MV2MV: Multi-View Image Translation via View-Consistent Diffusion Models

YOUCHENG CAI, University of Science and Technology of China, China
RUNSHI LI, University of Science and Technology of China, China
LIGANG LIU[*], University of Science and Technology of China, Laoshan Laboratory, China

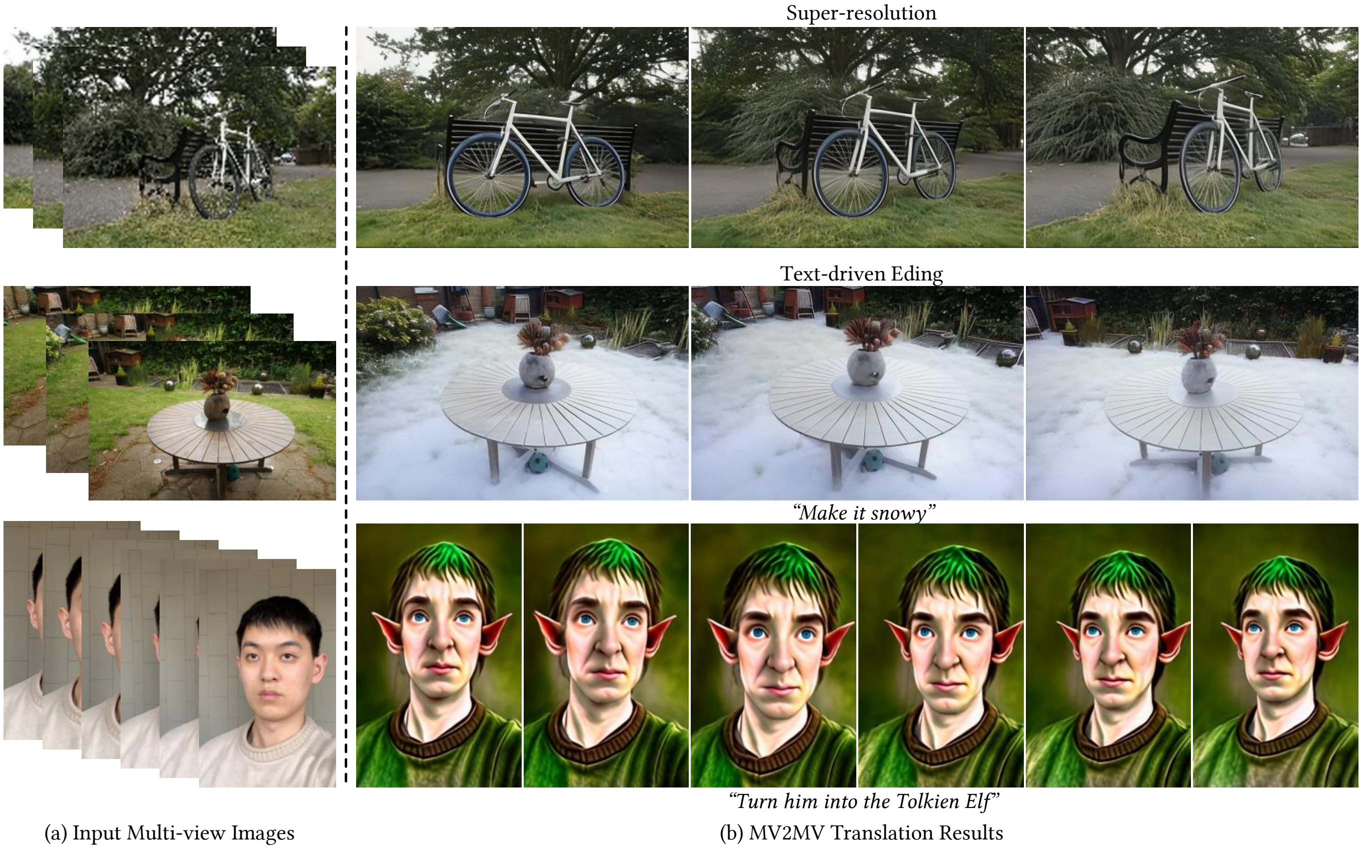
Super-resolution
Text-driven Eding
"Make it snowy"
"Turn him into the Tolkien Elf"
(a) Input Multi-view Images
(b) MV2MV Translation Results

Fig. 1. We present MV2MV, a unified multi-view image to multi-view image translation framework, enabling various multi-view image translation tasks such as super-resolution (top row), text-driven editing (2nd and 3rd rows), etc. Our method achieves high-quality results with fine details while maintaining view consistency. More results can be seen in the accompanying video.

Image translation has various applications in computer graphics and computer vision, aiming to transfer images from one domain to another. Thanks to the excellent generation capability of diffusion models, recent single-view image translation methods achieve realistic results. However, directly applying diffusion models for multi-view image translation remains challenging for two major obstacles: the need for paired training data and the limited view consistency. To overcome the obstacles, we present a first unified multi-view image to multi-view image translation framework based on diffusion models, called MV2MV. Firstly, we propose a novel self-supervised training strategy that exploits the success of off-the-shelf single-view image translators and the 3D Gaussian Splatting (3DGS) technique to generate pseudo ground truths as supervisory signals, leading to enhanced consistency and fine details. Additionally, we propose a latent multi-view consistency block, which utilizes the latent-3DGS as the underlying 3D representation to facilitate information exchange across multi-view images and inject 3D prior into the

[*] Corresponding author: Ligang Liu (lgliu@ustc.edu.cn)

Authors' Contact Information: Youcheng Cai, caiyoucheng@ustc.edu.cn, University of Science and Technology of China, Hefei, China; Runshi Li, University of Science and Technology of China, Hefei, China, stflrs@mail.ustc.edu.cn; Ligang Liu, University of Science and Technology of China, Laoshan Laboratory, Hefei, China, lgliu@ustc.edu.cn.

ACM 1557-7368/2024/9-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

diffusion model to enforce consistency. Finally, our approach simultaneously optimizes the diffusion model and 3DGS to achieve a better trade-off between consistency and realism. Extensive experiments across various translation tasks demonstrate that MV2MV outperforms task-specific specialists in both quantitative and qualitative.

CCS Concepts: • **Computing methodologies** → Computer vision; Image manipulation; Rendering; Point-based models.

Additional Key Words and Phrases: Image Editing, Diffusion Models, Gaussian Splatting

ACM Reference Format:
Youcheng Cai, Runshi Li, and Ligang Liu. 2024. MV2MV: Multi-View Image Translation via View-Consistent Diffusion Models. *ACM Trans. Graph.* 1, 1 (September 2024), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Image translation is a long-standing problem in computer graphics and computer vision [Isola et al. 2017; Parmar et al. 2023], which takes input images as the condition to output target images. Many problems can be considered as image-to-image translation, which transfers images from a source domain to a target domain while preserving the content representations [Pang et al. 2021], including super-resolution [Vavilala and Meyer 2021; Wang et al. 2020], deblurring [Lee and Cho 2013; Zhang et al. 2022], denoising [Chen et al. 2023a; Gu et al. 2024], editing [Brooks et al. 2023; Kawar et al. 2023], etc. Existing image translation methods usually focus on single-view images. While these methods produce promising results for single-view image processing in the respective tasks, they encounter difficulties when applied to multi-view image translation tasks. This is because simply performing frame-by-frame image translation poses 3D consistency issues that can lead to inconsistent geometry and appearance across different views.

A popular strategy is to use Neural Radiance Fields (NeRF), which is a continuous scene representation, to implement NeRF-to-NeRF translation as a means of indirectly achieving multi-view translation, such as NeRF-SR [Wang et al. 2022], Deblur-NeRF [Ma et al. 2022], NAN [Pearl et al. 2022] and Instruct-NeRF2NeRF [Haque et al. 2023], etc. However, these NeRF-based translation methods suffer from the following limitations: (1) the rendering resolution, which produces artifacts when the resolution diverges from those seen during training; (2) the rendering quality, as the training and rendering processes of NeRF inevitably result in information loss.

Recent diffusion models showcase formidable generative capabilities in the field of 2D image processing, and a range of methods have emerged to support versatile image translation [Parmar et al. 2023; Saharia et al. 2022; Zhang et al. 2023], which have achieved impressive translation results. This motivates us to raise an intriguing question: can we conduct unified multi-view image translation by leveraging diffusion models as well? Implementing multi-view image translation directly on the image domain is able to take better advantage of existing 2D generative priors, e.g., Stable Diffusion [Rombach et al. 2022], to achieve more flexible processing and obtain more realistic results. The challenges are two folds. The first one is the need for paired training data. Supervised learning with paired real-world data will greatly enhance the generalization ability of the model, which can effectively adapt to the complexity and variability of real scenarios. It is notable that collecting real-world high-quality/low-quality multi-view image pairs is often prohibitively expensive or unavailable. The second challenge is the difficulty of guaranteeing view consistency. The generative nature of diffusion models results in diverse and inconsistent content being inevitably generated for different views when translating the multi-view images individually.

To tackle the challenges above, we propose a unified multi-view image to multi-view image translation framework, called MV2MV, based on diffusion models for various multi-view image translation tasks such as super-resolution, denoising, deblurring and text-driven editing (see Fig. 1). Firstly, we introduce a novel self-supervised training strategy, called Consistent and Adversarial Supervision (CAS). Specifically, we first process multi-view images individually using off-the-shelf single-view image translators to obtain a set of high-quality outputs, and then feed them into 3D Gaussian Splatting (3DGS) to average out the inconsistencies and yield consistent outputs. These two outputs are regarded as pseudo ground truths serving as supervisory signals, and consistent loss and adversarial loss are introduced to effectively combine the advantages of the two pseudo ground truths to ensure both consistency and realism. Secondly, we propose a plug-in latent multi-view consistency block, named LAConsistNet, to construct our view-consistent diffusion model (VCDM). Specifically, the LAConsistNet block utilizes a latent-3DGS as the underlying 3D representation to ensure information exchange among multi-view images, thereby guaranteeing multi-view consistency. Finally, we introduce a joint optimization strategy by simultaneously training VCDM and 3DGS to ensure the consistency of the details derived from the adversarial loss, resulting in a better trade-off between consistency and realism.

To summarize, we provide the following contributions:

- A unified multi-view image to multi-view image translation framework processed on the image domain for various translation tasks.
- A generative multi-view diffusion model that enables better view consistency as well as high-quality detail.

As far as we know, our approach is the first unified framework for view-consistent multi-view image to multi-view image translation based on diffusion models. We conducted extensive experiments in terms of both qualitatively and quantitatively to validate the effectiveness of the proposed method. The experimental results have shown the superiority of our method in various multi-view image translation tasks, such as super-resolution, denoising, deblurring and text-driven editing. Our method not only generates images with richer details but also achieves remarkable improvements in view consistency.

2 RELATED WORKS

2.1 Single-view Image Translation

Most of the existing researches focus on single-view image translation tasks. Early image translation methods [Chen et al. 2023c; Iizuka et al. 2016; Li et al. 2023a,b] are typically reconstruction-based, where network architectures are designed based on assumed prior knowledge of image translation. While these methods may achieve impressive performance on specific datasets, their performance significantly deteriorates in real-world scenarios due to limited generalizability. To solve this issue, generative priors for image translation have been widely exploited in the form of generative adversarial networks (GANs) [Goodfellow et al. 2014]. For example, various models are developed for specific applications such as super-resolution [Ledig et al. 2017], style transfer [Kwon and Ye 2022], texture synthesis [Li and Wand 2016] and inpainting [Pathak et al. 2016]. In particular, Isola et al [Isola et al. 2017] propose a generic solution for image-to-image translation, named Pix2Pix, which explored myriad image-to-image translation tasks using GANs. While GANs are capable of generating more realistic perceptual details, they are unstable in the training stage and often suffer from unnatural visual artifacts.

Recently, Diffusion Model [Rombach et al. 2022; Song et al. 2020; Zhang et al. 2023] has demonstrated significant advantages in image translation tasks. For image super-resolution, DiffBIR [Lin et al. 2023] and StableSR [Wang et al. 2023b] leverage the generative ability of latent diffusion models to generate realistic images. CCSR [Sun et al. 2023] proposes a non-uniform sampling and early truncation strategy to improve the consistency of the generated content. For image deblurring, HiDiff [Chen et al. 2024] performs Diffusion Model in the latent space to generate a priori features for the deblurring process to recover exquisite images. Subsequently, AutoDIR [Jiang et al. 2023] establishes a unified framework with latent diffusion capable of handling multiple image degradations through joint training with various image restoration tasks. For image editing, InstructPix2Pix [Brooks et al. 2023] stands out by efficiently editing images following instructions, which leverages large pre-trained models in the language and image domains [Brown et al. 2020] to generate paired data for training.

2.2 NeRF-to-NeRF Translation

Single-view image translation methods cannot be directly applicable to multi-view image scenes due to the inherent 3D consistency of multi-view images. To address this problem, existing studies usually rely on 3D implicit fields of NeRF to achieve multi-view image translation indirectly through NeRF-to-NeRF translation.

For restoration tasks, several works [Chen et al. 2023b; Lee et al. 2023a,b; Wang et al. 2023c] have explored this task under specific types of degradation. For example, NeRF-SR [Wang et al. 2022] proposes a supersampling strategy that allows the rendering of a single pixel to be influenced by multiple rays, gathering more information for NeRF super-resolution. Deblur-NeRF [Ma et al. 2022] adopts an analysis-by-synthesis approach to simulate the blurring process, thereby making NeRF robust to blurry inputs. Dehazenerf [Li et al. 2023a] demonstrates successful multi-view haze removal using physically realistic terms that model atmospheric scattering. However, those methods can only deal with specific types of degradation, ignoring the generality of restoration. To overcome this limitation, RaFE [Wu et al. 2024] introduces GANs for NeRF generation to better accommodate the geometric and appearance inconsistencies present, which can apply to various types of degradations.

For editing tasks, EditNeRF [Liu et al. 2021] enables shape and appearance editing through learning the underlying part semantics. Instruct-NeRF2NeRF [Haque et al. 2023] implements the editing of NeRF using text instructions by editing 2D images individually with iterative updates. However, the edited image lacks multi-view consistency, making the method unstable and slow to converge. Later, ViCA-NeRF [Dong and Wang 2024] introduces geometric and learned regularization to explicitly propagate the editing information across different views, thus ensuring multi-view consistency. GenN2N [Liu et al. 2024] is a recent work that shares similarities with our motivation. It performs editing in the 2D domain and extends 2D editing to the 3D NeRF space by learning a generative model to depict the distribution of NeRF edits. However, in contrast to GenN2N, which focuses on NeRF-to-NeRF translation, our MV2MV framework trains a VCDM model that directly operates on multi-view images, enabling more flexible editing and leveraging the powerful prior of diffusion models to achieve realistic and resolution-unlimited results.

3 OVERVIEW

Given multi-view images, our goal is to achieve view-consistent multi-view image translation tasks, including the restoration tasks of super-resolution, denoising and deblurring (the top row of Fig. 1), as well as text-driven editing tasks (the 2nd and 3rd rows of Fig. 1).

3.1 Overview

An overview of our approach is shown in Fig. 2. To circumvent the requirements for ground truth images, we first propose a novel self-supervised training strategy of CAS (Sec 4.1), which leverages the knowledge from off-the-shelf single-view image translators well-trained on large 2D image datasets. Specifically, we process the multi-view images individually to generate high-quality translated images, and then feed them into the 3DGS to obtain geometrically consistent rendered images. Translated and rendered images are regarded as pseudo ground truths that provide supervision for our VCDM through consistent and adversarial loss. Secondly, the LAConsistNet utilizes a latent-3DGS as the underlying 3D representation to enable information exchange across multi-view images, where consistent feature maps are rendered by the latent-3DGS as input to LAConsistNet and each block of LAConsistNet is plugged into the corresponding decoder layer of the denoising UNet to enforce consistency (Sec 4.2). Finally, we introduce a joint optimization strategy that simultaneously trains VCDM and 3DGS to further enhance the trade-off between consistency and realism (Sec 4.3).

3.2 Preliminary

3.2.1 Stable Diffusion. The Stable Diffusion model [Rombach et al. 2022] is the backbone of our method, which is a large-scale text-to-image latent diffusion model. Diffusion models learn to generate data samples by employing a denoising sequence that estimates the score of the data distribution. To achieve better efficiency and stabilized training, Stable Diffusion pretrains a variational autoencoder (VAE) [Razavi et al. 2019] to compress an image x into a latent z. Subsequently, the forward and reverse processes are performed in the latent space. In the forward process, Gaussian noise $\epsilon\ (0, I)$ with variance $\beta_t \in (0, 1)$ is added to the encoded latent z to generate the

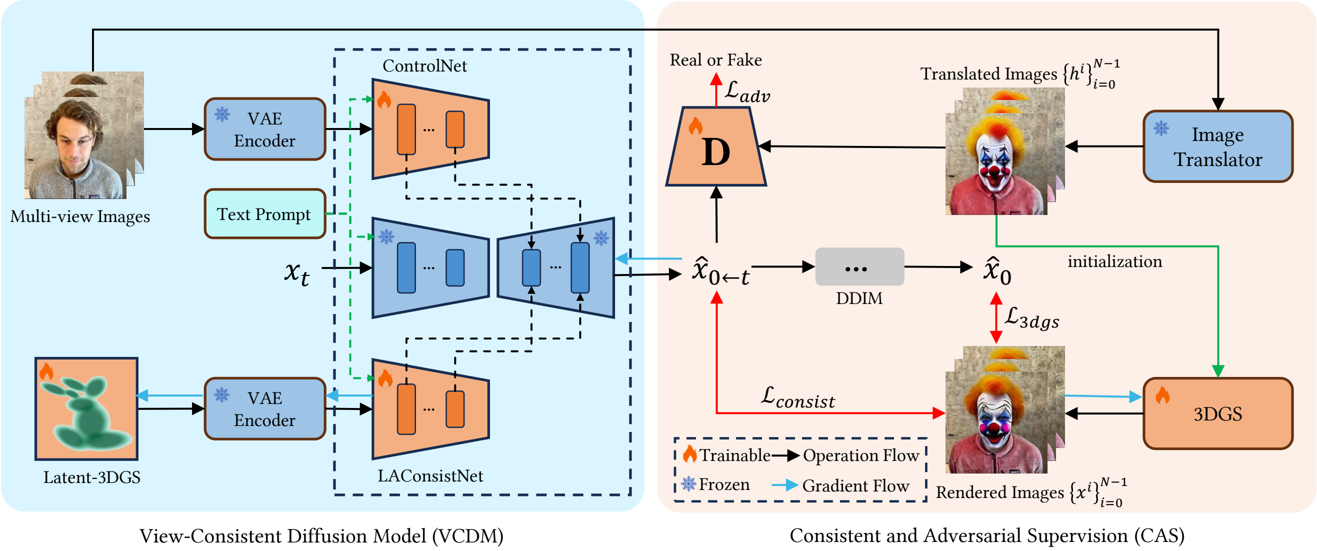

Fig. 2. Overview of MV2MV. Given multi-view images (top left), we utilize the CAS strategy (right) to train the proposed VCDM (left), which exploits the success of off-the-shelf single-view image translators and 3DGS to generate pseudo ground truths as supervisory signals. LAConsistNet (see details in Fig. 3) utilizes a latent-3DGS as the underlying 3D representation to enable information exchange across multi-view images, ensuring 3D consistency. The joint optimization strategy simultaneously optimizes VCDM and 3DGS to achieve a better trade-off between consistency and realism.

noisy latent at time t:

$$z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$. Then a denoising UNet ϵ_θ is trained to predict the added noise of the reverse process. The optimization of Stable Diffusion is specified as follows:

$$\mathcal{L}_{sd} = \mathbb{E}_{z,t,\epsilon} \left[\|\epsilon - \epsilon_\theta (z_t, t)\|_2^2 \right] \tag{2}$$

where t is uniformly sampled.

3.2.2 3D Gaussian Splatting. 3DGS [Kerbl et al. 2023] is an explicit 3D representation based on point clouds. It forgoes predicting density and color with neural networks, thus accelerating both training and rendering processes. Specifically, a set of 3D Gaussians with attributes of position p, rotation R, scale S, opacity o and Spherical Harmonic coefficients (SHs) is modelled to represent the scene. When rendering an image, 3D Gaussians are projected into the 2D plane by the following transformation:

$$\Sigma' = JW\Sigma W^T J^T \tag{3}$$

where W denotes the world-to-camera transformation matrix, J is the Jacobian of the affine approximation of the projective transformation and $\Sigma = RSS^T R^T$ is the covariance matrix. Next, a point-based rendering approach [Kopanas et al. 2022] computes the color C of a pixel by blending N ordered points overlapping the pixel:

$$C = \sum_{i \in N} c_i \sigma_i \prod_{j=1}^{i-1} \left(1 - \sigma_j\right) \tag{4}$$

where σ_i is computed by the Gaussian multiplied with the opacity o_i and c_i is the view-dependent color computed by SHs.

4 METHOD

4.1 Consistent and Adversarial Supervision

4.1.1 Pseudo Ground Truth Generation. Recently, the denoising diffusion probabilistic model has achieved great success in 2D image translation tasks due to its powerful prior knowledge and appearance generation capabilities. Therefore, we propose a self-supervised framework that leverages the success of existing diffusion-based image translation methods [Brooks et al. 2023; Chen et al. 2024; Sun et al. 2023] for generating pseudo ground truths as supervision signals to circumvent the requirements for ground truth images. While pseudo ground truths can also be obtained using non-diffusion methods, in practice, we prioritize methods with strong capabilities to generate high-quality and realistic texture details. In this paper, we employ several multi-view image translation tasks to demonstrate the adaptability of MV2MV: super-resolution, denoising, deblurring, and text-driven editing.

Given the source multi-view image set $\{I^i\}_{i=0}^{N-1}$, we first perform an existing image translator to produce a set of translated images $\{h^i\}_{i=0}^{N-1}$. Next, we use these images to directly optimize the 3DGS in order to enforce the production of geometrically consistent results $\{x^i\}_{i=0}^{N-1}$. Due to the generative nature of the image translator, $\{h^i\}_{i=0}^{N-1}$ exhibits significant multi-view inconsistency despite its good quality, whereas $\{x^i\}_{i=0}^{N-1}$ demonstrates good consistency but lacks clear details. These two results are regarded as pseudo ground truths. And then, we focus on achieving generative results that strike a balance between view consistency and high-quality through the proposed supervision scheme. Specifically, we propose the consistency loss and the adversarial loss to achieve these two optimization objectives. The former ensures that VCDM generates multi-view consistent results but may lack clear details, while the latter is employed to restore high-frequency details.

4.1.2 Consistent Loss. Given consistent multi-view images $\{x^i\}_{i=0}^{N-1}$, when training VCDM, we propose to directly minimize the discrepancy between the estimated result and x^i at each time step to ensure consistency across the multi-view setting. Specifically, we randomly select a timestep t and add noise to convert x^i to the noisy state x_t^i, and the consistent loss function $\mathcal{L}_{consist}$ is:

$$\mathcal{L}_{consist} = \left\|x^i - x^i_{0\leftarrow t}\right\|_2^2 \tag{5}$$

where $\hat{x}^i_{0\leftarrow t} = \left(x_t^i - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(x_t^i, t\right)\right)$ is the generated image estimated from predicted noise at timestep t. The consistent loss ensures that VCDM generates multi-view consistent results.

4.1.3 Adversarial Loss. The geometric and appearance inconsistencies among the $\{h^i\}_{i=0}^{N-1}$ preclude their direct use in supervising our VCDM models. Consequently, we aim to restore the high-frequency details by employing an adversarial training strategy. Previous research demonstrates the effectiveness of adversarial training in preventing blurred rendered images resulting from conflicts caused by viewpoint inconsistencies during image supervision from different viewpoints [Huang et al. 2020; Liu et al. 2024].

The denoising process plays a central role in the diffusion model, while adversarial training is crucial in GANs. Recently, several approaches [Sauer et al. 2023; Sun et al. 2023; Xiao et al. 2021; Xie et al. 2024] have been proposed that aim to improve the diffusion process through adversarial training. For example, CCSR [Sun et al. 2023] employs adversarial training to fine-tune the VAE decoder to enhance details. Adversarial Diffusion Distillation (ADD) [Sauer et al. 2023] uses a combination of adversarial training and score distillation to significantly accelerate inference speed. Here, we propose to combine the diffusion model and adversarial training to provide fine-grained information for multi-view images, where additional high-quality images can directly supply high-frequency priors for the diffusion model during training.

Specifically, we propose to minimize the distribution discrepancy between the VCDM result $\hat{x}^i_{0\leftarrow t}$ and the translated high-quality image h^i for supervising the parameters of the VCDM and discriminator. We adopt a saturate GAN loss with an image-level discriminator [Goodfellow et al. 2014]. The translated high-quality images $\{h^i\}_{i=0}^{N-1}$ are treated as the real samples $\mathbf{R}$, while the VCDM result $\{\hat{x}^i_{0\leftarrow t}\}_{i=0}^{N-1}$ are treated as the fake samples $\mathbf{F}$. Thus, the adversarial objective functions of the VCDM and the discriminator amounts to:

$$\begin{aligned}\mathcal{L}^G_{adv} &= E_{\mathbf{R}}[-\log D(\mathbf{R})] + E_{\mathbf{F}}[-\log(1-D(\mathbf{F}))]\\ \mathcal{L}^D_{adv} &= E_{\mathbf{F}}[-\log(D(\mathbf{F}))]\end{aligned} \tag{6}$$

Additionally, to introduce geometric-level constraints, we also incorporate perceptual loss to encourage the VCDM result $\hat{x}^i_{0\leftarrow t}$ to have a geometry similar to that of the translated high-quality images h^i:

$$\mathcal{L}_{geo} = \mathrm{LPIPS}(h^i, \hat{x}^i_{0\leftarrow t}) \tag{7}$$

where LPIPS$(\cdot,\cdot)$ refers to the Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018].

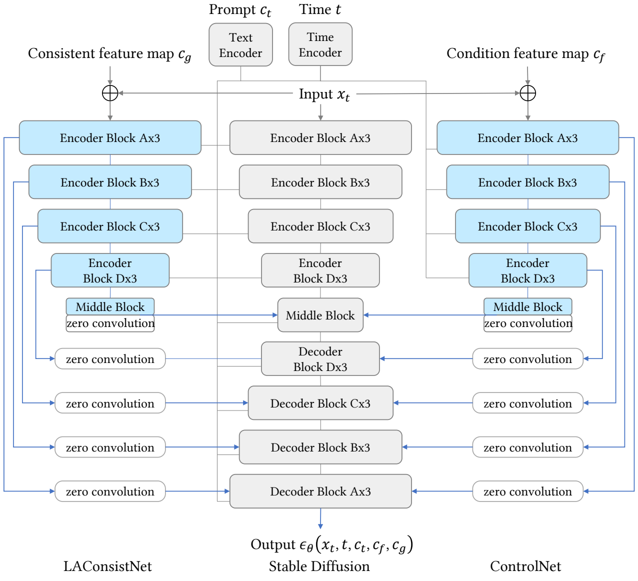

Fig. 3. Both LAConsistNet and ControlNet are integrated into the denoising UNet of Stable Diffusion to enforce view-consistent image translation, which corresponds to the dotted box in Fig. 2.

4.2 LAConsistNet Block

To explicitly enforce multi-view geometric consistency in our VCDM model, inspired by ConsistNet [Yang et al. 2023] and Syncdreamer [Liu et al. 2023], we propose a plug-in module, called LAConsistNet, to inject 3D prior as an additional condition into the blocks of the neural network, see Fig. 3 (corresponds to the dotted box in Fig. 2). ConsistNet uses view aggregation and ray aggregation modules to aggregate multi-view information, while SyncDreamer employs a 3D-aware feature attention mechanism to synchronize features across different views, all of which are based on an underlying 3D spatial feature volume. Differently, we propose to utilize Latent-3DGS as the underlying 3D representation modelling multi-view geometry principles, which is superior in two aspects: (1) it is a straightforward approach that can be directly initialized using the pre-trained 3DGS from Section 4.1; (2) it is memory- and efficiency-friendly, capable of handling large-scale scenes and applicable to different multi-view image transformation tasks.

Our backbone network is built upon ControlNet [Zhang et al. 2023] and pre-trained Stable Diffusion [Rombach et al. 2022]. The LAConsistNet is plugged into each encoder level of the denoising UNet [Ronneberger et al. 2015] to enforce 3D consistency. Specifically, the denoising UNet comprises an encoder and a decoder, each with 12 blocks and an intermediate block in between. Similar to ControlNet, we create a trainable copy of the 12 encoder blocks and 1 intermediate block to stabilize the diffusion process, and append the output to each decoder layer of UNet using zero convolution layers, i.e., 1 × 1 zero-initialized convolution. However, unlike ControlNet, which takes conditional images as input, our LAConsistNet introduces consistency constraints by utilizing 3D consistent feature maps rendered by Latent-3DGS.

To achieve fast training on pre-trained Stable Diffusion, the trainable structures of ControlNet and LAConsistNet are kept the same as that in [Zhang et al. 2023]. Additionally, we utilize the pre-trained 3DGS from Section 4.1 to initialize the Latent-3DGS as the initial 3D prior. Here, the latent features stored in 3D Gaussians are treated as trainable parameters, while the other attributes are frozen. Our Latent-3DGS, similar to Latent-NeRF [Metzer et al. 2023], serves as the underlying representation of the 3D scene that ensures the exchange of information between images from multiple viewpoints to achieve 3D consistency.

4.3 Joint Optimization VCDM and 3DGS

In our framework, the LAConsistNet and the consistent loss are proposed to ensure that VCDM generates multi-view consistent results, while the adversarial loss is introduced to enhance the fine details. However, the generation of details is accompanied by randomness, leading to variations in the high-frequency details generated for multi-view images, which adversely affect both training stability and view consistency. This motivated us to propose a joint optimization strategy that simultaneously optimizes VCDM and 3DGS, leveraging the view-consistent properties of 3DGS to capture coherent details across views. Therefore, high-frequency details that conform to view consistency constraints are retained, while inconsistent high-frequency details are eliminated. The results produced by 3DGS is used as the guidance for the VCDM to generates multi-view consistent fine details.

As shown in Fig. 2, the inference results $\{\hat{x}_0^i\}_{i=0}^{N-1}$ of VCDM with the denoising diffusion implicit model (DDIM) [Song et al. 2020] are used to optimize 3DGS. Subsequently, the rendering results $\{x^i\}_{i=0}^{N-1}$ of 3DGS are used to supervise VCDM by Eq.5, which forces the VCDM to generate consistent high-frequency details. Following [Kerbl et al. 2023], we optimize 3DGS with the loss functions below:

$$\mathcal{L}_{3dgs} = (1-\lambda)\left\|\hat{x}_0^i - x^i\right\|_1 + \lambda\,\mathrm{SSIM}\left(\hat{x}_0^i, x^i\right) \tag{8}$$

where SSIM$(\cdot,\cdot)$ refers to SSIM trem. It is worth noting that only a few timesteps are used during the inference period to minimize additional processing time. This approach is akin to a distillation operation, significantly enhancing recovery results and inference efficiency without compromising efficiency.

4.4 Inference

After the training stage, VCDM is capable of reasoning about high-quality and 3D-consistent multi-view images based on the translation target. We employ the same inference period as in the joint optimization process, ensuring consistency between training and inference. Our approach significantly reduces the sampling steps while maintaining satisfactory generation capabilities. In addition, to further ensure the consistency of the generated content, we employ the Non-Uniform Timestep Sampling approach from CCSR [Sun et al. 2023] to truncate the diffusion chain from x_{min} to x_{max}. Specifically, we set t_{min} and t_{max} in all our experiments to $T/3$ and $2T/3$ ($T = 15$).

5 EXPERIMENTS

The proposed MV2MV is a unified multi-view image to multi-view image translation framework that supports various image translation tasks. In our experiments, we demonstrate the effectiveness of MV2MV across a variety of tasks: super-resolution, denoising, deblurring, and text-driven editing. Our framework simply replaces different plug-and-play image translators to accomplish these tasks without any additional design. To demonstrate the ability of guaranteeing view consistency, we use the results trained and rendered by 3DGS for consistent and meaningful comparisons. An accompanying video is provided for dynamic qualitative comparisons.

5.1 Datasets

We utilized a diverse range of datasets to evaluate the performance of our model in different translation tasks. For super-resolution and denoising tasks, we conduct experiments on a complex real-world Mip-Nerf360 dataset [Barron et al. 2022], which contains 9 unbounded indoor and outdoor scenes. For the deblurring task, we perform experiments on the real-motion-blur dataset provided by Deblur-NeRF [Ma et al. 2022], comprising 10 real world scenes with camera motion. For the text-driven editing task, we conduct experiments on two forward-facing scenes of Face [Haque et al. 2023] and Fangzhou [Wang et al. 2023a], and the 360-degree scenes of Garden [Barron et al. 2022].

5.2 Implementation Details

Our method is implemented by using the PyTorch framework. Following [Kerbl et al. 2023], we train the 3DGS model for 30K iterations using the same Gaussian densification strategy. We utilize Stable Diffusion v2.1 [Rombach et al. 2022] with ControlNet [Sun et al. 2023; Zhang et al. 2023] as the generative prior. Our VCDM is fine-tuned for $2k$ iterations with a batch size of 1 and a learning rate of $1e^{-4}$ for each scene. The training phase takes about 2-3 hours on a single NVIDIA A40 GPU (48GB).

5.3 Metrics

For quantitative experiments, we utilize both reference and non-reference metrics to provide a comprehensive assessment for each method. Since the generative characteristic of VCDM, the details of results may not faithfully follow the ground truth. We employ metrics that correlate well with human visual perception. LPIPS [Zhang et al. 2018] and DISTS [Ding et al. 2020] serve as reference-based measures to evaluate the perceptual quality of results with respect to ground truths. NIQE [Zhang et al. 2015], MANIQA [Yang et al. 2022], and MUSIQ [Ke et al. 2021] are non-reference image quality assessment metrics designed to assess the fidelity of an image, and are closer to human perception. To better evaluate the performance of text-driven editing tasks, we also include CLIP Text Image Directional Similarity [Haque et al. 2023] and CLIP Direction Consistency [Pathak et al. 2016] as evaluation metrics, which measure the alignment of the performed edit with the text instruction and the temporal consistency of the performed edit across views.

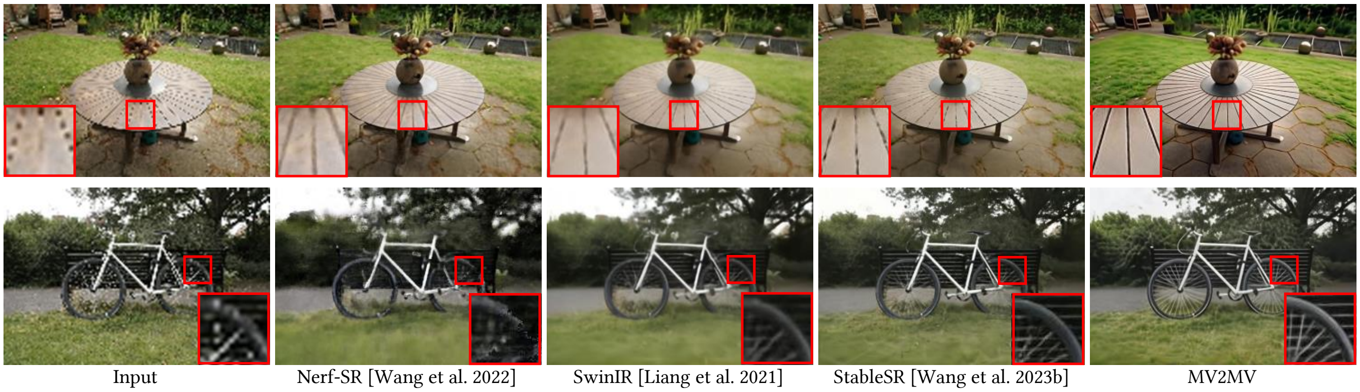
Input
Nerf-SR [Wang et al. 2022]
SwinIR [Liang et al. 2021]
StableSR [Wang et al. 2023b]
MV2MV

Fig. 4. Qualitative results on super-resolution. Our method is able to generate more realistic and more sharper details. Please zoom in for a better visualization.

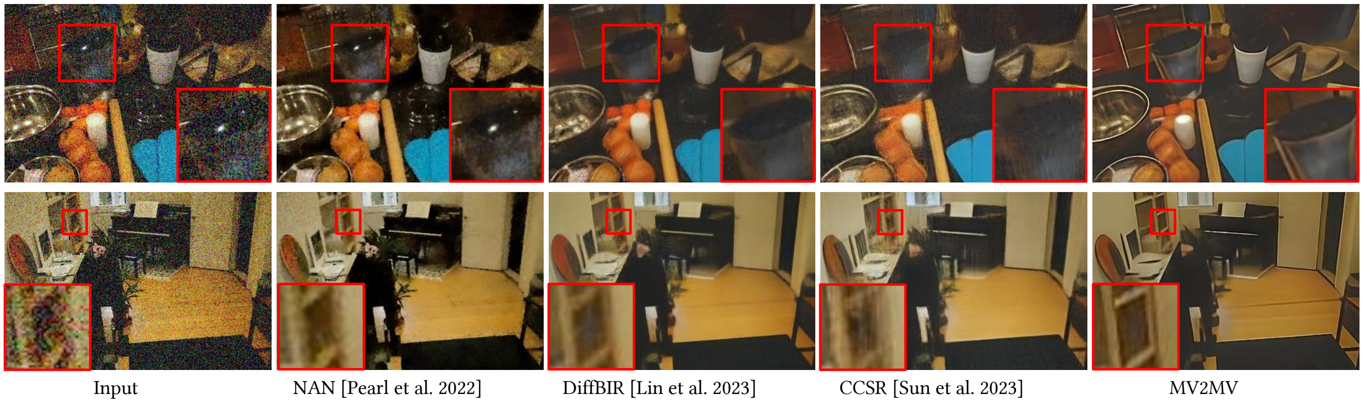

Fig. 5. Qualitative results on denoising. Our method effectively removes noise while restoring detailed texture. Please zoom in for a better visualization.

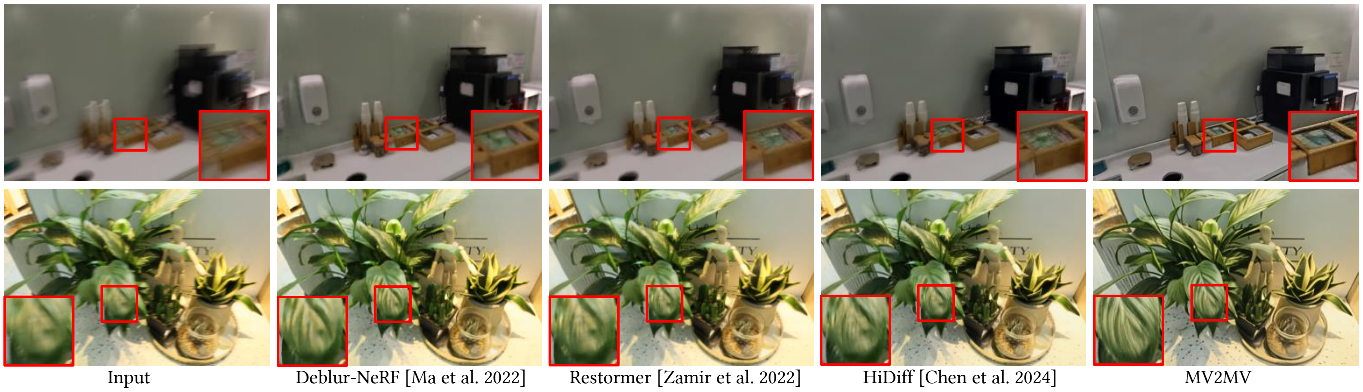
Input
Deblur-NeRF [Ma et al. 2022]
Restormer [Zamir et al. 2022]
HiDiff [Chen et al. 2024]
MV2MV

Fig. 6. Qualitative results on deblurring. Our method removes motion blur and generates detailed textures. Please zoom in for a better visualization.

5.4 Quantitative Results

5.4.1 Super-resolution. On the Mip-Nerf360 dataset, we use the image data downscaled by a factor of 8 as ground truths. In addition, we generate low-resolution images using bicubic interpolation with a scale factor of 4 to adapt to 4× super-resolution task. In this experiment, CCSR [Sun et al. 2023] as the image translator is used in our framework. We compare our MV2MV with several state-of-the-art methods, including the NeRF-based method: NeRF-SR [Wang et al. 2022], and 2D image super-resolution methods: SwinIR [Liang et al.

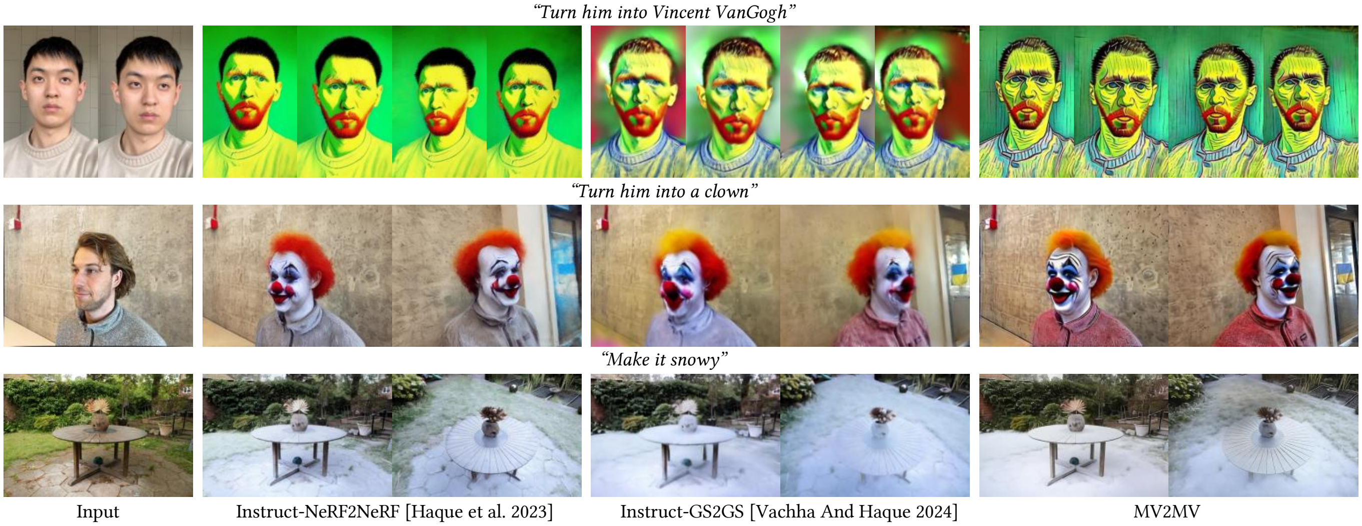
"Turn him into Vincent VanGogh"
"Turn him into a clown"
"Make it snowy"
Input
Instruct-NeRF2NeRF [Haque et al. 2023]
Instruct-GS2GS [Vachha And Haque 2024]
MV2MV

Fig. 7. Qualitative results on text-driven editing. Our method generates results that are more consistent and of better quality than previous state-of-the-art methods.

Table 1. Quantitative comparisons on the super-resolution task. The best result is highlighted in bold. Our method performs the best on non-reference perceptual metrics.

Methods	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	MUSIQ↑
NeRF-SR	0.370	0.205	4.192	0.428	39.346
Bicubic	0.358	0.216	5.808	0.279	26.946
SwinIR	0.404	0.230	5.813	0.351	30.546
BSRGAN	0.361	0.189	4.134	0.594	44.021
DiffBIR	0.393	0.212	3.754	0.571	40.485
CCSR	0.351	0.185	3.731	0.548	43.243
StableSR	**0.340**	**0.173**	3.565	0.621	49.241
MV2MV	0.358	0.188	**3.401**	**0.707**	**58.385**

Table 2. Quantitative results on denoising. The best result is highlighted in bold. Our method achieves the best scores in MANIQA and MUSIQ.

Methods	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	MUSIQ↑
NAN	**0.350**	**0.224**	**3.013**	0.563	49.121
SwinIR	0.478	0.280	6.676	0.422	34.314
CCSR	0.481	0.270	4.010	0.598	44.465
DiffBIR	0.468	0.264	4.436	0.620	44.029
MV2MV	0.466	0.262	3.858	**0.726**	**60.092**

2021], BSRGAN [Zhang et al. 2021], DiffBIR [Lin et al. 2023], CCSR [Sun et al. 2023], StableSR [Wang et al. 2023b]. Bicubic interpolation is included in our comparison as the baseline. Note that we use 2D image super-resolution methods for per-view processing and direct integration with 3DGS.

Table 3. Quantitative comparisons for deblurring. The best result is highlighted in bold. Our method performs the best on perceptual metrics.

Method	NIQE↓	MANIQA↑	MUSIQ↑
DeblurNeRF	4.756	0.352	54.686
BAD-NeRF	7.191	0.271	30.699
Restormer	3.919	0.355	60.289
HiDiff	3.814	0.401	67.346
MV2MV	**3.779**	**0.486**	**74.122**

We show the quantitative results in Tab. 1. Due to the stronger generation capability, MV2MV achieves the highest scores in NIQE, MANIQA and MUSIQ compared to all baselines, indicating better alignment with human visual perception. Furthermore, MV2MV achieves comparable scores in terms of LPIPS and DISTS, which demonstrate outstanding perceptual measures respect to ground truths. The qualitative comparisons are shown in Fig. 4. While diffusion-based methods of DiffBIR, CCSR and StableSR excel at generating realistic details when processing images individually, multi-view inconsistencies lead to reconstructions with varying degrees of blurring. By contrast, our MV2MV is able to produce view-consistent realistic details.

5.4.2 *Denoising.*

On the Mip-Nerf360 dataset, we follow the noise model used in [Mildenhall et al. 2018; Pearl et al. 2022] to obtain the noisy images according to the equation: $I_{noise}(p) \sim \mathcal{N}(I(p), \delta_r^2 + \delta_s^2 I(p))$, where p is an image coordinate, δ_r and δ_s are noise parameter and $\mathcal{N}$ represents the Gaussian distribution. In our experiments, we use the noise levels of 8 to get our noisy image and use CCSR [Sun et al. 2023] as the image translator. We compare with state-of-the-art methods of NAN [Pearl et al. 2022], SwinIR [Liang et al. 2021], DiffBIR [Lin et al. 2023] and CCSR [Sun et al. 2023].

Table 4. Quantitative results on text-driven editing. The best result is highlighted in bold. Our method achieves the best performance.

Method	CLIP Text-Image Direction Similarity↑	CLIP Direction Consistency ↑	NIQE↓	MANIQA↑	MUSIQ↑
Instruct-NeRF2NeRF	0.155	0.896	3.494	0.551	65.541
Instruct-GS2GS	0.149	0.941	5.163	0.344	42.327
MV2MV	**0.174**	**0.944**	**3.403**	**0.629**	**70.369**

The quantitative results are presented in Table 2. For the no-reference metrics, MV2MV achieves the best scores in MANIQA and MUSIQ, and the second-best scores in NIQE. In terms of reference metrics of LPIPS and DISTS, MV2MV maintains competitive measures and is only slightly inferior to NAN. This is due to the fact that the realistic details generated by diffusion-based methods may not match well with the ground truth, thus putting them at a disadvantage in reference metrics. Moreover, as shown in Fig. 5, MV2MV produces sharper results with clear details compared to other state-of-the-art methods.

5.4.3 Deblurring. For the deblurring task, we use HiDiff [Chen et al. 2024] as the image translator to recover high-quality images in our framework. We compare the deblurring performance with Deblur-NeRF [Ma et al. 2022], BAD-NeRF [Wang et al. 2023c] and single-view image deblurring methods (Restormer [Zamir et al. 2022], HiDiff [Chen et al. 2024]) combined with 3DGS. As the real-motion-blur dataset does not provide sharp images, we only report non-reference metrics in our experiment.

The quantitative results are shown in Table 3. MV2MV shows consistently improved MANIQA, MUSIQ and CLIPIQA scores compared to other methods, indicating that MV2MV is able to generate perceptually more realistic details. The qualitative results are shown in Fig. 6, demonstrating that MV2MV is more effective in removing motion blur while preserving fine image details than other methods.

5.4.4 Text-driven Editing. We achieve text-driven editing by using InstructPix2Pix [Brooks et al. 2023] as the image translator in our framework. We mainly compare our MV2MV with two recent state-of-the-art methods of Instruct-NeRF2NeRF [Haque et al. 2023] and Instruct-GS2GS [Vachha and Haque 2024], which employ an iterative updating mechanism to address a 3D inconsistency problem among different edits.

Although editing is a subjective task, we report the quantitative metrics of CLIP Text-Image Direction Similarity and CLIP Direction Consistency according to Instruct-NeRF2NeRF [Haque et al. 2023]. In addition, we report the non-reference perceptual metrics to evaluate the quality of text-driven editing results. Quantitative results are presented in Table 4. Our MV2MV outperforms other methods on all metrics, demonstrating its effectiveness in the text-driven editing task. The qualitative comparisons are depicted in Fig. 7. Our method generates more consistent and realistic results, while other methods suffer from blurry results due to the blending of inconsistent editing.

5.5 Ablation Studies

We conduct ablation studies to demonstrate the effectiveness of each component in MV2MV. Specifically, we test the following aspects of our model: the adversarial loss, the LAConsistNet block and the joint optimization strategy based on the super-resolution task on Mip-Nerf360 dataset. The quantitative results of the ablations are depicted in Table 5. To ensure a fair comparison, all results are rendered by 3DGS.

Table 5. Ablation studies on the proposed adversarial loss, LAConsistNet and the joint optimization strategy (JOS). The best results are highlighted in bold.

Methods	NIQE↓	MANIQA↑	MUSIQ↑
w/o Adversarial Loss	4.376	0.497	44.244
w/o LAConsistNet	4.089	0.667	54.243
w/o JOS	3.592	0.682	56.271
MV2MV	**3.401**	**0.707**	**58.385**

5.5.1 Effects of Consistent Loss. In our framework, VCDM requires $\{x^i\}_{i=0}^{N-1}$ as the direct supervisory signal based on the consistency loss, which cannot be removed. In fact, the direct outputs of the 2D image translator are the results obtained before applying the consistency loss, which exhibit inconsistencies across views.

5.5.2 Effects of Adversarial Loss. In this ablation study, we examine the influence of the adversarial loss, which is designed to guide our model in restoring high-frequency details. As shown in Table 5 and Fig. 8, the lack of adversarial loss results in blurry rendered images and leads to a significant decrease in perceptual metrics. This blurring can be attributed to the absence of high-quality image supervision. By incorporating the adversarial loss with the pseudo ground truths, our model is able to successfully generate the high-frequency details. For the discriminator, we follow the proposed design in [Goodfellow et al. 2014]. It is worth noting that different choices of discriminator networks may produce varying high-frequency details, which we will explore more thoroughly in future work. We believe that an improved discriminator network design leads to enhanced visual outcomes.

5.5.3 Effects of LAConsistNet. We demonstrate the effectiveness of the LAConsistNet block in guaranteeing view consistency, as shown in Table 5. Removing the LAConsistNet block results in a degradation of perceptual metrics due to the inconsistency of the multi-view images. As demonstrated in Fig. 9, we also present the results directly generated by VCDM. The introduction of the LAConsistNet block significantly enhances view consistency, resulting in high-quality rendering.

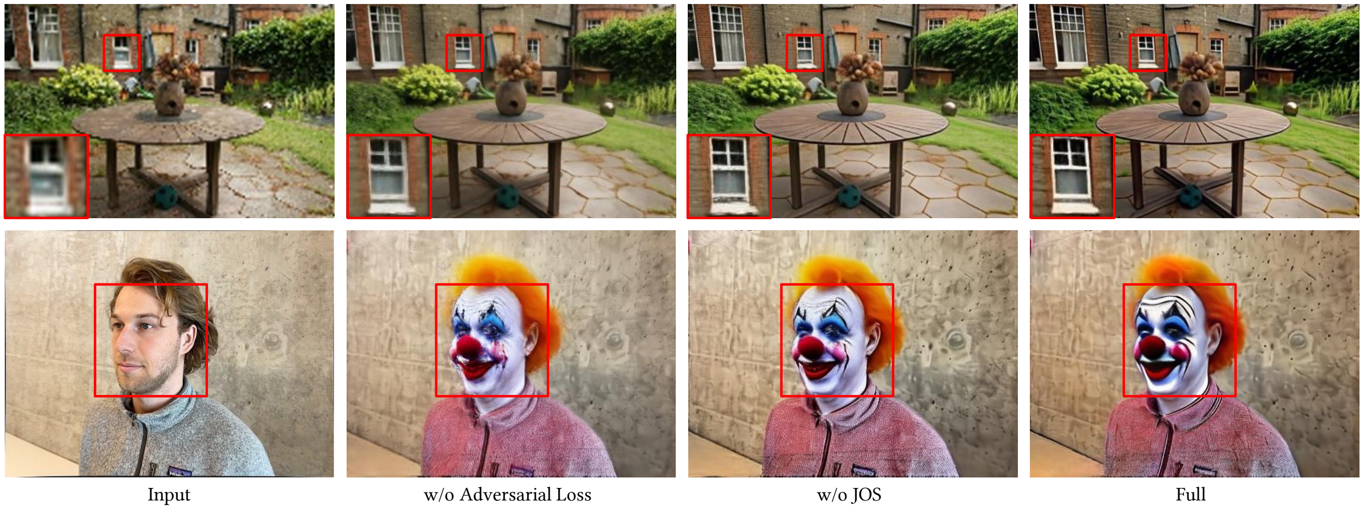

Fig. 8. Ablation of adversarial loss and joint optimization strategy (JOS).

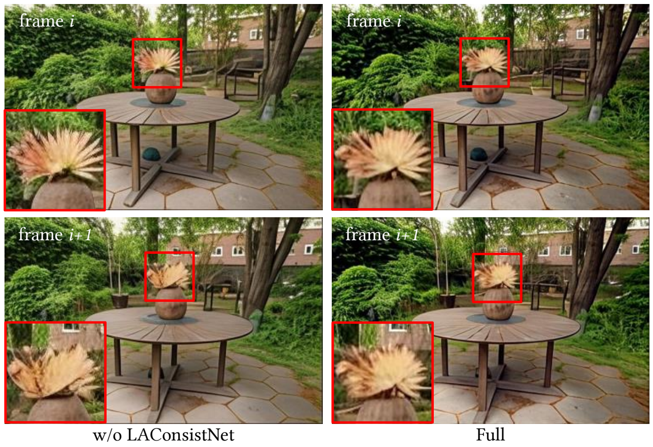

Fig. 9. Ablation of the LAConsistNet block.

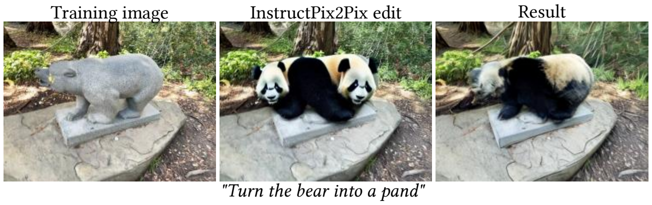

Fig. 10. Failure cases. When InstructPix2Pix produces incorrect edits, our method suffers from similar artifacts. Nevertheless, our method can still maintain consistency.

5.5.4 Effects of Joint Optimization Strategy. We examine the influence of the joint optimization strategy by removing 3DGS during training VCDM. Table 5 illustrates that all the metrics would have worsened without the joint optimization strategy, demonstrating its effectiveness in ensuring consistent and realistic detail generation. This can also be observed in the qualitative results in Fig. 8.

6 CONCLUSION

We propose MV2MV, a unified multi-view to multi-view translation framework that can handle various image translation tasks. Unlike previous NeRF-based translation approaches, our framework is built upon diffusion models and directly processes multi-view images in the image domain, taking better advantage of existing 2D generative priors to achieve more flexible processing and obtain more realistic results. To address the challenges of requiring training data and ensuring 3D consistency, we propose the consistent and adversarial supervision strategy and the LAConsistNet block, which can achieve high-quality and view-consistent results. Moreover, we introduce a joint optimization strategy that simultaneously optimizes the diffusion model and 3DGS to achieve a better trade-off between consistency and realism. Our experiments demonstrate that MV2MV outperforms existing state-of-the-art methods on various translation tasks. Our method represents a feasible path for achieving complete view-consistent multi-view translation.

Limitations and Future Work. Our method exhibits a couple of limitations. Firstly, the framework of MV2MV relies on off-the-shelf single-view image translators, thus naturally inheriting their limitations. As illustrated in Fig. 10, our method cannot perform well when InstructPix2Pix edits incorrectly. Previous methods also struggled in these cases. Additionally, since all training images in the editing task are edited only once by InstructPix2Pix, our method is adversely affected by inconsistencies caused by challenging cases of large viewpoint changes or dramatic changes with high uncertainty. This limitation arises from constraints inherent to 2D translators. With the fast development of the diffusion model, the choice of a better 2D translator can substantially alleviate this limitation. As an alternative, introducing additional 3D consistency constraints is also a good solution. For example, our framework can easily incorporate an iterative optimization strategy like Instruct-NeRF2NeRF [Haque et al. 2023], allowing us to obtain a feasible 3DGS as our initialization. We will try to solve this raised problem in the future. On the other hand, thanks to this design, our method can theoretically be generalized to more image translation tasks by simply replacing the off-the-shelf single-view image translator. Secondly, although our model achieves an acceptable trade-off between realism and consistency compared to previous diffusion-based methods, it may not completely guarantee 3D consistency due to the nature of probabilistic models. To completely guarantee 3D consistency, a potential solution would be to introduce stronger spatial-temporal constraints like SORA [OpenAI 2024] into MV2MV, thus allowing all images to be processed simultaneously.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive suggestions and comments. This work is supported by the National Key R&D Program of China (2022YFB3303400), the National Natural Science Foundationof China (62025207), and Laoshan Laboratory (No.LSKJ202300305).

REFERENCES

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5470–5479.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18392–18402.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

Wei-Ting Chen, Wang Yifan, Sy-Yen Kuo, and Gordon Wetzstein. 2023b. Dehazenerf: Multiple image haze removal and 3d shape reconstruction using neural radiance fields. *arXiv preprint arXiv:.11364* (2023).

Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. 2023a. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5896–5905.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. 2023c. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 12312–12321.

Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, and Xin Yuan. 2024. Hierarchical integration diffusion model for realistic image deblurring. *Advances in Neural Information Processing Systems* 36 (2024).

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis Machine Intelligence* 44, 5 (2020), 2567–2581.

Jiahua Dong and Yu-Xiong Wang. 2024. ViCA-NeRF: View-Consistency-Aware 3D Editing of Neural Radiance Fields. *Advances in Neural Information Processing Systems* 36 (2024).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27 (2014).

Jinjin Gu, Xianzheng Ma, Xiangtao Kong, Yu Qiao, and Chao Dong. 2024. Networks are slacking off: Understanding generalization problem in image deraining. *Advances in Neural Information Processing Systems* 36 (2024).

Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 19740–19750.

Jingwei Huang, Justus Thies, Angela Dai, Abhijit Kundu, Chiyu Jiang, Leonidas J Guibas, Matthias Nießner, and Thomas Funkhouser. 2020. Adversarial texture optimization from rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1559–1568.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics* 35, 4 (2016), 1–11.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1125–1134.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. 2023. Autodir: Automatic all-in-one image restoration with latent diffusion. *arXiv preprint arXiv:.10123* (2023).

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6007–6017.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 5148–5157.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.

Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. 2022. Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics* 41, 6 (2022), 1–15.

Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18062–18071.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, and Zehan Wang. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4681–4690.

Dogyoon Lee, Minhyeok Lee, Chajin Shin, and Sangyoun Lee. 2023a. Dp-nerf: Deblurred neural radiance field with physical scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12386–12396.

Dongwoo Lee, Jeongtaek Oh, Jaesung Rim, Sunghyun Cho, and Kyoung Mu Lee. 2023b. ExBluRF: Efficient Radiance Fields for Extreme Motion Blurred Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 17639–17648.

Seungyong Lee and Sunghyun Cho. 2013. Recent advances in image deblurring. *SIGGRAPH Asia Courses* (2013), 1–108.

Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision.* Springer, 702–716.

Haoying Li, Ziran Zhang, Tingting Jiang, Peng Luo, Huajun Feng, and Zhihai Xu. 2023a. Real-world deep local motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1314–1322.

Junyi Li, Zhilu Zhang, Xiaoyu Liu, Chaoyu Feng, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. 2023b. Spatially adaptive self-supervised learning for real-world image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9914–9924.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1833–1844.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. 2023. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:.15070* (2023).

Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. 2021. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 5773–5783.

Xiangyue Liu, Han Xue, Kunming Luo, Ping Tan, and Li Yi. 2024. GenN2N: Generative NeRF2NeRF Translation. *arXiv preprint arXiv:.02788* (2024).

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453* (2023).

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. 2022. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12861–12870.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12663–12673.

Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. 2018. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2502–2510.

OpenAI. 2024. Creating video from text. https://openai.com/

Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Transactions on Multimedia Chen. 2021. Image-to-image translation: Methods and applications. 24 (2021), 3859–3881.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings.* 1–11.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2536–2544.

Naama Pearl, Tali Treibitz, and Simon Korman. 2022. Nan: Noise-aware nerfs for burst-denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12672–12681.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems* 32 (2019).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted Intervention*. Springer, 234–241.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–10.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042* (2023).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. 2023. Improving the Stability of Diffusion Models for Content Consistent Super-Resolution. *arXiv preprint arXiv:.00877* (2023).

Cyrus Vachha and Ayaan Haque. 2024. Instruct-GS2GS: Editing 3D Gaussian Splats with Instructions. https://instruct-gs2gs.github.io/

Vaibhav Vavilala and Mark Meyer. 2021. *Towards Large-Scale Super Resolution Datasets via Learned Downsampling of Ray-Traced Renderings*. 1–2.

Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023a. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization Computer Graphics* (2023).

Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. 2022. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6445–6454.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2023b. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:.07015* (2023).

Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. 2023c. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4170–4179.

Zhihao Wang, Jian Chen, and Steven CH Hoi. 2020. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis Machine Intelligence* 43, 10 (2020), 3365–3387.

Zhongkai Wu, Ziyu Wan, Jing Zhang, Jing Liao, and Dong Xu. 2024. RaFE: Generative Radiance Fields Restoration. *arXiv preprint arXiv:.03654* (2024).

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. 2021. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804* (2021).

Rui Xie, Ying Tai, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. 2024. AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation. *arXiv preprint arXiv:2404.01717* (2024).

Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. 2023. Consistnet: Enforcing 3d consistency for multi-view images diffusion. *arXiv preprint arXiv:.10343* (2023).

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1191–1200.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5728–5739.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4791–4800.

Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. 2022. Deep image deblurring: A survey. *International Journal of Computer Vision* 130, 9 (2022), 2103–2130.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015), 2579–2591.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595.