SuperStyleNet: Deep Image Synthesis with Superpixel Based Style Encoder

Jonghyun Kim[1]
jhkim.ben@gmail.com
Gen Li[2]
li.gen@ed.ac.uk
Cheolkon Jung†[3]
zhengzk@xidian.edu.cn
Joongkyu Kim†[1]
jkkim@skku.edu

[1] Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, South Korea

[2] School of Informatics,
University of Edinburgh,
Edinburgh, UK

[3] School of Electronic Engineering,
Xidian University,
Xian, Shaanxi, China

Abstract

Existing methods for image synthesis utilized a style encoder based on stacks of convolutions and pooling layers to generate style codes from input images. However, the encoded vectors do not necessarily contain local information of the corresponding images since small-scale objects are tended to "wash away" through such downscaling procedures. In this paper, we propose deep image synthesis with superpixel based style encoder, named as SuperStyleNet. First, we directly extract the style codes from the original image based on superpixels to consider local objects. Second, we recover spatial relationships in vectorized style codes based on graphical analysis. Thus, the proposed network achieves high-quality image synthesis by mapping the style codes into semantic labels. Experimental results show that the proposed method outperforms state-of-the-art ones in terms of visual quality and quantitative measurements. Furthermore, we achieve elaborate spatial style editing by adjusting style codes. The codes are available at: https://github.com/BenjaminJonghyun/SuperStyleNet

1 Introduction

The goal of conditional image synthesis is to generate photo-realistic images conditioning on certain input data. Recent methods utilized neural networks to generate realistic images from other images, latent codes, edge maps, or pose key points [2, 5, 18, 20, 30, 38, 44, 46]. Especially, our interest lies in using semantic layouts to assist conditional image synthesis [6, 20, 33, 44], which generates photo-realistic images from semantic masks. In addition, it can manipulate semantic information in images, such as context generation and image editing, by controlling the segmentation masks. Although these methods generate high-quality images, it cannot freely control image styles since the segmentation masks are only fed to their networks. To achieve it, SPADE [32] and SelectionGAN [37] utilized a style image to extract style information by using an encoder network. Then, a decoder network conducts style reconstruction in semantic layouts by referring to the style information, and generates a synthesized image similar to the style image. However, the style information is characterized

† Corresponding authors

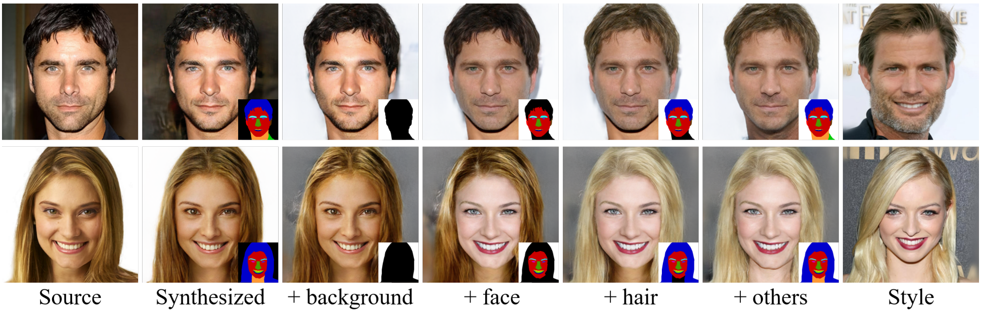

Figure 1: Face image synthesis by the proposed network. The synthesized images are generated with segmentation masks of source images and their style information. In other manipulated images, the style is controlled by replacing style codes in given segmentation masks from source to style images.

to represent large-scale objects since downscaling by convolutions and pooling layers tends to gradually "wash away" small-scale objects in feature maps. It indicates that the generator shows a bias toward large-scale objects rather than smaller ones during the learning process. Thus, it is difficult to conduct the style reconstruction of the local information. To equally contain local and global information, current methods [39, 50] modified the existing encoder or decoder to reconstruct local information from downscaled feature maps or encoded style vectors. However, undesirable effects appear in the results by reconstructing local information from high-level features (i.e., downscaled feature maps and encoded style vectors). In this case, there exists little style information of the small-scale objects in the high-level features. Thereby, it is difficult to expect that the local information can be well reconstructed when the decoder conducts style mapping in the local ones.

To tackle this issue, a straightforward strategy is to encode the style information from the original image while maintaining the original scale. However, it brings a large amount of parameters in the style encoder. Therefore, we propose Superpixel based Parameter-free Style Encoding (SPSE) to directly encode the original image into the style vector. We apply SLIC superpixel segmentation [1] to each segmentation mask for generating the style vector per semantic region. It can provide a different view from the existing style-encoding methods [22, 28, 34]. Furthermore, we analyze hidden representation of each node between nearest pixels to inject their spatial relationships to the style vector with a Graphical Self-Attention Strategy (GSAS) because compressing an image to the style vector based on superpixel wipes the spatial relationships off. Consequently, the proposed network is capable of synthesizing a high-quality image by considering both local and global information.

We provide extensive experiments to prove the effectiveness of the proposed method on challenging datasets: CelebAMask-HQ [26], Cityscapes [9] and CMP Facades [40]. We evaluate the performance of the proposed network in terms of various metrics. Compared with existing methods, our contributions can be summarized as follows:

- We propose superpixel based parameter-free style encoding, called as SPSE, to regularly encode local and global style information of the input image into the style vector with a parameter-free operation.

- We provide GSAS to compensate for loss of the spatial relationships between neighbor pixels of the encoded style vector. This strategy uses graph based the self-attention method to define them.

- The proposed network achieves better performance in image synthesis than state-of-the-art ones in terms of visual quality and quantitative measurements.

2 Related Work

Image-to-Image Translation is to learn a parametric mapping from input to output. Isola *et al.* [20] first proposed image-to-image translation with a conditional generative adversarial network (GAN) to translate the source to target domains. This task has been widely extended by many researchers. Huang *et al.* [17] and Alharbi *et al.* [2] proposed multi-modal image-to-image translation in an unsupervised way. Moreover, various methods [3, 7, 8] were proposed to conduct multi-domain translation. As a specific task of the image-to-image translation, semantic image synthesis [4, 13, 26, 32, 36, 50] has become much popular in terms of unconstrained image control. To be specific, SPADE [32] proposed spatially-adaptive denormalization to preserve semantic information of output images similar to input semantic layouts. Furthermore, SEAN [50] improved SPADE by regionally normalizing parameters to control the style of each semantic region individually. Although these methods showed outstanding performance on semantic image synthesis, there exists shortcomings in the process of style information encoding. The repeated downscalings with convolution layers in the style encoding erase features of local information. To overcome this shortcoming, the proposed method equally encodes local and global information into the style vector based on superpixels.

Style Encoding is essential in image synthesis to extract style information from reference images. Existing methods [12, 21] utilized a VGG [35] network pre-trained on image classification [10] to obtain style features. Unlike these methods, AdaIN [16] inferred affine parameters from the style image to map input parameters into the style space, which enabled arbitrary style transfer in real-time. On account of its easy application and outstanding performance, this method was adopted in various tasks [17, 22, 25]. Furthermore, Park *et al.* [32] and Zhu *et al.* [50] improved AdaIN to achieve semantic manipulation by normalizing the affine parameters in spatial or semantic regions, but neural networks are still indispensable for the style encoding. Compared with the aforementioned methods, the proposed method provides a new perspective in terms of non-parametric style encoding.

3 Proposed Method

We aim to regularly extract local and global style information per semantic mask from the reference image. The encoded style is scattered into its corresponding masks to synthesize photo-realistic images. To implement this concept, we first introduce the superpixel based parameter-free style encoding (SPSE) and the graphical self-attention strategy (GSAS). Then, we describe the overall network architecture of a superpixel based style encoding network, named as SuperStyleNet.

3.1 Superpixel based Parameter-free Style Encoding

A superpixel can be defined as a group of pixels that share similar visual characteristics. SLIC [1] is a representative superpixel method, which clusters pixels based on color similarity and proximity. To be specific, the clustering procedure is conducted in a five-dimensional space $[l\ a\ b\ x\ y]$, where $[l\ a\ b]$ is the pixel color vector in CIELAB color space and $[x\ y]$ is the pixel position. During the clustering, each pixel in the input image is assigned to k superpixels by calculating the $l2$ distance between each initial cluster center and its neighborhood.

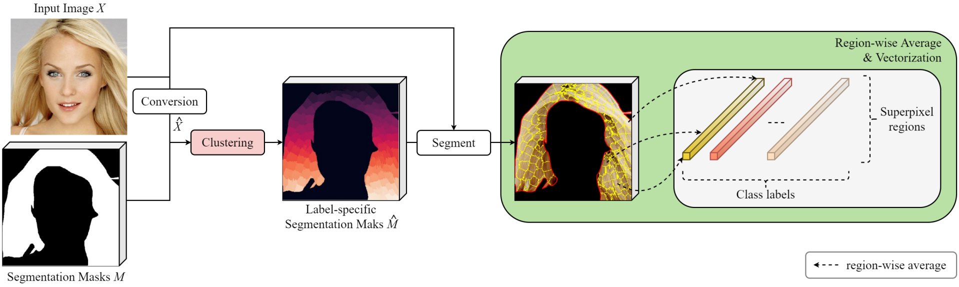

Figure 2: Illustration of the proposed Superpixel based Parameter-free Style Encoding (SPSE). To extract style codes of a specific semantic mask, we convert the input image into the five-dimensional space and cluster it in the semantic mask into superpixels. Thereafter, pixel values in each superpixel are averaged to obtain a style code.

Inspired by this method, we propose superpixel based parameter-free style encoding (SPSE) as illustrated in Fig. 2, which allocates an input image to k desired superpixels, and converts them to the style code in the color space.

Let $M \in \mathbb{I}^{H\times W\times L}$ be a semantic segmentation mask given the number of class labels L, where $\mathbb{I}$ is a set of integers consisting of either "1" or "0", and $H \times W$ is the image size. Before clustering the RGB input image $X \in \mathbb{R}^{H\times W\times 3}$, we convert it into the five-dimensional space $\hat{X} \in \mathbb{R}^{H\times W\times 5}$, and superpixel centers are initialized with an uniform distribution [19]. After that, we repeat the clustering in each segmentation mask by SLIC [1] to obtain superpixels per semantic label:

$$SP = \text{Clustering}(\hat{X}[M == 1]), \tag{1}$$

where $SP \in \mathbb{R}^{L\times K}$ is a set of superpixels in each class label with the number of superpixel centers K. Then, each superpixel region is converted as label-specific segmentation masks $\hat{M} \in \mathbb{I}^{H\times W\times L\times K}$. To obtain the style information from the original image X, we segment pixels from the input image referring to $\hat{M}$, and take average on pixel values per label-specific semantic region as follows:

$$SC = \mathbb{E}[X[\hat{M} == 1]], \tag{2}$$

where $SC \in \mathbb{R}^{L\times K\times 3}$ is style codes. Then, it is reshaped to $L \times 3K$, and interpolated with the desired length N of the style code in axis "1".

3.2 Graphical Self-Attention Strategy

Graph-based self-attention was proposed in [42], which computes the hidden representation of each node in graphs and is used in natural language processing (NLP) tasks. Due to its effective nodal analysis and attention mechanism, this method was adapted to interpret spatial representations, i.e., point cloud semantic segmentation [43], medical image segmentation [31], human pose estimation [49], and trajectory forecasting [24].

Similarly, we propose a graphical self-attention strategy (GSAS), as described in Fig. 3, to inject spatial hidden representations to the style vector. Let $\mathbf{a}_i = \{a_1, a_2, ..., a_N\}$ be the style vector given a specific class label, where N is the length of the style vector. In order to convert the style vector to the spatial domain, the style vector $\mathbf{a}$ is expanded as a $N \times N$ matrix $\mathbf{a}_{ij}$, where $i, j \in \{1, 2, ..., N\}$. After that, $\mathbf{a}_{ij}$ is concatenated with its transposed matrix $\mathbf{a}_{ij}^T$, then a 1×1 convolution is used to analyze relationships between style components in

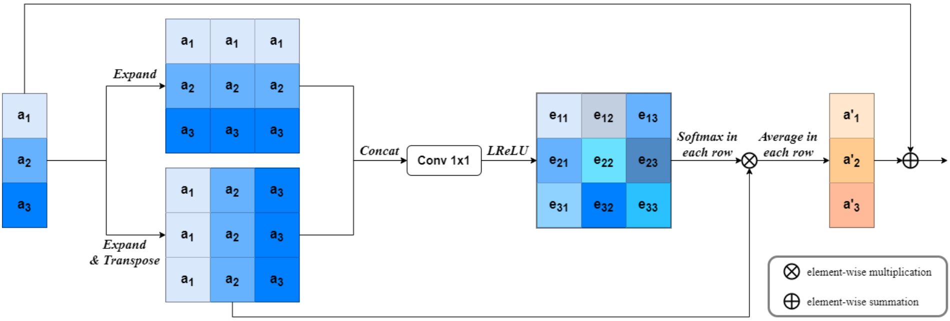

Figure 3: Flow of the proposed graphical self-attention strategy. For better understanding, we exemplify the style vector is set to be the length 3 given a specific class label. "Concat" and "LReLU" are notated as the concatenation and LeakyReLU operation, respectively.

the style vector and reduce the channel size to 1. Thereafter, we apply the LeakyReLU nonlinearity with $\alpha = 0.2$. Thus, correlation coefficients can be described as:

$$\mathbf{e}_{ij} = \text{LeakyReLU}(\mathbf{W}([\mathbf{a}_{ij} \parallel \mathbf{a}_{ij}^T])), \tag{3}$$

where $\mathbf{W}$ is a 1×1 convolution, and $\parallel$ represents concatenation operation. Following the Eq. 3, we obtain a correlation coefficient between i and j-th style components. To convert the coefficients to comparable scores across different style components, we normalize them across all j using the softmax function as follows:

$$\mathbf{s}_{ij} = \frac{exp(\mathbf{e}_{ij})}{\sum_{k \in N} exp(\mathbf{e}_{ik})}. \tag{4}$$

Then, we conduct pixel-wise multiplication between $\mathbf{a}_{ij}^T$ and $\mathbf{s}_{ij}$ and averaging across all j to obtain a new style vector $\mathbf{a}_i'$ considered all $i \neq j$-th style components. Before feeding the style vector to the generator, the element-wise summation is operated to aggregate both style vectors $\mathbf{a}_i$ and $\mathbf{a}_i'$.

3.3 Network Structure and Learning Objective

In this paper, we concentrate on semantic image synthesis and spatial style editing. To achieve them, we follow network structures in [32, 50] but exclude the style encoder as illustrated in Fig. 4, which employs several ResNet blocks [14] with upsampling layers. In each ResNet block, the affine parameters are learned by embedding SEAN [50] to scatter the style vector into the corresponding semantic layouts. Furthermore, we use the semantic layouts as the input of the generator, and feed the concatenation of the layouts and the synthesized images into the multi-scale discriminator [44].

With the help of SPSE, the embedding of the style encoder is not required since SPSE utilizes superpixels in the color domain to generate the style vector with non-parametric operation. Thus, the generator and discriminator excluding the style encoder are only considered in the training process. As the loss functions for the generator, we use perceptual loss [21], feature matching loss [44], and conditional adversarial loss [50]. For the discriminator, the hinge loss term [29, 47] is adopted. We provide further details of the network structure and objectives in the supplemental materials.

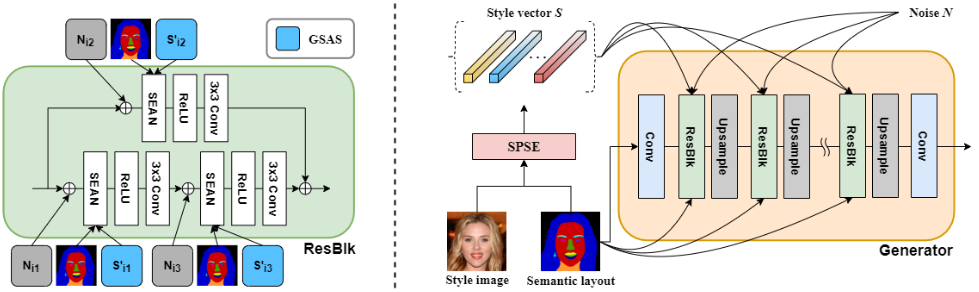

Figure 4: Illustration of the proposed SuperStyleNet. (*left*) Structure of a residual block embedded in SuperStyleNet. (*right*) Overview of SuperStyleNet, which contains a series of the residual blocks with 2× upsampling (nearest) layers.

4 Experiments

4.1 Implementation Details

In our experiments, following the experimental conditions in [50], the style length N is set to 512, and the class labels are decided by the number of class in datasets. We apply spectral normalization [29] to both the generator and discriminator, and a synchronized version of batch normalization is applied to SEAN layers of the residual block. Following a two time-scale update rule (TTUR) [15], the learning rates are set to 0.0001 and 0.0004 for the generator and discriminator respectively. Adam optimizer [23] is adopted with $\beta_1 = 0$ and $\beta_2 = 0.999$.

We train and test SuperStyleNet on the segmentation datasets: CelebAMask-HQ [26], Cityscapes [9] and CMP Facades [40]. 1) CelebAMask-HQ is a face image dataset containing 30,000 face images with 19 segmentation labels, which is split into 28,000 and 2000 images for train and test sets, respectively. 2) Cityscapes contains 3500 images annotated with 35 segmentation labels. For this dataset, the train and test set sizes are 3,000 and 500, respectively. 3) CMP Facades consists of 500 facade images with 12 segmentation labels. In this dataset, 400 and 100 images are utilized as train and test sets, respectively. All images are resized to 256×256 in both training and testing.

4.2 Evaluation Metrics

Following previous works [32, 44, 50], we perform semantic segmentation on synthesized images to quantify how well the predicted segments match ground-truth. Specifically, BiSeNet [45] is applied to the synthesized images to infer semantic segmentation results, and pixel-wise accuracy (pix acc) and mean intersection-over-union (mIoU) are utilized as its evaluation metrics. Furthermore, we compare SuperStyleNet with these state-of-the-art methods by adopting peak signal-to-noise ratio (PSNR), normalized root mean square error (NRMSE), Fréchet Inception Distance (FID) [15], and learned perceptual image patch similarity (LPIPS) [48].

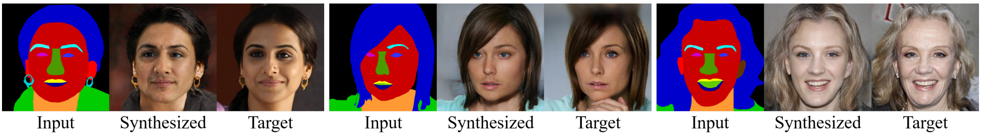

Figure 5: Failure cases of image synthesis. The first and second examples represent gender mismatch, while the last one is for age mismatch.

Table 1: Quantitative comparison on semantic segmentation and generation performance. Higher mIoU, higher pixel acc, and lower FID indicate better performance.

Method	CelebAMask-HQ			Cityscapes			CMP Facades		
	mIoU	pix acc	FID	mIoU	pix acc	FID	mIoU	pix acc	FID
Ground Truth	70.05	36.75	8.71	50.46	91.50	34.20	36.75	71.28	75.07
Pix2PixHD	73.15	95.22	27.45	49.21	91.18	104.39	40.83	72.19	156.91
SPADE	**74.55**	**95.72**	33.94	**55.02**	92.93	**51.18**	42.38	**75.48**	124.96
SEAN	72.63	95.28	**22.41**	52.52	92.53	52.62	41.41	74.93	127.38
Ours	73.89	**95.72**	25.49	53.37	**93.01**	60.45	**43.59**	75.45	**119.82**

Table 2: Quantitative comparison on reconstruction performance and perceptual similarity. Higher PSNR, lower NRMSE, and lower LPIPS indicate better performance.

Method	CelebAMask-HQ			Cityscapes			CMP Facades		
	PSNR	NRMSE	LPIPS	PSNR	NRMSE	LPIPS	PSNR	NRMSE	LPIPS
Pix2PixHD	15.78	0.451	0.359	16.25	0.447	0.393	12.23	0.507	0.441
SPADE	14.35	0.475	0.373	17.35	0.458	0.380	12.93	0.575	0.420
SEAN	**18.54**	**0.248**	0.274	20.08	0.348	**0.331**	14.36	0.459	0.402
Ours	18.22	0.263	**0.255**	**20.83**	**0.325**	0.349	**15.26**	**0.408**	**0.392**

4.3 Comparisons with State-of-the-art Methods

Quantitative comparisons. We quantitatively compare SuperStyleNet with the state-of-the-art ones on semantic segmentation, generation, reconstruction performance, and perceptual similarity. First, we select Pix2PixHD [44], SPADE [32], and SEAN [50] as current state-of-the-art methods for comparisons. In Tables 1, it is obvious that SuperStyleNet generally outperforms these state-of-the-art ones in maintaining segmentation masks. It indicates that all objects including local ones are successfully reconstructed and recognized as these correct classes. However, SuperStyleNet performs slightly worse than SEAN in CelebAMask-HQ in terms of reconstruction metrics, while SuperStyleNet shows better performance in both Cityscapes and CMP Facades as shown in Table 2. This is because SuperStyleNet infers personal characteristics (i.e., genders and ages) from segmentation masks due to the lack of the style encoding network. Therefore, SuperStyleNet yields failure cases as shown in Fig. 5, which causes degradation of the reconstruction performance. Overall, these results indicate that our SuperStyleNet structure is effective in preserving segmentation masks while generating high-fidelity synthesized images.

Qualitative comparisons. To validate the effectiveness of the proposed method in terms of visual quality, we compare SuperStyleNet with the aforementioned state-of-the-art methods. As shown in Fig. 6, SuperStyleNet generates high-quality synthesized images on all datasets compared with other methods. Concretely, SuperStyleNet is effective in reconstructing occluded local areas by glasses, hands, or characters on CelebAMask-HQ due to the help of SPSE. Furthermore, small-scale objects, such as human, car, and traffic lights, are well synthesized in Cityscapes. Despite injecting style information, SEAN suffers from color distortion when training on the small-scale dataset like CMP facades. In contrast, SuperStyleNet accurately recovers style information in each semantic layouts matched with its ground-truth. This is because SPSE directly utilizes color values in the source images as style information for style mapping. Overall, SuperStyleNet achieves better preservation of semantic layouts while generating realistic synthesized images compared with the state-of-the-art methods. More comparisons and synthesized images are shown in the supplementary material.

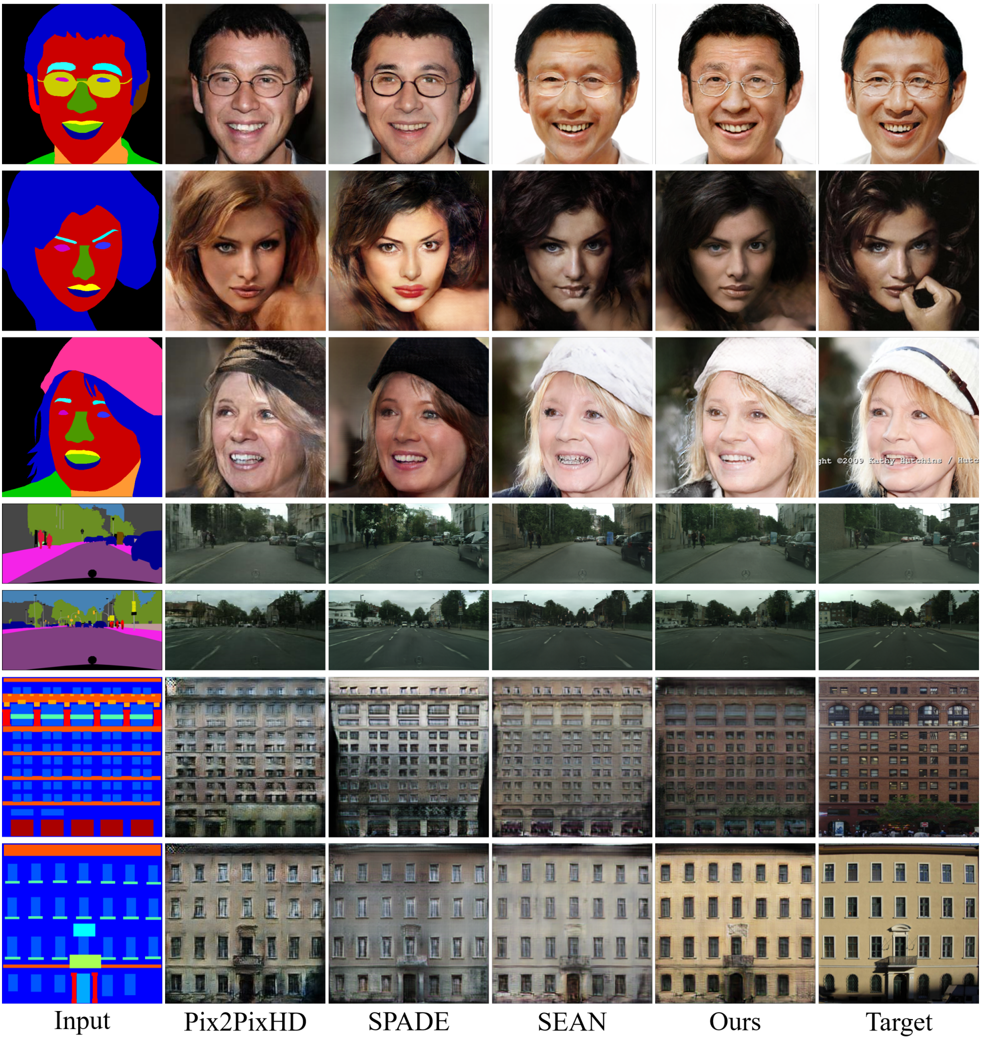

Figure 6: Qualitative comparison of semantic image synthesis with state-of-the-art methods on three datasets.

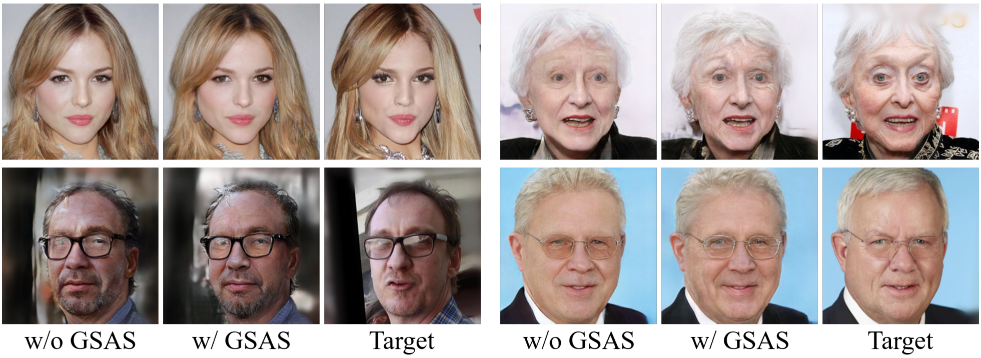

Figure 7: Effects of GSAS on the visual quality.

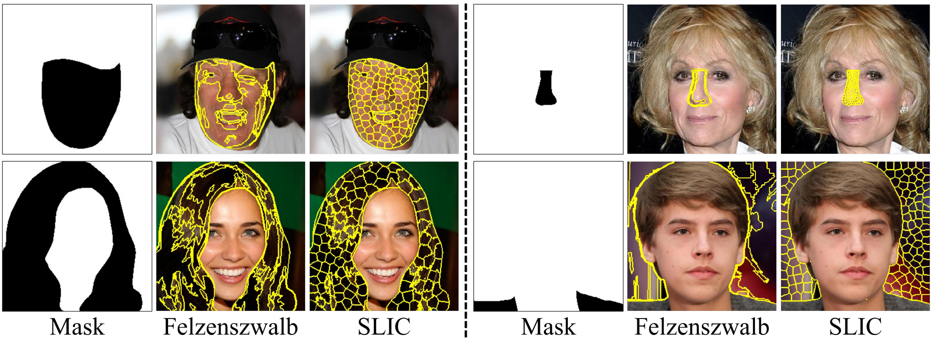

Figure 8: Visualization of superpixel segmentation using Felzenszwalb *et al.* [11] and SLIC [1]. Both methods cluster pixels in given semantic layouts into superpixels.

4.4 Ablation Study

Effectiveness of SuperStyleNet. For validating the effectiveness of SuperStyleNet, we adopt generator and discriminator of SEAN without a style encoding network as baseline that is regarded as the state-of-the-art model. SPSE substitutes for the style encoding network to extract style information from source images. It can be observed that the segmentation performance is improved from 72.63/95.28 to 73.91/95.58 in mIoU and pixel-wise accuracy on CelebAMask-HQ. Furthermore, its non-parametric operation allows SuperStyleNet to reduce 1.6M parameters while the style encoder increases the inference time around 3s/image due to the CPU processing. In addition, we provide quantitative and qualitative analyses to explore the effects of GSAS. When GSAS is embedded in our network, the reconstruction performance is improved from 18.15/0.285 to 18.22/0.263 in both PSNR and NRMSE on CelebAMask-HQ, while maintaining the segmentation performance. Furthermore, the visual quality is also enhanced as shown in Fig. 7. Specifically, artifacts and details are recovered in the synthesized images due to its inference of spatial representations.

Variations of SPSE. SPSE utilizes a superpixel algorithm to encode an input image into style vectors, thus it is mainly affected by the superpixel algorithm and its variations.

1) *Selection of the superpixel algorithm*: We consider that the number of superpixels should be controllable to obtain the same length of style codes extracted from the same categories in different images. However, conventional methods [11, 41] cannot control the number of superpixels. It incurs information imbalance of style codes in each object since style information is decided by the number of superpixels. Moreover, these methods irregularly encode pixels into superpixels as shown in Fig. 8. On the other hand, SLIC [1] and LSC [27] not only initialize superpixel centers with uniform distributions but also control the number of superpixels. Thereby, we adopt SLIC as the superpixel segmentation algorithm for SPSE, which shows the faster processing than LSC.

2) *The number of superpixel centers*: This factor determines the amount of information to represent encoded objects. It indicates that the larger number of superpixel centers k is capable of more containing details of objects and diverse color information. Therefore, the generator is able to better reconstruct synthesized images when k is larger. To fully encode the local ones into k-vectors, we set k to the maximum value that does not exceed the number of pixels on the smallest object across datasets. Thus, we empirically set k to 128. To validate it, we conduct experiments on CelebAMask-HQ by changing the parameter k. When k is reduced from 128 to 32, PSNR and mIoU decrease from 18.22 to 17.97 and 73.89 to 69.19. Furthermore, both performances are still lower although we increase k to 64.

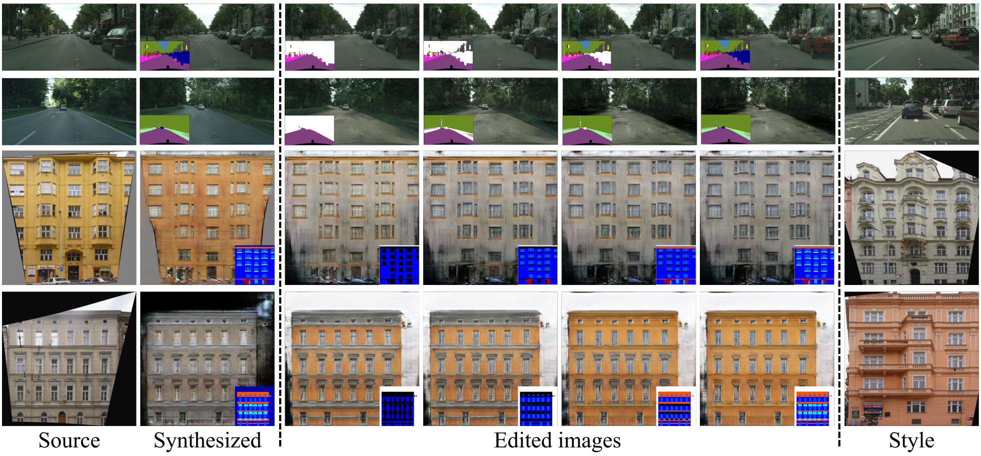

Figure 9: Image editing per semantic region. The style vectors are replaced from source to style images on given segmentation masks.

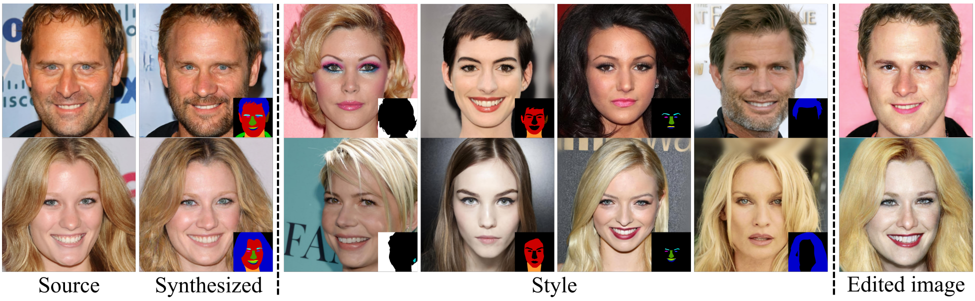

Figure 10: Style mixing with multiple style images on given segmentation masks.

4.5 Style Mixing

SuperStyleNet is capable of editing images per semantic region or mixing styles of multiple images. To achieve both image manipulations, we change input style vectors from source to style images on given segmentation masks while other vectors retain source ones as shown in 1 and 9. Furthermore, we mix multiple styles so as not to overlap semantic classes as illustrated in 10. As can be seen from the figures, the edited images maintain textures and structures of the source images while changing styles by referring to the style images with given segmentation masks.

5 Conclusion

We propose superpixel based parameter-free style encoding (SPSE) for image synthesis to evenly extract local and global style codes from the original image. To be specific, SPSE clusters pixels in a given image to yield superpixels as style codes per semantic region. Then, the graphical self-attention strategy (GSAS) interprets hidden representations between superpixels to capture spatial relationships. Consequently, SPSE and GSAS facilitate our network to generate high-quality synthesized images matched with target images while preserving semantic layouts. Furthermore, it benefits from reconstruction in occluded regions and small-scale objects. However, the proposed SuperStyleNet yields failure cases in the inference of personal characteristics. In our future work, we will investigate this problem.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1012159).

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] Yazeed Alharbi, Neil Smith, and Peter Wonka. Latent filter scaling for multimodal unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1458–1466, 2019.

[3] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2018.

[4] David Bau, Hendrik Strobelt, William Peebles, Bolei Zhou, Jun-Yan Zhu, Antonio Torralba, et al. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017.

[7] 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 Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[11] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[13] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] 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. *arXiv preprint arXiv:1706.08500*, 2017.

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[17] 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 172–189, 2018.

[18] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020.

[19] Benjamin Irving. maskslic: regional superpixel generation with application to local pathology characterisation in medical images. *arXiv preprint arXiv:1606.09518*, 2016.

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

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[22] 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, 2019.

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

[24] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *arXiv preprint arXiv:1907.03395*, 2019.

[25] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4422–4431, 2019.

[26] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[27] Zhengqin Li and Jiansheng Chen. Superpixel segmentation using linear spectral clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1356–1363, 2015.

[28] 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*, pages 10551–10560, 2019.

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

[30] Sanghyeon Na, Seungjoo Yoo, and Jaegul Choo. Miso: Mutual information loss with stochastic style representations for multimodal image-to-image translation. *arXiv preprint arXiv:1902.03938*, 2019.

[31] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[32] 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*, pages 2337–2346, 2019.

[33] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018.

[34] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020.

[37] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019.

[38] Hao Tang, Song Bai, Philip HS Torr, and Nicu Sebe. Bipartite graph reasoning gans for person image generation. *arXiv preprint arXiv:2008.04381*, 2020.

[39] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020.

[40] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *Proc. GCPR*, Saarbrucken, Germany, 2013.

[41] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *European conference on computer vision*, pages 705–718. Springer, 2008.

[42] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[43] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10296–10305, 2019.

[44] 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 Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[45] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *arXiv preprint arXiv:2004.02147*, 2020.

[46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5907–5915, 2017.

[47] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[49] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[50] 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*, pages 5104–5113, 2020.