Learning to Predict Convolutional Filters with Guidance for Conditional Image Generation

Lei Chen
chenleic@sfu.ca
Mengyao Zhai
mzhai@sfu.ca
Greg Mori
mori@cs.sfu.ca

Department of Computing Science,
Simon Fraser University,
Canada

Abstract

Various styles naturally exist in an image domain. To generate images with certain style, previous works would usually feed a style encoding as an input to the network. However, a fixed network may lack the capability to present different styles in the target domain precisely, and the style input may also lose its impact along the generation process. In this paper, we propose Guided Filter GAN for multi-modal image-to-image translation via guided filter generation, in which filters at convolutional and deconvolutional layers are constructed dynamically from the style representation from either a target domain image or random distribution. Compared to conventional treatment of style representations being network input, the proposed approach amplifies the guidance of the given style meanwhile enhances the capacity of with dynamic parameters to adapt to different styles. We demonstrate the effectiveness of our Guided Filter GAN on various image-to-image translation tasks, where the experimental results show our approach could precisely render a reference style onto the conditional image and generate images with high fidelity and large diversity in terms of FID and LPIPS metric.

1 Introduction

Image-to-image translation or conditional image generation is the task to convert an image from its original domain to a target domain. Numerous researches in generative deep models fall into the category and promising results have been demonstrated for this problem [9, 10, 14, 19, 23, 27, 35, 36]. Many practical applications can also be categorized as an sub problem in this vast field and thus have drawn focus from researchers as well, such as colorization [10, 17, 33], super-resolution [11, 16, 18] and style transfer [2, 8, 22, 26].

The boost in this field results largely from the success of deep generative models. Early models [10, 35] translate the conditional image to a deterministic target image. To allow for multi-modal generation, an intuitive way is to inject random noise as an extra input as in unconditional GANs [5]. However, this proves to have little effect in the conditional generation task [10, 36], implying the random noise is likely ignored. Therefore, researchers

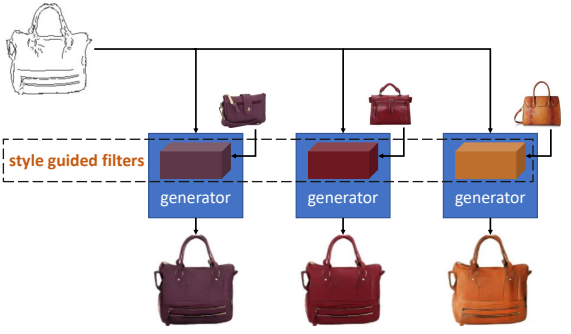

Figure 1: Guided Filter GAN (GFGAN). Convolutional filters are constructed dynamically from the style image or random noises, enforcing the guidance of certain style meanwhile allowing the model to adapt its weights for different styles.

turn to other architectures to achieve diversity. Instead of modelling the domain translation as a deterministic mapping for each image, it can be modelled as an conditional process under various instances of target styles. Following this path, several works [9, 19] propose to encode the styles in images from target domain to a latent space to replace the random noise as extra input to the generative network. These works manage to achieve multi-mode generations and bring in new possibility as they enable guided generations, where a reference image from target domain can be specified at generation to provide specific style information.

The output of a network is determined by the input and its weights. Previous works [9, 19] focus on exploring the input: style images are encoded into a latent vector and inputted to the decoder to allow for guided generation. In contrast, we focus on exploring both input and weights. If a style image is inputted to the model and the weights are dependent on the style, the information from the style is enforced into the output of the layer. When multiple layers are constructed in this manner, the style would be injected into the output repeatedly. Moreover, if the weights of a layer can adapt dynamically to the style, such flexibility is expected to enhance the capacity of the model to precisely present more possible styles compared to only a fixed set of parameters.

In this paper, we propose Guided Filter GAN (GFGAN) which dynamically generates filters at convolutional/deconvolutional layers from the style image using hypernetworks, allowing the network adapting its parameters to the given style. The guided convolutional layer would easily replace their conventional counterparts thus can fit easily into various existing conditional generation architectures. In this paper we take a conventional uni-modal architecture [10] as the underlying model and replace the corresponding layers with guided layers, which instantly enables its capability for multi-modal generation from simple style input. We apply GFGAN on various tasks of image-to-image translation. Experimental results demonstrate that GFGAN could successfully render the given style to the conditional input with accuracy and are able to produce generations with high fidelity and diversity. Superior performances compared against state-of-the-art approaches are observed in terms of FID and LPIPS.

2 Related Works

Image-to-image translation. Pioneering works [10, 20, 35] for image-to-image translations are deterministic: one output is generated for a given conditional image, thus lacking

diversity in the generated images. Naive attempts have been made to increase the generation variety by injecting random noise to the model, which usually is ignored by the generator [10, 36] and shows very limited ability in increasing diversity. Recent works have been proposed to tackle the problem. Multi-agent GAN [4] encourages diverse generations by learning multiple generators simultaneously. However, the number of outputs are limited by the number of generators trained. BicycleGAN [36] consists of two cycles and incorporates two GAN modules: cVAE-GAN that learns the cycle *image* $\rightarrow$ *random noise* $\rightarrow$ *image*, and cLR-GAN that learns the cycle *random noise* $\rightarrow$ *image* $\rightarrow$ *random noise*. MSGAN [24] proposed a regularization term to enforce the generations being different for different random noises with same conditional input. MUNIT [9] and DRIT [19] learns disentangled representations of images by assuming that image representation can be decomposed into a content code and a style code, and multimodal generation could be enabled by injecting different style codes with same content code to the model.

Style transfer. Style transfer is one important application of conditional image generation. The goal is to transfer the style suggested by a style image to an image while preserving its content. Most existing style transfer algorithms are confined to one certain style [11] or a certain set of pre-defined styles [2]. AdaIn [8] is a state-of-the-art algorithm in this field, in which the statistics of a given style images are learned and used to parameterize a instance normalization layer to align the corresponding statistics of the content images to adaptively transfer the styles while preserving the contents. Approaches proposed for multimodal generation are also capable of perform adaptive style transfer. However, since they cannot differentiate finer differences among images within same domain thus show limited ability in learning the styles.

Weights generation. Generating weights of networks has drawn many attentions since Hypernetworks [6], and researchers have adopted the idea addressing problems in various fields. Approaches [25, 28, 31, 32] have been proposed to adapt a model to new tasks. Parameter redundancy [3] and network compression [6] have also been explored by predicting network parameters. Classification has also seen the success of dynamic network, CondConv [30] learns a set of convolutional kernels and combine them differently according to the input. In generative models the idea is explored to help layout-to-image generation [21] where a semantic label map is used to predict convolutional kernels to generate intermediate features maps used for the eventual synthesis. BasisGAN [29] also adopts this idea to predict the parameters of generator from random noise thus multi-modal generation is enabled by sampling different set of random noises to produce different sets of network parameters. The focus of their work is to inject stochasticity directly to the network parameters to force diversity in the generation. To the contrast, our GFGAN focuses on constructing filters to adapt to different input styles and express them more precisely in the generation. StyleGAN [12] generates mean and variance for a set of adaptive instance normalization layers based on style code, these normalization layers help apply various attributes to the generated images. StyleGAN2 [13] improves upon the StyleGAN by modifying the adaptive instance normalization layers and introducing the weight modulation on the convolutional layers. Similar to BasisGAN, StyleGAN and StyleGAN2 are also proposed for unconditional generation and are not designed to generate based on given style images.

3 Approach

In this section, we introduce our Guided Filter GAN (GFGAN) that constructs the convolutional/deconvolutional filters in the network dynamically from the style representation either

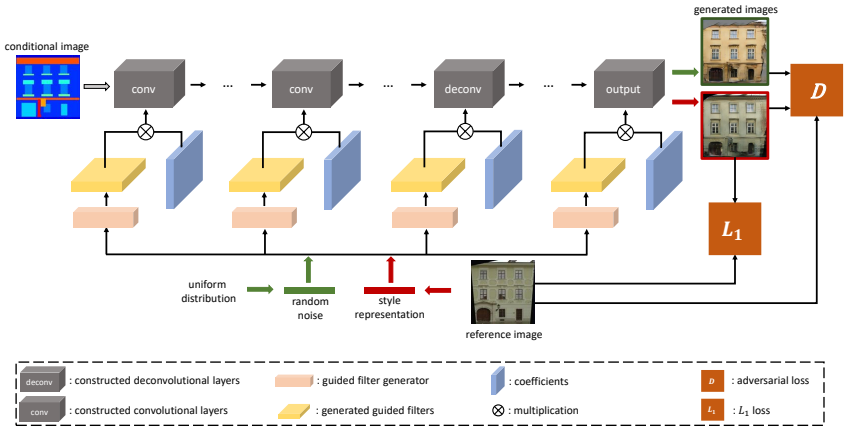

Figure 2: **Overview of Guided Filter GAN (GFGAN).** For a style representation that comes from either a target domain image or random distribution, the guided layers will firstly generate filter basis from the style representation and then linearly combines the basis into convolutional/deconvolutional filters with a set of learnable coefficients. With these guided layers the style information is repeatedly enforced into the output. Adversarial and reconstruction loss are utilized at training time.

extracted from target domain images or sampled from a simple distribution. We start by introducing the guided convolutional filters in Section 3.1. We then introduce how to learn the proposed model in Section 3.2. An overview of the model is illustrated in Figure 2.

3.1 Guided Convolutional Filter

Assume a style representation s that carries a specific style information from the target domain is present, the conditional generation relies heavily on how to incorporating the style representation in the generation process. As the output of a network is mostly determined by the input and its weights, different from previous works [9, 19] that focus on exploring style encoding as input, we focus on exploring both input and weights instead. We argue that if the weights of a layer are generated from the style input, the information from the style is enforced into the output of the layer. When multiple layers are constructed in this manner, the style would be injected into the output repeatedly.

Though it's intuitive to directly generate the filter weights all-together from the style representation, the weights are of a rather high dimension and generating all the weights directly would require even larger set of parameters and considerable extra computation time. To address the problem, we proposed to factorize convolutional filters into a set of filter basis and a coefficient matrix that linearly combines the basis. Moreover, instead of using a fixed filter basis, we choose to generate the filter basis for different styles while learn a set of universal coefficients for different styles at the layer. The dimension of the basis can be chosen so that generating these basis are less time consuming and more parameter efficient.

Consider the convolutional layer at l^{th} layer with N^l_{in} input channels and kernel size S^l. Each basis for this layer is set to the shape $S^l \times S^l \times N^l_{in}$ and a set of K^l basis is generated from the style representation. Thus the filter basis B^l is of shape $S^l \times S^l \times N^l_{in} \times K^l$. The basis B^l at layer l is constructed dynamically from the style representation s by the guided filter

generator G^l_{gf}.

$$B^l = G^l_{\text{gf}}(s). \tag{1}$$

Another set of coefficients W^l_c of shape $K^l \times N^l_{out}$ is learned to linearly combine the set of generated basis to construct the actual filter W^l_{gf} at l^{th} layer of shape $S^l \times S^l \times N^l_{in} \times N^l_{out}$.

$$W^l_{\text{gf}}(s) = \mathcal{R}^{-1}(\mathcal{R}(B^l) \cdot W^l_c), \tag{2}$$

where $\mathcal{R}$ denotes the operation to reshape B^l into shape $(S^l \times S^l \times N^l_{in}) \times K^l$ and $\mathcal{R}^{-1}$ denotes the inverse operation to reshape the multiplication result back to shape $S^l \times S^l \times N^l_{in} \times N^l_{out}$. The same construction process can be applied to deconvolutional layers in exact the same way.

3.2 Learning GFGAN

The proposed GFGAN can be incorporated in various conditional generation tasks by replacing the conventional convolutional/deconvolutional filters in the generators with our guided filters in the underlying models. In this work we demonstrate how it can work with an underlying pix2pix structure and trained with paired image data from two domains, while adaptation to most conditional generation models can be done in exactly the same way.

We assume the the presence style representation in the discussion above. In practice, style representations could come from various sources as encoded from target domain images or sampled from a distribution. Encoded style representation take the advantage from real target domain images while sampled random style representations grant the flexibility at deployment when no real images can be used as reference. We take the advantage of both by using style encoded from reference images as well as sampled from a uniform distribution at training time. This enables the network the ability to translate a conditional image either based on a reference image or randomly with a sampled style at test time.

For each training image pair (x, y) with same content from two different domains, we encode the target domain image y into style representation s_e,

$$s_e = E(y), \tag{3}$$

where E is a encoding module, and construct weights at all guided filter convolutional or deconvolutional layers $W_{\text{gf}}(s) = \{W^l_{\text{gf}}\}_l$. The loss for this training pair would consist of a reconstruction term as well as a adversarial term

$$\mathcal{L}_{pair} = \mathbb{E}_{x,y}|y - G(x; W_{\text{gf}}(E(y)))| \tag{4}$$

$$+ \mathbb{E}_{x,y}[\log(D(x))] + \mathbb{E}_{x,y}[\log(1 - D(G(x; W_{\text{gf}}(E(y)))))], \tag{5}$$

where G is the generator network and D is the discriminator network. And meanwhile we also sample random style representation s_r of same dimension from a uniform distribution, and apply adversarial loss to make sure generator constructed with weights $W_{\text{gf}}(s_r)$ would generate realistic images

$$\mathcal{L}_{random} = \mathbb{E}_{x,y}[\log(D(x))] + \mathbb{E}_{x,y,s_r}[\log(1 - D(G(x; W_{\text{gf}}(s_r))))], \tag{6}$$

The final objective would be a combination of the two objectives above,

$$\mathcal{J} = \arg\min_G \max_D \mathcal{L}_{pair} + \mathcal{L}_{random} \tag{7}$$

4 Experiments

We validate GFGAN on multiple image-to-image translation tasks, including *labels* $\rightarrow$ *building facades*, *maps* $\rightarrow$ *aerial photos*, *segmentations* $\rightarrow$ *street photos*, *edges* $\rightarrow$ *handbag photos* and *edges* $\rightarrow$ *shoe photos*. For each dataset, we explore two different generation settings: (a) reference generation setting where reference images from target domain are given to guide the translation and (b) random generation setting where no reference is provided.

Training Details. All the conditional generation models are trained on images of size 256×256. We implement our models with the Tensorflow [1] framework and all models are trained with Adam optimizer [15]. We do not use any network to encoder the style image but rather simply down-sample the reference image as the style representation for the filter construction. The guided filter generator network is a two layer MLP and the number of basis is set to 8 for all layers across all tasks. We adopt the Pix2Pix [10] with residue blocks as the underlying model for our GFGAN, where every convolutional layer and deconvolutional layer is replaced correspondingly with guided layer.

Quantitative Metrics. To quantitatively measure the generated images, the following metrics are used in the experiments, *Fréchet Inception Distance (FID)* [7] and *LPIPS* [34]. *FID* is an extensively used metric to compare the statistics of generated samples to real images from that domain. We use this metric to quantitatively evaluate the quality of generated images (lower FID indicates higher generation quality). *LPIPS* is the other metric adopted in our experiments which computes the distance of output images in the feature space given the same input [9, 19, 36]. It is a widely used metric to quantitatively evaluate the diversity (higher LPIPS indicates higher diversity) of generated images.

Baseline Models. We compare our GFGAN to the following state-of-the-art models that are also capable of generating multi-modal outputs: (a) *BicycleGAN* [36]: BicycleGAN learns to model the distribution of the latent representation of target domain images via two cycle losses which reconstructs the image and latent representation along domain translation respectively, with sampling from the latent representation it's capable to generate multi-modal outputs (b) *DRIT* [19]: DRIT learns to disentangle the content and style in an image into different encodings, and feed the style representation combined with the conditional image to the generator for generation and aside from extracting style from reference DRIT could also incorporate random noise style for multi-modal generation (c) *MUNIT* [9]: MUNIT also learns to disentagle style and content while takes a different way to utilize the disentagled representations compared to DRIT (d) *MSGAN* [24]: MSGAN introduces a mode-seeking loss to encourage the diversity in the generation.

4.1 Image-to-Image Translation

Labels $\rightarrow$ Building Facades The task is to translate segmentation maps of building facades into realistic photos maintaining the building structure. The results can be found in Table 1. Our GFGAN outperformed baseline approaches in both metrics, indicating the generated images are of higher fidelity and diversity. For reference generation, the comparison with DRIT and MUNIT can be found in Figure 3. GFGAN successfully recover the colors of the reference images meanwhile render the conditional image accordingly with photorealistic quality. The generations from DRIT show more artifacts and more importantly the colors of the reference images are not fully expressed and the diversity among these guided generations are less significant, MUNIT recovers the color better but the accuracy and quality is still not as good as GFGAN. Random generations with style sampled from the uniform

Figure 3: Visualizations of reference image guided generation. GFGAN is able to capture the style in the reference image and render it in a realistic way onto the conditional image for a high quality generation. DRIT in comparison does not express the color precisely and the generated images are less diverse and realistic. More visualizations can be found in the supplementary material.

distribution are included in Figure 4, even without the guidance from a real target domain image GFGAN is still able to generate various reasonable outputs.

	labels → building facades			
	Random		Reference	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
BicycleGAN	96.07	0.3013	-	-
MSGAN	90.71	0.3804	-	-
DRIT	120.71	0.1836	122.35	0.1091
MUNIT	119.92	0.2713	141.47	0.2303
Ours	77.3	0.3810	91.48	0.3691

Table 1: FID and LPIPS metrics for different models on *labels → building facades*.

Edges → Shoe Photos and **Edges → Handbag Photos** The tasks for these datasets are to translate edges of shoes and handbags into photos preserving the content. The quantitative results are in Table 2, where GFGAN also demonstrates its effectiveness in generating diverse outputs with high fidelity especially at the guided generation setting. Though for handbag photo generation DRIT shows better diversity in the unconstrained generation setting, GFGAN could generate more realistic images with its lead in FID. We also include qualitative results to compare the guided generation results. This time DRIT could also differentiate the input styles and render them differently in the generation, however, certain colors are not recovered accurately in the output. MUNIT again recovers the color better and generates more realistic images compared to DRIT. GFGAN also demonstrates its effectiveness in precisely transferring the styles from reference to the final output while maintaining a high quality, outperforming the other approaches significantly. Random generations are included in Figure 4.

	edges → shoe photos			
	Random		Reference	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
BicycleGAN	64.28	0.1519	-	-
DRIT	106.16	0.2040	99.8	0.1346
MUNIT	66.87	0.1273	64.23	0.1715
GFGAN	35.57	0.2088	39.77	0.1911
	edges → handbag photos			
	Random		Reference	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
BicycleGAN	56.21	0.2225	-	-
DRIT	79.73	0.2788	70.47	0.2399
MUNIT	49.86	0.1667	50.45	0.2144
GFGAN	40.83	0.2613	41.8	0.3017

Table 2: FID and LPIPS metrics for different models on *edges → shoe / handbag photos*.

Maps → Aerial Photos The task for this dataset is to translate a crop of map into a corresponding aerial image. Quantitative results can be found in Table 3. GFGAN achieves best FID indicating better image quality. DRIT shows a stronger diversity in the unconstrained generation setting, suggesting it's sacrificing part of fidelity to express more extreme styles and the same situation also happens to MSGAN with higher LPIPS but worse FID.

	maps → aerial photos			
	Random		Reference	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
BicycleGAN	104.56	0.1039	-	-
MSGAN	141.91	0.4915	-	-
DRIT	123.45	0.3115	133.05	0.2265
MUNIT	121.99	0.0996	126.56	0.0986
GFGAN	96.37	0.2409	116.75	0.2186

Table 3: FID and LPIPS metrics for different models on *maps → aerial photos*.

Segmentations → Street Photos For this task we train the models on Cityscapes dataset to translate segmentions into street photos at a car view. The quantitative results can be found in Table 4 where GFGAN take the lead in both image quality and diversity for both settings.

	segmentations → street photos			
	Random		Reference	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
BicycleGAN	119.51	0.2082	-	-
DRIT	81.98	0.2755	130.47	0.1466
MUNIT	59.90	0.1743	57.43	0.2457
GFGAN	35.13	0.3964	35.22	0.3306

Table 4: FID and LPIPS metrics for different models on *segmentations → street photos*.

Overall the GFGAN achieves best quality and most of the time highest diversity as well across various image-to-image translation tasks. Qualitative results on guided generation further demonstrate its strength in its accuracy to render a given reference image onto the conditional image meanwhile generating realistic images in the target domain. All these results validate our intuition that constructing convolutional filters from style representations

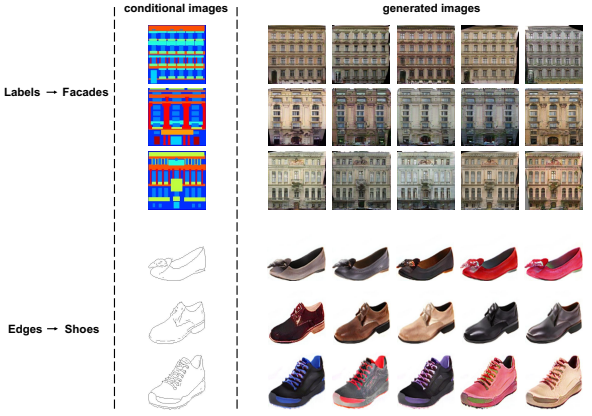

Figure 4: Visualizations of random sampled images. The style representations are sampled from a uniform distribution. GFGAN can generate diverse images with high fidelity through simple sampling of style representation with the guided filters. More visualizations can be found in the supplementary material.

helps with the guidance of style information over the output and brings in extra capacity for the model to adapt to various styles with tailored parameters.

4.2 Ablation Study

Ablation study on number of basis. The number of basis K serves to balance the parameter size and the model capacity. To better understand the influence of the number of basis on the quality and diversity of generated images, we conduct experiments on *labels* $\rightarrow$ *building facades* with different choices of basis number K as in Table 5.

	Random		Reference	
# filter basis	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
Fixed Filter	128.95	0.0	-	-
4	87.68	0.3515	93.22	0.3425
8	77.3	0.3810	91.48	0.3691
16	76.47	0.3711	95.84	0.3641
32	73.02	0.3755	93.21	0.3753

Table 5: Ablation study on number of filter basis on task *labels* $\rightarrow$ *building facades*.

We first include the results with conventional fixed convolutional layers, and the model thus becomes the vanilla Pixel2Pixel, the image quality is significantly worse than when we use the guided filters and there's also no diversity in the generated images. When the number of basis increase from 4 to 8, an overall improvement of performance is observed. While more basis filters does not bring significant improvement. This may result from the increased difficulty in learning the model with the extra complexity in the guided filter generation. We also include some qualitative results with 4 and 8 basis in Fig 5. Using more filter basis help with rendering the style more smoothly and accurately onto the conditional image, it also helps improve image quality.

Effect of guided filters on different layers. In the above experiments we replace every convolutional and deconvolutional layer in the underlying Pix2Pix network with our guided filter layer. It would be desirable to know what role these layers are playing in the generation

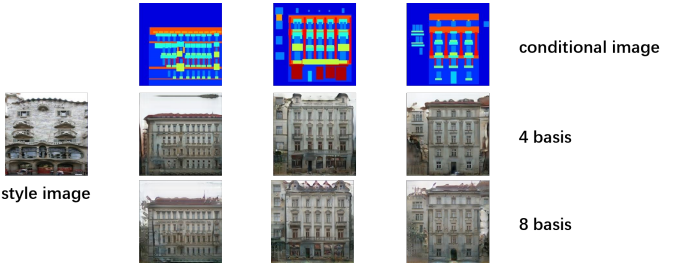

Figure 5: Reference based generations with 4 and 8 basis resepectively.

process. To explore the effect of guided filters at different layers, we train two variants of GFGAN with guided filters in part of convolutional/deconvolutional layers on the facades dataset. In the first variant we replace the filters in the convolutional layers till the end of last residue block with guided filter layers and denote it as Top GFGAN. In the other variant we leave all those convolutional layers untouched while replace the rest convolutional and deconvolutional layers with guided filter layers which we denote as Bottom GFGAN. The quantitative results can be found in Table 6. The Top GFGAN results in a slightly worse performance than the full model. Even if the final layers are not guided layers, the model still performs well in fidelity and diversity, this is suggesting that the style information enforced earlier in the network parameters is still taking effect at later layers. The Bottom GFGAN model incorporate less portion of layers of guided filter, it shows a satisfying guided generation quality while outperformed by other models in random generation quality as well as both diversity by a lot. This is indicating that the guided filters at final layers does effectively enhance the capacity, but is still not sufficient to express all the different styles faithfully by themselves.

	Random		Reference	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
Top GFGAN	77.97	0.3662	95.22	0.3652
Bottom GFGAN	93.58	0.2215	93.06	0.1496
GFGAN	77.3	0.381	91.48	0.3691

Table 6: Ablation study results on the influence of guided filters at different layers on the task *labels → building facades*.

5 Conclusion

In this paper, we propose GFGAN with guided filter generation for image-to-image translation. The convolutional filters are constructed dynamically according to a style representation that can either be encoded from a target domain image or sampled from a random distribution. The proposed approach is generic as the guided filter generation process can be directly applied to most existing generative networks by simply replacing the convolutional filters with corresponding guided filters. By conducting experiments on various image-to-image translation tasks, we demonstrate the effectiveness of the guided filters for better generation quality as well faithfully express various styles for diversity. Quantitative measurements shows a superior performance over state-of-the-art methods on both image fidelity and diversity of our model, and qualitative results demonstrate GFGAN could render reference style onto the conditional image accurately.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

[2] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando De Freitas. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[4] Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[6] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *International Conference on Learning Representations (ICLR)*, 2017.

[7] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2018.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision (ECCV)*, 2018.

[20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[21] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[22] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[24] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German Conference on Pattern Recognition (GCPR)*, 2016.

[27] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations (ICLR)*, 2017.

[28] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Ze Wang, Xiuyuan Cheng, Guillermo Sapiro, and Qiang Qiu. Stochastic conditional generative networks with basis decomposition. In *International Conference on Learning Representations (ICLR)*, 2019.

[30] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[31] Mengyao Zhai, Lei Chen, Jiawei He, Megha Nawhal, Frederick Tung, and Greg Mori. Piggyback gan: Efficient lifelong learning for image conditioned generation. In *European Conference on Computer Vision (ECCV)*, 2020.

[32] Mengyao Zhai, Lei Chen, and Greg Mori. Hyper-lifelonggan: Scalable lifelong learning for image conditioned generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[33] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[36] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.