

Appendix to: PanoStyle

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
magus@hbku.edu.qa

Abstract

In this Appendix to "PanoStyle: Semantic, Geometry-Aware and Shading Independent Photorealistic Style Transfer for Indoor Panoramic Scenes", we provide additional details regarding our methodology. These include a detailed review of our network architecture, the loss functions involved, as well as additional details regarding the results and evaluation.

1. Additional Details: Methodology

1.1. Network Architecture

Our network architecture follows that of "SEAN" [10]. Unlike SEAN, however, we add a geometry loss and albedo disentanglement to the pipeline. Our full setup is discussed in the main paper.

Generator Our generator's design architecture, illustrated in Fig. 2, is based on residual blocks (Res-Blks) [1], which are a crucial component of Residual Neural Network (ResNet) architectures. They were designed to help address the problem of vanishing gradients in very deep neural networks by allowing gradients to flow directly through a shortcut connection. In a Residual Block, the input to the block is added to the output of one or more convolutional layers, allowing the network to learn the difference between the input and the desired output rather than having to learn the entire output from scratch. This makes it easier for the network to learn complex mappings and can improve the performance of deep neural networks. Our generator, similar to the one of He *et al.* [1], is composed of several Res-Blks, followed by a nearest neighbor upsampling layer. Notably, we only integrate the style codes (SC) into the first six Res-Blks, while the remaining inputs are integrated into all Res-Blks.

Discriminator The discriminator plays an essential role in adversarial training as it aims to distinguish between real

and generated images/depths. The architecture of our discriminator, as shown in Fig. 2, is based on the approach of Zhu *et al.* [10], called "SEAN", which employs two multi-scale discriminators with instance normalization (IN) [7] and Leaky ReLU (LReLU). IN helps in reducing internal covariate shift and improving the generalization of the network. Additionally, we incorporated LReLU activation functions, which prevent the vanishing gradient problem during backpropagation. To further stabilize the training, similar to Zhu *et al.* [10] and Park *et al.* [4], we applied spectral normalization [3] across all convolutional layers of the discriminator.

1.2. Other Loss Functions

The loss functions in this appendix have previously been used in SEAN [10]. To SEAN's setup, we add a geometry loss (discussed in the main paper).

Conditional adversarial loss Let E denote the style encoder, G the generator, and D_1, D_2 two discriminators at different scales [8]. Additionally, let R represent a particular style image, and M represent the corresponding segmentation mask for that image. To specify the conditional adversarial learning aspect of our loss function, we express it as follows.

$$\min_{E, G} \max_{D_1, D_2} \sum_{k=1}^2 L_{\text{GAN}}(E, G, D_k). \quad (1)$$

Specifically, L_{GAN} is built with the hinge loss [5], where $\mathbb{E}[\cdot]$ denotes the expectation operator:

$$L_{\text{GAN}} = \mathbb{E}[\max(0, 1 - D_k(R; M))] + \mathbb{E}[\max(0, 1 + D_k(G(SC, M), M))], \quad (2)$$

where SC is the style codes of R extracted by E under the guidance of M ,

$$SC = E(R, M) \quad (3)$$

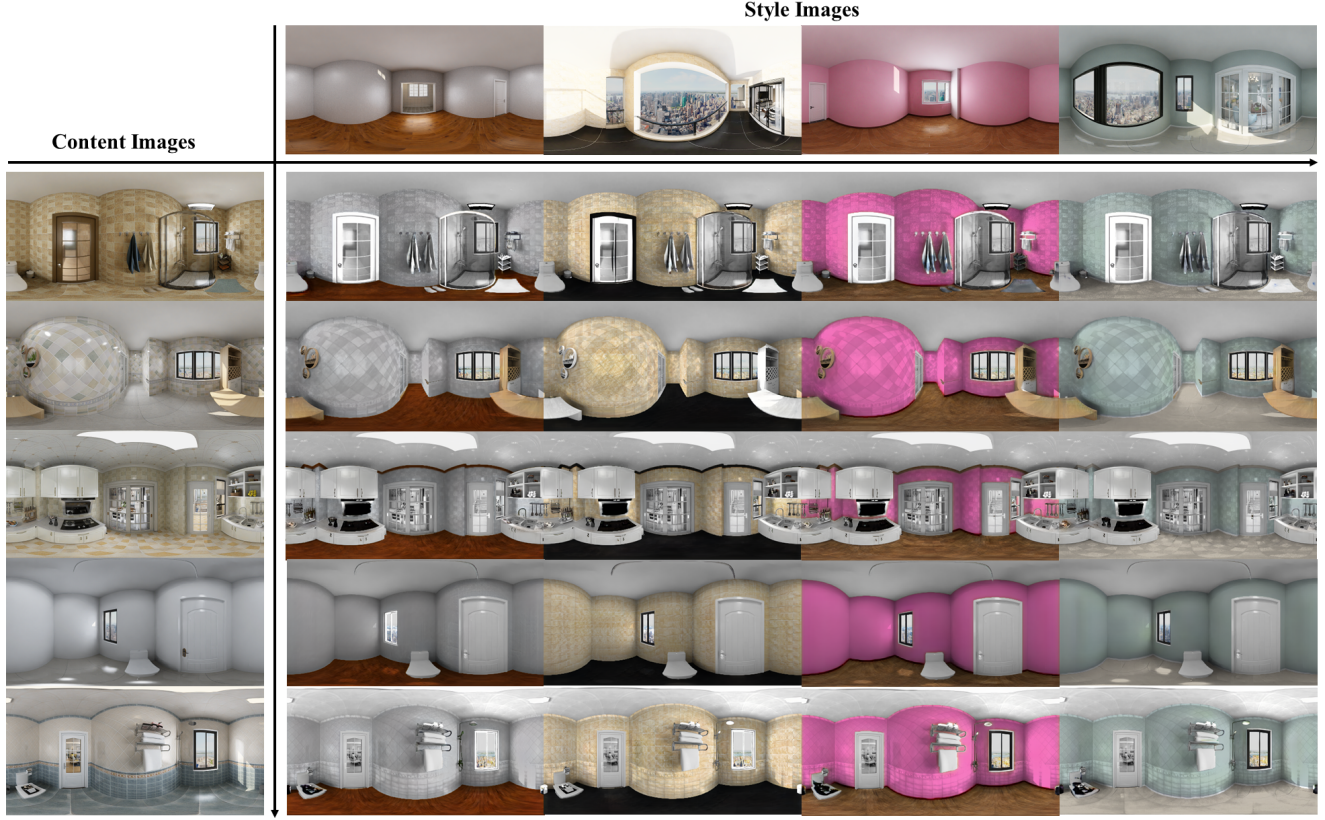


Figure 1. We propose the first geometry-aware and shading-independent, photorealistic and semantic style transfer method for indoor panoramic scenes. The first column displays the content images, while the remaining four columns represent different editing scenarios. The top of each column shows the image that provides new style information. The results of successive edits are presented in rows two to six. The four edits mainly focus on modifying the floor, wall, and ceiling, and in some cases, include other elements such as doors, windows, towels, and toilets, if the style to be applied is meaningful.

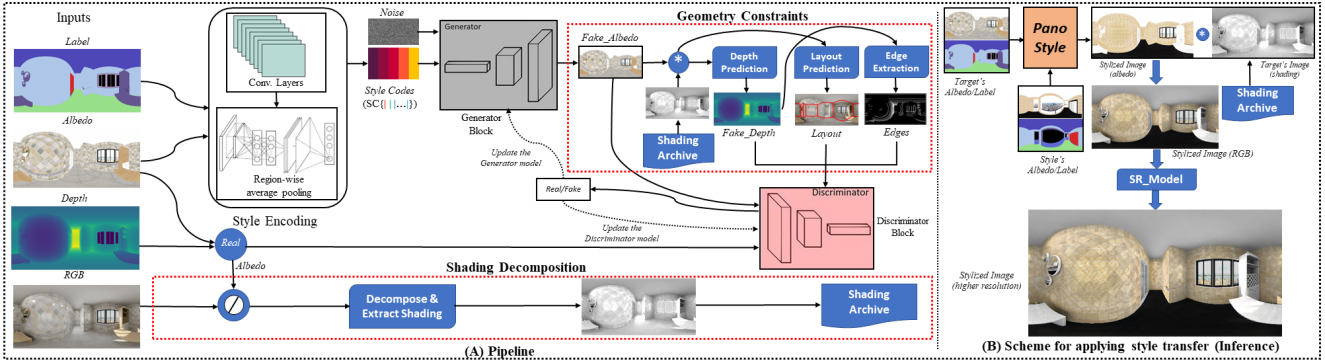


Figure 2. **PanoStyle framework.** A: the PanoStyle architecture is a generative adversarial model exploiting semantic-aware style encoding, integrated with a shading decomposition component and a set of geometry constraints in the discriminator. B: the scene generation process applies PanoStyle on target and style reflectance and semantic signals, computes the stylized RGB through the application of target shading signal, and applies super-resolution to increase details.



Figure 3. We visually compare external-style semantic image synthesis results on the Structured3D data set. We evaluate SEAN [10], DANST [2], and our own method (with and without all geometry constraints). The top row shows content images, while the bottom row displays style images applied in each column.

Feature matching loss Consider a discriminator D_k with a total of T layers. Let $D_k^{(i)}$ and N_i represent the output feature maps and the number of elements in the i^{th} layer of D_k , respectively. We use the term \mathcal{L}_{FM} to denote the feature matching loss,

$$\mathcal{L}_{\text{FM}} = \mathbb{E} \left[\sum_{i=1}^T \frac{1}{N_i} \left(\left\| D_k^{(i)}(R, M) - D_k^{(i)}(G(SC, M), M) \right\|_1 \right) \right]. \quad (4)$$

Perceptual loss Let N be the total number of layers used to calculate the perceptual loss. Let $F^{(i)}$ represent the output feature maps of the i^{th} layer of the VGG network [6], and let M_i denote the number of elements in $F^{(i)}$. The perceptual loss is given by

$$\mathcal{L}_{\text{percept}} = \mathbb{E} \left[\sum_{i=1}^N \frac{1}{M_i} \left\| F^{(i)}(R) - F^{(i)}(G(SC, M)) \right\|_1 \right]. \quad (5)$$

2. Additional Details: Results and Evaluation

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:

1. ArtFID [9]—a recently introduced quantitative evaluation metric for neural style transfer. It measures the distance between feature representations of real images and stylized images in a feature space of a pre-trained image classification model. The lower the ArtFID score, the closer the stylized image is to the real image.
2. Peak Signal-to-Noise Ratio (PSNR)—a well-known metric that measures the similarity between two images by computing the ratio between the maximum possible value of the signal and the noise. The higher the PSNR value, the greater the similarity between the two images.
3. Weighted Peak Signal-to-Noise Ratio (W-PSNR)—an extension of the PSNR metric that takes into account the spatial frequency distribution of the image. This

Table 1. Our research investigates the performance of various geometry losses in the context of the Structured3D data set. Through rigorous experimentation, we have evaluated five different geometry losses and observed that the Weighted Mean Squared Error (WMSE) consistently outperforms other loss functions across all the utilized metrics (PSNR, W-PSNR, SSIM, and ArtFID); hence, WMSE was adopted as the main loss function in this paper.

Geometry Loss Function	PSNR \uparrow	W-PSNR \uparrow	SSIM \uparrow	ArtFID \downarrow
Weighted MSE Loss	20.2289	26.6986	0.7922	7.9094
berHu Loss	19.8984	26.6324	0.7775	8.6273
Weighted berHu Loss	19.6020	26.6010	0.7772	8.8459
MSE Loss	19.5156	26.5769	0.7738	8.6749
Sobel Loss	19.5595	26.5059	0.7829	9.4896

Table 2. Original Versus Reconstructed RGB. In this table, we present the results of comparing the reconstructed RGB images \hat{I}_{rgb} with the original images I_{rgb} . We report the metrics including PSNR, W-PSNR, and SSIM. The obtained average structural similarity of 0.92 and ArtFID of 4.23 indicate that the reconstructed images closely resemble the original ones.

Data Set	PSNR \uparrow	W-PSNR \uparrow	SSIM \uparrow	ArtFID \downarrow
Structured3D	25.7097	28.3574	0.9203	4.2331

Table 3. Ablation study on the Structured3D data set. We report the metrics including PSNR, W-PSNR, SSIM, and ArtFID. Our method outperforms the state-of-the-art in all reported metrics (see also our video).

Method	PSNR \uparrow	W-PSNR \uparrow	SSIM \uparrow	ArtFID \downarrow
SEAN	18.9147	24.6233	0.6326	10.7989
DANST	10.1023	23.0925	0.5742	23.5553
Ours	20.2289	26.6986	0.7922	7.9094
Ours (w/o layout)	19.8277	26.4822	0.7857	8.9405
Ours (w/o depth)	19.5962	26.4607	0.7847	9.1599
Ours (w/o geom.)	19.3729	26.4582	0.7715	9.3261
Ours (w/o shad.)	18.9932	24.6819	0.6522	10.4410

metric assigns different weights to different spatial frequencies in order to account for the fact that some frequency components may be more important than others for a particular image.

4. Structural Similarity (SSIM)—a metric that compares the structural similarities between two images by measuring the differences in luminance, contrast, and structure. The higher the SSIM value, the greater the similarity between the two images in terms of their structural features.

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.

2.1. Quantitative and qualitative Assessment

One of the key design choices in our approach is how to incorporate different geometry loss functions at the discriminator stage. To address this, we compared five variants for encoding geometry losses: Weighted-MSE, berHu, weighted-berHu, MSE, and Sobel losses. MSE and berHu

losses correspond to mean-square error and inverse Huber losses, respectively. To ensure a consistent and fair comparison, we conducted a quantitative evaluation on uniformly selected samples across all methods. Our findings indicate that the Weighted MSE loss outperforms the other four variants, leading us to adopt it as the primary loss function in this paper (refer also to Table 1). We also visually confirmed that the Weighted-MSE loss function generally results in better visual quality, especially for challenging inputs. The berHu loss and weighted berHu loss closely follow in terms of performance, ranking as the second and third best, respectively. Furthermore, the MSE loss performs slightly better than the Sobel loss in terms of W-PSNR and ArtFID, although it exhibits slightly lower performance than the Sobel loss in terms of PSNR and SSIM. Importantly, all five variants outperform other state-of-the-art competitors across all the utilized metrics (SSIM, PSNR, W-PSNR, and ArtFID). Therefore, we further conclude that the incorporation of both geometry losses, constraints, and shading decomposition significantly impacts our network’s performance.

Moreover, in order to preliminarily evaluate the ef-



Figure 4. Visual comparison of SEAN [10] and our method on the external image style transfer results using Structured3D data set. Note: Here, the style is applied to all components of the content image. Also note that lack of color in regions is due to oversaturation of albedo in the input data.

effectiveness of our approximated shading decomposition scheme on the Structured3D data set, we compare the reconstructed RGB images \hat{I}_{rgb} generated using our scheme with the original images I_{rgb} from the data set. Table 2 presents the results of this comparison, including metrics

such as PSNR, W-PSNR, and SSIM. The average scores obtained, particularly for the structural similarity (0.92) and ArtFID (4.23) metrics, indicate a close resemblance between the reconstructed images and the originals.

Furthermore, in Fig. 1, we present an extensive example

that demonstrates the efficacy of our style transfer approach (PanoStyle). This showcases the application of various style images to a specific target image, highlighting the versatility and effectiveness of our method. In Fig. 4, we present an extensive visual comparison of SEAN [10] and our method for external image style transfer using the Structured3D data set. In contrast, Fig. 3 provides a comprehensive visual comparison of external-style semantic image synthesis results, including not only SEAN and our method, but also DANST [2]. In addition, our method is evaluated with and without all geometry constraints. The visual results in both figures unequivocally demonstrate that our approach surpasses state-of-the-art competitors by generating realistic and visually appealing images. This is further supported by the numerical results demonstrated in Table 3, which includes not only PSNR, SSIM, and ArtFID metrics but also the W-PSNR metric. These results affirm the effectiveness and robustness of our approach.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, New York, NY, USA, 2016. IEEE/CVF. 1
- [2] 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. 3, 6
- [3] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957, 2018. 1
- [4] 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. 1
- [5] L. Rosasco, E. D. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004. 1
- [6] K. Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR)*, pages 1–14, Online, 2015. OpenReview. aXiv Preprint, arXiv:1409.1556. 3
- [7] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016. 1
- [8] 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. 1
- [9] 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. 3
- [10] 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, 3, 5, 6