2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

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 [35]. 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 robot perception tasks. 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 framework is shown in Fig. 1.

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

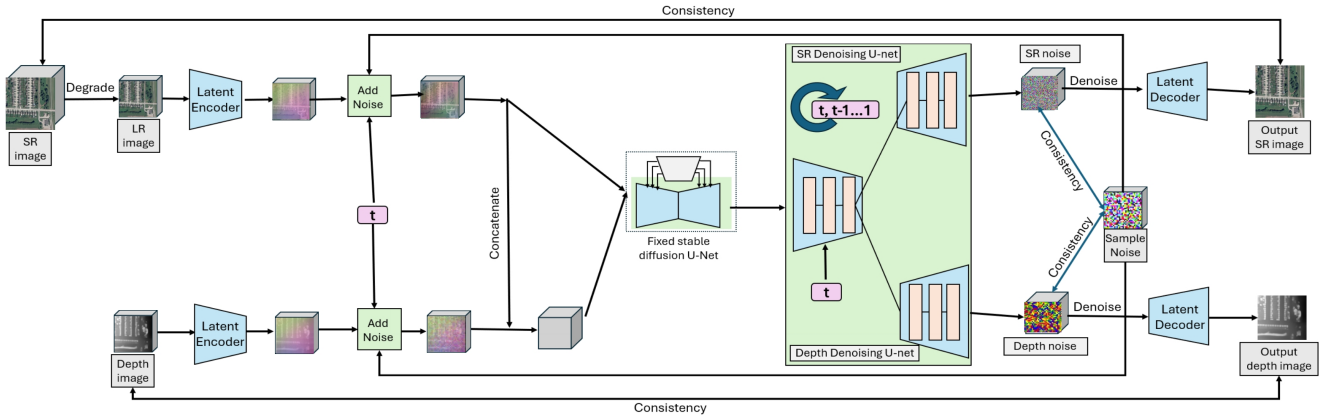

Fig. 1. The framework for our simultaneous super-resolution and depth estimation algorithm targeting satellite images. Gaussian noise is added to the low-resolution satellite images and the depth images. The noisy images are input to the latent encoder to learn the latent features. A fixed stable diffusion 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.

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 [26], [25], [1], [27] or recursive learning [23]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 [33] and Restormer[34] leverage the Transformer's ability to model long-range dependencies and are pre-trained on large-scale datasets like ImageNet [4] and COCO[11] 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) [10], 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.

Super-Resolution via Repeated Refinement (SR3) [21] exemplifies the application of diffusion models to enhance image resolution. 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) [20], which operate within a lower-dimensional latent space. 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) [9] leverages cross-attention to incorporate contextual information from high-resolution reference images.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, which be used for many applications such as crop status estimation [13] and localization [16], [12]. Traditional methods like Structure from Motion (SfM) [22], simultaneous localization and mapping (SLAM) [15], and Multi-View Stereo (MVS) [31] 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 [32] and volumetric reconstruction [28]. For example, methods such as DeepMVS [2] combine deep learning with MVS for enhanced accuracy. Hybrid approaches that integrate geometric constraints with neural networks are also gaining traction [14]. Additionally, Neural Radiance Fields (NeRF) [18] and Transformer-based architectures [30] offer innovative solutions by modeling 3D scenes with photorealistic details and effectively capturing global context.

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) [7], 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 [17] 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 [17]. Hybrid approaches that integrate diffusion models with other deep learning techniques have also been investigated. For example, integrating diffusion 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 [19]. 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 [24]. 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. The entire framework involves super-resolution, depth estimation, and joint training, described in the following subsections.

A. Joint Training Framework

Training images will be input to the image latent encoder to obtain features for the diffusion model. Both degraded super-resolution images and the corresponding depth maps are input into the network, for both SR resolution and depth estimation tasks. The diffusion models are learned in latent feature space. As depth estimation also involves RGB images as the input to learn the image content and structure, the latent features from RGB images are also input to the depth estimation branch. Therefore, RGB latent features are jointly learned with the depth latent features for depth estimation.

Besides feature level joint learning, 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. This network design enables network-level joint learning.

The joint learning scheme involves both diffusion optimization and consistency constraint. In the diffusion process, the diffusion models estimate the noise during the inverse process. Time t is input to the diffusion denoising U-net. The noise output from the diffusion models ideally should be the same as the noise input during the forward process. The Diffusion U-net feature outputs are forwarded to super-resolution latent decoder and depth estimation latent decoder for super-resolution images and depth maps, which we build consistency with the original input of the framework to further optimize the network.

B. 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 [20] and [29], 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 SR 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 $t \in \{1, \ldots, T\}$, resulting in noisy SR images s_t according to the equation below:

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

where $\epsilon \sim \mathcal{N}(0, I)$ and $\overline{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$, with $\{\beta_1, \ldots, \beta_T\}$ 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 SR 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 SR 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 SR image s_T, the final SR 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 [20], 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 SR image into a latent space.

Diffusion models sometimes exhibit color shifts, as noted in previous studies [3] and [29]. To resolve this problem, we perform color normalization on the generated image by matching its mean and variance to those of the low-resolution (LR) input. In particular, let P denote the LR input and $\hat{Q}$ the high-resolution (HR) image that is generated. The resulting color-corrected output, Q, is subsequently calculated as:

$$Q_c = \frac{\hat{Q}_c - \mu_c^{\hat{Q}}}{\sigma_c^{\hat{Q}}} \cdot \sigma_c^P + \mu_c^P \tag{3}$$

where $c \in \{r, g, b\}$ denotes the color channel, and $\mu_c^{\hat{Q}}$ and $\sigma_c^{\hat{Q}}$ (or μ_c^P and σ_c^P) represent the mean and standard deviation computed from the c-th channel of $\hat{Q}$ (or P), respectively. While channel matching for pixel-level color correction enhances color fidelity, this method's effectiveness may be constrained by the absence of pixel-wise precision. The underlying limitation stems from its reliance on global statistics, such as the mean and variance of each channel, which overlooks fine-grained pixel-level semantic details. To address this limitation and improve visual outcomes in specific scenarios, we introduce a wavelet-based approach to color correction. This method leverages the fact that color information is primarily found in the low-frequency components, whereas most degradations occur in the high-frequency domain. By incorporating the low-frequency content from the input image, we enhance color fidelity without noticeably affecting overall visual quality. Specifically, for an image I, we apply wavelet decomposition to isolate its high-frequency component H_i and low-frequency component L_i at the i-th scale, as defined below:

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

where $L_0 = I$, C_i represents the convolution operator with a dilation factor of 2^i, and k is the convolutional kernel. By denoting the l-th low-frequency and high-frequency components of P (or $\hat{Q}$) as L_l^P and H_l^P (or $L_l^{\hat{Q}}$ and $H_l^{\hat{Q}}$), the target high-resolution output Q can be expressed as:

$$Q = H_l^{\hat{Q}} + L_l^P \tag{5}$$

The low-frequency component $L_l^{\hat{Q}}$ of $\hat{Q}$ is substituted with L_l^P to mitigate color bias. For simplicity, we apply pixel-domain color correction as the default approach. While the outcomes generated by our approach are visually striking, they may differ from the ground truth due to the intrinsic randomness of the diffusion model. Furthermore, we employ a Controllable Feature Wrapping (CFW) module, as proposed in CodeFormer [36] and [29], providing adjustable control over the trade-off between fidelity and realism.

$$I_f = I_d + C(I_e, I_d; \theta) \times w \tag{6}$$

where $C(\cdot; \theta)$ represents convolutional layers with trainable parameters θ. In this setup, a smaller w emphasizes the generative capabilities of Stable Diffusion, producing highly realistic outputs even under severe degradations. On the other hand, a larger w strengthens structural guidance from the low-resolution (LR) image, thereby improving fidelity. With the predicted super-resolution images, we also introduce appearance loss, which is to make the output super-resolution images to be consistent with the original super-resolution images before degradation. During the training process, we want to maintain the SR image generated by the diffusion model to match with the original input SR image by combining the Structural Similarity Index Metric (SSIM) structure [6] and L1 as photometric image loss $L_{SRconsistency}$.

$$\begin{aligned} L_{SRconsistency} = \frac{1}{N} \sum_{i,j} \alpha \frac{1 - SSIM(I_{ij}, \widetilde{I}_{ij})}{2} \\ + (1 - \alpha) \left\| I_{i,j}^l - \widetilde{I}_{i,j}^l \right\|_1 \end{aligned} \tag{7}$$

where N is the number of pixels, I_{ij} and $\widetilde{I}_{ij}$ respectively represent input original SR image and its output SR image. And I_{ij} is i th row and j th column pixel in the original input image. The simplified SSIM with 3 × 3 block filter is used. And let α = 0.85.

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 divide the larger image into multiple overlapping patches, processing each one independently. This approach allows for effective enhancement across images of varying resolutions.

C. Depth Estimation Diffusion

Similarly, we estimate the scene depth through conditional denoising diffusion as [8]. Our model is trained to model the conditional probability distribution $D(d \mid x)$, where the depth map $d \in \mathbb{R}^{W \times H}$ is predicted based on an RGB image $x \in \mathbb{R}^{W \times H \times 3}$. In the forward diffusion process, starting with the original depth map $d_0 := d$, Gaussian noise is progressively added over a sequence of timesteps $t \in \{1, \ldots, T\}$, producing noisy depth maps d_t as described by the equation:

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

where $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise, $\overline{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$, and $\{\beta_1, \ldots, \beta_T\}$ defines the noise variance schedule. During the reverse process, the model $\epsilon_\theta(\cdot)$, parameterized by θ, iteratively reduces the noise from the noisy depth map d_t to recover the previous depth map d_{t-1}.

Training involves optimizing the parameters θ by adding noise to pairs of input RGB images x and their corresponding depth maps d, drawn from the dataset. Noise ϵ is introduced at a random timestep t, and the model estimates the noise $\hat{\epsilon} = \epsilon_\theta(d_t, x, t)$. The loss function that is minimized during training is given by:

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

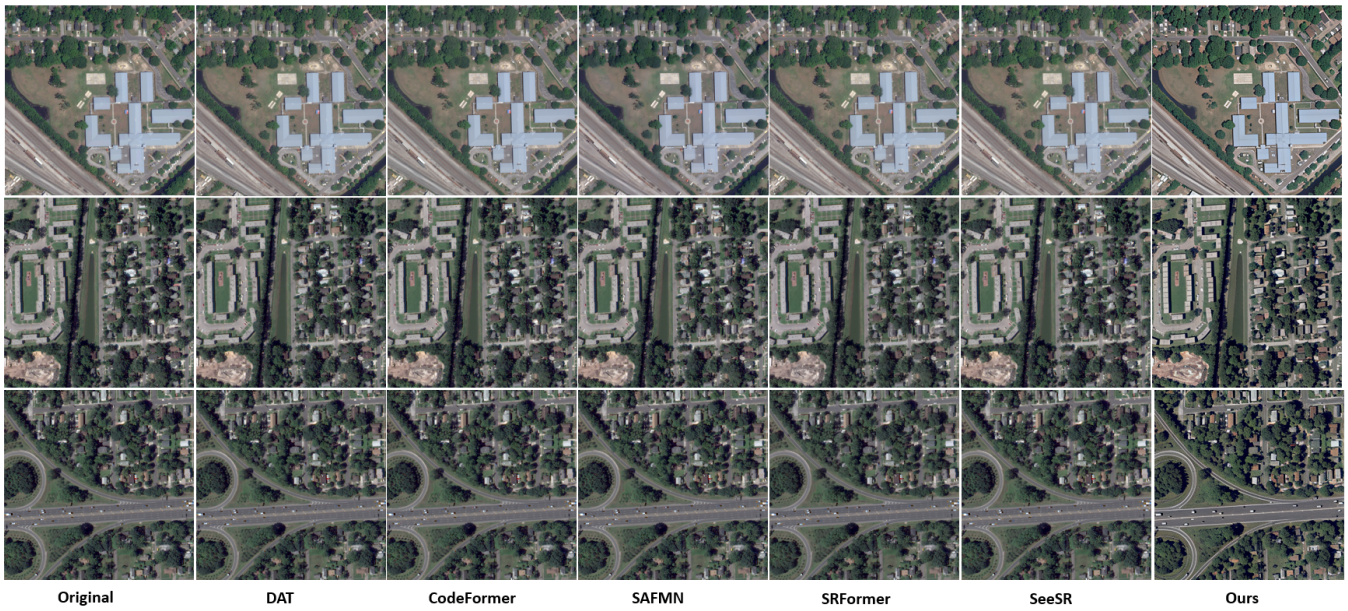

Fig. 2. Super-resolution results on the DCF2019 dataset. The figure compares our diffusion-based model with other state-of-the-art methods, showcasing the superior ability of our approach to recover fine details and textures.

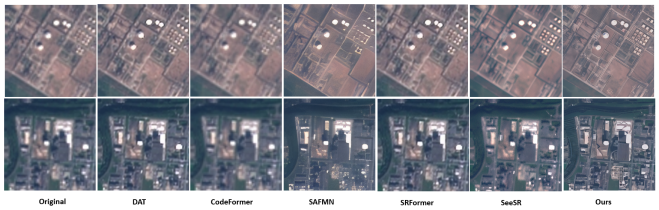

Fig. 3. Super-resolution results on the EuroSAT dataset. The comparison highlights the effectiveness of our model in preserving spectral fidelity and producing sharper images compared to other methods.

An initial noisy depth map d_T sampled from Gaussian noise is input during the inference stage. The trained denoising network $\epsilon_\theta(d_t, x, t)$ is then applied iteratively to progressively refine the noisy depth map and reconstruct the clean depth map d_0.

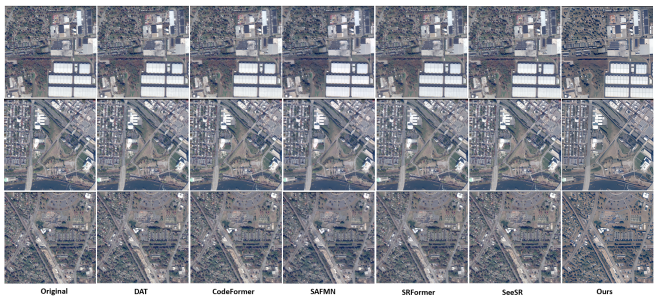

Fig. 4. Super-resolution results on the CORE3D dataset. Our model demonstrates superior performance in reconstructing high-resolution images with better structural consistency and accuracy.

To enable efficient training, we still employ a pre-trained Latent Diffusion Model (LDM), specifically Stable Diffusion v2 [20] as super-resolution.We make use of the pre-trained Variational Autoencoder (VAE) from Stable Diffusion to map both the RGB image and the depth map into a shared latent space. This latent encoding is crucial for training the denoising model. Since the VAE was initially developed for RGB images, the depth map is expanded by duplicating it across three channels, thereby mimicking the structure of an RGB input. Additionally, the depth map is normalized to maintain affine invariance. The VAE efficiently reconstructs the depth map with minimal loss, verifying its ability to represent depth data accurately.

In order to condition the latent denoiser $\epsilon_\theta(z(d)_t, z(x), t)$ on the input RGB image x, we concatenate the latent representations of both the depth map and the image into a unified input, denoted by $z_t = \text{cat}(z(d)_t, z(x))$. Consequently, the denoising network's input channels are doubled to accommodate this combined latent space. The first layer is carefully adjusted to ensure consistent activation levels across the expanded input space.

The ground truth depth maps undergo normalization to ensure they predominantly lie within the range of $[-1, 1]$, corresponding to the input range expected by the VAE. This transformation guarantees that the depth representations are affine-invariant, regardless of variations in the underlying data distribution. The normalization process is defined by the following equation:

$$\tilde{d} = \left(\frac{d - d_2}{d_{95} - d_5} - 0.5 \right) \times 2, \tag{10}$$

where d_5 and d_{95} represent the 5% and 95% percentiles of the depth values in d. This normalization step is critical for ensuring that the model focuses solely on estimating affine-invariant depth.

Building on prior methodologies that employed non-Gaussian noise models or modified schedules, we introduce a multi-scale noise strategy combined with an annealing schedule to further optimize training efficiency. Multi-resolution noise is constructed by layering Gaussian noise across different scales, and as the diffusion process advances, the annealing schedule gradually shifts towards conventional Gaussian noise.

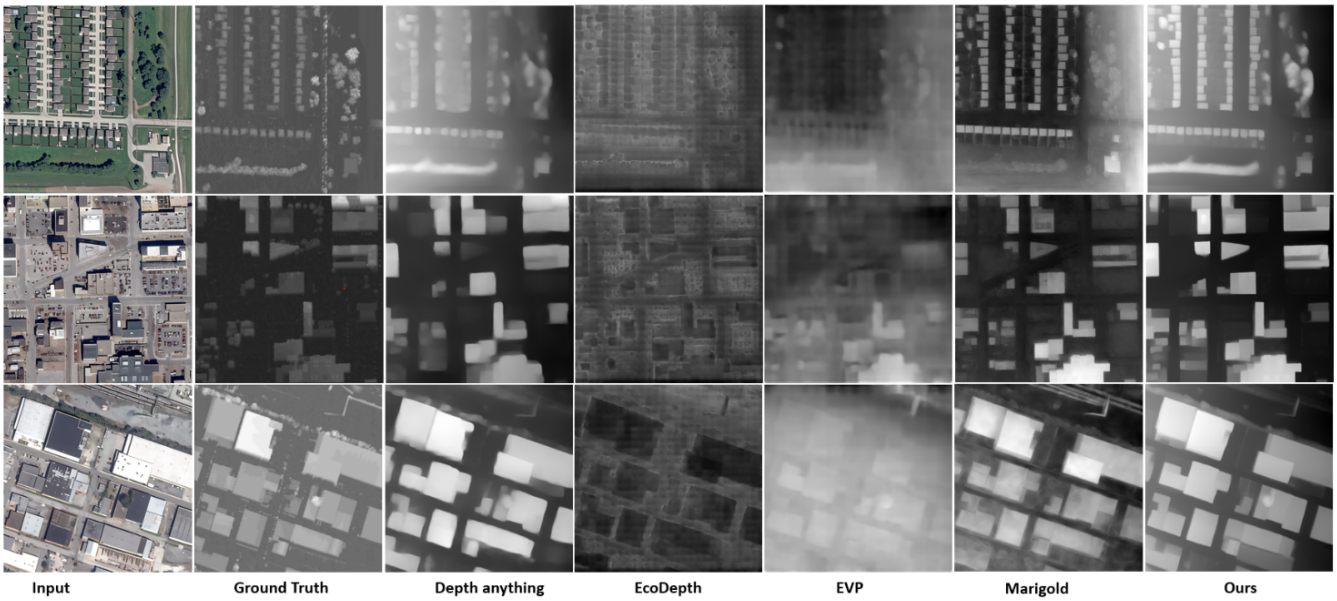

Fig. 5. Depth estimation results on the DCF2019 dataset. The comparison shows that our diffusion-based model achieves lower error rates and higher accuracy in depth prediction compared with existing methods.

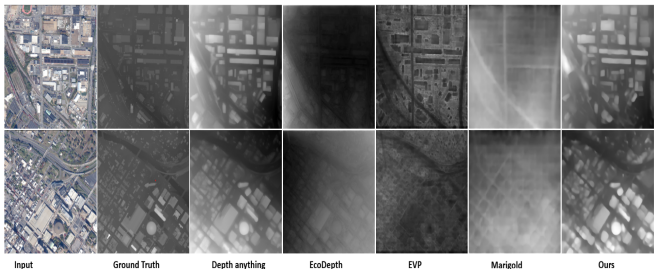

Fig. 6. Depth estimation results on the CORE3D dataset. This figure illustrates the robustness of our approach in handling complex geometric structures and varying terrains, outperforming other state-of-the-art models.

To generate predictions, the input image is encoded into the latent space, while the depth latent is initialized with Gaussian noise. The depth map is progressively denoised according to the fine-tuning schedule, utilizing the DDIM (Denoising Diffusion Implicit Models) technique for non-Markovian sampling to expedite inference. The final depth map is produced by decoding the latent representation and averaging the output across channels. The final depth map is enforced to be consistent with the depth ground truth, as the following equation.

$$L_{Dconsistency} = \frac{\alpha}{2}(1 - SSIM(d_{est}, d_{gt})) + (1-\alpha)(||d_{est} - d_{gt}||_1) \quad (11)$$

Here d_{est} refers to the estimated depth from the diffusion model and d_{gt} is the ground truth. Through L2 loss and SSIM loss, the depth value difference and the structural information are learned. Due to the inherent stochastic nature of the inference phase, predictions can vary depending on the initial noise sample. To mitigate this variance, we introduce a test-time ensembling technique, where multiple predictions are aggregated for enhanced depth estimation. Each prediction is adjusted via joint optimization of scale and shift parameters, and the final depth map is derived by computing the median across the ensemble of 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 re-

finement 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 pre-training 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.

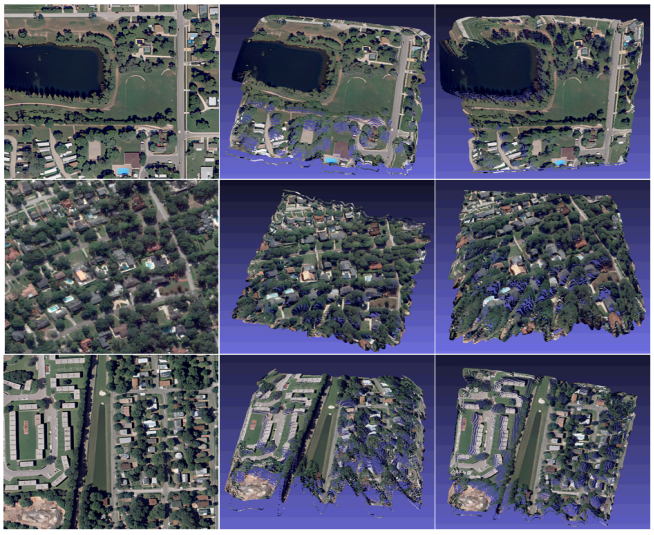

Fig. 7. 3D model experimental results, showcasing different angles of the reconstructed model. This figure highlights the capability of our approach to accurately capture and reconstruct 3D structures from multiple perspectives, demonstrating superior detail preservation across various angles.

Dataset	Method	PSNR (dB)	SSIM	SAM	LPIPS	FID
DCF2019	Ours	**29.87**	**0.912**	**3.15**	**0.112**	**36.4**
	SeeSR	28.92	0.901	3.34	0.126	39.2
	SAFMN	28.56	0.896	3.48	0.131	44.8
	DAT	28.12	0.893	3.61	0.140	47.6
	SRFormer	28.74	0.898	3.42	0.128	48.7
	CodeFormer	28.33	0.890	3.53	0.135	46.2
EuroSAT	Ours	**31.12**	**0.925**	**2.92**	**0.098**	**54.8**
	SeeSR	30.24	0.913	3.10	0.109	61.1
	SAFMN	29.87	0.910	3.24	0.114	63.4
	DAT	29.45	0.905	3.31	0.122	68.6
	SRFormer	30.02	0.911	3.17	0.111	69.5
	CodeFormer	29.66	0.906	3.28	0.117	65.9
CORE3D	Ours	**30.48**	**0.918**	**3.07**	**0.105**	**35.1**
	SeeSR	29.56	0.907	3.22	0.116	41.5
	SAFMN	29.14	0.903	3.35	0.121	42.8
	DAT	28.78	0.898	3.41	0.129	48.9
	SRFormer	29.38	0.905	3.27	0.119	40.3
	CodeFormer	28.95	0.899	3.39	0.124	45.7

TABLE I

SUPER-RESOLUTION 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.

Dataset	Method	AbsRel ↓	RMSE ↓	Log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DCF2019	Ours	**0.085**	**4.12**	**0.036**	**0.927**	**0.974**	**0.992**
	Depth-anything	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
	EVP	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**
	Depth-anything	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
	EVP	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

In the domain of super-resolution, our model consistently outperformed existing methods like SeeSR, SAFMN, DAT, SRFormer, and CodeFormer, as Fig. 2, Fig. 3, Fig. 4 and Table I. Specifically, it achieved higher PSNR and SSIM scores across all datasets, with an average improvement of 0.95 dB in PSNR and a 0.012 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. Our model also has the lowest FID value. 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 as Fig. 6 and Fig. 5. As shown in Table II, our method 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, lakes, 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 noise inputs and original ground truth. The validation on multiple satellite datasets demonstrates the effectiveness in generating high-quality 3D models and depth maps of our framework.

Acknowledgement

The work is under support of NIH Grant Number 1R15EB034519-01A1 and NSF Grant Number 2346790.

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*, 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 *CVPR*, pages 248–255, 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] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

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

[8] 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.

[9] G. Li, W. Xing, L. Zhao, Z. Lan, J. Sun, Z. Zhang, Q. Zhang, H. Lin, and Z. Lin. Self-reference image super-resolution via pre-trained diffusion large model and window adjustable transformer. In *MM*, pages 7981–7992, 2023.

[10] 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.

[11] 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 *ECCV*, 2014.

[12] Guoyu Lu. Image-based localization for self-driving vehicles based on online network adjustment in a dynamic scope. In *IJCNN*, 2022.

[13] Guoyu Lu. Bird-view 3d reconstruction for crops with repeated textures. In *IEEE IROS*, pages 4263–4270, 2023.

[14] Guoyu Lu. Deep unsupervised visual odometry via bundle adjusted pose graph optimization. In *IEEE ICRA*, pages 6131–6137, 2023.

[15] Guoyu Lu. Slam based on camera-2d lidar fusion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[16] Guoyu Lu, Yan Yan, Li Ren, Jingkuan Song, Nicu Sebe, and Chandra Kambhamettu. Localize me anywhere, anytime: a multi-task point-retrieval approach. In *IEEE ICCV*, pages 2434–2442, 2015.

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

[18] 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*, pages 99–106, 2021.

[19] 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.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

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

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

[23] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, pages 3147–3155, 2017.

[24] 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.

[25] Chunwei Tian, Yong Xu, Wangmeng Zuo, Chia-Wen Lin, and David Zhang. Asymmetric cnn for image superresolution. *IEEE TSMC-S*, 52(6):3718–3730, 2021.

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

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

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

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

[30] Xianfeng Wu, Xinyi Liu, Junfei Wang, Xianzu Wu, Zhongyuan Lai, Jing Zhou, and Xia Liu. Transformer-based point cloud classification. In *ISAIR*, pages 218–225. Springer, 2022.

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

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

[33] 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.

[34] 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*, 2022.

[35] 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.

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