Scene-Level Appearance Transfer with Semantic Correspondences

Anonymous Authors

Figure 1. Given an interior design image (style image) and a 3D scene captured by video or multi-view images, ReStyle3D first transfers the appearance based on semantic correspondences to a single view, then lifts the stylization to multiple viewpoints using 3D-aware style lifting, achieving multi-view consistent appearance transfer with fine-grained details. Project page: restyle3d.github.io.

Abstract

We introduce ReStyle3D, a novel framework for scene-level appearance transfer from a single style image to a real-world scene represented by multiple views. The method combines explicit semantic correspondences with multi-view consistency to achieve precise and coherent stylization. Unlike conventional stylization methods that apply a reference style globally, ReStyle3D uses open-vocabulary segmentation to establish dense, instance-level correspondences between the style and real-world images. This ensures that each object is stylized with semantically matched textures. ReStyle3D first transfers the style to a single view using a training-free semantic-attention mechanism in a diffusion model. It then lifts the stylization to additional views via a learned warp-and-refine network guided by monocular depth and pixel-wise correspondences. Experiments show that ReStyle3D consistently outperforms prior methods in structure preservation, perceptual style similarity, and multi-view coherence. User studies further validate its ability to produce photo-realistic, semantically faithful results. Our code, pretrained models, and dataset will be publicly released, to support new applications in interior design, virtual staging, and 3D-consistent stylization.

1. Introduction

Generative diffusion models have recently spurred significant advances in image stylization and broader generative applications, enabling the seamless synthesis or editing of images with remarkable visual fidelity. While existing image stylization approaches [7, 30] often excel at transferring well-known artistic styles (*e.g.*, Van Gogh paintings) onto photographs, they fall short when it comes to practical and realistic style applications, such as virtual staging or professional interior decoration, where transferring the style of one image (*style* image) to another (*source* image) entails transferring the individual appearance of objects (Fig. 1).

These methods tend to treat the style image globally, ignoring the semantic correspondence between individual objects or regions in the images. This coarsely aligned stylization not only misrepresents object appearances but also fails to adapt fine-grained textures to semantically matched regions (*e.g.*, transferring *couch* textures only to *couches*). This is crucial for real-world use cases where style is defined by the unique characteristics (*e.g.*, color, material, shape) of design elements (*i.e.*, furniture, decor, lighting, and accessories) that give it its signature look [44]. Another line of work pursues *semantic correspondence* for transferring object appearances [5, 70]. While these methods show promise in aligning single objects or small regions via deep

feature matching, they typically operate at low spatial resolutions (often 64×64) and therefore struggle to handle complex scenes with strong perspective and multiple object instances. Extending them to scene-level stylization remains a challenging problem due to both semantic and geometric complexity.

Moreover, when a scene is represented by multiple images (*e.g.*, for larger coverage), ensuring multi-view *consistency* in scene-level appearance transfer further complicates the task. Existing multi-view image editing methods [10, 16, 35, 45] commonly require known camera poses and an existing 3D scene representation (*e.g.*, a neural radiance field [39] or 3D Gaussian splatting [28]), which needs a dense set of input views and considerable compute time. These methods struggle with sparse or casually captured views, and their specialized 3D pipelines hinder plug-and-play use. A pixel-space approach preserving geometric cues without heavy 3D modeling is preferable but remains under explored. We propose **ReStyle3D**, a novel framework for scene-level appearance transfer that cobines semantic correspondence and multi-view consistency, addressing limitations of 2D stylization and 3D-based editing methods. *Our key insight* is that the inherent but implicit semantic correspondences from pretrained diffusion models or vision transformers (*e.g.*, StableDiffusion [49] and DINO [4, 42]) are insufficient for fine-grained, scene-level appearance transfer, especially when different objects or viewpoints are involved. We tackle this by explicitly matching open-vocabulary panoptic segmentation predictions between the style and source images, while ensuring that unmatched parts of the scene still receive a global style harmonization. This open-vocabulary labeling (with no predefined semantic categories) helps us robustly align semantically corresponding regions even in cluttered indoor scenes. By integrating these explicit correspondences into the attention mechanism of a diffusion process, we achieve more accurate and flexible stylization of multi-object scenes.

To further ensure *3D awareness* and view-to-view consistency, we adopt a two-stage pipeline. First, we achieve *training-free* semantic appearance transfer in a single view by injecting our correspondence-informed attention into a pretrained diffusion model. Second, a warp-and-refine diffusion network that efficiently propagates the stylized appearance to additional views in an auto-regressive manner, guided by monocular depth and pixel-level optical flows. Our method does not require explicit pose or 3D modeling, and we show that the final stylized frames are fully compatible with off-the-shelf 3D reconstruction tools, enabling complete 3D visualizations and consistent multi-view stylization with minimal overhead. In summary, our contributions are as follows:

- We introduce *SceneTransfer*, a new task of transferring multi-object appearance from a single style image to a 3D scene captured in multi-view images.
- We propose ReStyle3D, a two-stage pipeline that (*i*) adapts a pretrained diffusion model with *semantic attention* for instance-level stylization, and (*ii*) trains a warp-and-refine novel-view synthesis module to propagate the style across all views, maintaining global consistency.
- We create the *SceneTransfer* benchmark with 25 interior design images and 31 indoor scenes (243 style-scene pairs) from different categories (*e.g.* bedroom, living room, and kitchen). Our results show strong improvements in structure preservation, style fidelity, and cross-view coherence.

2. Related Work

Image Stylization aims to transfer artistic styles to images while preserving structural content. Early CNN-based methods [11, 17, 24] laid the groundwork by capturing style and content representations. With the advent of diffusion models [23, 49], recent approaches leverage pretrained architectures and textual guidance for high-quality stylization [7, 14, 30, 31, 59, 65, 73]. InST [73] employs textual inversion to encode styles in dedicated text embeddings, achieving flexible transfer. StyleDiffusion [31] further refines style-content separation through a CLIP-based disentanglement loss applied during fine-tuning. StyleID [7] adapts self-attention in pretrained diffusion models to incorporate artistic styles without additional training. While these methods produce compelling results, they focus on overall style transfer without explicitly modeling semantic correspondences. In contrast, we attempt to inject explicit semantic matching in stylization, thereby enabling precise style transfer according to semantically matching regions.

Semantic Correspondence. Foundational works and recent innovations have shaped the evolution of semantic correspondence. SIFT-Flow [33] pioneered dense image alignment with handcrafted SIFT descriptors [37]. Self-supervised vision transformers like DINO [4] and DINO-V2 [8, 42] improved feature representation for semantic matching without labeled data [56, 57]. Recent methods, such as [19, 70], DIFT [53], cross-image-attention [1] and [18], integrate diffusion models with these transformers, achieving superior zero-shot correspondence. Techniques like Deep Functional Maps [5] further refine correspondences by enforcing global structural consistency, demonstrating the potential of advanced representations in addressing correspondence challenges. The development of these techniques enables the extraction of semantic correspondences using intermediate representations.

Attention-based Control in Diffusion Models. The attention modules in pretrained diffusion models are essential

in controlling the generated content, allowing various image editing tasks through attention mask manipulation. Prompt-to-Prompt [20] pioneered text-based local editing by manipulating cross-attention between text prompts and image regions. Similarly, Plug-and-play [58] leverages the original image's spatial features and self-attention maps to preserve spatial layout while generating text-guided edited images. Epstein et al. [12] introduced Diffusion Self-Guidance, a zero-shot approach that leverages internal representations for fine-grained control over object attributes. While these methods focus on text-to-image attention control, recent works like Generative Rendering [3] explore cross-image attention by injecting 4D correspondences from meshes into attention for stylized video generation. In contrast, we propose a direct image-to-image semantic attention mechanism that transfers appearances across all semantic categories simultaneously through explicit correspondence masks, enabling efficient and accurate scene-level stylization without text prompts or 3D priors.

Diffusion-based Novel-View Synthesis of 3D scenes typically requires inferring and synthesizing new regions that are either unobserved or occluded in the original viewpoint. A paradigm in prior work [29, 32, 47, 62] is the warp-and-refine approach: estimate a depth map from the input image, warp the image to the desired viewpoint, and then fill in occluded or missing areas through a learned refinement stage. More recent research [26, 48, 68] avoids explicit depth-based warping by directly training generative models that handle view synthesis in a single feed-forward pass. Another line of work [2, 6, 9, 40, 43, 50–52, 54, 55, 69] integrates diffusion models such as StableDiffusion [49], making it possible to extrapolate plausible new views that are far from the input image for in-the-wild contents. ReconX [34] and ViewCrafter [69] both harness powerful video diffusion models combined with coarse 3D structure guidance to mitigate sparse-view ambiguities, achieving improved 3D consistency for novel-view synthesis. Motivated by recent success in the warp-and-refine paradigm [50], we adopt a similar strategy but with a focus on style lifting, incorporating historical frames through adaptive blending to consistently propagate our style transfers across multiple views.

3. ReStyle3D

We present ReStyle3D, a framework for fine-grained *appearance transfer* from a style image $\mathbf{I}_{style} \in \mathbb{R}^{H\times W\times 3}$, to a 3D scene captured by *unposed* multi-view images or video $\mathcal{X}_{src} := \{\mathbf{I}^i_{src} \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^N$. Specifically, ReStyle3D aims to transfer the appearance of each region in $\mathbf{I}_{style}$ to its semantically corresponding region in $\mathcal{X}_{src}$, while maintaining multi-view consistency across all images. We assume spatial overlap between two consecutive frames in $\mathcal{X}_{src}$.

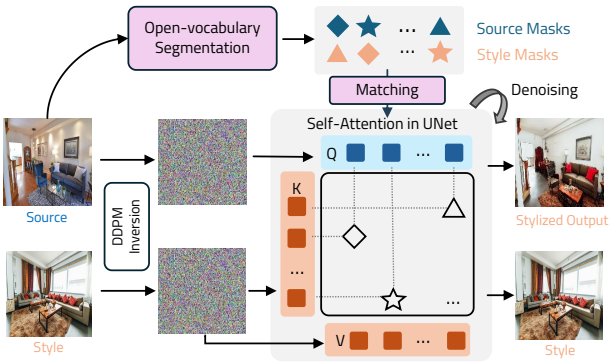

Figure 2. **Semantic Appearance Transfer**. The style and source images are first noised back to step T using DDPM inversion [25]. During the generation of the stylized output, the extended self-attention layer transfers style information from the style to the output latent. This process is further guided by a semantic matching mask, which allows for precise control.

3.1. Preliminaries

Diffusion models progressively add noise to an image $\mathbf{I}_0$ sampled from a data distribution $p_{\text{data}}(\mathbf{I})$, transforming it into Gaussian noise $\mathbf{I}_T$ over T steps, following a variance schedule $\{\alpha_t\}_{t=1}^T$:

$$p(\mathbf{I}_t \mid \mathbf{I}_0) = \mathcal{N}(\mathbf{I}_t;\ \sqrt{\alpha_t}\,\mathbf{I}_0, 1-\alpha_t\mathbf{I}), \tag{1}$$

where $\mathbf{I}_t$ represents the noisy image at timestep t. The *reverse* process is performed by a denoising model $\epsilon_\theta(\cdot)$ that gradually removes noise from $\mathbf{I}_t$ to obtain cleaner $\mathbf{I}_{t-1}$. Here θ is the learnable parameters of the denoising model. During training, the denoising model is trained to remove noise following the objective function [23]:

$$\mathcal{L} = \mathbb{E}_{\mathbf{I}_0,t\sim\mathcal{U}(T),\epsilon\sim\mathcal{N}(0,I)}||\hat{\epsilon}_\theta - \epsilon||_2^2, \tag{2}$$

where $\hat{\epsilon}_\theta = \hat{\epsilon}_\theta(\mathbf{I}_t, t, c)$, and c is an optional input condition such as text, image mask, or depth information. At inference stage, a clean image $\mathbf{I} := \mathbf{I}_0$ is reconstructed from a randomly sampled Gaussian noise $\mathbf{I}_T \sim \mathcal{N}(0, I)$ through an iterative noise-removal process. The cornerstone of modern image-based diffusion models is the latent diffusion model [49] (LDM), where the diffusion process is brought to the latent space [13] of a variational autoencoder (VAE). This approach is significantly more efficient compared to working directly in the pixel space.

Attention layers are fundamental building blocks in LDM. Given an intermediate feature map $F \in \mathbb{R}^{L\times d_h}$, where L denotes the feature length and d_h represents the feature dimension, the attention layer captures the interactions between all pairs of features through query-key-value

operations:

$$\phi = \text{softmax}\left(\frac{Q' \cdot K'^T}{\sqrt{d_h}}\right) \cdot V' \\ Q' = Q \cdot W_q, \quad K' = K \cdot W_k, \quad V' = V \cdot W_v, \tag{3}$$

where ϕ is the updated feature map, Q', K', and V' are linearly projected representations of the inputs via W_q, W_k, and W_v, respectively. In self-attention, the key, query, and value originate from the same feature map, enabling context exchange within the same domain. For cross attention, the key and value come from a different source, facilitating information exchange across domains. In ReStyle3D, we tailor the self-attention layers specifically for semantic appearance transfer and keep the cross-attention unchanged.

3.2. Appearance Transfer via Semantic Matching

To transfer the appearance of $\mathbf{I}_{style}$ to $\mathbf{I}_{src}$, prior attempts also employing diffusion models [1, 5, 70] have primarily focused on single objects, and struggle with scene-level transfer involving multiple instances. Our key observation is that the implicit semantic correspondences in foundation models [42, 49] are insufficient for more complex multi-instance semantic matching. To address this limitation, ReStyle3D explicitly establishes and leverages semantic correspondences throughout the transfer process.

Open-vocabulary Semantic Matching. We leverage the open-vocabulary panoptic segmentation model ODISE [63] for semantic matching. For a given input image, ODISE generates segmentation maps $\mathcal{M} \in \{1, \ldots, C\}^{H \times W}$, assigning each pixel to one of C semantic categories. These maps enable semantic correspondences between the style and source images (detailed below). By matching open-vocabulary semantic predictions, ReStyle3D is not limited by predefined semantic categories in a scene. The correspondences are injected into the diffusion process to guide appearance transfer between matched regions.

Injecting Semantic Correspondences in Self-attention. ReStyle3D enables training-free style transfer by extending the self-attention layer of a pretrained diffusion model (Fig. 2). This approach injects style information from $\mathbf{I}_{style}$ into $\mathbf{I}_{src}$ while preserving its structure. Specifically, we first encode both the style and source images into the latent space of Stable Diffusion [49], producing $\mathbf{z}_0^{style}$ and $\mathbf{z}_0^{src}$. These latent representations are then inverted to Gaussian noise, $\mathbf{z}_T^{style}$ and $\mathbf{z}_T^{src}$, using edit-friendly DDPM inversion [25]. To enhance structural preservation and mitigate LDM's over-saturation artifacts, we incorporate monocular depth estimates [64] of the input images through a depth-conditioned ControlNet [71] during the inversion process. The stylized image latent is then initialized as $\mathbf{z}_T^{out} = \mathbf{z}_T^{src}$.

Next, we transfer the style from $\mathbf{z}_T^{style}$ to $\mathbf{z}_T^{out}$ by denoising them along parallel paths [1]. At each de-noising step t, we extract style features (K_{style}, V_{style}) and query features Q_{out} from individual self-attention layers. The semantic-guided attention for the output feature ϕ_{out} is computed by combining the attention features with the attention mask M as follows:

$$\phi_{out} = \text{softmax}\left(\frac{Q_{out} \cdot K_{style}^T}{\sqrt{d_h}} \odot M\right) \cdot V_{style}, \tag{4}$$

where $\odot$ denotes element-wise multiplication and $\phi_{out} \in \mathbb{R}^{d^2 \times d_h}$ is passed to the next layer after self-attention.

To obtain the attention mask $M \in \mathbb{R}^{d^2 \times d^2}$, we flatten and bilinearly downsample the semantic masks $\mathcal{M}_{style}$ and $\mathcal{M}_{src}$ to match the resolution of attention feature maps, which is $d \times d$. The attention mask is defined as $M(i, j) = 1$ if the i-th region in the source and the j-th region in the style image share the same semantic class; otherwise, $M(i, j) = 0$. This formulation ensures that each region in the output image samples its appearance solely from semantically corresponding regions in the style image. For example, a chair in the source image is only cross-attended to its counterpart in the style image, inheriting its appearance. If multiple instances in the style image share the same semantic class, attention is distributed across them based on sampling weights determined by softmax attention scores. This mechanism naturally extends to support user-specified correspondences. Regions without semantic matches attend to the entire style image to preserve global harmony. Although semantic attention effectively transfers appearance, it may compromise the realism and structure of the stylized output, requiring further refinement.

Guidance and Refinement. We draw inspiration from [1, 22] and incorporate classifier-free guidance (CFG) combined with semantic and depth-conditioned generation. At each denoising step t, we compute three noise predictions: ϵ_t, ϵ_t^d and ϵ_t^s. Here, ϵ_t^s represents the predicted noise from the semantic attention path, ϵ_t^d is obtained from the depth-conditioned ControlNet [71], and ϵ_t is the unconditional noise prediction. The final noise prediction is then calculated as follows:

$$\hat{\epsilon}_t = (1 - \alpha)\epsilon_t + \alpha(\lambda_s \epsilon_t^s + \lambda_d \epsilon_t^d), \tag{5}$$

where λ_s and λ_d are the respective guidance weights ($\lambda_s + \lambda_d = 1$) for semantic and depth guidance. $(1 - \alpha)$ is the classifier-free guidance scale, which balances conditional and unconditional predictions, improving image realism.

To enhance image quality, we employ a two-stage refinement process. First, we upscale the initial stylized image from 512×512 to 1024×1024 resolution. Then, following SDEdit [38], we add high-frequency noise to this upscaled

Figure 3. **Appearance Transfer Multi-view Inconsistency**. When stylizing each view separately, we observe inconsistencies (red arrows) due to high variance in generative modeling.

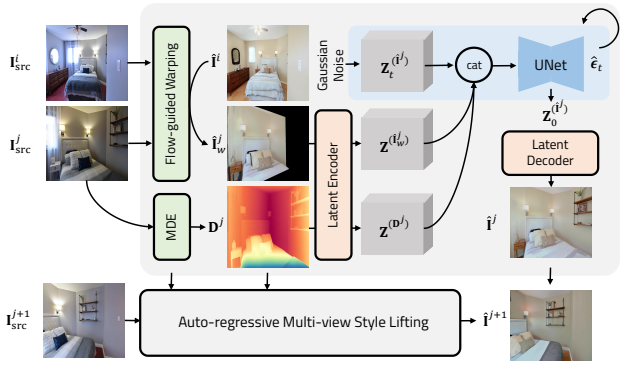

Figure 4. **Multi-view Style Lifting**. Stereo correspondences are extracted from the original image pair $(\mathbf{I}^i_{src}, \mathbf{I}^j_{src})$ and used to warp the stylized image $\hat{\mathbf{I}}^i$ to the second image, $\mathbf{I}^j_w$. To address missing pixels from warping, we train a warp-and-refine model to complete the stylized image $\hat{\mathbf{I}}^j$. This model is applied across multiple views within our auto-regressive framework.

image and denoise it for 100 steps with SDXL [46]. This refinement process enhances local details while maintaining the overall style, producing our final output $\hat{\mathbf{I}}_{src}$.

3.3. Multi-view Consistent Appearance Transfer

Although our semantic attention module effectively transfers appearance for a single view, independently applying it to each view may cause inconsistent artifacts (see Fig. 3). Therefore, we develop an approach to transfer the appearance from the stylized image $\hat{\mathbf{I}}^i_{src}$ to all remaining views while maintaining multi-view consistency.

Flow-guided Style Warping. Given a pair of source images $(\mathbf{I}^i_{src}, \mathbf{I}^j_{src})$, we first leverage a stereo matching method DUSt3R [60] to extract the dense point correspondence and the camera intrinsics. Using these, the optical flow $\mathbf{W}_{i\to j} \in \mathbb{R}^{H\times W\times 2}$ is calculated by projecting the pointmaps of i-th image to the j-th image. Next, given the optical flow and the stylized i-th image $\hat{\mathbf{I}}^i_{src}$, we employ softmax splatting [41] to obtain the initial stylized image $\hat{\mathbf{I}}^j_w$ and its warping mask $\mathbf{M}^j_w$, which indicates missing pixels in the j-th frame after forward warping.

Learning View-to-View Style Transfer. Given the source image $\mathbf{I}^j_{src}$ and its initial stylized version $\hat{\mathbf{I}}^j_w$, we train a 2-view warp-and-refine model $\hat{\epsilon}_\theta = \hat{\epsilon}_\theta(\mathbf{z}_t, t, c)$ to generate a complete and consistent stylized image following conditions c: the initial stylized image, the inpainting mask, and the monocular depth map $\mathbf{D}^j$ of the source image $\mathbf{I}^j_{src}$ (Fig. 4). The final condition $c = \text{concat}(\mathbf{z}^{(\hat{\mathbf{I}}^j_w)}, \mathbf{z}^{(\mathbf{M}^j_w)}, \mathbf{z}^{(\mathbf{D}^j)})$, $\mathbf{z}^*$ denotes individual latent representations. To harness the power of a pretrained diffusion model [46], like Marigold [27], we modify the input channels of its initial convolution layer to accommodate additional conditions and zero-initializing the additional weights. Following Eq. (2), we train the model using quadruplets of the warped and incomplete image, depth map, mask, and the clean and complete image. The model simultaneously learns to complete missing pixels and refine all pixels to address warping artifacts.

Auto-regressive Multi-view Stylization. We propose an auto-regressive approach to extend two-view stylization to handle multiple views or even videos, ensuring global coherence across the scene (Fig. 4). Stylizing the j-th frame using only the previous frame $(j-1)$ can lead to inconsistencies with earlier frames while warping all historical frames could produce blurry outputs. Instead, we warp the stylized frame $(j-1)$ along with two randomly selected historical frames. In overlapping regions, where multiple pixels are warped to the same location, we adopt an exponential weighted averaging to blend pixels, prioritizing pixels from frame $(j-1)$. This adaptive weighting maintains temporal consistency and preserves sharp details in the resulting warped image $\hat{\mathbf{I}}^{1:j-1}_w$. Finally, our model refines the output, producing a fully stylized frame.

4. Experiments

Implementation Details. We base our semantic attention module on Stable Diffusion 1.5 [49] and the refinement and 2-view warp-and-refine model on SDXL [46]. To train our two-view warp-and-refine model (Sec. 3.3), we use 4 NVIDIA A100 40GB GPUs with an effective batch size of 256 for 20K iterations, using the AdamW optimizer [36] with learning rate 10^{-4}. We randomly drop out half of the text prompt during training to make our model agnostic to text conditions. The model is trained on a dataset with 57K house tour images featuring 57 different houses/apartments.

4.1. Evaluation Setting

Dataset. Our *SceneTransfer* benchmark comprises 31 distinct indoor scenes captured as short video clips, totaling

Table 1. **Quantitative comparison of ReStyle3D and baseline methods on 2D appearance transfer**. Our method achieves the best overall performance for both structure preservation and percetual similarity, benefiting from its explicit semantic guidance and refinement.

Method	Depth Metrics (**Structure**)			Perceptual Similarity (**Style**)			Avg. Rank
	AbsRel↓	SqRel↓	δ_1↑	DINO↑	CLIP↑	DreamSim↓	
Cross-Image-Attn. [1]	22.47	7.944	5.78	0.553	0.709	0.414	4.8
IP-Adatper SDXL [66]	9.38	1.847	79.29	0.570	0.752	0.371	3.0
StyleID [7]	11.25	2.59	**93.44**	0.546	0.741	0.332	3.2
ReStyle3D (Ours w/o refinement)	11.30	2.65	89.11	**0.586**	0.778	0.319	2.5
ReStyle3D (Ours w/ refinement)	**8.34**	**1.67**	88.45	0.584	**0.783**	**0.316**	**1.5**

Figure 5. **Image appearance transfer results**. Our method enables precise appearance transfer between semantically corresponding elements, evidenced by the green rug and glass table (first row), textured cabinet (second row), and bedsheets (third row). Unlike baselines that either apply global style transfer or fail to preserve structure, ReStyle3D maintains both semantic fidelity and structural integrity.

15,778 frames across multiple room categories, including living rooms, kitchens, and bedrooms, all disjoint from our training data. To evaluate stylization capabilities, we curated a set of 25 interior design reference images, enabling 243 unique style-scene combinations. Evaluation is performed on 1,109 keyframes sampled from these clips. For more details on data, please refer to the supplementary material (Supp.).

Evaluation Metrics. We evaluate multiple different aspects of our pipeline. First, we assess the appearance transfer performance using source images on two aspects: structure preservation and style transfer quality. For structure preservation, we compare depth maps predicted by DepthAnythingV2 [64] between stylized and original images using standard metrics: Absolute Relative Error (AbsRel), $\delta 1$ accuracy, and Squared Relative Error (SqRel), following established protocols [27, 64]. For style transfer quality, we measure perceptual similarity between the stylized output and the style image using DINOv2 [42], CLIP, and DreamSim [15] scores. We evaluate this task on the stylized source images of each scene. Next, we evaluate our two-view lifting model (Sec. 3.3). We assess its warp-and-refine quality using PSNR, SSIM [61], and LPIPS [72] while also reporting FID [21] to quantify the realism of generated frames under challenging viewpoint extrapolation. We evaluate using pairs of the source images per scene and their warped projections on the rest of the frames in each scene—we exclude pairs without correspondences. We do not use any stylization to train or evaluate since there is no ground truth. To evaluate global consistency, we leverage DUSt3R [60] to extract poses by aligning point maps from stylized sequences and compute cumulative error curve (AUC) by comparing recovered camera

poses against those from original images.

4.2. Results

Image Appearance Transfer. We compare with three state-of-the-art methods on image-conditioned stylization and appearance transfer: Cross Image Attention [1], IP-Adapter [66], and StyleID [7]. For a fair comparison, we add depth ControlNet [71] to SDXL IP-Adapter [66] and use the style image as the image prompt. As shown in Tab. 1, our method achieves superior performance on both structure preservation and style transfer metrics. Notably, our explicit semantic attention mechanism in the diffusion UNet enhances the perceptual similarity between stylized outputs and style images, as evidenced by better DINO, CLIP, and DreamSim scores. The refinement step further improves structure preservation, reducing AbsRel from 11.30 to 8.34 and SqRel from 2.65 to 1.67. Qualitative comparisons (Figs. 5 and 7) reveal the limitations of existing approaches. Cross Image Attention effectively captures style textures but fails to maintain scene structure due to the lack of semantic guidance. IP-Adapter SDXL preserves overall structure but struggles with local detail transfer, as it compresses style information into a global feature vector. Although StyleID achieves the second-best performance, its results tend to preserve high-frequency details from the source image while applying style changes more globally, demonstrating limited capability in fine-grained appearance transfer.

We conduct a user study with 27 participants who were shown examples of a source and style image with outputs from four methods. Participants selected the result that best preserved the structure while faithfully transferring the style. Out of 252 evaluations (Tab. 2), ReStyle3D was the most preferred (42.4%), demonstrating its effectiveness in balancing structure preservation and appearance transfer under human perception.

Table 2. **Image Appearance Transfer User Study.** We show user preference rates (%) for different methods, where participants selected the result that best preserved the original scene structure while closely matching the reference style. ReStyle3D achieves the highest preference rate.

Method	ReStyle3D (Ours)	Cross Image Attn.	IP-Adapter	StyleID
Preferred Rate (%)	42.4	16.3	4.4	36.9

Two-view NVS. We compare our approach to: *i)* SDXL inpainting model [46] with depth-conditioned ControlNet [71], *ii)* GenWarp [50], an image-based diffusion model for single view NVS, and *iii)* ViewCrafter [69], a video-diffusion model for NVS. Note that the proposed task differs from traditional NVS as it leverages geometry information from the *novel view* itself. We employ DUSt3R [60] to extract the correspondences and provide the initial warped image as input to all methods. ReStyle3D outperforms across all metrics, achieving a superior reconstruction ability as evidenced by the best PSNR, SSIM, and LPIPS metrics (*cf*. Tab. 3,). Additionally, it exhibits strong capability in extending style to unseen regions, evidenced by the lowest FID score (Fig. 8). Notably, the second best method ViewCrafter [69], requires a predefined camera trajectory as input to video diffusion and runs 10× slower than ours.

Table 3. **Results on two-view novel-view synthesis.** ReStyle3D achieves the highest scores on all metrics, indicating more accurate view synthesis and visually pleasing outputs compared to existing methods.

Method	Res.	PSNR↑	SSIM↑	LPIPS↓	FID↓
GenWarp [50]	512^2	13.503	0.465	0.435	59.965
SDXL Inpainting [46]		16.228	0.535	0.389	89.502
ViewCrafter [69]		17.178	0.594	0.278	56.127
ReStyle3D (Ours)		**18.614**	**0.677**	**0.246**	**34.138**
GenWarp [50]	1024^2	13.491	0.565	0.440	60.540
SDXL Inpainting [46]		16.153	0.565	0.426	89.537
ViewCrafter [69]		17.137	0.652	0.317	57.898
ReStyle3D (Ours)		**18.568**	**0.711**	**0.283**	**35.721**

Table 4. **Pose deviation from real-world estimates.** We measure the fraction of camera poses within certain rotation (at $5°$, $10°$, $15°$) and translation (at $1,\text{cm}$, $2,\text{cm}$, $5,\text{cm}$) error thresholds, reporting area-under-curve (AUC) values. ReStyle3D achieves significantly higher AUC in both, showing superior multi-view geometric consistency vs. existing methods.

Method	Rotation AUC↑			Translation AUC↑		
	@5°	@10°	@15°	@1cm	@2cm	@5cm
GenWarp [50]	25.89	46.70	58.89	58.38	59.39	70.05
SDXL Inpainting [46]	34.52	52.79	66.50	61.42	65.99	74.11
ViewCrafter [69]	37.56	55.33	68.53	60.91	65.99	77.16
ReStyle3D (Ours)	**52.79**	**69.54**	**79.70**	**66.50**	**77.66**	**83.25**

Multi-view Consistency Evaluation. We further evaluate the multi-view consistency of the stylized results through a proxy task. Specifically, we input the original and stylized images to DUSt3R [60] and estimate the camera poses, separately. By evaluating the agreement with the poses from the original images, we analyze whether the geometry is preserved in the stylized images. The results are presented in Tab. 4. Enabled by our adaptive auto-regressive approach, which effectively mitigates inconsistencies while preserving image sharpness, our method significantly outperforms the baselines on both rotation and translation metrics. Figs. 8 and 9 show multi-view transfer results, including the 3D reconstruction of stylized outputs with estimated camera poses, demonstrating both geometric and stylistic consistency despite camera motion and multiple objects.

Ablation Study. In Tab. 5(a), we run ReStyle3D without our guidance strategy and observe significant degradation in structure preservation (AbsRel↑ from 8.34 to 16.72). In (b),

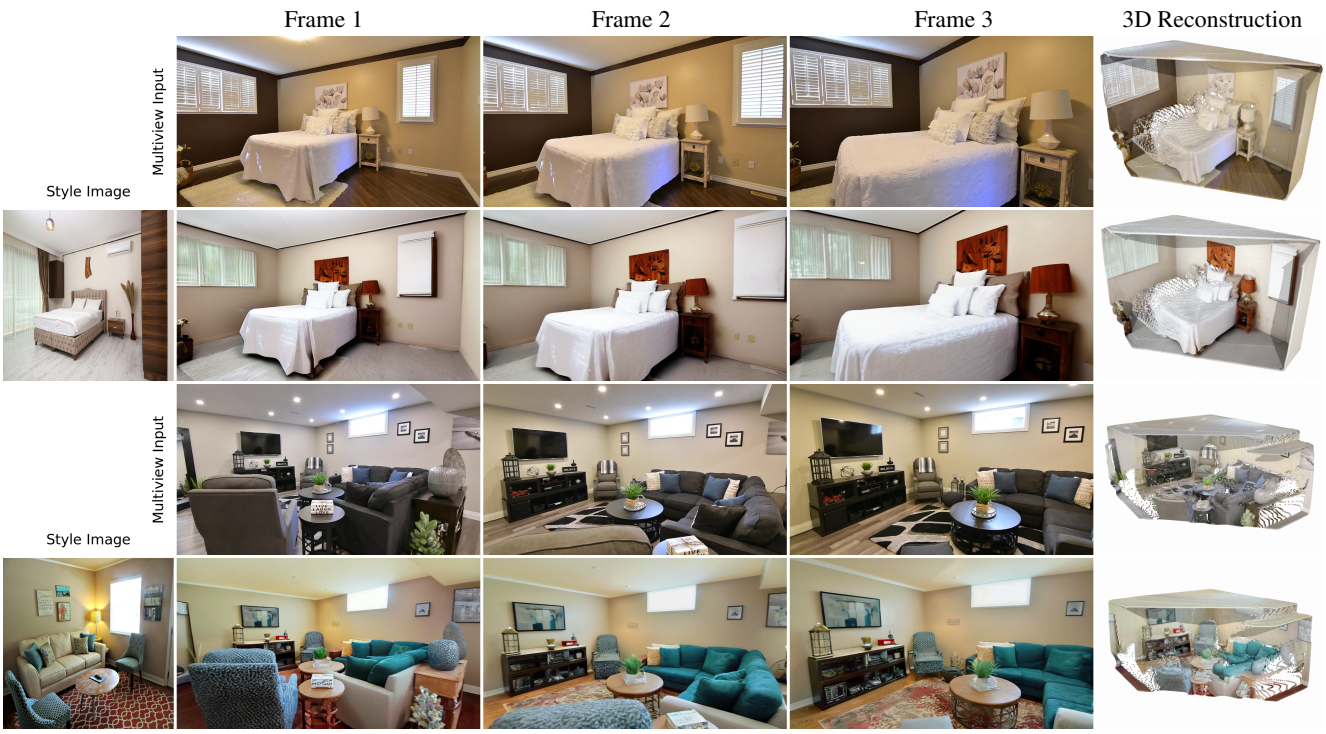

Figure 6. **Results on Video/Multi-view Appearance Transfer of ReStyle3D**. We show the style images, three frames stylized by ReStyle3D, followed by a 3D reconstruction of these outputs using an off-the-shelf pipeline. Despite challenging camera motion and multiple objects in the scene, our method preserves consistent geometry and seamlessly transfers the reference style across all frames.

Table 5. **Ablation Study.** We separately remove the guidance strategy and the semantic attention module to evaluate their impact on both structure preservation and style fidelity. Removing either significantly degrades performance, highlighting the importance of both components in achieving robust scene geometry and perceptually faithful stylization.

	AbsRel↓	SqRel↓	δ_1 ↑
Ours w/o guidance	16.72	4.36	67.46
Ours w/ guidance	**8.34**	**1.67**	**88.45**

(a) Ablation on Guidance

	DINO↑	CLIP↑	DreamSim↓
Ours w/o Sem. Attn.	0.492	0.682	0.419
Ours w/ Sem. Attn.	**0.584**	**0.783**	**0.316**

(b) Ablation on Semantic Attention

removing semantic attention hurts performance on perceptual similarity *w.r.t.* style image, showing that both components are crucial for semantic-accurate style transfer while maintaining structural integrity.

5. Conclusion

We presented ReStyle3D, a framework for semantic appearance transfer from a design image to multi-view scenes. Our two-stage approach combines training-free semantic attention in diffusion models with a warp-and-refine network to ensure geometric consistency across views. Experiments and user studies on our *SceneTransfer* benchmark confirm that ReStyle3D surpasses existing methods in semantic fidelity, structure preservation, and multi-view coherence. By using open-vocabulary segmentation and off-the-shelf reconstruction models, ReStyle3D avoids assumptions about scene semantics or geometry, making it suitable for real-world interior design and virtual staging. While focused on indoor scenes, its principles could apply to other domains. Limitations and more implementation details are discussed in the supplementary material.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *SIGGRAPH*, 2024. 2, 4, 6, 7

[2] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *ICCV*, 2023. 3

[3] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Huang, Tuanfeng Wang, and Gordon. Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In *CVPR*, 2024. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[5] Xinle Cheng, Congyue Deng, Adam Harley, Yixin Zhu, and Leonidas Guibas. Zero-shot image feature consensus with deep functional maps. In *ECCV*, 2024. 1, 2, 4

[6] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. In *arXiv*, 2023. 3

[7] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, 2024. 1, 2, 6, 7

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2

[9] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *SIGGRAPH*, 2024. 3

[10] Jiahua Dong and Yu-Xiong Wang. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. In *NeurIPS*, 2023. 2

[11] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017. 2

[12] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023. 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[14] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in Style. In *ICCV*, 2023. 2

[15] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 6

[16] Haruo Fujiwara, Yusuke Mukuta, and Tatsuya Harada. Style-nerf2nerf: 3d style transfer from style-aligned multi-view images. In *SIGGRAPH Asia*, 2024. 2

[17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[18] Sooyeon Go, Kyungmook Choi, Minjung Shin, and Youngjung Uh. Eye-for-an-eye: Appearance transfer with semantic correspondence in diffusion models, 2024. 2

[19] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *NeurIPS*, 2023. 2

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *arXiv*, 2022. 3

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. In *NeuIPS*, 2017. 6

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 4

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[25] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *CVPR*, 2024. 3, 4

[26] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *arXiv*, 2024. 3

[27] 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*, 2024. 5, 6

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[29] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. In *arXiv*, 2022. 3

[30] Shaoxu Li. Diffstyler: Diffusion-based localized image style transfer. In *arXiv*, 2024. 1, 2

[31] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. In *arXiv*, 2023. 2

[32] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 3

[33] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. In *TPAMI*, 2011. 2

[34] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. In *arXiv*, 2024. 3

[35] Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. Stylegaussian: Instant 3d style transfer with gaussian splatting. In *SIGGRAPH Asia*, 2024. 2

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[37] David G. Lowe. Distinctive image features from scale-invariant keypoints. 2004. 2

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 4

[39] 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*, 2021. 2

[40] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *CVPR*, 2024. 3

[41] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, 2020. 5

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, et al. DINOv2: Learning robust visual features without supervision. In *TMLR*, 2024. 2, 4, 6

[43] Hao Ouyang, Tiancheng Sun, Stephen Lombardi, and Kathryn Heal. Text2immersion: Generative immersive scene with 3d gaussians. In *Arxiv*, 2023. 3

[44] Bo Hyeon Park and Kyung Hoon Hyun. Analysis of pairings of colors and materials of furnishings in interior design with a data-driven framework. *Journal of Computational Design and Engineering*, 9(6):2419–2438, 2022. 1

[45] Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De La Torre. Consolidating attention features for multi-view image editing. In *SIGGRAPH Asia*, 2024. 2

[46] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 5, 7

[47] Chris Rockwell, David F. Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*, 2021. 3

[48] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *ICCV*, 2021. 3

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5

[50] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. In *NeurIPS*, 2024. 3, 7

[51] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. In *3DV*, 2025.

[52] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. In *arXiv*, 2024. 3

[53] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023. 2

[54] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023. 3

[55] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 3

[56] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *CVPR*, 2022. 2

[57] Narek Tumanyan, Omer Bar-Tal, Shir Amir, Shai Bagon, and Tali Dekel. Disentangling structure and appearance in vit feature space. *ACM Trans. Graph.*, 2023. 2

[58] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 3

[59] Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sýkora. Diffusion image analogies. In *SIGGRAPH*, 2023. 2

[60] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 5, 6, 7, 14

[61] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *TIP*, 2004. 6

[62] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 3

[63] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 4, 14

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 4, 6, 14

[65] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *ICCV*, 2023. 2

[66] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. In *arXiv*, 2023. 6, 7

[67] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *CVPR*, 2023. 14

[68] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[69] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. In *arXiv*, 2024. 3, 7

[70] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. In *NeurIPS*, 2023. 1, 2, 4

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4, 7

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[73] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023. 2

Figure 7. **Additional results on 2D appearance transfer**. Each example shows the source image, the reference style image, and the stylized outputs. While the baseline methods either disrupt scene structure or misalign local style details, ReStyle3D consistently preserves geometric fidelity and correctly maps the reference appearance to each semantic region. Subtle details like furniture textures and decorative elements are accurately adapted to match the style.

Figure 8. **Results on two-view NVS with warp-and-refine**. Given a single input view and a target viewpoint, each method attempts to synthesize the target frame by warping and refining the source image. ReStyle3D recovers more accurate geometry and fewer artifacts, while also preserving finer scene details. By contrast, baseline methods struggle with consistent edge alignment and realism, showing noticeable artifacts and incomplete regions.

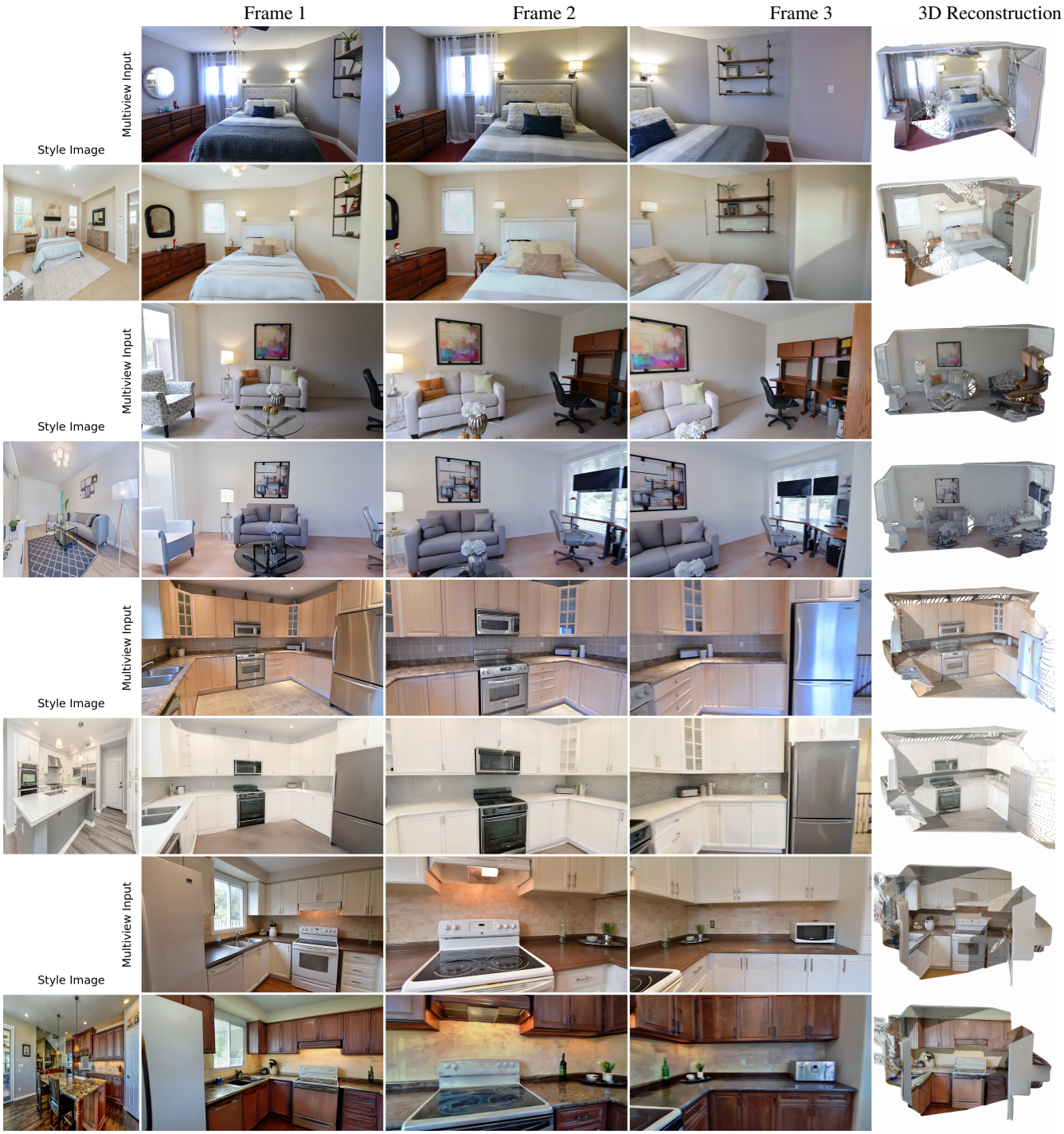

Figure 9. **Additional Results on Video/Multiview Appearance Transfer**. We showcase three frames from a new indoor sequence stylized by ReStyle3D, followed by a 3D reconstruction of these stylized images using an off-the-shelf algorithm. Despite dynamic viewpoint changes and scene complexity, ReStyle3D consistently enforces semantic correspondences and preserves geometric integrity across all frames, enabling high-quality multi-view edits for practical applications such as interior design or virtual staging.

A. Qualitative Ablation Study

We further validate our design choices with additional qualitative results in Fig. 10, comparing frame selection strategies for our warp-and-refine model. Warping only the last frame can cause inconsistencies, *e.g.*, the painting on the wall appears and disappears due to occlusion and incorrect correspondence estimates and occlusions (Fig. 10 (a)). Warping all past frames improves consistency but introduces conflicts, as overlapping pixel projections in the same location can result in smeared pixels and confuse the model about whether to refine or preserve them(Fig. 10 (b)). Our strategy can mitigate the disadvantages of both approaches, yielding more coherent multi-view stylization (Fig. 10 (c)).

We also validate the need for monocular depth condition to fully leverage two-view geometry in multi-view appearance transfer. As the stereo estimate from DUSt3R [60] implicitly contains geometry information, the additional monocular depth condition could be deemed unnecessary. To this end, we train a model without monocular depth estimation (MDE) that is solely conditioned on warped images, to ablate its effect. Fig. 11 shows that without MDE, the model still effectively learns the task, but loses fidelity in local details. With pixel-aligned MDE control, the model corrects wrong projections from two-view geometry based on the high-resolution depth map improving accuracy in areas like the door handle, chairs, and wine bottles.

B. Qualitative Video Results

We present additional results that demonstrate our method's versatility in handling diverse styles and scene contents. Our supplemental video best illustrates this versatility, showing how a single style image can transform multiple indoor scenes into a cohesive appearance, as well as how one scene can be re-imagined across various styles. The video format particularly highlights our method's ability to maintain visual consistency across multiple viewpoints, a key advantage that is difficult to convey through static images.

C. Data Curation

We curated the training data for our warp-and-refine model using 57 hourly-long 2K house tour videos. We first extract keyframes from all the videos at every 10^{th} frame to get 57K sparse multi-view images. Then we divide the keyframes into 10-image chunks to emulate different overlap ratios in real-world scenarios. In each window, we use DUSt3R [60] to compute the dense stereo between the first frame in the window and all the rest frames and get 9 warped images and their warping masks in each window. We also use Depth Anything V2 [64] to extract the monocular depth for the 9^{th} image in the window. Then each training sample is generated as (warped image, monocular depth, mask, and clean target image) and used for training.

We would like to note that both the warp-and-refine training data and the *SceneTransfer* benchmark data are derived from a collection of house tour videos, which are part of a separate submission to another conference. These videos will be made publicly available through that paper, along with all relevant metadata (e.g., video IDs, start-end timestamps) to enable full replication and encourage future research and comparisons. The selection of this dataset was intentional. Unlike existing indoor scene datasets [67**?**], which are typically designed for 3D reconstruction and feature imagery that remains close to surfaces and lacks spatial context, our house tour videos were collected for real-estate purposes. They offer stable, smooth trajectories with views that are further from surfaces, making them more suitable for stylization tasks.

The style images in the *SceneTransfer* benchmark were manually collected from *Pexels*, which provides copyright-free interior images.

D. Style Images and Segmentation

Fig 12 displays several style images from our *SceneTransfer* benchmark alongside their open-vocabulary segmentation overlays. Each row pairs a real indoor photograph (kitchen, living room, or bedroom) with its corresponding color-coded semantic masks, as predicted by ODISE [63]. The segmentation assigns each pixel to a semantic category (e.g., "wall", "sofa", "cabinet", "table")[1], enabling instance-level alignment between the target scene and the reference style images. In the ReStyle3D pipeline, these masks form the basis for establishing precise semantic correspondences. By ensuring that each object region in the target image only "borrows" style cues from its from its corresponding instance in the reference image, our framework preserves the scene layout while selectively transferring appearance. This capability is particularly valuable in multi-object, complex scenes like the ones shown here, where various furniture and architectural elements must be stylized coherently while maintaining their original spatial relationships.

E. Limitations and Future Work

In this work we focused exclusively on indoor scenes for interior design applications, without exploring other types of scenes, such as outdoor or dynamic environments. While our method effectively transfers appearances, it lacks strong disentanglement between color, texture, and material prop-

[1]The open vocabulary method was queried with a list of common semantic labels. In our experiments we used the semantic list of the ODISE model.

Figure 10. **Comparison between different auto-regression strategies.** The rows represent two non-sequential views used in multi-view appearance transfer, while the columns represent different frame selection strategies. (a) *Single* warps only the previous view, failing to preserve details like the painting on the wall present in the top image, as pixel information is lost. (b) *All* warps all past frames to the current frame, improving consistency but introducing smearing effects, demonstrated by the red arrows (highlighted by red arrows) due to overlapping pixel projections with large color value differences originating from lighting changes. (c) *Ours*, the proposed strategy in ReStyle3D, achieves the most consistent and clean multi-view stylization.

Figure 11. **Qualitative comparison between two different conditional models.** We train a warp-and-refine model without the monocular depth (MDE) condition. Without MDE supervision, the model struggles to correct local alignment issues (*e.g.*, the door handle, wine bottles, and chairs), resulting in noticeable geometric distortions and texture artifacts. In contrast, incorporating MDE enables the model to leverage pixel-aligned depth cues and more faithfully reconstruct fine details, yielding sharper and more consistent multi-view results.

erties, and struggles with significant lighting changes. Future work could address these limitations by improving the transfer of appearances for smaller objects in the scene—currently overlooked due to downsampling of semantic masks— and developing finer control mechanism for appearance transfer, such as material or texture.

Figure 12. **Style Image Samples.** We showcase exemplar images of various designs across various styles and types of rooms from our *SceneTransfer* benchmark style image collection (top rows), paired with their open-vocabulary semantic segmentation masks (bottom rows). These images span a variety of interior spaces–including kitchens, living rooms, and bedrooms–and illustrate the core challenge of ReStyle3D: transferring the appearance of each semantically matched region (e.g., cabinets, couches, and beds) from a style image to real-world 3D scenes. Our method leverages these semantic masks to ensure each object is stylized with the correct instance-level textures while maintaining photo-realism and multi-view consistency.