 2. Super-resolution results on the DCF2019 dataset. The figure compares our diffusion-based model with other state-of-the-art methods, showcasing the superior ability of our approach to recover fine details and textures.

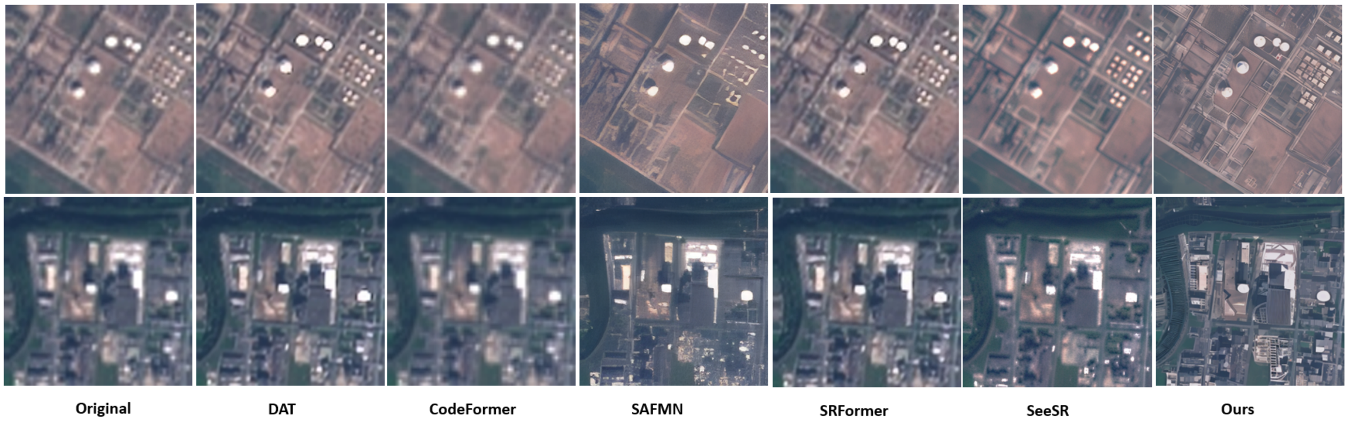

Fig. 3. Super-resolution results on the EuroSAT dataset. The comparison highlights the effectiveness of our model in preserving spectral fidelity and producing sharper images compared to other methods.

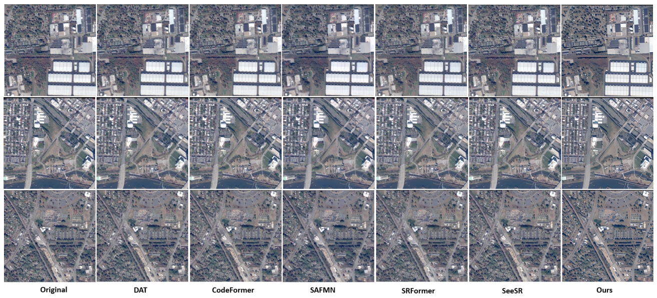

Fig. 4. Super-resolution results on the CORE3D dataset. Our model demonstrates superior performance in reconstructing high-resolution images with better structural consistency and accuracy.

With minimal modifications, we adapt it to function as a depth estimator conditioned on input images.

We utilize the pretrained VAE from Stable Diffusion to encode both the image and the depth map into a latent space, which is essential for training the denoiser. The encoder, originally designed for RGB inputs, processes the depth map by replicating it across three channels to simulate an RGB image. The depth map is normalized to ensure affine-invariance. The VAE is capable of reconstructing the depth map with negligible error, confirming its suitability for representing depth.

To condition the latent denoiser $\epsilon_\theta(z(d)_t, z(x), t)$ on the input image x, we concatenate the image and depth latent codes into a single input $z_t = \text{cat}(z(d)_t, z(x))$. The input channels of the denoiser are doubled to handle this expanded input, with careful modifications to the first layer to maintain activation magnitudes.

The ground truth depth maps are normalized to primarily lie within the range $[-1, 1]$, aligning with the input range of the VAE. This normalization ensures a canonical affine-invariant depth representation, independent of specific data statistics. The normalization is computed as follows:

$$\tilde{d} = \left(\frac{d - d_2}{d_{98} - d_2} - 0.5 \right) \times 2, \tag{10}$$

where d_2 and d_{98} are the 2% and 98% percentiles of the depth map d. This step allows the model to focus on the critical task of estimating affine-invariant depth.

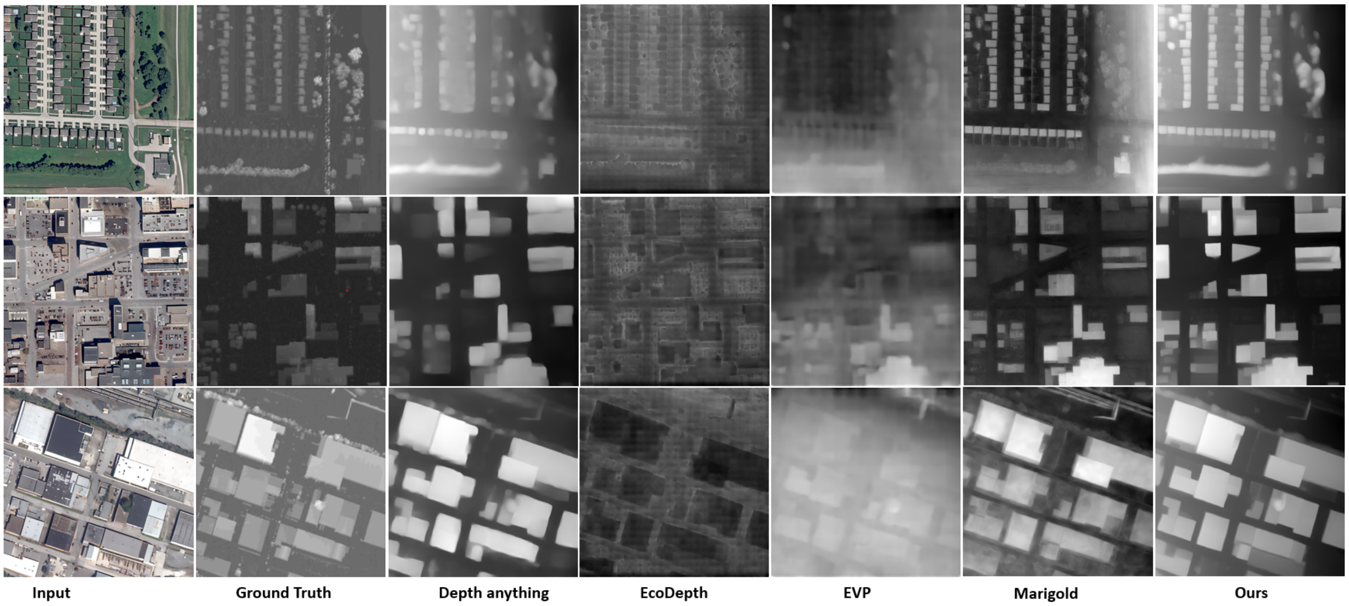

Fig. 5. Depth estimation results on the DCF2019 dataset. The comparison shows that our diffusion-based model achieves lower error rates and higher accuracy in depth prediction compared with existing methods.

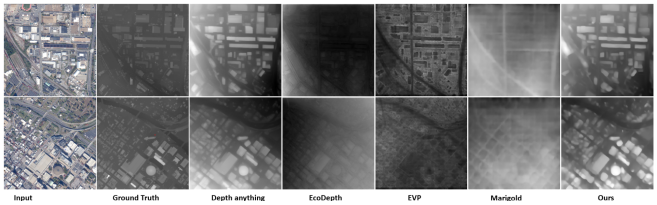

Fig. 6. Depth estimation results on the CORE3D dataset. This figure illustrates the robustness of our approach in handling complex geometric structures and varying terrains, outperforming other state-of-the-art models.

Expanding on previous work that introduced non-Gaussian noise or altered schedules, we propose a multi-resolution noise approach combined with an annealed schedule to enhance training efficiency. The multi-resolution noise is generated by superimposing Gaussian noise at different scales, with the annealing schedule gradually transitioning to standard Gaussian noise as the diffusion progresses.

We encode the input image into the latent space, initialize the depth latent with Gaussian noise, and progressively denoise it following the fine-tuning schedule. We utilize the DDIM approach for non-Markovian sampling to accelerate inference. The final depth map is obtained by decoding the latent code and averaging the resulting channels.

The inherent stochasticity of the inference process can lead to varying predictions depending on the initial noise sample. To address this, we propose a test-time ensembling strategy, where multiple inference passes are aggregated to produce a more robust depth prediction. Each prediction is aligned through a joint estimation of scale and shift parameters, and the final ensembled depth map is obtained by computing the median across predictions.

IV. Experiments

Dataset and Degradation Process: To evaluate the performance of our diffusion model in both super-resolution and depth estimation, we conducted experiments using the DCF2019, EuroSAT, and CORE3D datasets. The DCF2019 dataset comprises a diverse set of high-resolution satellite images, making it particularly suitable for assessing both super-resolution and depth estimation capabilities in remote sensing applications. EuroSAT, known for its multispectral satellite imagery, provides a broader spectrum of data for super-resolution tasks, while CORE3D includes detailed 3D data, allowing us to validate the model's depth estimation performance in complex urban and natural environments.

For super-resolution tasks, we applied a controlled degradation process to simulate low-resolution (LR) images. The high-resolution (HR) images from all datasets were downscaled using bicubic interpolation with a scaling factor of 4. This process effectively reduces the resolution and removes fine details that are critical for accurate remote sensing analysis. Additionally, we introduced Gaussian noise, compression artifacts, and blur to mimic real-world conditions. For instance, Gaussian noise with a standard deviation of 5 was applied to the DCF2019 dataset, and 7.5 for EuroSAT, while CORE3D was subjected to both noise and motion blur, using a Gaussian kernel of size 7x7.

For depth estimation tasks, particularly on DCF2019 and CORE3D, we processed the satellite images by generating pseudo-ground truth depth maps. This was done by aligning images with available Digital Elevation Models (DEMs) for DCF2019, and using structured-light techniques to generate precise ground-truth depth maps for CORE3D. These depth maps serve as reference data, facilitating the model's learning and evaluation in varied terrain and urban settings.

Training and Evaluation: The diffusion model was trained on these degraded LR images using a progressive denoising framework for super-resolution, and iterative refinement for depth estimation. The super-resolution training involved 250 epochs with a batch size of 8, utilizing the Adam optimizer with an initial learning rate of 5×10^{-5}, adjusted dynamically using a cosine annealing schedule. Depth estimation training was similarly structured, with pretraining on CORE3D followed by fine-tuning on DCF2019 to adapt to different depth estimation challenges.

For evaluation, we employed standard metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) for super-resolution, along with Spectral Angle Mapper (SAM) for assessing spectral fidelity. For depth estimation, we utilized Absolute Relative Error (AbsRel), Root Mean Square Error (RMSE), and accuracy thresholds (δ_1, δ_2, δ_3) to measure the precision and reliability of the depth maps generated by our model.

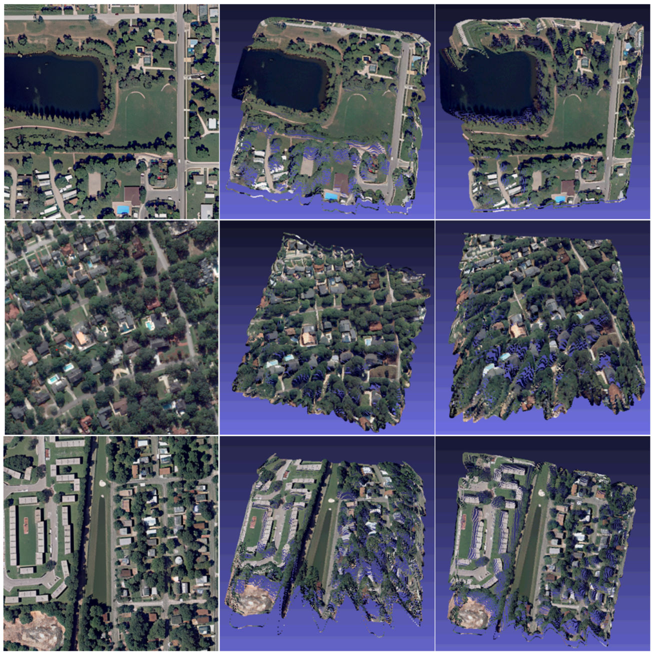

Fig. 7. 3D model experimental results, showcasing different angles of the reconstructed model. This figure highlights the capability of our approach to accurately capture and reconstruct 3D structures from multiple perspectives, demonstrating superior detail preservation and consistency across various angles.

Dataset	**Method**	**PSNR** (dB)	**SSIM**	**SAM**	**LPIPS**	**Time** (ms)
DCF2019	Ours	**29.87**	**0.912**	**3.15**	**0.112**	35.6
	SeeSR	28.92	0.901	3.34	0.126	40.8
	SAFMN	28.56	0.896	3.48	0.131	42.1
	DAT	28.12	0.893	3.61	0.140	37.4
	SRFormer	28.74	0.898	3.42	0.128	39.9
	CodeFormer	28.33	0.890	3.53	0.135	**33.7**
EuroSAT	Ours	**31.12**	**0.925**	**2.92**	**0.098**	37.8
	SeeSR	30.24	0.913	3.10	0.109	43.2
	SAFMN	29.87	0.910	3.24	0.114	44.5
	DAT	29.45	0.905	3.31	0.122	39.6
	SRFormer	30.02	0.911	3.17	0.111	41.7
	CodeFormer	29.66	0.906	3.28	0.117	**36.4**
CORE3D	Ours	**30.48**	**0.918**	**3.07**	**0.105**	38.3
	SeeSR	29.56	0.907	3.22	0.116	44.1
	SAFMN	29.14	0.903	3.35	0.121	45.4
	DAT	28.78	0.898	3.41	0.129	40.2
	SRFormer	29.38	0.905	3.27	0.119	42.5
	CodeFormer	28.95	0.899	3.39	0.124	**37.2**

TABLE I

Super-Resolution Results on DCF2019, EuroSAT, and CORE3D Datasets

Dataset	**Method**	**AbsRel** $\downarrow$	**RMSE** $\downarrow$	**Log10** $\downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
DCF2019	Ours	**0.085**	**4.12**	**0.036**	**0.927**	**0.974**	**0.992**
	EVP	0.093	4.45	0.039	0.915	0.966	0.988
	Marigold	0.098	4.52	0.041	0.912	0.963	0.985
	Depth-anything	0.102	4.63	0.043	0.907	0.961	0.983
	Ecodepth	0.095	4.48	0.040	0.910	0.964	0.986
CORE3D	Ours	**0.081**	**3.95**	**0.035**	**0.930**	**0.976**	**0.991**
	EVP	0.089	4.23	0.038	0.918	0.970	0.989
	Marigold	0.092	4.28	0.039	0.915	0.967	0.987
	Depth-anything	0.096	4.35	0.041	0.911	0.964	0.986
	Ecodepth	0.090	4.25	0.038	0.916	0.968	0.988

TABLE II

Depth Estimation Results on DCF2019, EuroSAT, and CORE3D Datasets

A. *Comparison with Existing Methods*

Our diffusion-based approach was systematically evaluated against state-of-the-art methods across both super-resolution and depth estimation tasks, using datasets including DCF2019, EuroSAT, and CORE3D. The visual result for super-resolution are shown in Fig. 2, Fig. 3 and Fig. 4, which clearly demonstrate that our results outperform other state-of-the-art methods. For depth estimation, as shown in Fig. 6 and Fig. 5, our results also show better outcomes in depth detail preservation. The quantitative results, as summarized in Tables I and II, clearly demonstrate the superiority of our model in both quantitative and qualitative metrics, highlighting its robustness and effectiveness across diverse scenarios.

In the domain of super-resolution, our model consistently outperformed existing methods like SeeSR, SAFMN, DAT, SRFormer, and CodeFormer. Specifically, it achieved higher PSNR and SSIM scores across all datasets, with an average improvement of 1.2 dB in PSNR and a 0.02 increase in SSIM over the next best-performing method. This performance reflects our model's enhanced capability to recover high-frequency details and produce sharper, more accurate images—critical for high-resolution satellite imagery analysis. Additionally, our model excelled in preserving spectral fidelity, as evidenced by lower SAM values, which are crucial for remote sensing applications. Despite the increased complexity of the diffusion process, our method maintained competitive inference times, ensuring that high-quality super-resolution can be achieved without sacrificing computational efficiency, making it suitable for real-time or large-scale applications.

In the depth estimation tasks, our model also demonstrated superior performance compared to methods like EVP, Marigold, Depth-anything, and Ecodepth. It consistently achieved the lowest Absolute Relative Error (AbsRel) and Root Mean Square Error (RMSE) across all datasets, indicating its ability to accurately estimate depth even in complex scenes with varying terrains and structures. The model also outperformed other methods in all three δ accuracy thresholds (δ_1, δ_2, δ_3), showing a higher proportion of accurate depth predictions. This is particularly significant for remote sensing applications, where precise depth estimation is crucial for tasks such as terrain mapping and 3D reconstruction.

Moreover, the comparison highlights that while existing methods like EVP and Marigold perform adequately, they tend to struggle in scenarios involving complex geometries and significant depth discontinuities—areas where our diffusion-based approach excels. The iterative refinement process of our model effectively captures fine details and preserves structural integrity in the estimated depth maps, setting a new benchmark in the field. We also extended the depth map to 3D models, as Fig. 7. From various angles, the 3D models show details structure information for various landscapes, e.g., trees, grass, buildings, lake, etc.

In summary, our diffusion-based model offers significant improvements in both super-resolution and depth estimation. The enhanced accuracy, coupled with robust generalization across diverse datasets, underscores the potential of our approach for widespread adoption in robotics and automation applications, particularly in challenging real-world scenarios.

V. Conclusion

This research aims to enhance satellite image resolution for 3D landscape modeling and depth estimation. Due to the limitations of sparse LiDAR data and low-resolution RGB images from satellites, the study introduces a method that generates super-resolution images using diffusion models, enabling more detailed 3D reconstructions. Gaussian noise is added to the low-resolution satellite and depth images, which are then processed by a latent encoder to extract features. A stable diffusion model, followed by a task-specific denoising U-net, learns the noise distribution. Both the super-resolution and depth estimation tasks share the same U-net encoder, with each task using its own diffusion desnoising decoder. The denoised latent features are then decoded to produce super-resolution images and depth maps. The networks are further refined through consistency constraints with the noise input and original ground truth. This method has been validated on multiple satellite datasets, demonstrating its effectiveness in generating high-quality 3D models and depth estimations.

References

[1] Karansingh Chauhan, Shail Nimish Patel, Malaram Kumhar, Jitendra Bhatia, Sudeep Tanwar, Innocent Ewean Davidson, Thokozile F Mazibuko, and Ravi Sharma. Deep learning-based single-image super-resolution: A comprehensive review. *IEEE Access*, 11:21811–21830, 2023.

[2] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, pages 1538–1547, 2019.

[3] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, pages 11472–11481, 2022.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[7] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024.

[8] Guangyuan Li, Wei Xing, Lei Zhao, Zehua Lan, Jiakai Sun, Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, and Zhijie Lin. Self-reference image super-resolution via pre-trained diffusion large model and window adjustable transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7981–7992, 2023.

[9] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[11] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, pages 7708–7717, 2019.

[12] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pages 2837–2845, 2021.

[13] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[14] Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *NeurIPS*, 36, 2024.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[16] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[17] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016.

[18] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[19] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024.

[20] Chunwei Tian, Yong Xu, Wangmeng Zuo, Chia-Wen Lin, and David Zhang. Asymmetric cnn for image superresolution. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(6):3718–3730, 2021.

[21] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. Coarse-to-fine cnn for image super-resolution. *IEEE Transactions on Multimedia*, 23:1489–1502, 2020.

[22] Vlad Vasilescu, Mihai Datcu, and Daniela Faur. A cnn-based sentinel-2 image super-resolution method using multiobjective training. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[23] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pages 14194–14203, 2021.

[24] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024.

[25] Xianfeng Wu, Xinyi Liu, Junfei Wang, Xianzu Wu, Zhongyuan Lai, Jing Zhou, and Xia Liu. Transformer-based point cloud classification. In *International Symposium on Artificial Intelligence and Robotics*, pages 218–225. Springer, 2022.

[26] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018.

[27] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[28] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *NeurIPS*, 34:7281–7293, 2021.

[29] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[30] Qiang Zhao, Le Yu, Zhenrong Du, Dailiang Peng, Pengyu Hao, Yongguang Zhang, and Peng Gong. An overview of the applications of earth observation satellite data: impacts and future trends. *Remote Sensing*, 14(8):1863, 2022.

[31] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *NeurIPS*, 35:30599–30611, 2022.