Semantic Image Synthesis with Photorealistic Image Stylization

Phan Pham Thanh Tuyen pptttuyen18@apcs.vn
Nguyen Thao Ninh ntninh18@apcs.vn
Truong Thuy Quyen ttquyen18@apcs.vn
Nguyen Trung Hau nthau18@apcs.vn
Phan Ho Nguyen Bao phnbao18@apcs.vn
University of Science - Viet Nam National University Ho Chi Minh City
Advanced Program in Computer Science

Abstract

Generating photorealistic images from semantic layouts is the purpose of semantic image synthesis. Spatially-Adaptive Normalization (SPADE, also known as GauGAN) is a recent approach on this task, which aims to convert an input segmentation mask to a photorealistic image while performing the affine transformation in the normalization layers. SPADE can either better preserve semantic information against common normalization layers, or produce results with much better visual quality and fewer visible artifacts, especially for diverse scenes in the COCO-Stuff and ADE20K datasets. However, the number of styles that SPADE is capable of generating is limited to only one for each image. To improve, we present the combined method of SPADE and Photorealistic Image Stylization (PIS) which is created from a closed-form solution. PIS can transfer the style of a reference photo to a content photo with the constraint that the stylized photo should remain photorealistic. As a consequence, the combination of the two methods allows users to quickly create a set of diverse style images given a semantic layout and desirable style image.

1. Introduction

Most human beings can inherently visualize things based upon description, while it is a challenge to reproduce this ability using technology - something like a machine that can interpret natural language to image or even translate image to image. Toward issues of interpreting natural language to an image [35, 28], this challenge involves some problems including natural language processing, image processing or fusion mechanism (for fusing natural language and image). Image to image translation also has analogous issues consisting of approaches to render image (RGB, gradient, edge map, semantic label map, etc), the resolution of an image, styles after translating, etc. Image synthesis is a process of generating new images from some form of description, which is, in other words, a kind of image to image translation. In this paper, we concentrate on image synthesis with segmentation input data.

One of the famous methods for deep generative models to learn to synthesize images is Generative Adversarial Nets (GANs) [6] which has become one of the most popular research areas. One of the main researches is conditional GANs (cGANs). The task of cGANs is to generate images that have certain conditions or attributes. Multiple valuable frameworks exist in many different forms based on the type of input data. In this work, we focus on segmentation input to synthesize images. Although this is not the first time this problem is encountered [3, 11, 19, 31], Spatially-Adaptive Normalization (SPADE) [27] is initially a semantic image synthesis model that can produce photorealistic outputs for diverse scenes including indoor, outdoor, landscape, and street scenes.

The previous normalization layers [12, 31] tend to leave out information contained in the input semantic masks. SPADE, which provides conditionally normalization layer, is proposed to address this problem. This layer modulates the activation using input semantic layouts through a spatially-adaptive, learned transformation and can effectively propagate the semantic information throughout the network. After deeper inspection on their results, we found that SPADE uses only one style code to manipulate the entire style of an image hence it generates only one style for each image.

The most recently proposed method [42] uses semantic region-adaptive normalization (SEAN) that can deal with the style problem of SPADE through semantic regions. However, the method requires more time to operate than the combination of SPADE and another stylization method. Moreover, changing the general style of an image costs a lot of effort. To overcome this problem, we present Photorealistic Image Stylization (PIS), which uses a closed-form solution that can transfer the style of a reference photo to a content photo with the constraint that the stylized photo should remain photorealistic. The crucial characteristic of PIS is that it is much faster and more preferred by human subjects as compared to those by the competing methods

[5, 21].

We experiment PIS on challenging datasets including the COCO-Stuff [2] and the ADE20K [38, 37]. The results show that the combination of SPADE and PIS has better performance as compared to the previous approach. Besides, our method has simple implementation since it only needs a segmentation of image and desired style image to create a new stylized image. Depending on different style images, our method can generate various photorealistic stylization outputs.

The main contributions of our work are:

- We introduce a fast method, SPADE with PIS, to change the general style after synthesizing images. The method helps generate diverse photorealistic stylization outputs just from the segmentation of images.
- We validate the method on several datasets such as COCO-Stuff, ADE20K to show its effective results.

2. Related Work

Deep generative modeling is a method to synthesize images. Generative adversarial networks (GANs)[6] and Variational [15] Autoencoder (VAE) are two well-known methods in the field using this model. SPADE is built on GANs but focuses more on the conditional image synthesis task. Conditional image synthesis has been investigated in many researches with diverse input data type such as image synthesis given category labels [1, 22, 23, 24, 26] or images generated from texts [8, 28, 32, 35]. Another form to which many researchers appertain, is image-to-image translation where both input and output are images [9, 11, 14, 40, 41]. Learning-based methods run faster during test time and produce more realistic results than non-parametric methods [4, 7, 13] as confirmed by authors of SPADE [27], therefore Spatially-Adaptive Normalization mainly focuses on converting segmentation masks into photorealistic images. Unconditional normalization layers plays an important role in modern deep networks [10, 30]. In SPADE the normalization layer is labelled as unconditional as they are independent of external data, in contrast to the conditional normalization layers.

Photorealistic Image Stylization (PIS) is another image-to-image translation method[11, 31, 20, 29]. Its main task is to learn to translate an image from one domain to another. PIS does not need a training dataset of content and style images for learning the translation function, and can be considered as a special kind of image-to-image translation. One of the closest method to PIS is the work of Yijun Li [17]. This method consists of 2 main steps: stylizing the content image with an image carrying style and flattening them so that the content and style are complement.

Our model's principal task is to combine two processes in order to provide less human work. SPADE uses only one style for all images whereas PIS requires a real content image and a style image to stylize them. To address this problem, we assume that the output of SPADE is the content image to continue stylizing it in PIS process. Our model uses the pretrained models so that it can save training time, datasets and work as well as these tasks.

3. Spatially-Adaptive Normalization with Photorealistic Image Stylization

Our method consists of two steps as illustrated in Figure 1. Our model inputs consist of a semantic label image and a image carrying style . After briefly reviewing SPADE and PIS methods, we will discuss our combination method.

3.1. Spatially-Adaptive Normalization

With the SPADE [27], there is no need to feed the segmentation map to the first layer of the generator, since the learned modulation parameters have encoded enough information about the label layout. The segmentation mask in the SPADE Generator is fed through spatially adaptive modulation without normalization. Only activations from the previous layer are normalized. Hence, the SPADE generator can preserve better semantic information. Specifically, while normalization layers such as the InstanceNorm [30] are essential pieces in almost all the state-of-the-art conditional image synthesis models [31], they tend to leave out semantic information when applied to a uniform or flat segmentation masks. On the other hand, SPADE enjoys the benefit of normalization without losing the semantic input information.

3.2. Photorealistic Image Stylization

The algorithm of PIS consists of two steps stylization and photorealistic smoothing via two-step mapping function:

$$\mathcal{F}_2(\mathcal{F}_1(\mathcal{I}_C, \mathcal{I}_S), \mathcal{I}_C) \tag{1}$$

where $\mathcal{I}_S$ and $\mathcal{I}_C$ are the style photo and the content photo, respectively; $\mathcal{F}_1$ and $\mathcal{F}_2$ are the first step in stylization transform and the second step in photorealistic smoothing.

In the stylization step, the algorithm is based on Whitening and Coloring Transforms (WCT) [16]. The key idea behind the WCT is to directly match feature correlations of the content image to those of the style image via the two projections. However, WCT generates structural artifacts (e.g., distortions on object boundaries) for photorealistic image stylization which lead to the approach of PhotoWCT. Due to the loss of spatial information of WST and the inspiration by the success of unpooling layer [36, 34, 25] in preserving spatial information, the PhotoWCT replaces the upsampling layers in the WCT with unpooling layers. The formulation of PhotoWCT is shown below:

$$\mathcal{Y} = \mathcal{F}_1(\mathcal{I}_C, \mathcal{I}_S) \tag{2}$$

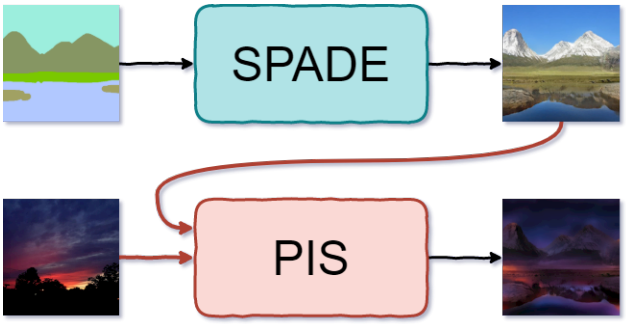

Figure 1. Our combined method of SPADE and PIS. The semantic image is synthesized through SPADE creating the content image. The content image is stylized by PIS with the image carrying style.

The PhotoWCT-stylized result still barely looks photorealistic since semantically similar regions are often stylized inconsistently. Due to this problem, the next step is to employ the pixel affinities in the content photo to smooth the PhotoWCT-stylized result. Motivated by the graph-based ranking algorithms [39, 33], the smoothing step can be written as a function mapping with closed-form solution given by:

$$\mathcal{R}^* = \mathcal{F}_2(\mathcal{Y}, \mathcal{I}_C) \tag{3}$$

where the optimal solution $\mathcal{R}$ is the smoothed version of Y based on the pairwise pixel affinities, which encourages consistent stylization within semantically similar regions; $\mathcal{Y}$ is the PhotoWCT-stylized result.

3.3. The combination of SPADE and PIS

Despite great work in image synthesis, SPADE uses only one style code to manipulate the entire style of an image, which is not sufficient for high-quality synthesis or detailed control. Owing to this drawback, we present a simple source code to implement consecutively the two methods. We execute both SPADE and PIS on the pre-trained model by writing a short script to use segmentation masks as input data. After finishing the SPADE's process, we use its results as content images. PIS will generate a photorealistic stylized image by receiving the synthetic image and the desired style image. The final results show the smooth stylized images which preserve both semantic and style information. Our method helps generate diverse photorealistic stylization outputs just from the segmentation of images. SPADE with PIS is also a simple method to change the general style after synthesizing images.

4. Experiments

4.1. Implementation details

We use Python to write three scripts which invoke SPADE, PIS and SPADE + PIS. The SPADE script implements SPADE and saves the results in the input folder of PIS. The PIS script gets the directory of style image, converts type and size of both semantic image and style image. Finally, the SPADE + PIS script invokes consecutively SPADE and PIS. Since SPADE requires NVIDIA DGX1 machine with 8 V100 GPUs to generate semantic images, we use Google Colab with GPU runtime type. To synthesize a photorealistic image our model spends time in each station mentioned in the table ... as for statistical evidence.

4.2. Datasets

We conduct experiments on the following datasets.

- COCO-Stuff [2] is derived from the COCO dataset [18]. It has 118,000 training images and 5,000 validation images captured from diverse scenes. It has 182 semantic classes. The experiment on this dataset is shown in Figure 2.

- ADE20K [37, 38] consists of 20,210 training and 2,000 validation images. Similarly to the COCO, the dataset contains challenging scenes with 150 semantic classes. The experiment on this dataset is shown in Figure 3.

4.3. Research and discussion

We achieved results on the two datasets COCO-Stuff and ADE20K shown in Figure 2 and Figure 3.

Overall, a content image is generated by SPADE from a label map and is then applied with a desired style from the user. Going into detail of Figure 1, the semantic labels describing a scene of mountains and rivers are converted into realistic details at SPADE stage. The result image is then proceeded to the next stage - PIS. At this stage the global appearances of the output image can be controlled by choosing an external style carrying image. As observed from the figure, the overall mood of the content photo is altered such that it agrees with the given style - changing from blue sky and white mountains into dark sky and black mountains.

At first glance, our model obtains visually pleasing results. However, a closer look reveals that the generated output contain noticeable artifacts, e.g., in the sixth case in Figure 3, the model fails at SPADE stage where the bridge is mistaken as a pile of soil. There are also cases where the model fails to transfer the style due to mismatch between style image and content image. This can be fixed from user's side by choosing a style image that has components similar to those in the content image.

5. Conclusion

In this paper we propose an approach for generating stylized photorealistic images given a label set as input. It

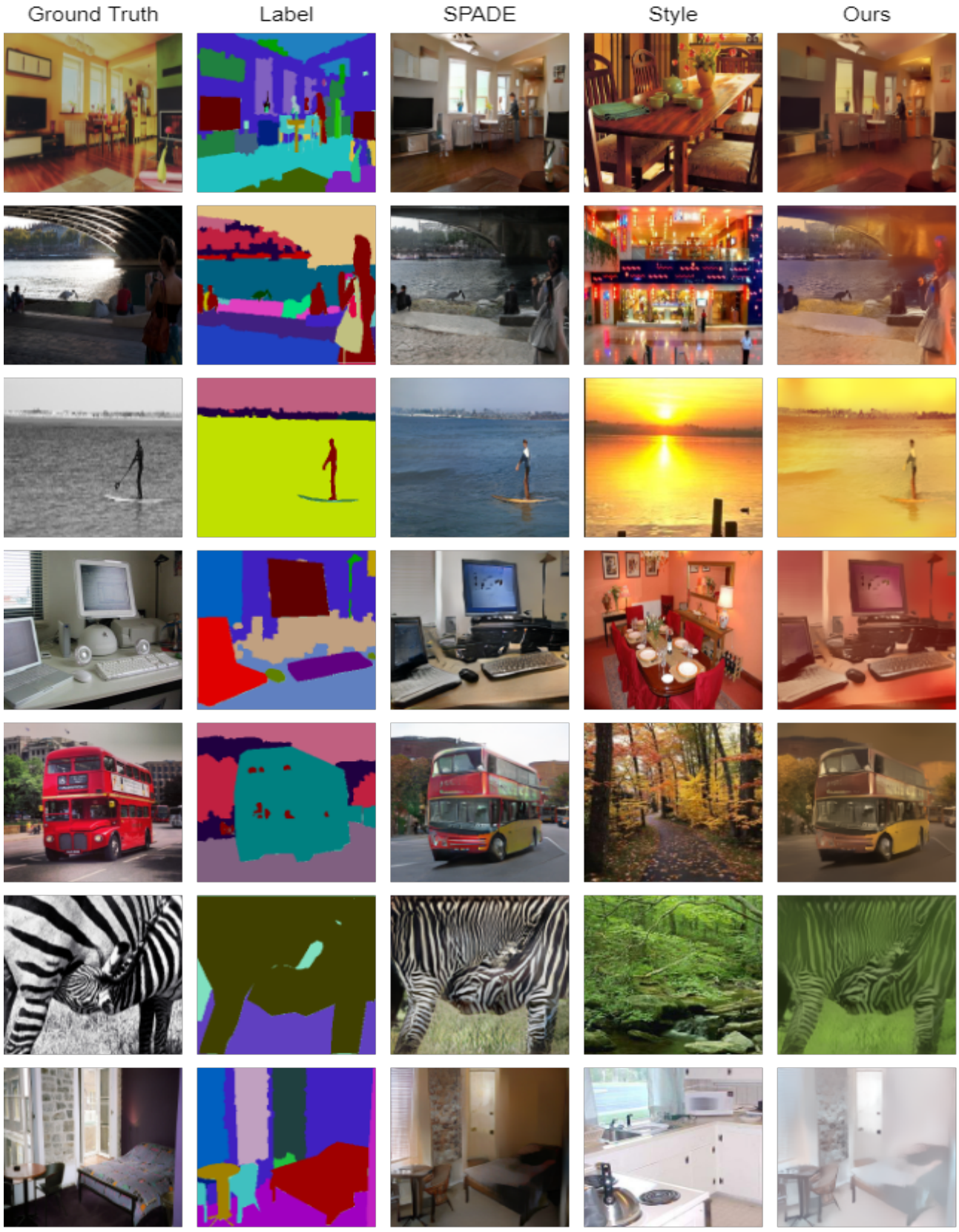

Figure 2. The SPADE + PIS outputs, which were tested on COCO-Stuff dataset, are generated by the combination of SPADE results and arbitrary style images.

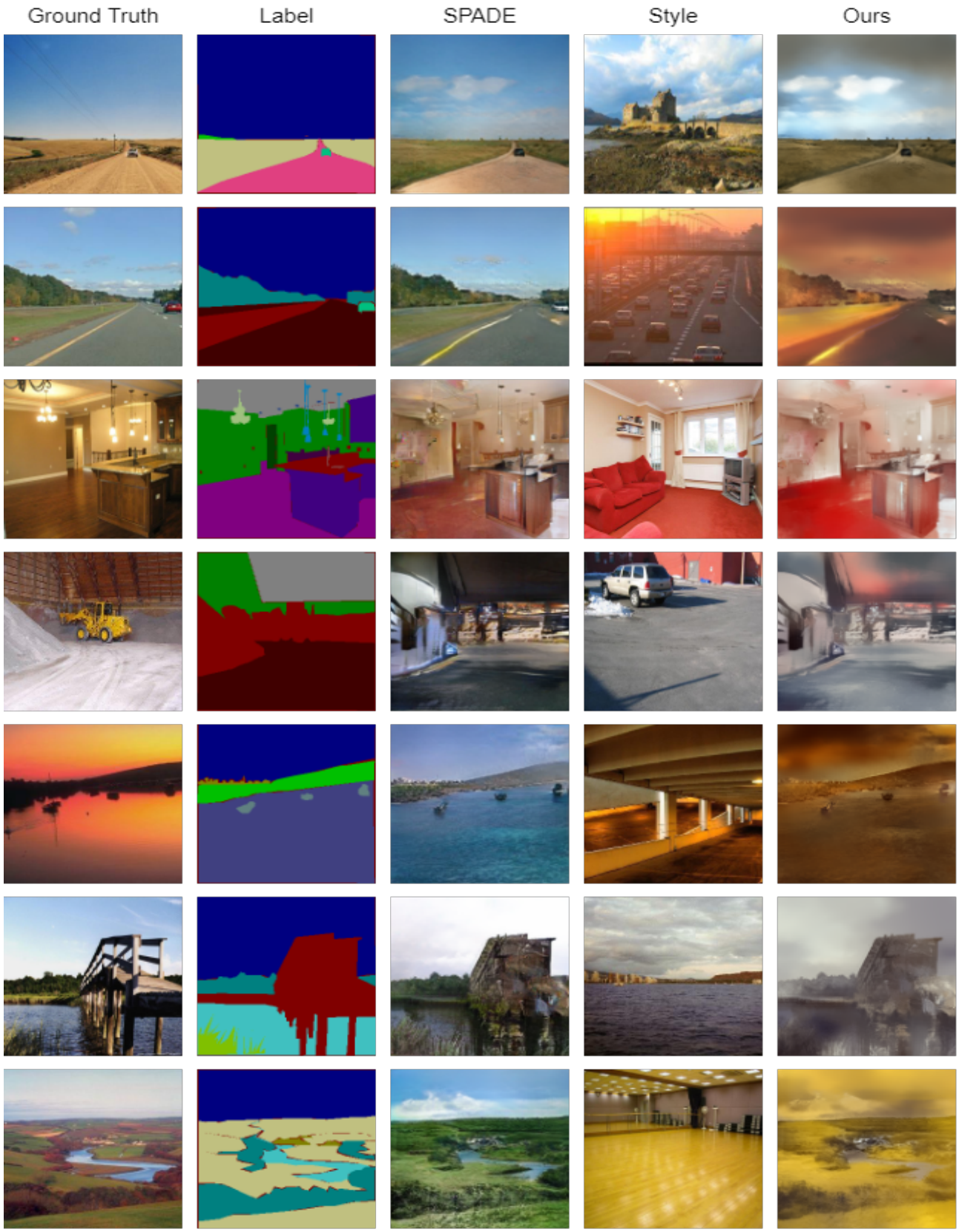

Figure 3. The SPADE + PIS outputs, which were tested on ADE20K dataset, are generated by the combination of SPADE results and arbitrary style images.

comprises of two stages: a synthesis step using Spatially-Adaptive Normalization (SPADE) and a stylization step using Photorealistic Image Stylization (PIS). By combining the two stated methods, we present a small step towards a quick interactive tool for users to generate realistic results with desired atmostphere.

References

[1] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[2] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 2, 3

[3] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 1

[4] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: Internet image montage. *ACM transactions on graphics (TOG)*, 28(5):1–10, 2009. 2

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[7] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 2

[8] S. Hong, D. Yang, J. Choi, and H. Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018. 2

[9] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 2

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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. 1, 2

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[13] M. Johnson, G. J. Brostow, J. Shotton, O. Arandjelovic, V. Kwatra, and R. Cipolla. Semantic photo synthesis. In *Computer Graphics Forum*, volume 25, pages 407–413. Wiley Online Library, 2006. 2

[14] L. Karacan, Z. Akata, A. Erdem, and E. Erdem. Manipulating attributes of natural scenes via hallucination. *arXiv preprint arXiv:1808.07413*, 2018. 2

[15] D. P. Kingma and M. Welling. Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, 2014. 2

[16] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, pages 386–396, 2017. 2

[17] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018. 2

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[19] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 1

[20] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 2

[21] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017. 2

[22] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018. 2

[23] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[24] T. Miyato and M. Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 2

[25] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[26] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017. 2

[27] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1, 2

[28] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 1, 2

[29] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 2

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. 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. 1, 2

[32] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. 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*, pages 1316–1324, 2018. 2

[33] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 3

[34] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[35] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 1, 2

[36] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2015. 2

[37] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016. 2, 3

[38] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[39] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *Advances in neural information processing systems*, pages 169–176, 2004. 3

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[41] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 2

[42] P. Zhu, R. Abdal, Y. Qin, and P. Wonka. Sean: Image synthesis with semantic region-adaptive normalization, 2019. 1