SEMANTICGAN: GENERATIVE ADVERSARIAL NETWORKS FOR SEMANTIC IMAGE TO PHOTO-REALISTIC IMAGE TRANSLATION

Junling Liu[♮] *Yuexian Zou*[♮,♯] *Dongming Yang*[♮]

[♮] ADSPLAB, School of ECE, Peking University, Shenzhen, China
[♯]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Generative Adversarial Networks (GANs) have shown remarkable success in **S**emantic label map to **P**hoto-realistic image **T**ranslation (S2PT) task. However, the results of the state-of-the-art approaches are often limited to blurriness and artifacts, and still far from realistic, since these methods lack effective semantic constrains to preserve the semantic information and ignore the structural correlations between the textures. To address those problems, we propose a SemanticGAN to synthesize high resolution image with fine details and realistic textures from the semantic label map. Specifically, we propose a Semantic Information Preserved Loss (SIPL) to maintain semantic information in the process of the generation via a segmentation model. Furthermore, we develop a novel generator to obtain the correlations between the image textures using newly-designed Correlated Residual Block (CRB). Experiments evaluated on Cityscapes dataset show that SemanticGAN outperforms many recent state-of-the-art methods in terms of qualitative and quantitative performance.

***Index Terms*—** Generative adversarial networks, S2PT, Semantic information preserved loss, Structural correlations between images textures, Correlated residual blocks

1. INTRODUCTION

Photo-realistic image synthesizing conditioned on semantic layouts has been an emerging research area in computer vision and computer graphics. Although traditional graphics algorithms excel at the task, building virtual environments is expensive and time-consuming since they have to model every aspect of the world explicitly such as geometry, materials, and light transport.

Recently, lots of researches have make significant progresses in S2PT task. These methods use GANs [1] to learn the mapping from input semantic layouts to output photographic images. Isola [2] leverage GAN in a conditional setting for image-to-image translation problems. Wang [3] uses multi-scale generators and discriminators to synthesize high-resolution photo-realistic images. However, these methods lack effective semantic constrains to maintain semantic information in the process of generation, which will cause blurriness and artifacts, as shown in Fig. 1. What's more, these methods ignore the correlations between the textures, which will cause the overlap and mess between the textures of an object, like the bus windows and buildings generated by Pix2pixHD in Fig. 1.

To solve the challenges mentioned above, we propose a novel conditional generative adversarial network together with a novel generator using newly-designed Correated Residual Block and Semantic Information Preserved Loss for S2PT task. We first introduce a semantic segmentation network which takes the synthesized image as input and produce a label map, and calculate the cross entropy between label map and input semantic image as SIPL, to cooperate with adversarial loss and perceptual loss. It explicitly enables SemanticGAN to synthesize photo-realistic images which could still maintain the semantic information. Furthermore, taking advantage of the Res2Net module [4], we develop a novel generator structure with newly-designed residual blocks named Correlated Residual Blocks (CRB) to explore the correlations between image textures effectively. By using group convolutions and channel shuffle, the CRB is able to promote sufficient receptive field and gain adequate structural correlated information, and therefore helps synthesize photographic images with more realistic textures.

The remainder of the paper is organized as follows: in Section 2, the proposed network architecture is described. Experimental results and conclusion are given in Section 3, and Section 4, respectively.

(a) Ground Truth (b) Pix2pixHD (c) SemanticGAN

Fig. 1. (a) An original image and its corresponding semantic label map. (b) The synthesized image generated by Pix2pixHD and feed it into a segmentation model to get corresponding semantic label map. The results of Pix2pixHD are still far from realistic, and the semantic information has been lost (see from the yellow boxes). (c) With the proposed SIPL, SemanticGAN can preserve the semantic information to guide the generator to synthesize photo-realistic images.

2. METHODOLOGY

In this section, Semantic Generative Adversarial Network (SemanticGAN) is proposed for S2PT. Firstly, we explain our Semantic Information Preserved Loss. Then we introduce our newly-designed Correlated Residual Block (CRB) and the whole framework of SemanticGAN. An overview of this method can be observed in Fig. 2.

This paper was partially supported by National Engineering Laboratory for Video Technology - Shenzhen Division, Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing).Special acknowledgements are given to Aoto-PKUSZ Joint Lab for its support.

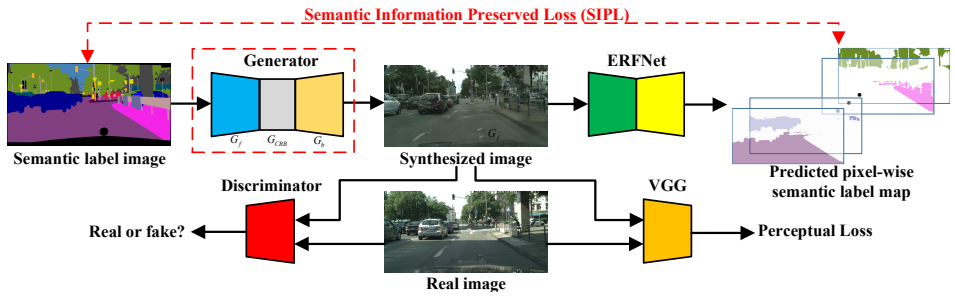

Fig. 2. The overall architecture of our proposed SemanticGAN for S2PT task.

2.1. Semantic Information Preserved Loss

In recent years, GANs have emerged as a powerful generative model and have been widely used in previous works. However, as illustrated in Fig. 1, it can be clearly observed that the semantic information of synthesized images cannot be well preserved, resulting in the blurriness and artifacts. It is mainly because using the adversarial loss only is not able to maintain the semantic information effectively.

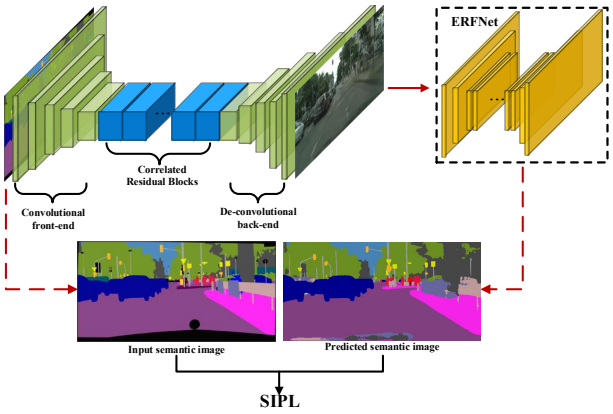

Fig. 3. The illustration of SIPL calculation.

Therefore, we propose a new loss function called Semantic Information Preserved Loss (SIPL) to maintain the pixel-wise sematic information of the synthesized images. The motivation is that if the generated images are realistic enough, segmentation model trained on real images would be able to segment the synthesized image correctly as well, so we could use the segmentation model to assign semantic labels to each pixel in synthesized images, and constraint the generation by enforcing the predicted semantic images of generated pictures to be similar with the input semantic label maps, as illustrated in Fig. 3, we use ERFNet [5] as the segmentation model.

The SIPL is computed by a pixel-wise log-softmax over the final feature map of ERFNet combined with the cross entropy loss function. The log-softmax is defined as:

$$p_k(\mathrm{x}) = \log\left(\frac{\exp\left(a_k(\mathrm{x})\right)}{\left(\sum_{i=1}^{K}\exp\left(a_i(\mathrm{x})\right)\right)}\right) \tag{1}$$

where $a_k(\mathrm{x})$ denotes the activation in feature channel k at the pixel position x, K is the number of classes. The cross entropy then penalizes at each position the deviation of $p_{\ell(\mathrm{x})}(\mathrm{x})$ from 1. Specifically, the SIPL is as followed:

$$\mathcal{L}_{\mathrm{SIPL}} = \sum_{\mathbf{x}\in\Omega}\log\left(p_{\ell(\mathbf{x})}(\mathbf{x})\right) \tag{2}$$

where $\ell : \Omega \rightarrow \{1, \ldots, K\}$ is the true label of each pixel.

2.2. Correlated Residual Block

The encoder-decoder architecture has been widely used in GANs. However, such a network could not obtain the structural correlation between image textures effectively due to the limitation of computational constraints, and cause the overlap and mess of the textures synthesized by generator. As shown in Fig. 1, the textures of the bus windows and buildings are overlapped and disordered. Therefore, we propose a Correlated Residual Block (CRB) to expand the reception field further to obtain the structural correlation between textures while maintaining a similar computational load.

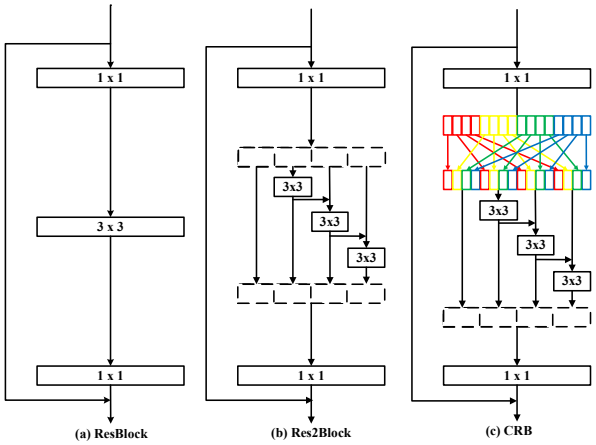

Fig. 4. Comparison of original ResBlock, Res2Block and proposed CRB.

The structure of CRB is inspired by the work [4] in which a Res2Block is proposed to represent multi-scale features at a granular level and increase the range of receptive fields for each network layer. However, the features extract by different groups in Res2Block only related to the inputs within the group and the features calculated by the previous convolution group. If we allow group convolution to obtain input data from different groups, the input and output channels will be fully related.

Specifically, we introduce Channel shuffle [6] into Res2Block and propose a newly-designed Correlated Residual Block to obtain

the structural correlations between image textures. For the feature map generated from the 1x1 convolution, we first divide the channels in each group into several subgroups, then feed each group in the 3x3 convolution with different subgroups, as shown in Fig. 4.

2.3. Semantic Generative Adversarial Network

We follow the work of Pix2pixHD which utilizes the cGAN[7] for S2PT. As shown in Fig. 2, SemanticGAN consists of a newly-designed generator G, multi-scale discriminators D and a segmentation model. The framework aims to model the conditional distribution of real images given the input semantic label maps via the following minimax game:

$$\min_G \max_D \mathcal{L}_{\mathrm{GAN}}(G, D) \tag{3}$$

where $\mathcal{L}_{\mathrm{GAN}}(G, D)$ is given by:

$$\mathbb{E}_{(\mathbf{s},\mathbf{x})}[\log D(\mathbf{s}, \mathbf{x})] + \mathbb{E}_{\mathbf{s}}[\log(1 - D(\mathbf{s}, G(\mathbf{s}))] \tag{4}$$

where $\mathbf{s}$ are semantic label maps and $\mathbf{x}$ are corresponding natural photos.

The generator has three components: a convolutional front-end G_f, a set of CRBs G_{CRB}, and a de-convolutional back-end G_b. Note that G_b is a mirrored version of G_f. A semantic label map of resolution 1024 $\times$ 512 is passed through three components to produce an 1024 $\times$ 512 photographic image with structural correlated textures.

We use the multi-scale discriminators proposed in Pix2pixHD. The discriminators have an identical network structure, and downsample the real and synthesized high-resolution images by a factor of 2 and 4 to create an image pyramid of three scales. The improved adversarial loss is then calculated as:

$$\mathcal{L}_{\text{improved}}(G, D_k) = \mathbb{E}_{(s,x)} \sum_{i=1}^{T} \frac{1}{N_i} \left[\left\| D_k^{(i)}(\mathbf{s}, \mathbf{x}) - D_k^{(i)}(\mathbf{s}, G(\mathbf{s})) \right\|_1 \right] \tag{5}$$

where $D_k^{(i)}$ denotes the ith layer of discriminator D_k, T is the total number of layers and N_i denotes the number of elements in each layer.

To ensure that the generated image and its ground truth are similar in high-level feature representation, we introduce the perceptual loss [8]:

$$\mathcal{L}_{\text{perceptual}} = \sum_{i=1}^{N} \frac{1}{M_i} \left[\left\| F^{(i)}(\mathbf{x}) - F^{(i)}(G(\mathbf{s})) \right\|_1 \right] \tag{6}$$

where $F^{(i)}$ denotes the ith layer with M_i elements of the VGG network [9].

By combining above losses, we can achieve our full loss:

$$\min_G \left(\lambda_1 \left(\max_{D_1,D_2,D_3} \sum_{k=1,2,3} \mathcal{L}_{\mathrm{GAN}}(G, D_k) \right) + \lambda_2 \mathcal{L}_{\mathrm{SIPL}} + \lambda_3 \sum_{k=1,2,3} \mathcal{L}_{\mathrm{improved}}(G, D_k) + \lambda_4 \mathcal{L}_{\mathrm{perceptual}} \right) \tag{7}$$

where λ_1,λ_2,λ_3,λ_4 controls the importance of the four terms. In our experiments, λ_1 and λ_2 is set to 1, λ_3 and λ_4 is set to 10.

3. EXPERIMENTS

We conduct experiments on Cityscapes dataset [10]. The Cityscapes dataset consists of a large and diverse set of stereo video sequences recorded in streets from different cities in Germany and neighbouring countries. We use 2975 training images from the Cityscapes training set with image size 1024 $\times$ 512 and 500 testing images from the Cityscapes validation set with the same image size.

3.1. Implementation details

We follow the naming convention used in Johnson [8]. Let $c7s1$-k denote a 7$\times$7 Convolution-InstanceNorm-ReLU layer with k filters and stride 1. dk denotes a 3$\times$3 Convolution-InstanceNorm-ReLU layer with k filters and stride 2. Ck denotes a CRB with k filters. uk denotes a 3$\times$3 fractional-strided-Convolution-InstanceNorm-ReLU layer with k filters, and stride $\frac{1}{2}$. Our generator network:

$c7s1$-64, $d128$, $d256$, $d512$, $d1024$, $C1024$, $C1024$, $C1024$, $C1024$, $C1024$, $C1024$, $C1024$, $C1024$, $u512$, $u256$, $u128$, $u64$, $c7s1$-3

For the discriminator networks, we use 70$\times$70 PatchGAN [2]. Let Dk denote a 4$\times$4 Convolution-InstanceNorm-LeakyReLU layer with k filters and stride 2. After the last layer, we apply a convolution to produce a 1 dimensional output. We use leaky ReLUs with slope 0.2. All three discriminators have the identical architecture as follows:

$D64$-$D128$-$D256$-$D512$

Our model was trained on a NVIDIA GTX 1080TI GPU. Adam [11] was used for optimization. The initial learning rate was set to 0.0002 for the first 100 epochs and linearly decay the rate to zero over the next 100 epochs. Weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. Similar to Pix2pixHD, the instance map is concatenated with semantic label map as the input for further improving the quality of synthesized images.

3.2. Ablation study

To quantify the quality of our results, we introduce Fréchet Inception Distance (FID) [12] for evaluating. FID calculates the Wasserstein-2 distance [13] between the generated images and the real images in the feature space of an Inception-v3 network [14]. Lower FID values mean closer distances between synthetic and real data distributions. In SemanticGAN, we use Pix2pixHD as baseline, and make two main modifications that contribute to the overall effectiveness: 1) propose SIPL to maintain semantic information, 2) design CRB to obtain the structural correlation between textures. We compute the FID scores under different configurations to quantify the contribution of different configurations to overall effectiveness.

Table 1. Ablation study of the proposed SIPL and CRB

Modification	Usage					
	a	b	c	d	e	f
Baseline	✓	✓	✓	✓	✓	✓
Res2Block		✓		✓		
Proposed SIPL			✓	✓		✓
Proposed CRB					✓	✓
FID	56.920	55.489	56.453	55.098	55.187	**54.775**

In Tab. 1, we could find that only replacing residual blocks [15] in baseline with Res2Block could get 1.4 decrease comparing with

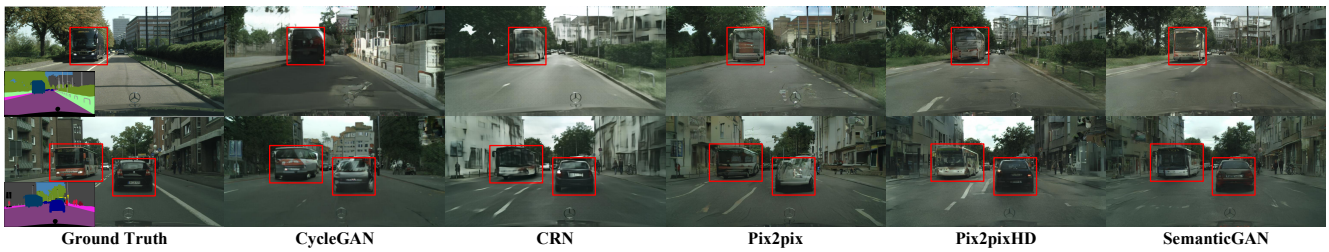

Fig. 5. Comparison results with CycleGAN, Pix2pix, CRN, Pix2pixHD on the Cityscapes dataset. Our results have finer details in the synthesized cars, the trees, the buildings, etc. Please zoom in for details.

column a and b, and we use this as an improved baseline. Then we conduct our ablation studies of proposed SIPL and CRB on the original baseline and improved baseline. Using SIPL in the training process decrease the FID around 0.4 comparing with column a and c, b and d, e and f. By observing column d and f, we could get extra 0.3 reduction using the newly-designed CRB to replace Res2Block in the generator. In the end, our proposed SemanticGAN could get 2.2 decrease totally. Through the above analysis we could draw a conclusion that both modifications are critical to the final effectiveness of the proposed model.

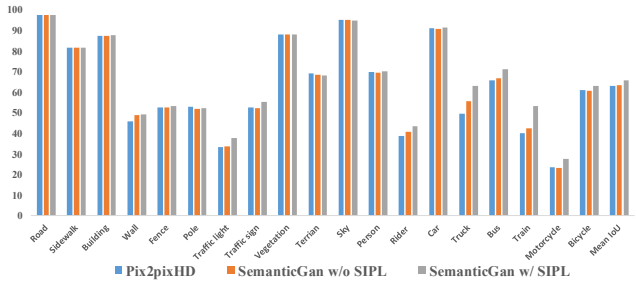

Fig. 6. The IoU of the synthesized images of Pix2pixHD, SemanticGAN with and without proposed SIPL.

Furthermore, we perform semantic segmentation on the synthesized images and score how well the predicted segments match the input semantic label maps, which has been widely used in recent works [3, 16, 17]. Fig. 6 reports the calculated intersection-over-union (IoU) of different methods. As can be seen, our SemanticGAN with SIPL outperforms the version without SIPL and Pix2pixHD by a large margin in many objects such as the traffic light, the truck, the bus, the train, etc. The mean IoU in the last groups of columns are 63.1%, 63.6% and 65.8%, respectively. We therefore conclude that training SemanticGAN using SIPL can maintain more semantic information and improve the qualities of the synthesized images.

Fig. 7. Comparison results with Pix2pixHD, SemanticGAN with and without proposed CRB. Please zoom in for details

Moreover, we study the importance of the proposed CRB qualitively. Fig. 7 presents some synthetic images from different methods. The results of Pix2pixHD have sharp edges but possess obvious artifacts and noise, and the textures within a small area are messy and uncorrelated (see from the red boxes). The introduction of CRB can solve this problem to a large extend. The results of our SemanticGAN demonstrate that using CRB in generator helps obtain the structural correlated information between textures and therefore guide the generator to synthesize satisfactory textures.

3.3. Comparison against State-of-the-arts

We compare our method with four state-of-the-art algorithms: CycleGAN [18], Pix2pix [2], CRN [19] and Pix2pixHD [3]. We train CycleGAN, Pix2pix, Pix2pixHD models on high-resolution images with the default setting. We produce the high-resolution CRN images via the authors publicly available model. The FID of SemanticGAN and other methods are shown in Tab. 2. It is obviously that the FID of SemanticGAN is lower than the state-of-the-art methods which means the synthesized images produced by our algorithm is more similar to the real images.

Table 2. The FID for our proposed SemanticGAN and state-of-the-art models on the Cityscapes dataset

	CycleGAN [18]	CRN [19]	Pix2pix [2]	Pix2pixHD [3]	SemanticGAN
FID	75.252	70.595	62.203	56.920	**54.775**

For qualitative evaluation, Fig. 5 shows examples generated by our SemanticGAN and the state-of-the-art models. It can be noted that the introduction of adversarial loss improves the visual performance over the regression-based architecture CRN but leads to artifacts. The introduction of our SIPL and CRB in SemanticGAN are able to tackle the artifacts and blurriness better, and achieve the best performance. Our results have finer details and more realistic textures in large scale objects such as the bus, the buildings, etc.

4. CONCLUSION

In this paper, we proposed a conditional GAN-based method called SematicGAN for S2PT task. In our network, a Semantic Information Preserved Loss (SIPL) is proposed to maintain semantic information. Furthermore, we design a novel generator using Correlated Residual Block (CRB) to obtain the structural correlations between the image textures. Detailed experiments and comparisons are performed on Cityscapes Dataset to demonstrate that our method significantly outperforms many recent state-of-the-art methods.

5. REFERENCES

[1] 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*, 2014, pp. 2672–2680.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[3] 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*, 2018, pp. 8798–8807.

[4] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr, "Res2net: A new multi-scale backbone architecture," *arXiv preprint arXiv:1904.01169*, 2019.

[5] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[6] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

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

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

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

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[11] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 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," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[13] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] Dim P Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba, "How to make a pizza: Learning a compositional layer-based gan model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8002–8011.

[17] Shuangting Liu, Jiaqi Zhang, Yuxin Chen, Yifan Liu, Zengchang Qin, and Tao Wan, "Pixel level data augmentation for semantic image segmentation using generative adversarial networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1902–1906.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

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