PanoStyle: Semantic, Geometry-Aware and Shading Independent Photorealistic Style Transfer for Indoor Panoramic Scenes

M. Tukur, A. Ur Rehman
ICT, CSE, HBKU
Doha, Qatar

G. Pintore, E. Gobbetti
CRS4
Cagliari, Italy

J. Schneider, M. Agus
ICT, CSE, HBKU
Doha, Qatar
{jeschneider|magus}@hbku.edu.qa

Abstract

While current style transfer models have achieved impressive results for the application of artistic style to generic images, they face challenges in achieving photorealistic performances on indoor scenes, especially the ones represented by panoramic images. Moreover, existing models overlook the unique characteristics of indoor panoramas, which possess particular geometry and semantic properties. To address these limitations, we propose the first geometry-aware and shading-independent, photorealistic and semantic style transfer method for indoor panoramic scenes. Our approach extends semantic-aware generative adversarial architecture capabilities by introducing two novel strategies to account the geometric characteristics of indoor scenes and to enhance performance. Firstly, we incorporate strong geometry losses that use layout and depth inference at the training stage to enforce shape consistency between generated and ground truth scenes. Secondly, we apply a shading decomposition scheme to extract the albedo and normalized shading signal from the original scenes, and we apply the style transfer on albedo instead of full RGB images, thereby preventing shading-related bleeding issues. On top of that, we apply super-resolution to the resulting scenes to improve image quality and yield fine details. We evaluate our model's performance on public domain synthetic data sets. Our proposed architecture outperforms state-of-the-art style transfer models in terms of perceptual and accuracy metrics, achieving a 26.76% lower ArtFID, a 6.95% higher PSNR, and a 25.23% higher SSIM. The visual results show that our method is effective in producing realistic and visually pleasing indoor scenes.

1. Introduction

With the recent advances in deep learning, style transfer has become a popular technique in the field of computer graphics/vision. It enables the transfer of styles from one image to another while maintaining the content of the original image. This technique has been applied in various fields, including photography, advertising, and digital art. This technology has seen significant improvements over the last years, and, by now, style transfer can be used efficiently for automatic creation of compelling images through the application of artistic styles to casual photography [55, 6], or for the generation of high-quality, photo-realistic images related to specific categories like low-resolution portraits [59] or natural scenes [4].

Simultaneously, panoramic or spherical images, capable of capturing complete environments in a single shot, have become increasingly popular, particularly for representing indoor scenes in applications like virtual staging [57]. In this context, photo-realistic style transfer methods, if they were availabile, would offer a significant advantage for automatically creating immersive indoor environments. However, despite recent advances, even the most up-to-date style transfer approaches have yet to effectively handle indoor panoramic images. As a result, the generated images exhibit noticeable artifacts that limit their practical usability in virtual reality or architectural design applications.

In general, the application of style transfer to indoor panoramic images presents several significant challenges, including the following.

- *Complex illumination patterns*: Indoor environments consist of various reflective surfaces, diverse lighting conditions (both direct and indirect), and occlusions, leading to intricate illumination patterns. Style transfer architectures often struggle to accurately interpret and preserve these complex illumination patterns during the transfer process.

- *Incorporating 3D characteristics*: Automatic processing methods need to incorporate knowledge about the specific 3D characteristics of indoor scenes, such as their layout and clutter. Style transfer algorithms should consider the spatial relationships between objects and surfaces within

the panoramic image to ensure coherent and realistic style transfer results.

- *Equirectangular projection distortions*: Panoramic indoor images typically use equirectangular projections, which introduce distortions due to the transformation from the spherical to the planar domain. Style transfer methods need to account for these distortions to maintain the visual quality and consistency of the transferred styles across the entire panoramic image.
- *High-resolution requirements*: To enable immersive applications and professional editing purposes, panoramic images need to have a high resolution. Meeting these resolution requirements poses a challenge as current style transfer technologies may struggle to handle the computational demands and memory constraints associated with high-resolution panoramic images.

This paper addresses all these challenges by proposing the first semantic, geometry-aware and shading-independent architecture dubbed *PanoStyle*. Our architecture can generate photo-realistic indoor panoramic images in a controlled fashion (cf. Fig. 1). Our architecture is a custom-built evolution of classical style-based generative adversarial network [34, 59] for panoramic indoor scenes and incorporates the following technical contributions.

- *Shading-independent style encoding:* We propose to model the illumination of an indoor scene according to intrinsic image decomposition, and we approximate shading as a normalized neutral modulation signal; we encode styles by considering reflectance in a way that prevents intricate lighting patterns from creating bleeding artifacts in the generative process (Sec. 3.1).
- *Geometry-aware discrimination:* We design specific losses for supporting the adversarial process that take into account the particular 3D characteristics of the indoor scenes, in terms of clutter, layout, and edges. We achieve this by using state of the art models for depth and layout inference, and we compensate the equirectangular projection through area-based weighting. Finally, we exploit the latent features extracted during depth and layout prediction for deriving geometry style losses enforcing global and local similarity (Sec. 3.2).

Furthermore, we improve the resolution of the generated images by applying GAN-based super-resolution models [47]. To evaluate the effectiveness of our approach, we provide an extensive empirical evaluation on the synthetic Structured3D data set [56], using a wide range of quantitative and qualitative metrics. This data set provides a suitable benchmark for testing our method's performance, as it has comprehensive ground truth annotations for indoor panoramas. *PanoStyle* significantly improves the quality of generating photo-realistic panoramic images compared to state-of-the-art methods [59, 17], as demonstrated by clear improvements in quantitative and perceptual metrics (such as SSIM, PSNR, and ArtFID score [49] (see also Table 1) and visual inspection (see also Figs. 4 and 5). The resulting stylized panoramic images can be visualized in 3D using various 3D-Viewers [43], and can be used for VR and metaverse-based applications. This has significant implications for industries such as real estate, furniture retail, and interior design, where users can visualize the edited space in a more immersive and realistic manner. Despite the significant advantages, however, the architecture also has a significant drawback limiting the applicability to real world scenarios: it depends on the availability of high quality annotated data, in the form of semantic segmentation of input images and pre-computed reflectance signals.

Figure 1. We present the first geometry-aware and shading-independent, photorealistic and semantic style transfer method for indoor panoramic scenes. Left column: Content images. Remaining columns: different editing scenarios. Top row: target style images. Remaining rows: successive edits (floor, wall, ceiling). Some edits also include doors, windows, towels, and toilets.

2. Related work

Generative Adversarial Networks (GANs) have emerged as a fundamental tool in the field of Neural Style Transfer (NST) and have demonstrated remarkable efficacy in image synthesis tasks. The inception of GANs in 2014 [10] marked a groundbreaking development. Since then, they have been utilized in various domains, including image inpainting [54], image editing [1], and texture generation [8]. Constant advancements in GAN architecture [34], loss function design [29], and regularization techniques [30] have contributed to the production of increasingly realistic and high-quality images. Recently, Ruder *et al*. [39] adapted the original artistic image style transfer method to spherical images based on energy minimization [9] . One notable example of the progress in GANs is the StyleGAN model [21], which enables the generation of remarkably detailed and high-fidelity human facial images. Traditional GANs suffer from a lack of user control as they rely on noise vectors as input. To address this limitation, condi-

tional GANs (cGANs) [31] were conceived, allowing users to provide conditioning data to the generator and regulate the synthesis process. By incorporating conditional information such as class labels [33], textual descriptions [52], or reference images [45], cGANs can produce tailored and contextualized outputs, expanding the creative possibilities and enhancing the user experience.

Image-conditional GANs have proven to be versatile solutions for various image-to-image translation challenges. The effectiveness of employing image-conditional GANs was showcased in the seminal work of Isola *et al*. [18], demonstrating their applicability in diverse scenarios, including unsupervised learning [58], few-shot learning [27], high-resolution image synthesis [45], multi-modal image synthesis [15], and multi-domain image synthesis [5]. Among these applications, semantic image synthesis stands out as an invaluable form of image-to-image translation, as it enables users to manipulate the input semantic layout image and exercise effortless control over the synthesis process. For what concerns photorealistic style transfer, recently various solutions have been proposed: for example, Li *et al*. [25] developed a method with a stylization step and a smoothing step for ensuring spatial consistency, Yoo *et al*. [53] propose a wavelet correction scheme based on whitening and coloring transforms (WCT2) preserving the structural information and statistical properties of VGG feature space during stylization, and Xia *et al*. [50] propose a feed-forward neural network learning local edge-aware affine transforms in a way to enforce the photorealism constraint. However, all these methods do not consider the semantic content and the geometry content of the scene. The conventional network architecture [45, 18] employed in GANs consists of stacked convolutional, normalization, and non-linearity layers. However, this architecture tends to diminish the information present in input semantic masks, which can negatively affect the quality of synthesized images. To overcome this challenge, Spatially-Adaptive Normalization (SPADE) was introduced [34], leveraging conditional normalization layers to effectively propagate semantic information throughout the network. By modulating activations based on input semantic layouts through a spatially adaptive, learned transformation, SPADE ensures enhanced information preservation and utilization. Furthermore, in order to exert individual control over the style of each semantic region within an image, Zhu *et al*. [59] proposed Semantic Region-Adaptive Normalization (SEAN). While SEAN demonstrated excellent performance on facial images, its application to indoor panoramic images revealed limitations. These limitations stem from the neglect of factors such as depth preservation and coherence of details, which are considered crucial for assessing the visual quality of NST algorithm results [19]. In this work, we build on top of the architecture proposed by Zhu *et al*. [59] by incorporating Spectral Norm [32] in both the generator and discriminator modules together with a shading independent style encoding module operating on reflectance signals.

The preservation of both the depth and the structural characteristics is crucial for generating aesthetically pleasing stylized images. Consequently, some researchers [28, 3] have focused on enhancing structural characteristics, particularly for images with faces or varying object depths, by adjusting the amount of retained structure during stylization. Integrating depth preservation as an additional loss function has also been explored, utilizing deep residual convolutional networks and advanced depth prediction networks [17]. However, preserving both depth and semantic information simultaneously remains a challenge. Moreover, an innovative technique called StyleMesh [14] revolutionizes mesh stylization in room-scale interior scenes by extending style transfer into the immersive 3D domain. It overcomes limitations of traditional methods by leveraging mesh depth and surface normals, employing a meticulously designed loss calculation methodology. However, this method is designed for perspective images and relyies on intrinsic and extrinsic camera estimation. It is, thus, not adaptable to panoramic images. Instead, we preserve the depth and structural characteristics of panoramic indoor scenes by designing geometry constraints that take into account the characteristics of panoramic images, the scene content, and the layout of the represented environments.

Apart from the aforementioned studies, the realm of panoramic depth estimation has seen various research endeavors, including approaches using deformable convolutional filters [42], specialized encoder-decoder networks [60], and perspective views in cubemap formats [44]. Other methods involve encoding panoramic inputs into vertical slices [37, 41], leveraging stereomatching concepts [26], and utilizing transformer architectures to align/merge depth maps [24]. Contributions also include techniques for stitching perspective depth maps [38] and computing high-resolution depths for panoramas [36]. All of these advancements aim to improve accuracy and efficiency in panoramic depth estimation.

In our work, we exploit the most advanced pre-trained depth and layout inference architectures [37, 40] to constrain the generation process such that it preserves the geometric details of the indoor scene, and we incorporate a normalized shading decomposition component to preserve the lighting information. In this way, our architecture is able to produce high-quality photo-realistic panoramic images. For what concerns reflectance models and albedo disentanglement, they have been exploited successfully for applications related to makeup simulation [23], but, to our knowledge, the concept of intrinsics image decomposition has never been applied to panoramic indoor scenes.

3. Methodology

Figure 2 depicts PanoStyle's architecture. It is composed by a generative adversarial architecture exploiting semantic-aware style encoding [59], in which we integrate the following novel components.

1. A normalized shading decomposition component for extracting neutral illumination signals from the original scene to be used in the discriminator process and in the application of style transfer (Sec. 3.1);

2. A geometry consistency component exploiting the particular features of indoor scenes through custom losses that consider the geometry content of the scene in terms of depth, layout, and edge signals to be applied as additional constraints in the discriminator network of the GAN architecture (Sec. 3.2).

The training process of PanoStyle takes as input an RGB image of indoor scenes, together with precomputed semantic content, original depth, and reflectance (albedo) signals. Once trained, our PanoStyle model can be used for generating novel indoor scenes by applying different style images to indoor scenes in a controlled fashion through the interactive modification of semantic signals: casual users can adaptively select which parts of the style scene are applied to the original scene (Sec. 3.3).

3.1. Shading-independent encoding

Normalized shading decomposition. Figure 2 provides a schematic overview of our shading decomposition component. We base our method on classical intrinsic image decomposition, that is, the process of extracting the underlying components of an image, namely reflectance (albedo) and shading (illumination) [2]. The reflectance component represents the actual color (or albedo) of an object, unaffected by illumination or camera viewpoint. On the other hand, the shading component includes various photometric effects like direct light, ambient light (reflections within and between objects), and shadows. Employing intrinsic images instead of raw RGB images has proven to offer advantages in various computer vision applications, ranging from fabric re-colorization [51] to 3D shape reconstruction [11]. In our case, we start with an original RGB image I_{rgb} and the available reflectance signal I_{alb} to define a normalized illumination modulation scalar signal I_{shad} as the Euclidean norm of the Hadamard division between I_{rgb} and I_{alb}:

$$I_{\text{shad}} := \max\left(\left\|I_{\text{rgb}} \oslash I_{\text{alb}}\right\|_2, 1\right), \tag{1}$$

in a way to neutralize all secondary effects that would provide color artifacts during the application of different reflectance signals to the same scene. In this way, the application of the illumination signal would lead to an approximated shaded scene in the form of

$$\hat{I}_{\text{rgb}} = I_{\text{shad}} \cdot I_{\text{alb}}. \tag{2}$$

We carried out a preliminary analysis of the effects of this first approximation shading model using the Structured3D data set, and we found that the difference with respect to the ground truth signals are perceptually negligible (see Sec. 4). We represent the shading signal as 16-bit resolution single channel images, and we apply this intrinsic image decomposition scheme during pre-processing to generate a shading archive to be used during the application of the style transfer.

Shading-independent style encoder. The style encoder in the GAN architecture is composed of a bottleneck convolutional neural network (CNN) and a region-wise average pooling layer (see also Fig. 2). It takes the style image (albedo) and its corresponding segmentation mask as inputs and produces style codes as outputs. To avoid bleeding issues that can arise from shading effects, especially in panoramic indoor scenes, we use albedo instead of RGB. In this way, we ensure that the generated images are not affected by variations in lighting conditions. This is of particular importance in the context of photo-realistic style transfer, since we aim to preserve the style of the input image while transferring it to a new scene.

Moreover, by utilizing a bottleneck CNN architecture (as also used by SEAN [59]), we increase the receptive field while managing computational cost. The resulting gradual dimension reduction is important for compact style codes and, in more general frameworks than ours (left for future work), allows to apply loss functions at different image resolutions. In addition, the latent space can be further regularized for better control of the generation process. The region-wise average pooling layer further enhances the performance of the style encoder by allowing it to capture spatial information about the image at a higher level of abstraction.

3.2. Indoor geometry consistency

Our geometry consistency component is integrated in the discriminator network of the semantic-aware generative adversarial architecture [59, 34], in form of losses added to the classical discriminator components (the conditional adversarial loss, the feature matching loss [45], and the perceptual loss [20]). To this end, we introduce the following geometry constraints that enforce the consistency of the geometry content created during the training process by the generator architecture.

1. A loss contribution $\mathcal{L}_{\text{depth}}$ to enforce the geometric and latent consistency of the depth signal of the generated scene;

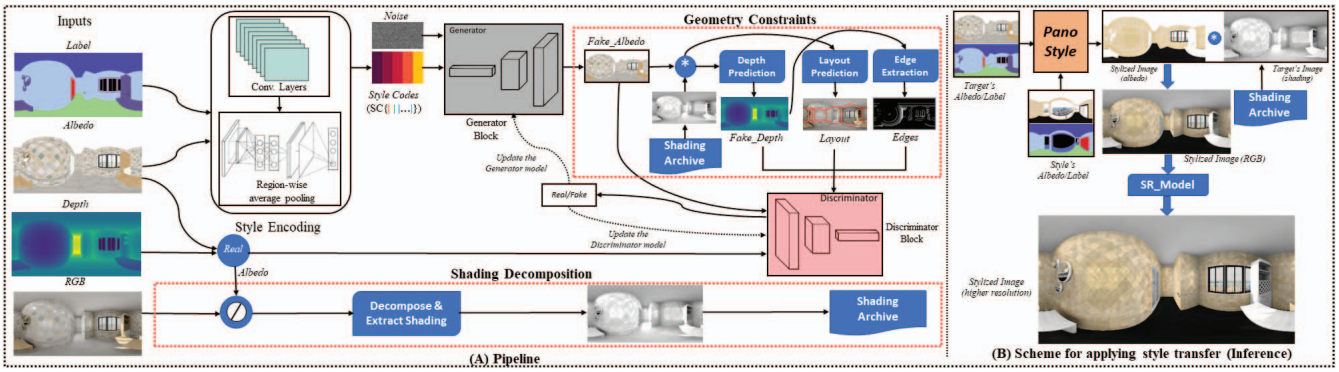

Figure 2. **PanoStyle framework.** (A) PanoStyle uses a GAN model exploiting semantic-aware style encoding, integrated with a shading decomposition component and geometry constraints in the discriminator. (B) PanoStyle applied to target, style reflectance, and semantic signals to generate a new scene. Using a target shading signal followed by super-resolution to enhance details results in the stylized image.

2. A contribution $\mathcal{L}_{\text{layout}}$ to enforce the geometric and latent consistency of the layout of the generated scene;

3. A contribution $\mathcal{L}_{\text{edge}}$ to enforce the geometric consistency of the edges of the generated scene.

Concerning the metrics involved, we applied standard L_1 losses in all cases but we also performed experiments with other metrics like berHu (inverse Huber) [7] without obtaining significant advantages.

Depth consistency. In order to enforce depth consistency of the generated scenes, we exploit the capabilities of state-of-the-art depth inference models (in our experiments we considered Slicenet [37]) to predict the depth signal of generated fake scenes D^{G} to be compared to the original real scenes D^{R}. We use the tensors D^{G} and D^{R} to compute a pure geometry loss with a weighting scheme that considers the change in area when uniformly distributed samples from the equirectangular domain are mapped to the spherical surface representing the scene:

$$\mathcal{L}_{\text{depth}}^{\text{geo}} = \sum_{ij} w_{ij} \left\| D_{ij}^{\text{G}} - D_{ij}^{\text{R}} \right\|_1 , \tag{3}$$

where the spherical weights are computed in a way to penalize pixels close to the poles of the sphere:

$$w_{ij} = \cos\left(\frac{\pi}{N} \left(j + \frac{1}{2} - \frac{N}{2} \right) \right) , \tag{4}$$

N being the height in pixels of the panoramic images.

Moreover, from the depth inference model, we extract the latent features F_n of both the generated image and ground truth depth for computing additional loss components incorporating the philosophy of style-transfer [1], for preserving the global content of the scene:

$$\mathcal{L}_{\text{depth}}^{\text{glob}} = \sum_{n} \left\| F_n(D^{\text{G}}) - F_n(D^{\text{R}}) \right\|_1 \tag{5}$$

Similarly, we consider an objective function for enforcing local similarity, acting as a sort of *style* loss, based on the *Gram matrix* function of the same latent features F_n:

$$\mathcal{L}_{\text{depth}}^{\text{loc}} = \sum_{n} \Big\| K_n \Big(F_n \left(D^{\text{G}} \right)^T F_n \left(D^{\text{G}} \right) - F_n \left(D^{\text{R}} \right)^T F_n \left(D^{\text{R}} \right) \Big) \Big\|_1 , \tag{6}$$

where K_n is the Gram matrix normalization factor $1/(s \times l)$ for the n^{th} layer. The complete depth consistency loss is a weighted sum of the three aforementioned components:

$$\mathcal{L}_{\text{depth}} = \lambda_{\text{d1}} \mathcal{L}_{\text{depth}}^{\text{geo}} + \lambda_{\text{d2}} \mathcal{L}_{\text{depth}}^{\text{glob}} + \lambda_{\text{d3}} \mathcal{L}_{\text{depth}}^{\text{loc}} \tag{7}$$

Layout consistency. For enforcing layout consistency, we designed a set of losses representing similar concepts considered for deriving depth constraints: pure geometric consistency, and local and global similarity in the latent feature space. In this case, we use the popular HorizonNet model [40] for predicting indoor scene layouts in the form of 1D parameterized functions for both the floor and ceiling, and the three loss contributions are computed on predicted layouts for the original scene L^{R} and for the generated image L^{G}, and the corresponding latent features H_n in similar fashion as for depth objective functions (see Eqs. 5 and 6), leading to the weighted layout consistency loss

$$\mathcal{L}_{\text{layout}} = \lambda_{\text{l1}} \mathcal{L}_{\text{layout}}^{\text{geo}} + \lambda_{\text{l2}} \mathcal{L}_{\text{layout}}^{\text{glob}} + \lambda_{\text{l3}} \mathcal{L}_{\text{layout}}^{\text{loc}} . \tag{8}$$

Edge consistency. For constraining the edge content of the generated scene to be consistent with the original image, we approximate the gradient components of the generated and real depth signals D^{G} and D^{R} through a convolution with Sobel filters of width 3 for the horizontal and vertical derivatives that we use for an additional geometric loss related to edge content [35],

$$\mathcal{L}_{\text{edge}}^{\text{geo}} = \lambda_{\text{e}} \left\| \nabla \left(D^{\text{R}} - D^{\text{G}} \right) \right\|_1 . \tag{9}$$

Figure 3. Comparison between original RGB image and prediction using our shading decomposition system (Structured3D data set). We report PSNR, W-PSNR, and SSIM. Depicted are visual comparisons, color-mapped differences, and accompanying shading/albedo signals.

3.3. Semantic image synthesis

The semantic-aware style transfer task involves generating a realistic image based on a given semantic layout that contains pixel-level labels. This problem is inherently challenging, especially for indoor panoramic images, since multiple feasible images may correspond to a single semantic layout. Consequently, in order to extract styles specific to each region, PanoStyle's generator employs the style encoder network described in Sec. 3.1, to simultaneously distill the corresponding style code from every semantic region of an input image (see also Fig. 2 A). The style encoder produces a style matrix, denoted as SC, which has dimensions of $512\times$ the number of labels (representing the number of semantic regions present in the input image). Each column in the matrix (SC) corresponds to the style code of a specific semantic region. To focus solely on style-related information and eliminates irrelevant details regarding shapes of semantic regions, the per-region style encoder utilizes a bottleneck structure. Intermediate feature maps, consisting of 512 channels, generated by the convolution layers network block, undergo region-wise average pooling. This pooling operation reduces the feature maps into a collection of 512-dimensional vectors. It is important to note that we handle cases where the number of semantic labels in the data set may differ. In such instances, we set the columns corresponding to non-existent regions in an input image to zero, ensuring consistency and accurate representation. Thus, our proposed model enables the user to control both semantic and style during image synthesis. The semantic information, such as the presence of a cabinet in an indoor environment, can be controlled via a label map. The style can be controlled on a region-by-region basis, as demonstrated in the generated stylized image of Fig. 2(B)) in which the style is applied to cabinet, floor, wall, ceiling only, and of Fig. 7, where the style is applied to only floor, wall, and ceiling. In Fig. 6 the style was applied to all semantic classes. Additionally, our model supports style crossover, allowing multiple style images to be applied to a single content image, as illustrated in Fig. 8. Furthermore, our model being trained on albedo necessitates multiplication of the resulting generated images with the shading information that was initially extracted from the original RGB and albedo frames to generate a stylized RGB image. Finally, we apply a super-resolution model [46] on the generated stylized RGB image, which results in a more visually appealing, immersive, and photo-realistic image with fewer artifacts (cf. Fig. 2(B)).

4. Results

Implementation and training details. Our PanoStyle framework has been implemented in Python on top of the PyTorch library. For the generative adversarial architecture, the learning rates for the generator and discriminator are set to 0.0001 and 0.0004, respectively [12], and we employ the ADAM optimizer [22] with $\beta_1 = 0$ and $\beta_2 = 0.999$. All experiments are performed on a single NVIDIA Quadro RTX 8000 equipped with 48GB GPU. Due to PanoStyle working on high-resolution images with a size of 1024×512 pixels and on a single GPU, we have opted for a batch size of 2. Moreover, in all the experiments detailed in the paper, we trained for 100 epochs and we estimated a training time of 3.6 seconds per image per epoch. We use the following parameters for the geometry consistency losses: $\lambda_{\text{d1}} = 0.1$, $\lambda_{\text{d2}} = 10^5$, and $\lambda_{\text{d3}} = 10^9$ for depth consistency, $\lambda_{\text{l1}} = 10^5$, $\lambda_{\text{l2}} = 10^5$, and $\lambda_{\text{l3}} = 10^9$ for layout consistency, and $\lambda_{\text{e}} = 10^9$ for the edge constraint.

Table 1. Quantitative Comparison and Ablation Study of Self-Style Semantic Image Synthesis Results on the Structured3D data set. We conduct a comparison between SEAN [59], DANST [17], and our approach, PanoStyle. To ensure a fair comparison with our competitors, the superresolution has not been applied to both our method and the competing approaches.

Method	PSNR ↑	SSIM ↑	ArtFID ↓
SEAN	18.9147	0.6326	10.7989
DANST	10.1023	0.5742	23.5553
Ours	**20.2289**	**0.7922**	**7.9094**
Ours(w/o layout)	19.8277	0.7715	9.3261
Ours(w/o depth)	19.5962	0.7857	8.9405
Ours(w/o geometry)	19.3729	0.7847	9.1599
Ours(w/o shading)	18.9932	0.6522	10.4410

Data sets. Our experiments make use of the Structured3D data set, which comprises over 18,000 RGB-D images sourced from 3,500 different scenes (house designs). This data set offers an ideal benchmark for testing the efficacy of

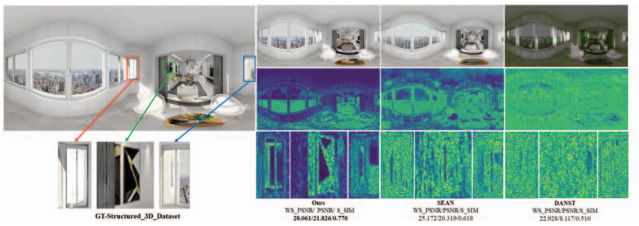

Figure 4. Qualitative comparisons of results (Structured3D data set) using SEAN [59], DANST [17], and ours, PanoStyle. We report PSNR and SSIM. We show the comparison between ground truth and predicted stylized images and color-mapped (viridis) L_2 distance between ground-truth and generated stylized images.

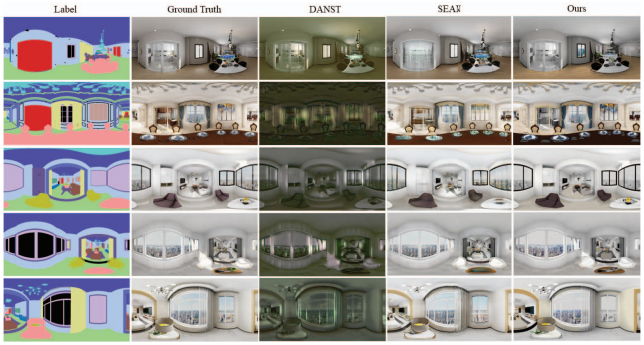

Figure 5. Visual comparison of the self-style semantic image synthesis results on the Structured3D data set. We compare SEAN [59], DANST [17], and our method, PanoStyle. For a fair comparison, the superresolution is applied to both our method and the competing approaches.

Figure 6. Visual comparison of SEAN [59] and our method on external image style transfer results using Structured3D data set. The style is applied to all components of the content image.

our proposed method due to its comprehensive ground truth annotations for indoor panoramas, including semantic, depth, normal, albedo, and RGB. The data set includes 41 region categories, such as wall, floor, cabinet, chair, sofa, and table, among others. To support the training and inference process, we developed several pre-processing scripts: a script to pre-compute shading images from the albedo and RGB frames, a script to extract label images from color-mapped semantic signals, and a clustering technique that groups similar images together based on their label association. This approach reduces the number of training iterations required and the inference time by

Figure 7. Visual comparison of SEAN [59] and our method on external image style transfer results using the Structured3D data set. The style is applied only to floor, wall, and ceiling.

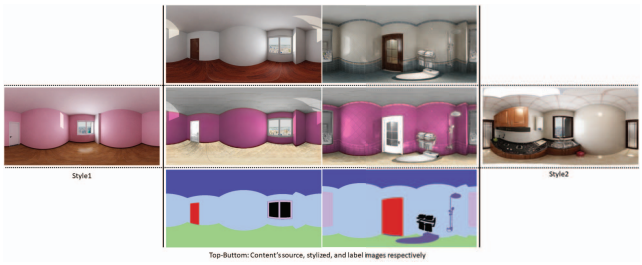

Figure 8. Style-crossover. Our method can select different styles from different reference images for different regions of an input image. Here, a scene is reconstructed by using two different style images (Style_1: wall and door, Style_2: floor and ceiling).

clustering images containing similar semantic classes (e.g., kitchen with kitchens, toilet with toilets, bedroom with bedrooms, living room with living-rooms, etc.).

Performance metrics. In order to assess the performance of our method, we utilize well-established metrics commonly employed in the field of computer vision and image processing to compare our results with state-of-the-art methods. These metrics include ArtFID [49], Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) [48]. By utilizing these well-established metrics, we can provide a comprehensive evaluation of the performance of our method and compare it to state-of-the-art approaches in a quantitative and objective manner.

Quantitative assessment. For a preliminary assessment of the validity of the approximated shading decomposition on the Structured3D data set, we performed a comparison between the reconstructed RGB images, $\hat{I}_{\text{rgb}}$, and the original ones, I_{rgb}, and we obtained an average structural similarity of 0.92 and ArtFID of 4.23, meaning that the reconstructed images are perceived very similar to the

original ones (see also Fig. 3). To ensure a fair comparison with our competitors, we have employed quantitative comparisons by assessing the reconstruction performance on uniformly selected samples that were used across all methods. It is important to emphasize that the results presented in Table 1 reflect a scenario in which super-resolution has not been applied to both our method and the competing approaches. This approach guarantees a more equitable comparison. Since SEAN is our main baseline, both our method and SEAN were trained under the same setup configuration, while other competitors were trained using their recommended training conditions. The metrics are reported in Tab. 1 (refer also to the video in the supplemental materials). As can be seen, our method outperforms current leading methods on all metrics used (SSIM, PSNR, and ArtFID). In our analysis, we have selected SEAN [59] as the current best state-of-the-art method, and DANST [16] as the second-best competitor method. We have also conducted a visual inspection of the results and found that ArtFID [49], a recently introduced quantitative metric for neural style transfer, is the most indicative of visual quality of the results. Based on our findings, we can confidently state that our framework has outperformed the state-of-the-art on all the reported metrics (PSNR, SSIM, and ArtFID).

Qualitative assessment. In this section, we present the qualitative results of our proposed method on selected samples from the Structured3D data set. We demonstrate the effectiveness of our method by showcasing the generated images using different style transfer techniques. Fig. 4 shows a colormapped comparison between methods, while Fig. 5 compares the RGB generated by application of self-style transfer. On the other side, Fig. 6 and 7 demonstrate style transfer results on the application of external scenes. Our method outperforms SEAN and DANST [17] in terms of visual quality, as shown in Fig. 4. We attribute the success of our method to its suitability for learning more of the variability in the data while preserving the global structure of the original content. We achieve this by incorporating strong geometry losses at the discriminator stage and by training the network using albedo instead of RGB. We refer the reader to the accompanying video for more results as visual inspection is critical for generative modeling. Additionally, our per-region style encoding enables new image editing operations on panoramas, such as iterative panoramic image editing with per-region style control (see also Figs. 1 and 7) as well as style crossover (see Fig. 8).

Ablation study. To evaluate the effectiveness of our design choices, we conducted an ablation study on different versions of our model (summarized in Tab. 1). Our analysis shows that all our variations improve results compared to our competitors across all utilized metrics. To validate this finding, we trained using the following scenarios. i) Shading decomposition without geometry constraint losses; ii) Geometry constraint losses without shading decomposition; iii) Shading decomposition and geometry constraint losses without depth consistency; iv) Shading decomposition and geometry constraint losses without layout consistency. The numerical results support the significance of both shading decomposition and geometry constraints in generating improved outcomes, emphasizing the importance of their combined incorporation.

5. Conclusions and future work

We have presented PanoStyle, the first semantic, geometry-aware, and shading-independent photo-realistic style transfer method for indoor panoramic scenes. Our approach addresses the limitations of existing models by incorporating strong geometry losses based on layout and depth inference to ensure shape consistency, and by applying a shading decomposition scheme that prevents color bleeding problems. Additionally, we employ super-resolution techniques to enhance image quality and capture fine details. Experimental evaluations on public domain synthetic data sets demonstrate that our proposed architecture outperforms state-of-the-art style transfer models. The visual results further validate the effectiveness of our method in producing realistic and visually pleasing images that maintain the semantic, global structure, layout, edge, and depth information of the input scenes. Our work paves the way for future research in style transfer for indoor panoramic scenes and opens up new opportunities for applications in the metaverse, virtual reality, real-estates, furniture retails, interior design, and architecture. Despite the promising results, our proposed system contains still limitations, that we plan to address in the future:

- **Network for extracting multiple signals from the same image:** Currently, we use the Structured3D synthetic data set which includes nearly all the necessary ground-truth signals required for training and evaluating our approach. We plan to investigate technologies for extracting the required signals from one single captured panorama [13];
- **Develop accurate shading models:** we plan to investigate networks able to infer different direct and indirect shading components, depending on reflectance, geometry content, and illumination information [57].

Acknowledgments This publication was made possible by NPRP-Standard (NPRP-S) 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility, of the authors.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 4432–4441, New York, NY, USA, 2019. IEEE/CVF. 2, 5

[2] Anil S. Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. ShadingNet: Image intrinsics by fine-grained shading decomposition. *Int. J. Comput. Vision*, 129(8):2445—-2473, aug 2021. 4

[3] Ming-Ming Cheng, Xiao-Chang Liu, Jie Wang, Shao-Ping Lu, Yu-Kun Lai, and Paul L Rosin. Structure-preserving neural style transfer. *IEEE Trans. Image Processing*, 29:909–920, 2019. 3

[4] Tai-Yin Chiu and Danna Gurari. PCA-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7844–7853, New York, NY, USA, June 2022. IEEE/CVF. 1

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, New York, NY, USA, 2018. IEEE/CVF. 3

[6] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11326–11336, New York, NY, USA, June 2022. IEEE/CVF. 1

[7] Zhicheng Fang, Xiaoran Chen, Yuhua Chen, and Luc Van Gool. Towards good practice for cnn-based monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1091–1100, New York, NY, USA, 2020. IEEE/CVF. 5

[8] Anna Frühstück, Ibraheem Alhashim, and Peter Wonka. TileGAN: Synthesis of large-scale non-homogeneous textures. *ACM Trans. Graphics*, 38(4):1–11, 2019. 2

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, June 2016. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 1–9, New York, NY, USA, 2014. Curran Associates, Inc. 2

[11] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision*, 128(4):835–854, 2020. 4

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30:1–12, 2017. 6

[13] Steven Hickson, Karthik Raveendran, and Irfan Essa. Sharing decoders: Network fission for multi-task pixel prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3771–3780, New York, NY, USA, 2022. IEEE/CVF. 8

[14] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3D scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6198–6208, New York, NY, USA, 2022. IEEE/CVF. 3

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 179–196, Cham, Switzerland, 2018. Springer. 3

[16] Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR*, 139:4487–4499, 2021. 8

[17] Eleftherios Ioannou and Steve Maddock. Depth-aware neural style transfer using instance normalization. In *Computer Graphics & Visual Computing (CGVC)*, pages 1–8, Eindhoven, The Netherlands, 2022. The Eurographics Association. 2, 3, 6, 7, 8

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, New York, NY, USA, 2017. IEEE/CVF. 3

[19] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style trans-

fer: A review. *IEEE Trans. Visualization and Computer Graphics*, 26(11):3365–3385, 2019. 3

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *14*th *European Conference on Computer Vision (ECCV)*, pages 694–711, Cham, Switzerland, 2016. Springer. 4

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, New York, NY, USA, 2019. IEEE/CVF. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6

[23] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4621–4629, June 2015. 3

[24] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2801–2810, New York, NY, USA, 2022. IEEE. 3

[25] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *CVF European Conference on Computer Vision (ECCV)*, pages 453–468, September 2018. 3

[26] Yuyan Li, Zhixin Yan, Ye Duan, and Liu Ren. Panodepth: A two-stage approach for monocular omnidirectional depth estimation. In *International Conference on 3D Vision (3DV)*, pages 648–658, New York, NY, USA, 2021. IEEE. 3

[27] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10551–10560, New York, NY, USA, 2019. IEEE/CVF. 3

[28] Xiao-Chang Liu, Ming-Ming Cheng, Yu-Kun Lai, and Paul L. Rosin. Depth-aware neural style transfer. In *Proceedings of the Symposium on Non-photorealistic Animation and Rendering*, pages 1–10, New York, NY, USA, 2017. ACM. 3

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 2794–2802, New York, NY, USA, 2017. IEEE/CVF. 2

[30] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? *International Conference on Machine Learning, PMLR*, 80:3481–3490, 2018. 2

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv Preprint, arXiv:1411.1784, 2014. 3

[32] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957, 2018. 3

[33] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. arXiv Preprint, arXiv:1802.05637, 2018. 3

[34] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, New York, NY, USA, 2019. IEEE/CVF. 2, 3, 4

[35] Sandip Paul, Bhuvan Jhamb, D. Mishra, and M. S. Kumar. Edge loss functions for deep-learning depth-map. *Machine Learning with Applications*, 7:#100218, 2022. 5

[36] Chi-Han Peng and Jiayao Zhang. High-resolution depth estimation for 360° panoramas through perspective and panoramic depth images registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3116–3125, New York, NY, USA, 2023. IEEE/CVF. 3

[37] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. SliceNet: Deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11536–11545, New York, NY, USA, 2021. IEEE/CVF. 3, 5

[38] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3762–3772, New York, NY, USA, 2022. IEEE/CVF. 3

[39] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos and spherical images. *International Journal of Computer Vision*, 126(11):1199–1219, 2018. 2

[40] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056, New York, NY, USA, June 2019. IEEE/CVF. 3, 5

[41] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2573–2582, New York, NY, USA, 2021. IEEE. 3

[42] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, Cham, Switzerland, 2018. Springer. 3

[43] M Tukur, G Pintore, E Gobbetti, J Schneider, and M Agus. Spider: Spherical indoor depth renderer. In *STAG: Smart Tools and Applications in Graphics*, pages 131–138, Eindhoven, The Netherlands, 2022. The Eurographics Association. 2

[44] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 462–471, New York, NY, USA, 2020. IEEE/CVF. 3

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, New York, NY, USA, 2018. IEEE/CVF. 3, 4

[46] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 1905–1914, New York, NY, USA, 2021. IEEE/CVF. 6

[47] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, pages 63–79, Cham, Switzerland, September 2018. Springer. 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 7

[49] Matthias Wright and Björn Ommer. ArtFID: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pages 560–576, Cham, Switzerland, 2022. Springer. 2, 7, 8

[50] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. In *European Conference on Computer Vision (ECCV)*, pages 327–342, 2020. 3

[51] Chen Xu, Yu Han, George Baciu, and Min Li. Fabric image recolorization based on intrinsic image decomposition. *Textile Research Journal*, 89(17):3617–3631, 2019. 4

[52] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324, New York, NY, USA, 2018. IEEE/CVF. 3

[53] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9036–9045, October 2019. 3

[54] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5505–5514, New York, NY, USA, 2018. IEEE/CVF. 2

[55] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. ACM. 1

[56] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *European Conference on Computer Vision (ECCV)*, pages 519–535, Cham, Switzerland, 2020. Springer. 2

[57] Tiancheng Zhi, Bowei Chen, Ivaylo Boyadzhiev, Sing Bing Kang, Martial Hebert, and Srinivasa G. Narasimhan. Semantically supervised appearance decomposition for virtual staging from a single panorama. *ACM Trans. Graph.*, 41(4), jul 2022. 1, 8

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference*

on Computer Vision (CVPR), pages 2223–2232, New York, NY, USA, 2017. IEEE/CVF. 3

[59] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5104–5113, New York, NY, USA, 2020. IEEE. 1, 2, 3, 4, 6, 7, 8

[60] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, Cham, Switzerland, 2018. Springer. 3