Style-Guided Inference of Transformer for High-resolution Image Synthesis

Jonghwa Yim   Minjae Kim

Vision AI Lab., AI Center, NCSOFT

South Korea

{jonghwayim, minjaekim}@ncsoft.com

Abstract

Transformer is eminently suitable for auto-regressive image synthesis which predicts discrete value from the past values recursively to make up full image. Especially, combined with vector quantised latent representation, the state-of-the-art auto-regressive transformer displays realistic high-resolution images. However, sampling the latent code from discrete probability distribution makes the output unpredictable. Therefore, it requires to generate lots of diverse samples to acquire desired outputs. To alleviate the process of generating lots of samples repetitively, in this article, we propose to take a desired output, a style image, as an additional condition without re-training the transformer. To this end, our method transfers the style to a probability constraint to re-balance the prior, thereby specifying the target distribution instead of the original prior. Thus, generated samples from the re-balanced prior have similar styles to reference style. In practice, we can choose either an image or a category of images as an additional condition. In our qualitative assessment, we show that styles of majority of outputs are similar to the input style.

1. Introduction

Image generation is a task that learns a mapping from a prior density to real image distribution. Recent works of deep generative models developed for this task derive from two major branches; one is Generative Adversarial Nets (GAN) [11] which is a pioneering research of implicit density models; the other is explicit density models which include Variational Auto-Encoder (VAE) [17] and Autoregressive generative models. Until recently, studies derive from GAN and VAE show distinguished achievement in image quality, however, they yet have a couple of limitations; first, they have a difficulty achieving high resolution and satisfactory output quality due to inherent limitations in either loss design or structurally restricted receptive field; second, the training process is relatively unstable.

To improve the problems of GAN and VAE based meth-

ods, another direction of research heads towards more direct formulation. To formulate tractable density, Pixel-RNN [45] proposes an auto-regressive architecture to compute the global relationship explicitly and model the image components recursively. Thanks to the advantage in viewing global context, this research stably produces locally and globally coherent results. Following research, Conditional PixelCNN [31], introduces a condition in representing the target distribution and displays qualitatively coherent results to given labels. Despite solving the two major problems in GAN and VAE, its computation complexity is proportional to the square of the output size, hampering extending the size of an architecture and output resolution.

Recently, a transformer architecture from previous research [46] provides clue for the above scalability problem in a tractable density model. By replacing recursive computation for long range relationships to transformer with masked self-attention module, the study achieves both of full receptive field and computational efficiency. Thus, following researches start adopting transformer instead of the recursive architecture. The research [34] introduces transformer in image generation task and shows comparable or even better performance than recursive neural network (RNN) based architectures with a better computational efficiency. However, the limitation of these methods lies in insufficient quality compared to GAN-based methods since modeling low-level vision, pixel values in most cases, is qualitatively inaccurate compared to GAN-based methods.

Instead of predicting low-level vision, predicting quantised latent representation becomes a substitute. The first of this appears in VAE approaches. The previous researches [44], [36] introduce vector quantisation to latent representation in VAE, called Vector Quantised-Variational AutoEncoder (VQ-VAE), thereby the latent prior becomes discrete and categorical. Inspired by this study, the recent study [10] adopts VQ-VAE to represent latent representation while the transformer infers quantized code instead of pixel value. Hence, the output of the transformer is a prior, probability density over discrete codebook. After the transformer infers the prior, sampling a code from the prior prob-

ability creates a sequence of codes which is deterministically translated to an image. Because a quantised latent representation relieves the transformer from modeling the complex low-level vision, instead it can focus more on learning the global context. With the advantage of efficient role allocation, this study displays high-resolution and natural-looking outputs. However, what would be generated from prior sampling is unknown until it generates and decodes the full sequence of quantised codes, which is much more time consuming as the target resolution is getting higher.

To resolve the unpredictable behavior of sampling from prior density, our proposal confers controllability on the above research. We propose a novel method to guide prior sampling by adopting a style image as a conditional input. To this end, we reformulate conditional generative model to posterior expression where conditional input is a given observation. Therefore, our method guides the inference of pre-trained transformer using a conditional input without re-training the model. Our method inherits the merits of the previous transformer-based generative approach, creating high-resolution photo-realistic output images. Besides, our method accepts conditional input image and generates outputs of similar styles to the given condition.

Our method can work both on a non-conditional and conditional pre-trained transformer, and its output is visually well persuasive, especially in landscape scenes where we can feed a semantic label map as a condition. In the evaluation section, we show experiments of a single image and multiple images as input styles. In the case of multiple images, we perform sampling and averaging the distribution of latent features over the entire images.

2. Related Works

GAN-based Image Synthesis. Over the last years, the field of image synthesis has been massively researched. Since the debut of Generative Adversarial Network [11], this field has progressed to higher resolution and better quality than ever before. However, the original GAN holds a few drawbacks; the training process is unstable, enlarging the resolution is limited by the architectural design and machine to run the model, and the output quality is not yet that of the real world. Recently, a few cornerstone researches break the limitations and enable to synthesize the real-world scene. The pioneering research, PG-GAN [15], suggests GAN generating higher resolution by progressively stacking more layers and consequently enlarging the resolution. Meanwhile, to alleviate the instability in training process, we can find various former researches including Wasserstain GAN [2], PGGAN [15], SN-GAN [25], techniques for training GANs [38], LSGAN [22], and BE-GAN [3]. Improving the architecture is another branch of research; DCGAN [35] embraces convolutional architecture, SA-GAN [52] adds self-attention blocks to its back-

bone, and BigGAN [4] composes many previously-existing modules such as self-attention, spectral normalization, and a discriminator with projection [26]. Lastly, StyleGAN [16] is a foundation in its way. It proposes disentangled latent priors from mapping network instead of random noise for attribute control, stochastic variation, and better output quality.

Conditional Generative Model. Conditional image generation refers to the task that generates images conditioning on a particular input data. For example, conditioning on a class label [24], [27], [4], [23], [26] allows the generated samples to be in the class. Another case of conditional generative models [37], [54], [13], [49] is based on text to synthesize images. Recently, many researches [47], [20], [32], [9], [43], [55], [56], [57], [42], [41] use semantic segmentation map to indicate what to draw. SPADE [32] suggests a spaital-varying normalization layer and becomes a significant backbone for the following researches. CC-FPSE [20] applies spatially-varying convolutional kernels which are based on the semantic label map to fully utilize semantic layout information to generate high-quality and semantically high-fidelity images. SEAN [56] displays per-semantic styling by applying region-adaptive normalization. CLADE [42] proposes class-adaptive normalization layer. INADE [41] presents an instance-adaptive modulation sampling and improves diversity. Our work also extends the series of conditional generative model, except that it suggests guided inference to follow a given condition instead of architecturally accepting conditional input with additional parameters and requiring full training process.

Autoregressive Generative Model. Autoregressive model formulates a tractable density function as a generative approach. Due to its simplicity and directness, it is broadly used in image and audio generation. In image generation, autoregressive deep generative model predicts the probability over discrete pixel value regressively using the architectures called PixelRNN and PixelCNN [45], [31], [6]. In audio generation, WaveNet [30] introduces Pixel-CNN architecture to the 1D audio data by applying dilated causal convolution layer, followed by many researches [48], [1], [39], [29] that improve the quality of synthesis. Afterwards, a transformer architecture [46] is newly introduced in autoregressive generative method [34] replacing the former RNN based architecture. Transformer becomes prevalent and also appears in GAN based approaches [52], [14]. Recently, the state-of-the-art study, Taming Transformer [10], adopts vector quantisation introduced in VQ-VAE [44], [36] and predicts probability density over the entries of a quantised latent vector instead of direct prediction of a discrete pixel value. Thanks to lessen the burden of transformer from predicting low-level complex vision feature, this work generates photo-realistic and high-resolution scenes.
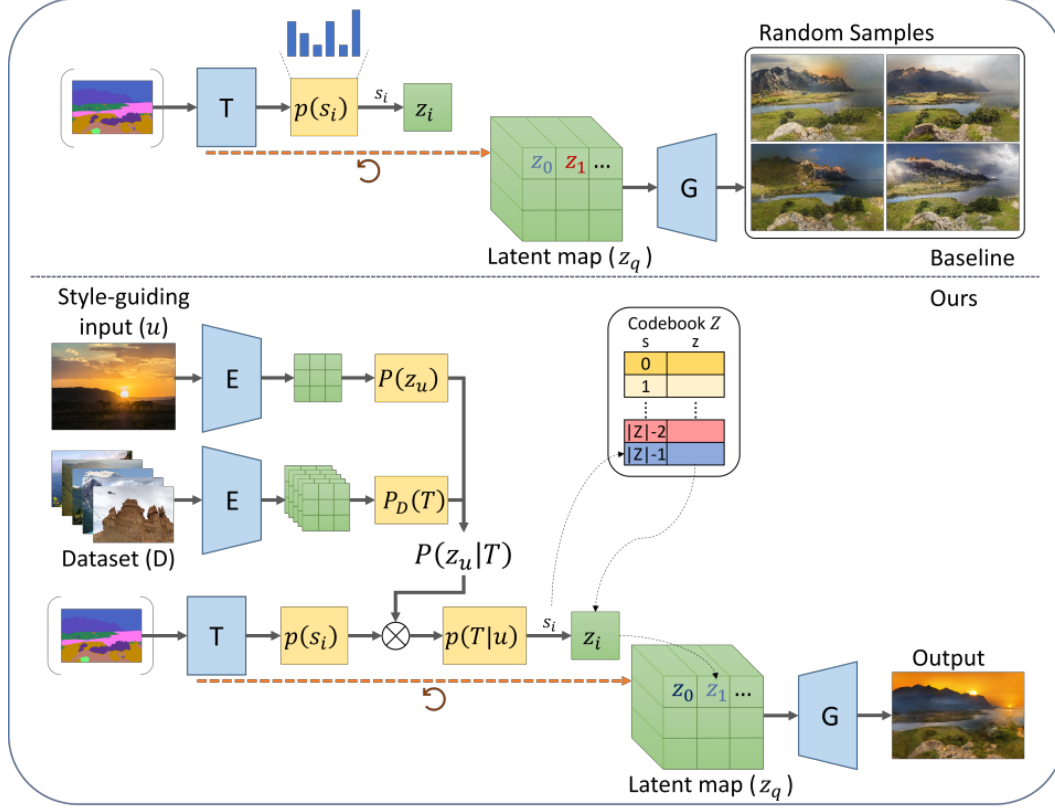
Figure 1. **System overview of image-generative process of baseline method [10] (top) and our method (bottom).** The major difference from the baseline work is applying style likelihood $P(z_u|T)$ in the middle of the process to apply an additional condition u. $P(z_u)$ and $P_D(T)$ are probability distributions of all indices in latent maps. Latent map for $P(z_u)$ is from conditional image u. Latent maps for $P_D(T)$ are from images in training dataset D. Encoder E, decoder G, and transformer T are pre-trained. In case of landscape synthesis, baseline method accepts semantic segmentation map as a conditional input, while in case of face synthesis, baseline is unconditional.

Arbitrary Photorealistic Style Transfer. Arbitrary photorealistic style transfer is partially in line with our method in that it stylizes the output, though it requires a real input image. Some recent studies [21], [18], [19], [40], [51], [50], [33], [7] introduce various methods to transfer style from a single image, but they demand a large dataset for training. Besides, they often yield undesirable transfer of color and texture due to the nature of the neural network.

3. Method

3.1. Baseline Work

The upper half of Figure 1 overviews the baseline method [10]. The baseline model is possible to be both unconditional and conditional. Specifically, it is conditional upon a semantic label map in the case of landscape synthesis. The training process of the baseline is a two-stage. In the first stage, encoder E and decoder G are trained. While in the second stage, transformer T is trained, and E is omitted later. Z is a discrete codebook of size $|Z|$. Transformer model T auto-regressively outputs a probability distribution

$p(s_i)$ over indices s in Z, where an index s_i is random-sampled. Latent code z_i is a paired entry of s_i in Z. z_i fills the i-th entry in the latent map z_q. As a result of an auto-regressive process, z_q is formed and fed into G to synthesize the output. In this process, random sampling from $p(s_i)$ makes the output unpredictable. Thus it is hard to find desired styles such as u from random samples.

3.2. Style-guided Inference of Transformer

Our method aims for high-resolution conditionally guided image generation without re-training the transformer. On top of the latest progress of transformer [10], which incorporates an auto-regressive prediction and a quantized prior, our method will be a missing-piece to guide the inference of transformer conditionally and to stylize the output. To this end, we present a method to control the probability distribution over an auto-regressive latent prediction.

The bottom half of Figure 1 indicates an overview of the proposed method. $P(z_u)$ is a probability distribution of indices in the latent map of style image u, and $P_D(T)$ is the model's probability distribution of all indices in latent

maps of images sampled from the training dataset D. A style likelihood $P(z_u|T)$ is drawn from $P(z_u)$ and $P_D(T)$. Then, it is repetitively multiplied to $p(s_i)$ to form a posterior $P(T|z_u)$. (Throughout the paper, we write $p(s_i)$ as $P(T)$ to denote a model's prior probability distribution.) From $P(T|z_u)$, a style-guided index s_i is sampled, and $z^{(i)}$ is the paired entry of s_i.

To be specific, if we condition $P(T)$, the inference of pre-trained model, on a style input u, then conditional probability is written as $P(T|u)$. Previously, to make T represent this distribution, T needed supervised learning on a style-labeled dataset. Our method, however, proposes the conditionally guided-inference of pre-trained transformer without learning such condition. Without re-train, we represent the given problem to a style likelihood and a model's prior: (Note that $\odot$ and $\oslash$ we use in this paper are element-wise multiplication and division of two vectors; i-th entries of $A \odot B$ and $A \oslash B$ are $A_i \times B_i$ and A_i / B_i each)

$$P(T|u) \propto P(u|T)P(T). \tag{1}$$

Instead of u, we replace it with the probability distribution of latent codes, $P(z_u)$, by ignoring the sequence of latent map z_u and expressing all indices in z_u as a probability distribution. (Note that the generator part is deterministic, and the specific sequence of z_u restores the correct u.) However, ignoring the sequence makes $P(z_u)$ ill-posed to rebuild u. Instead, we rely on the last term, $P(T)$, to build the sequence as this term is recursive. In the same way we get $P(z_u)$, we also obtain the probability distribution of T of the training dataset D, $P_D(T)$, by counting indices in latent maps from images in D. Then, we can view the likelihood term is proportional to the events of latent indices of conditional input u divided by model's general event counts: $P(z_u) \oslash P_D(T)$. Accordingly, the problem is rewritten into

$$P(z_u|T) \odot P(T) \propto \{P(z_u) \oslash P_D(T)\} \odot P(T), \tag{2}$$

where $P(\cdot) \in \mathcal{R}^{|Z|}$, probability over the codebook indices.

Practically, instead of calculating all latent maps from images in D to get $P_D(T)$, we perform Monte-Carlo sampling and take the average, $1/K \sum_{k=1}^{K} P(z_{x_k})$, for all K number of randomly sampled x from the training dataset D. Hence, the Eq. (2) becomes

$$P(z_u|T)P(T) \propto P(z_u) \oslash \{\frac{1}{K}\sum_{k=1}^{K} P(z_{x_k})\} \odot P(T) \tag{3}$$

where $\sum$ is element-wise summation.

Consequently, as the auto-regressive generative process continues, $P(z)$, the distribution of sampled indices from $P(T|u)$, is getting closer to $P(z_u)$, making the style of the generated sample similar to u.

In practice, to reinforce accuracy lost from dropping the spatial and sequential information when representing $P(u)$

as $P(z_u)$, we can use a semantic label map or coordinates in latent map. For example, when we have a semantic label map as a conditional input, we use it to count $P(z_u|T)$ partially by semantic regions: $P(z_{u,j}) \oslash P_D(T_j)$ for j labeled semantic region. (So we have J number of style likelihoods where J is the number of semantic labels.) In this way, we can accurately transfer the style of each semantic to a corresponding semantic. In the following chapter, we show experiments on single style image and a categorized set of style images. In case of a categorized set of style images, we can compute the average style likelihood.

4. Experiments

In this section, firstly, we design experiments on a landscape images because these have visually distinguishable styles like snowy or rocky mountain. To prepare a landscape dataset for training, we collect 80k images from various sources, including Flickr, Pixabay, and Google search. Then, we generate pseudo-semantic labels using the well-known semantic segmentation method, DeepLabv3 [5] with ResNeSt-101 [53] backbone. As a base architecture for image generation, we configure an autoregressive transformer conditioning on a semantic label map presented by the previous paper [10]. To get the model's probability $P_D(T)$ over the training dataset, we randomly select $K = 700$ number of images from dataset. Since our method can accept one or more style inputs, firstly, we show experiments on a single style image, then on a style-categorized set of multiple images sequentially for diversity. Note that we can also choose different source of style for each semantic.

In the first experiment shown in Figure 2, we experiment with our method conditioned on a single style image. For exact modeling of style likelihood probability distribution from a style image, we divide regions by semantic labels and extract the style likelihood regionally. Then, we apply each distribution of semantic labels to the target semantic label, which we call semantic-aware stylization. For this experiment, we choose 11 different styles for the sky from the Flickr dataset and two different styles for the mountain from randomly synthesized samples. Then, style from one of the sky images and one of the mountain images are applied to the sky and mountain semantics each. Thus, the results have stylized sky and mountain in each image. (It is also possible to stylize more semantic labels with different style images.) As a conditional input, we prepare three types of semantic label maps, (a), (b), and (c). (It can be inferred from result images.) We repeat the guided-synthesis four times and pick the best one for each result image.

In the following experiment shown in Figure 3, we perform experiment on categorized set of style images. At first, we search images by keywords - "mountain" + "rock", "mountain" + "snow", "night", "sunset", and "winter" - in Pixabay website. Then we choose 30 to 50 images for each
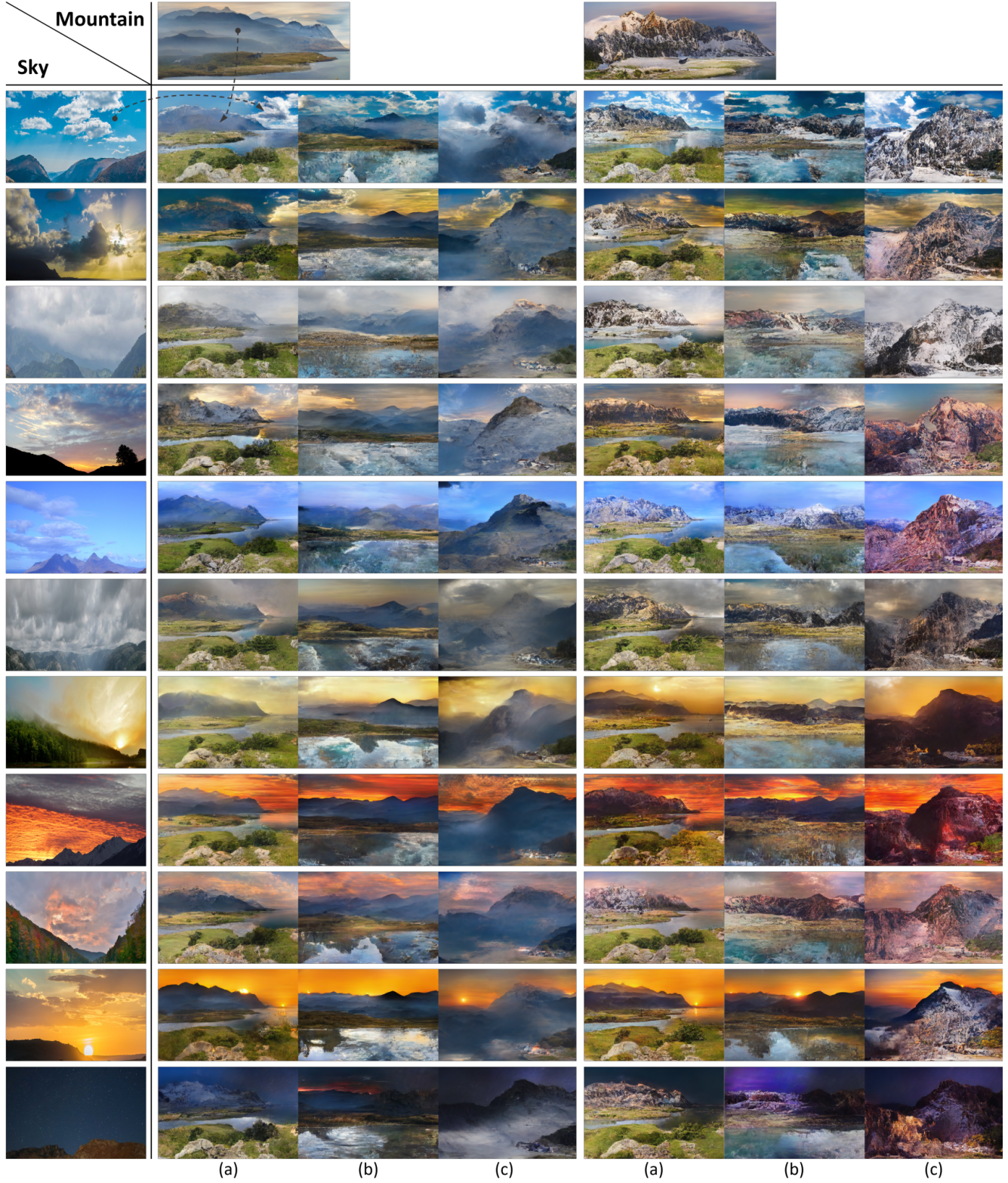
Figure 2. **Qualitative results of our proposed method.** Images on the left are style references of the sky, and on the top are styles of the mountain. Style of one of the sky references is applied to where semantic label matches the sky, while one of the mountain references is applied to where the semantic label matches the mountain. Three different semantic types (a), (b), and (c) are used. Thus, each cell is a mixture of semantically one of (a), (b), and (c), sky from sky reference on the left, and mountain from mountain reference on the top.

(a). Mountain + Rock

(b). Mountain + Snow

(c). Night

(d). Sunset

(e). Winter

(f). Green

Figure 3. **Qualitative results for categorized styles.** We create six categories (a-f) and collect around 30 to 50 images each by keyword-search and manual inspection. Those images are used as style images to get average probabilities for categorized styles. Then we generate images of each category four times

Methods	Babies	Sunglasses
Random sampling	155.30	94.21
[28]	74.39	42.13
Ours	**49.96**	**39.15**

Table 1. **Quantitative evaluation on FFHQ.** FID scores on two categories of FFHQ dataset are measured. Our work synthesizes images that are closer to the category of given guidance than the state-of-the-art competitor [28]. "Random sampling" means random sampling using unconditional baseline [10].

out learning such condition. To achieve a similar result using any existing method, we have a few workarounds. Firstly, there are a few researches on few-shot learning of image generative model, however, they requires hundreds of samples and hardly generate plausible samples on complex scenes. Instead, we can stylize the randomly generated samples using an arbitrary style transfer method. Since our method can apply style conditions by semantic regions, we compare it with one of the state-of-the-art arbitrary style transfer, WCT2 [51], especially which can perform a semantic-aware style transfer. Also we prepare recently well-known texture transferring method, SwappingAE [33]. Although the style transfer method produces reasonable outputs in some cases, it does not consistently generate pleasant outputs in most cases. Furthermore, style transfer approach still yields unmatched or ruined color and texture, which is not desirable in high-resolution realistic scene generation. SwappingAE is excellent at transferring tones and texture details. However, it fails to maintain semantic contents. Moreover, it tends to follow the spatial scene compositions of style image. (For example, it alters mountain to sky in the first two images, river to ground in the second last, and ground to river in the last image.) Detailed results are shown in Figure 4.

Our work uses pseudo-segmentation map which is already available for semantic image synthesis of our baseline work [10]. Having a semantic segmentation map improves extracting the exact per-semantic style and lets us partially stylize the output. On the contrary, in cases of unconditional face synthesis, we can use aligned dataset and collect the probability vectors by spatial locations. To perform a quantitative assessment of an unconditional model, we measure Frechet Inception Distance (FID) [12] score on Flickr-Faces-HQ (FFHQ) [16] dataset and compare with the state-of-the-art few shot image generation [28]. To use as a guidance, we randomly select 10 images of baby and sunglasses each. Then we calculate FID using 2492 and 2683 images of baby and sunglasses each which are the same test sets made public by [28]. We use the same face synthesis model that the baseline work [10] released. Quantitative results of guided-inference of unconditional transformer are given in Table 1. One limitation is, however, our proposed

category. Additionally, category "green" is collected by manually selecting images of green fields. After collecting images by style category, we get the likelihood probabilities for every semantic labels from semantic regions in all style images. Therefore, results in Figure 3 well reflects the common styles of collected images in the category. In this figure, we repeat the inference four times for each category and contain all of them.

To the best of our knowledge, this is the first conditional image generation using pre-trained generative model with-
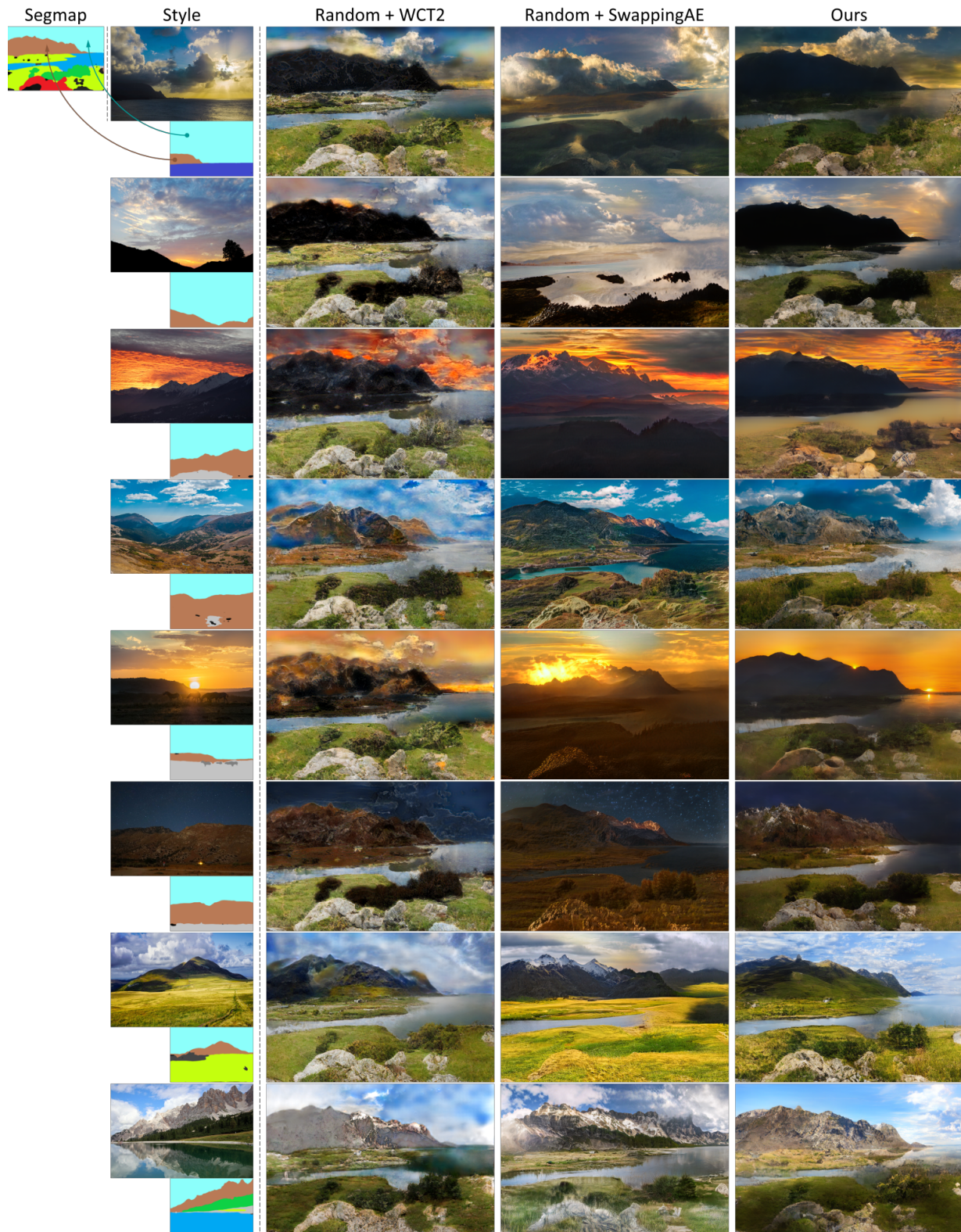
Figure 4. **Qualitative comparison with arbitrary style transfer.** Given a style images, our semantic-aware conditional generation shows plausible outputs in every semantic regions. "Segmap" is a semantic segmentation map to feed as a conditional input to baseline and our method for semantic-aware stylization. For comparison, we run semantic-aware style transfer, WCT2 [51], and texture swapping, SwappingAE [33] (official "Flickr Mountains" pretrained model), on randomly generated samples from the baseline method

Figure 5. **FFHQ babies/sunglasses-guided samples using our method.** (Top): Babies-guided random samples. (Bottom): Sunglasses-guided random samples.

Styles	Random sampling	Ours
Abstract Expressionism	94.24	**80.82**
Art Nouveau Modern	65.73	**58.13**
Baroque	83.02	**68.63**
Cubism	96.18	**69.45**
Naive Art Primitivism	79.25	**70.58**
Northern Renaissance	84.42	**80.21**
Rococo	96.79	**82.26**

Table 2. **Ablation study of unconditional image synthesis on WikiArt dataset [8].** Unlike above experiments, in this experiment, our method is not provided any spatial aid for style likelihood. FID scores on the styles of the art by the history are measured. We only count styles that have more than 2000 images and less than 4500 images to prevent biased random sampling. "Baseline" stands for random sampling-based synthesis of baseline transformer, while "Ours" does for our style-guided synthesis.

method is not able to stylize out of learnt domain because our method guides the inference of transformer only, not re-training the model. Thus, in this experiment, we exclude "Sketches" category from the experiment shown in the competitor [28].

5. Limitations

Our method provides a style image to guide the inference of transformer introduced in our baseline work [10]. To achieve similar effect, most of the recent studies require a re-training process and thousands of style images as well, and this makes recent studies harder to apply to the cases where only one or dozens of style images are provided. Our paper is unique in that, instead of requiring a re-training process, it only needs a single or a few style images to get the output of the desired style, while it inherits advantages of the pre-trained backbone. However, since it does not include a re-training process, it cannot generate samples out of the distribution of the training dataset. Therefore, as shown in Table 1, it is impossible to generate "Sketches" style output using the model trained on real face images.

To model the ill-posed $P(u|T)$ with the known variables, we use probability distributions of $P(z_u)$ and $P_D(T)$. To improve accuracy in the process, we use semantic label maps in case of landscape synthesis and coordinates of a latent map in case of aligned face synthesis. However, if we aggregate probability distribution over all entries of a latent map without spatial separation, it generates a limited result. To show this limitation, where we claim the minimal performance, we carry out an ablation study on WikiArt

dataset [8]. In this experiment of unconditional image synthesis, we collect style vectors from the entire region of the image. As shown in Table 2, FID scores using our method are lower than random sampling-based synthesis. However, it is not dramatic gab like proven in Table 1.

6. Conclusion and Future Work

We believe that this is the first study on the guided inference of transformer to accept an additional arbitrary condition to stylize the output without requiring any parameter nor training procedure. To this end, we interpret probability distribution conditioned on style to a product of style distribution and a prior from transformer. In the experiment section, we prove our method on landscape dataset where styles are diverse and distinguishable. Compared to random generation of the original architecture [10], our method can generate scenes of an arbitrary style that are rare from the original method such as night scene. As an additional arbitrary condition, our work can use a single style image as well as a categorized set of style images. Our work is also similar to style transfer in a way that it stylizes the output image. However, our method uses the pretrained architecture and lets it describe target condition without re-training nor requiring any additional architecture, and thus our method does not need a condition-labeled dataset either.

Since our approach is built upon autoregressive latent code prediction, our work is currently applied on a limited previous work. In the future, we therefore plan to build a general conditional autoregressive module to support an arbitrary conditional input. In addition, our work relies on transformer to build the proper sequence, not the relative layout or sequence of latent codes of style image. Therefore, for more accurate spatial application of style, we will work on space-aware style-guided inference of transformer in the near future.

References

[1] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[3] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. In *International Conference on Machine Learning*, pages 864–872. PMLR, 2018.

[7] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–143, 2021.

[8] Cs-Chan. Artgan/wikiart dataset. https://github.com/cs-chan/ArtGAN/tree/master/WikiArt Dataset.

[9] Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, and Bryan Catanzaro. Panoptic-based image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8070–8079, 2020.

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[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, 2017.

[13] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak 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.

[14] Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. In *International Conference on Machine Learning*, pages 4487–4499. PMLR, 2021.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[16] 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.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[18] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[20] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 570–580, 2019.

[21] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017.

[22] 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 international conference on computer vision*, pages 2794–2802, 2017.

[23] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pages 3481–3490. PMLR, 2018.

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[25] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[26] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.

[27] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651. PMLR, 2017.

[28] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021.

[29] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Ed-

ward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International Conference on Machine Learning*, pages 3918–3926. PMLR, 2018.

[30] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125, 2016.

[31] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4797–4805, 2016.

[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] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020.

[34] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[36] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14866–14876, 2019.

[37] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2234–2242, 2016.

[39] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[40] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatarnet: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018.

[41] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021.

[42] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[43] Hao Tang, Song Bai, and Nicu Sebe. Dual attention gans for semantic image synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1994–2002, 2020.

[44] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318, 2017.

[45] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[47] 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.

[48] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Interspeech 2017*, pages 4006–4010, 2017.

[49] 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*, pages 1316–1324, 2018.

[50] Jonghwa Yim, Jisung Yoo, Won-joon Do, Beomsu Kim, and Jihwan Choe. Filter style transfer between photos. In *European Conference on Computer Vision*, pages 103–119. Springer, 2020.

[51] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9036–9045, 2019.

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

[53] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2736–2746, 2022.

[54] 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 international conference on computer vision*, pages 5907–5915, 2017.

[55] Haitian Zheng, Haofu Liao, Lele Chen, Wei Xiong, Tianlang Chen, and Jiebo Luo. Example-guided image synthesis using masked spatial-channel attention and self-supervision. In *European Conference on Computer Vision*, pages 422–439. Springer, 2020.

[56] 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.

[57] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5467–5476, 2020.