Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation

Hao Tang[1] Dan Xu[2] Yan Yan[3] Philip H. S. Torr[2] Nicu Sebe[1]
[1]University of Trento [2]University of Oxford [3]Texas State University

Abstract

In this paper, we address the task of semantic-guided scene generation. One open challenge in scene generation is the difficulty of the generation of small objects and detailed local texture, which has been widely observed in global image-level generation methods. To tackle this issue, in this work we consider learning the scene generation in a local context, and correspondingly design a local class-specific generative network with semantic maps as a guidance, which separately constructs and learns sub-generators concentrating on the generation of different classes, and is able to provide more scene details. To learn more discriminative class-specific feature representations for the local generation, a novel classification module is also proposed. To combine the advantage of both the global image-level and the local class-specific generation, a joint generation network is designed with an attention fusion module and a dual-discriminator structure embedded. Extensive experiments on two scene image generation tasks show superior generation performance of the proposed model. The state-of-the-art results are established by large margins on both tasks and on challenging public benchmarks. The source code and trained models are available at https://github.com/Ha0Tang/LGGAN.

1. Introduction

In this work, we focus on semantic-guided scene generation, which is a hot research topic covering several main-stream research directions, including cross-view image translation [21, 52, 36, 37, 43, 38] and semantic image synthesis [48, 8, 34, 32]. The cross-view image translation task proposed in [36] is essentially an ill-posed problem due to the large ambiguity in the generation if only a single RGB image is given as input. To alleviate this problem, recent works such as SelectionGAN [43] try to generate the target image based on an image of the scene and several novel semantic maps, as shown in Fig. 1 (bottom). Adding a semantic map allows the model to learn the correspondences in the target view with appropriate object relations and transformations. On the other side, the semantic image synthesis task aims to generate a photo-realistic image from a semantic map [48, 8, 34, 32], as shown in Fig. 1 (top). Recently, Park et al. [32] propose a spatially-adaptive normalization for synthesizing photo-realistic images given an input semantic map. With the useful semantic information, existing methods on both tasks achieved promising performance in scene generation.

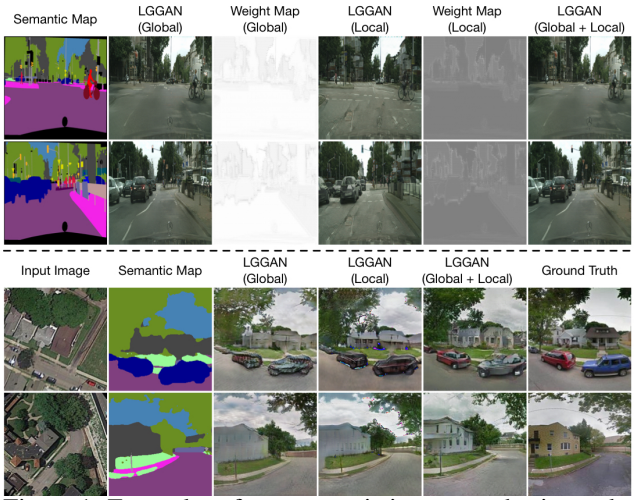

Figure 1: Examples of our semantic image synthesis results on the Cityscapes dataset (top) and our cross-view image translation results on the Dayton dataset (bottom) with different settings of the proposed LGGAN.

However, one can still observe unsatisfying perspectives, especially on the generation of local scene structure and details as well as small scale objects, which we believe are mainly due to several reasons. First, existing methods on both tasks are mostly based on a global image-level generation, which accepts a semantic map containing several object classes and aims to generate the appearance of all the different classes, by using the same network design or using shared network parameters. In this case, all the classes are treated equally by the network. While different semantic classes have distinct properties, specific network learning for different semantic classes intuitively would benefit the complex multi-class generation. Second, we observe that

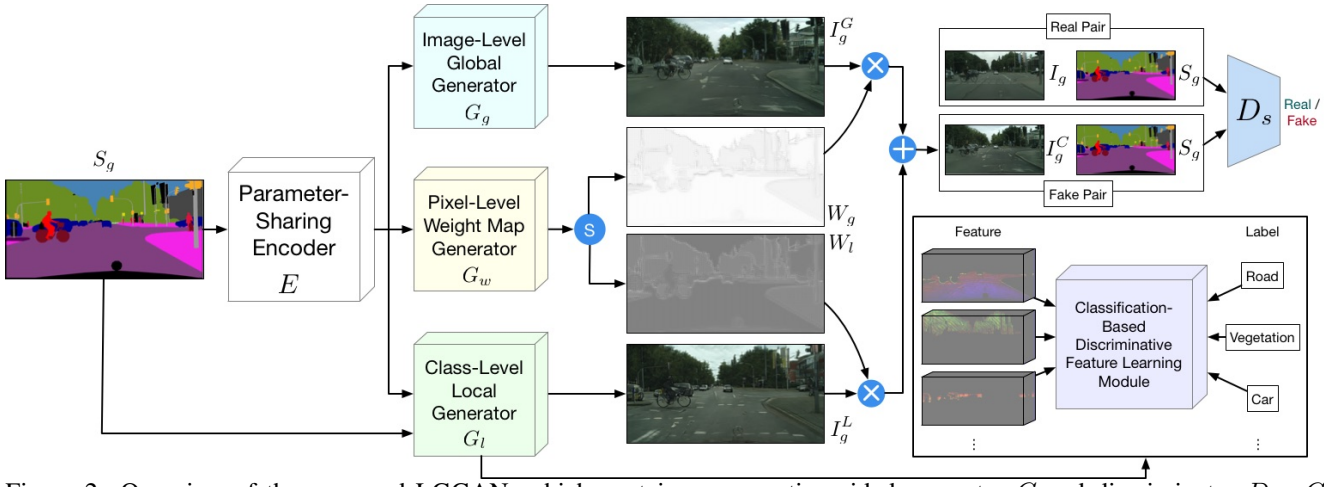

Figure 2: Overview of the proposed LGGAN, which contains a semantic-guided generator G and discriminator D_s. G consists of a parameter-sharing encoder E, an image-level global generator G_g, a class-level local generator G_l and a weight map generator G_w. The global generator and local generator are automatically combined by two learned weight maps from the weight map generator to reconstruct the target image. D_s tries to distinguish the generated images from two modality spaces, i.e., image space and semantic space. Moreover, to learn a more discriminative class-specific feature representation, a novel classification module is proposed. All of these components are trained in an end-to-end fashion so that the local generation and the global generation can benefit from each other. The symbols $\oplus$, $\otimes$ and Ⓢ denote element-wise addition, element-wise multiplication and channel-wise Softmax, respectively.

the number of training samples of different scene classes is imbalanced. For instance, for the Dayton dataset [46], the cars and buses only occupy less than 2% with respect to all pixels in the training data, which naturally makes the model learning be dominated by the classes with the larger number of training samples. Third, the size of objects in different scene classes is diverse. As shown in the first row of Fig. 1, larger-scale object classes such as road, sky usually occupy bigger area of the image than smaller-scale classes such as pole and traffic light. Since the convolutional network usually shares the parameters at different convolutional positions, the larger-scale object classes would thus take advantage during the learning, further increasing the difficult in generating well the small-scale object classes.

To tackle these issues, a straightforward consideration would be to model the generation of different scene classes specifically in a local context. By so doing, each class could have its own generation network structure or parameters, thus greatly avoiding the learning of a biased generation space. To achieve this goal, in this paper we design a novel class-specific generation network. It consists of several sub-generators for different scene classes with a shared encoded feature map. The input semantic map is utilized as the guidance to obtain feature maps corresponding to each class spatially, which are then used to produce a separate generation for different class regions.

Due to the highly complementary properties of global and local generation, a Local class-specific and Global image-level Generative Adversarial Network (LGGAN) is proposed to combine the advantage of these two. It mainly contains three network branches (see Fig. 2). The first branch is the image-level global generator, which learns a global appearance distribution using the input, and the second branch is the proposed class-specific local generator, which aims to generate different objects classes separately using semantic-guided class-specific feature filtering. Finally, the fusion weight-map generation branch learns two pixel-level weight maps which are used to fuse the local and global sub-networks in a weighted-combination of their final generation results. The proposed LGGAN can be jointly trained in an end-to-end fashion to make the local and global generation benefit each other in the optimization.

Overall, the contributions of this paper are as follows:

- We explore scene generation from the local context, which we believe is beneficial to generate richer scene details compared with the existing global image-level generation methods. A new local class-specific generative structure has been designed for this purpose. It can effectively handle the generation of small objects and scene details which are common difficulties encountered by the global-based generation.
- We propose a novel global and local generative adversarial network design able to take into account both the global and local contexts. To stabilize the optimization of the proposed joint network structure, a fusion weight-map generator and a dual-discriminator are introduced. Moreover, to learn discriminative class-specific feature representations, a novel classification module is proposed.
- Experiments for cross-view image translation on the Dayton [46] and CVUSA [49] datasets, and semantic im-

age synthesis on the Cityscapes [11] and ADE20K [56] datasets demonstrate the effectiveness of the proposed LGGAN framework, and show significantly better results compared with state-of-the-art methods on both tasks.

2. Related Work

Generative Adversarial Networks (GANs) [15] have been widely used for image generation [23, 53, 7, 24, 17, 13, 40, 27, 39]. A vanilla GAN has two important components, i.e., a generator and a discriminator. The goal of the generator is to generate photo-realistic images from a noise vector, while the discriminator is trying to distinguish between the real and the generated image. To synthesize user-specific images, Conditional GAN (CGAN) [29] has been proposed. A CGAN combines a vanilla GAN and an external information, such as class labels [30, 31, 9], text descriptions [35, 54], object keypoint [35], human body/hand skeleton [1, 42, 3, 59], conditional images [58, 21], semantic maps [48, 43, 32, 47], scene graphs [22, 55, 2] and attention maps [53, 28, 41].

Global and Local Generation in GANs. Modelling global and local information in GANs to generate better results has been used in various generative tasks [19, 20, 26, 25, 33, 16]. For instance, Huang et al. [19] propose TPGAN for frontal view synthesis by simultaneously perceiving global structures and local details. Gu et al. [16] propose MaskGAN for face editing by separately learning every face component, e.g., mouth and eye. However, these methods are only applied to face-related tasks such as face rotation or face editing, where all the domains have large overlap and similarity. However, we propose a new local and global image generation framework design for a more challenging scene generation task, and the local context modeling is based on semantic-guided class-specific generation, which is not explored by any existing works.

Scene Generation. Scene generation tasks are a hot topic as each image can be parsed into distinctive semantic objects [6, 2, 45, 14, 4, 5]. In this paper, we mainly focus on two scene generation tasks, i.e., cross-view image translation [52, 36, 37, 43] and semantic image synthesis [48, 8, 34, 32]. Most existing works on cross-view image translation have been conducted to synthesize novel views of the same objects [12, 57, 44, 10]. Moreover, several works deal with image translation problems with drastically different views and generate a novel scene from a given different scene [52, 36, 37, 43]. For instance, Tang et al. [43] propose SelectionGAN to solve the cross-view image translation task using semantic maps and CGAN models. On the other side, the semantic image synthesis task aims to generate a photo-realistic image from a semantic map [48, 8, 34, 32]. For example, Park et al. propose GauGAN [32], which achieves the best results on this task.

With the semantic maps as guidance, existing approaches on both tasks achieve promising performance. However, we still observe that the results produced these global image-level generation methods are often unsatisfactory, especially on detailed local texture. In contrast, our proposed approach focuses on generating more realistic global structure/layout and local texture details. Both local and global generation branches are jointly learned in an end-to-end fashion that aims at using the mutually improved benefits from each other.

3. The Proposed LGGAN

We start by presenting the details of the proposed Local class-specific and Global image-level GANs (LGGAN). An illustration of the overall framework is shown in Fig. 2. The generation module mainly consists of three parts, i.e., a semantic-guided class-specific generator modelling the local context, an image-level generator modelling the global layout, and a weight-map generator for fusing the local and the global generators. We first introduce the used backbone structure, and then present the design of the proposed local and global generation networks.

3.1. The Backbone Encoding Network Structure

Semantic-Guided Generation. In this paper, we mainly focus on two tasks, i.e., semantic image synthesis and cross-view image translation. For the former, we follow GauGAN [32] and use the semantic map S_g as the input of the backbone encoder E, as shown in Fig. 2. For the latter, we follow SelectionGAN [43] and concatenate the input image I_a and a novel semantic map S_g as the input of the backbone encoder E. By so doing, the semantic maps act as priors to guide the model to learn the generation of another domain.

Parameter-Sharing Encoder. As we have three different branches for three different generators, the encoder E is sharing parameters to all the three branches to make a compact backbone network. The gradients from all the three branches contribute together to the learning of the encoder. We believe that in this way, the encoder can learn both local and global information and the correspondence between them. Then the encoded deep representations from the input S_g can be represented as $E(S_g)$, as shown in Fig. 2.

3.2. The LGGAN Structure

Class-Specific Local Generation Network. As shown in Fig. 1 and discussed in the introduction, the issue of training data imbalance between different classes and size difference between scene objects makes it extremely difficult in generation of small object classes and scene details. To overcome this limitation, we propose a novel local class-specific generation network design. It separately constructs a generator for each semantic class and thus is able to largely avoid the interference from the large object classes in the joint optimization. Each sub-generation branch has independent net-

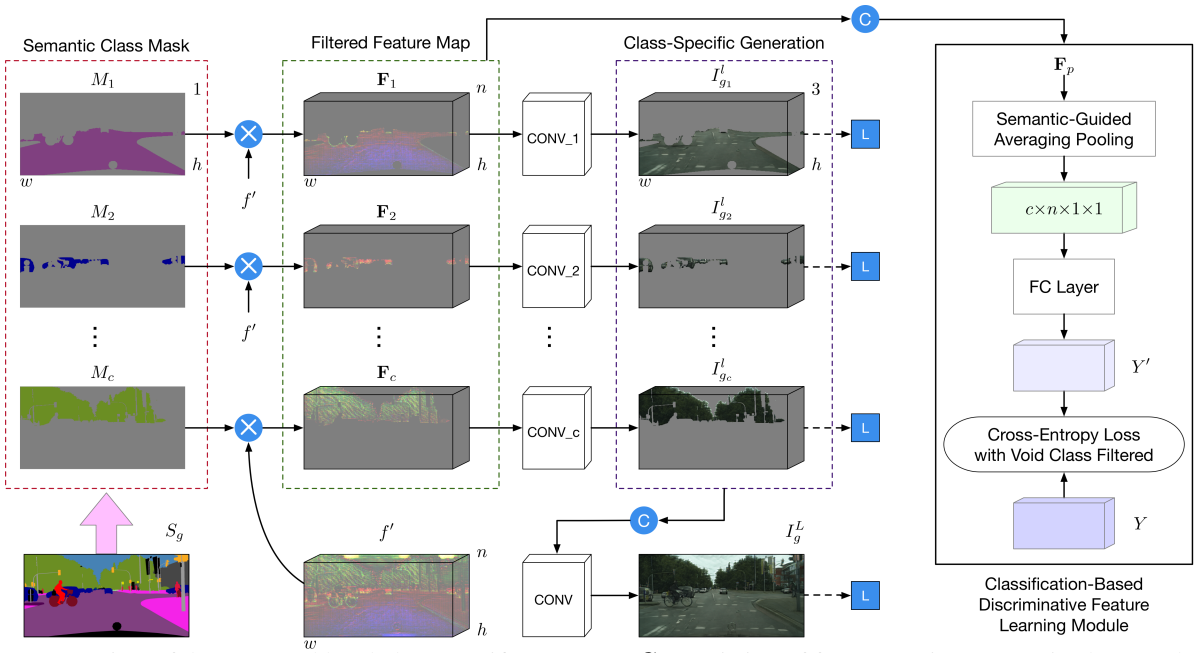

Figure 3: Overview of the proposed local class-specific generator G_l consisting of four steps, i.e., semantic class mask calculation, class-specific feature map filtering, classification-based discriminative feature learning and class-specific generation. A cross-entropy loss with void class filtered is applied at each class feature representation for learning a more discriminative class-specific feature representation. A semantic-mask guided pixel-wise $L1$ loss is applied at the end for class-level reconstruction. The symbols $\otimes$ and $\copyright$ denote element-wise multiplication and channel-wise concatenation, respectively.

work parameters and concentrates on a specific class, being therefore capable of effectively producing similar generation quality for different classes and yielding richer local scene details.

The overview of the local generation network G_l is illustrated in Fig. 3. The encoded features $E(S_g)$ are first fed into two consecutive deconvolutional layers to increase the spatial size with the number of channels reduced two times. Then the scaled feature map f' is multiplied by the semantic mask of each class, i.e., M_i, to obtain a filtered class-specific feature map for each one. The mask-guided feature filtering operation can be written as:

$$\mathbf{F}_i = M_i * f', \quad i = 1, 2, ..., c, \tag{1}$$

where c is the number of semantic classes. Then the filtered feature map $\mathbf{F}_i$ is fed into several convolutional layers for the corresponding i-th class and generate an output image $I^l_{g_i}$. For better learning each class, we utilize a semantic-mask guided pixel-wise $L1$ reconstruction loss, which can be expressed as follows:

$$\mathcal{L}^{local}_{\text{L1}} = \sum_{i=1}^{c} \mathbb{E}_{I_g, I^l_{g_i}}[||I_g * M_i - I^l_{g_i}||_1]. \tag{2}$$

The final output I^L_g from the local generation network can be obtained in two ways. The first one is performing an element-wise addition of all the class-specific outputs:

$$I^L_g = I^l_{g_1} \oplus I^l_{g_2} \oplus \cdots \oplus I^l_{g_c}. \tag{3}$$

The second one is performing a convolutional operation on all the class-specific outputs, as shown in Fig. 3,

$$I^L_g = \text{Conv}(\text{Concat}(I^l_{g_1}, I^l_{g_2}, \cdots, I^l_{g_c})), \tag{4}$$

where $\text{Concat}(\cdot)$ and $\text{Conv}(\cdot)$ denote channel-wise concatenation and convolutional operation, respectively.

Class-Specific Discriminative Feature Learning. We observe that the filtered feature map $\mathbf{F}_i$ is not able to produce very discriminative class-specific generations, leading to similar generation results for some classes, especially for small-scale object classes. In order to have more diverse generation for different object classes, we propose a novel classification-based feature learning module to learn more discriminative class-specific feature representations, as shown in Fig. 3. One input sample of the module is a pack of feature maps produced from different local generation branches, i.e., $\{\mathbf{F}_1, ..., \mathbf{F}_c\}$. First, the packed feature map $\mathbf{F}_p \in \mathbb{R}^{c \times n \times h \times w}$ (with n, h, w as the number of feature map channels, height and width, respectively) is fed into a semantic-guided averaging pooling layer, and we obtain a pooled feature map with dimension of $c \times n \times 1 \times 1$. Then the pooled feature map is connected with a fully connected layer to predict classification probability of the c object classes of the scene. Since some object classes may not exist in the input semantic mask sample, the features from the local branches corresponding to the void classes should not contribute to the classification loss. Therefore,

Figure 4: Qualitative comparison of different methods in a2g direction on the Dayton dataset.

Table 1: Quantitative evaluation of the Dayton dataset in the a2g direction. For all metrics except KL score, higher is better. $(*)$ Inception Score for real (ground truth) data is 3.8319, 2.5753 and 3.9222 for all, top-1 and top-5 setups, respectively.

Method	Accuracy (%)				Inception Score*			SSIM	PSNR	SD	KL
	Top-1		Top-5		All	Top-1	Top-5				
Pix2pix [21]	6.80	9.15	23.55	27.00	2.8515	1.9342	2.9083	0.4180	17.6291	19.2821	38.26 ± 1.88
X-SO [37]	27.56	41.15	57.96	73.20	2.9459	2.0963	2.9980	0.4772	19.6203	19.2939	7.20 ± 1.37
X-Fork [36]	30.00	48.68	61.57	78.84	3.0720	2.2402	3.0932	0.4963	19.8928	19.4533	6.00 ± 1.28
X-Seq [36]	30.16	49.85	62.59	80.70	2.7384	2.1304	2.7674	0.5031	20.2803	19.5258	5.93 ± 1.32
Pix2pix++ [21]	32.06	54.70	63.19	81.01	3.1709	2.1200	3.2001	0.4871	21.6675	18.8504	5.49 ± 1.25
X-Fork++ [36]	34.67	59.14	66.37	84.70	3.0737	2.1508	3.0893	0.4982	21.7260	18.9402	4.59 ± 1.16
X-Seq++ [36]	31.58	51.67	65.21	82.48	3.1703	2.2185	3.2444	0.4912	21.7659	18.9265	4.94 ± 1.18
SelectionGAN [43]	42.11	68.12	77.74	92.89	3.0613	2.2707	3.1336	**0.5938**	**23.8874**	**20.0174**	2.74 ± 0.86
LGGAN (Ours)	**48.17**	**79.35**	**81.14**	**94.91**	**3.3994**	**2.3478**	**3.4261**	0.5457	22.9949	19.6145	**2.18 ± 0.74**

we filter the final cross-entropy loss by multiplying it with a void class indicator for each input sample. The indicator is an one hot vector $H=\{H_i\}_{i=1}^{c}$ with $H_i=1$ for a valid class and $H_i=0$ for a void class. Then, the Cross-Entropy (CE) loss is defined as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{m=1}^{c} H_m \sum_{i=1}^{c} 1\{Y(i)=i\}\log(f(\mathbf{F}_i)) \qquad (5)$$

where $1\{\cdot\}$ is an indicator function, i.e., having a return 1 if $Y(i)=i$ else 0. $f(\cdot)$ is a classification function which produces a prediction probability given an input feature map $\mathbf{F}_i$. Y is a label set of all the object classes.

Image-Level Global Generation Network. Similar to the local generation branch, $E(S_g)$ is also fed into the global generation sub-network G_g for global image-level generation, as shown in Fig. 2. Global generation is capable to capture the global structure information or layout of the input images. Thus, the global result I_g^G can be obtained through a feed-forward computation: $I_g^G=G_g(E(S_g))$. Besides the proposed G_g, many existing global generator architectures can also be used with the proposed local generator G_l, making the proposed framework very flexible.

Pixel-Level Fusion Weight-Map Generation Network. In order to better combine the local and the global generation sub-networks, we further propose a pixel-level weight map generator G_w, which aims at predicting pixel-wise weights for fusing the global generation I_g^G and the local generation I_g^L. In our implementation, G_g consists of two Transpose Convolution→InstanceNorm→ReLU blocks and one Convolution→InstanceNorm→ReLU block. The number of the output channels for these three block are 128, 64 and 2, respectively. The kernel sizes are 3×3 with stride 2 except for the last layer which has a kernel size of 1×1 with stride 1 for dense prediction. We predict a two-channel weight map W_f using the following calculation:

$$W_f = \text{Softmax}(G_w(E(S_g))), \qquad (6)$$

where $\text{Softmax}(\cdot)$ denotes a channel-wise softmax function used for normalization, i.e., the sum of the weight values at the same pixel position is equal to 1. By so doing, we can guarantee that information from the combination would not explode. W_f is sliced to have a weight map W_g for the global branch and a weight map W_l for the local branch. The fused final generation result is calculated as follows:

$$I_g^C = I_g^G \otimes W_g + I_g^L \otimes W_l, \qquad (7)$$

where $\otimes$ is an element-wise multiplication operation. In this way, the pixel-level weights predicted from G_w directly operate on the output of G_g and G_l. Moreover, generators G_w, G_g and G_l affect and contribute to each other in the model optimization.

Dual-Discriminator. To exploit the prior domain knowledge, i.e., the semantic map, we extend the single domain vanilla discriminator [15] to a cross domain structure and

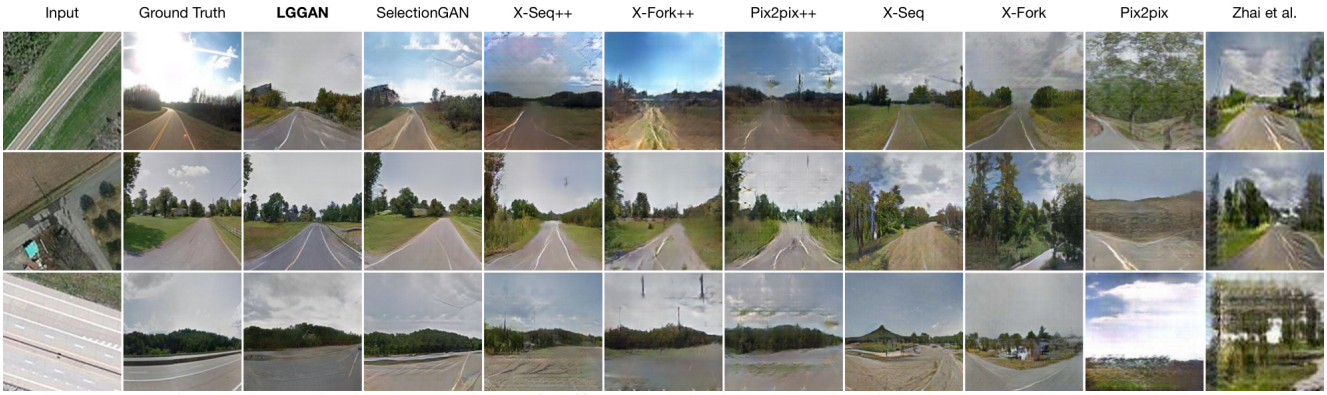

Figure 5: Qualitative comparison of different methods in a2g direction on the CVUSA dataset.

Table 2: Quantitative evaluation of the CVUSA dataset in a2g direction. For all metrics except KL score, higher is better. $(*)$ Inception Score for real (ground truth) data is 4.8741, 3.2959 and 4.9943 for all, top-1 and top-5 setups, respectively.

Method	Accuracy (%)				Inception Score*			SSIM	PSNR	SD	KL
	Top-1		Top-5		All	Top-1	Top-5				
Zhai et al. [52]	13.97	14.03	42.09	52.29	1.8434	1.5171	1.8666	0.4147	17.4886	16.6184	27.43 ± 1.63
Pix2pix [21]	7.33	9.25	25.81	32.67	3.2771	2.2219	3.4312	0.3923	17.6578	18.5239	59.81 ± 2.12
X-SO [37]	0.29	0.21	6.14	9.08	1.7575	1.4145	1.7791	0.3451	17.6201	16.9919	414.25 ± 2.37
X-Fork [36]	20.58	31.24	50.51	63.66	3.4432	2.5447	3.5567	0.4356	19.0509	18.6706	11.71 ± 1.55
X-Seq [36]	15.98	24.14	42.91	54.41	3.8151	2.6738	**4.0077**	0.4231	18.8067	18.4378	15.52 ± 1.73
Pix2pix++ [21]	26.45	41.87	57.26	72.87	3.2592	2.4175	3.5078	0.4617	21.5739	18.9044	9.47 ± 1.69
X-Fork++ [36]	31.03	49.65	64.47	81.16	3.3758	2.5375	3.5711	0.4769	21.6504	18.9856	7.18 ± 1.56
X-Seq++ [36]	34.69	54.61	67.12	83.46	3.3919	2.5474	3.4858	0.4740	21.6733	18.9907	5.19 ± 1.31
SelectionGAN [43]	41.52	65.51	74.32	89.66	3.8074	2.7181	3.9197	**0.5323**	**23.1466**	19.6100	2.96 ± 0.97
LGGAN (Ours)	**44.75**	**70.68**	**78.76**	**93.40**	**3.9180**	**2.8383**	3.9878	0.5238	22.5766	**19.7440**	$\mathbf{2.55 \pm 0.95}$

we refer to it as the semantic-guided discriminator D_s, as shown in Fig. 2. It employs the input semantic map S_g and the generated image I_g^C (or the real image I_g) as input:

$$\begin{aligned}\mathcal{L}_{\text{CGAN}}(G, D_s) =& \mathbb{E}_{S_g, I_g}\left[\log D_s(S_g, I_g)\right] + \\ & \mathbb{E}_{S_g, I_g^C}\left[\log(1 - D_s(S_g, I_g^C))\right],\end{aligned} \tag{8}$$

which aims to preserve scene layout and capture the local-aware information.

For the cross-view image translation task, we also propose another image-guided discriminator D_i, which takes the conditional image I_a and the final generated image I_g^C (or the ground-truth image I_g) as input:

$$\begin{aligned}\mathcal{L}_{\text{CGAN}}(G, D_i) =& \mathbb{E}_{I_a, I_g}\left[\log D_i(I_a, I_g)\right] + \\ & \mathbb{E}_{I_a, I_g^C}\left[\log(1 - D_i(I_a, I_g^C))\right].\end{aligned} \tag{9}$$

In this case, the total loss of our Dual-Discriminator D is $\mathcal{L}_{\text{CGAN}} = \mathcal{L}_{\text{CGAN}}(G, D_i) + \mathcal{L}_{\text{CGAN}}(G, D_s)$.

4. Experiments

The proposed LGGAN can be applied to different generative tasks such as the cross-view image translation [43] and the semantic image synthesis [32]. In this section we present experimental results and analysis on both tasks.

4.1. Results on Cross-View Image Translation

Datasets. We follow [43, 36] and perform the cross-view image translation experiments on the Dayton [46] and CVUSA datasets [49]. The Dayton dataset contains 76,048 images with a train/test split of 55,000/21,048 pairs. The CVUSA dataset consists of 35,532/8,884 image pairs in train/test split.

Evaluation Metric. Similarly to [36, 37, 43], we employ Inception Score (IS), Accuracy (Acc.), KL Divergence Score (KL) to evaluate the proposed model. These three metrics evaluate the distance between two different distributions from a high-level feature space. We also employ pixel-level similarity metrics to evaluate our method, i.e., Structural-Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Sharpness Difference (SD).

State-of-the-Art Comparisons. We compare our LGGAN with several recently proposed state-of-the-art methods, i.e., Zhai et al. [52], Pix2pix [21], X-SO [37], X-Fork [36] and X-Seq [36]. The comparison results are shown in Tables 1 and 2. We can observe that LGGAN consistently outperforms the competing methods on all metrics.

To study the effectiveness of LGGAN, we conduct experiments with the methods using semantic maps and RGB images as input, including Pix2pix++ [21], X-Fork++ [36], X-Seq++ [36] and SelectionGAN [43]. We implement Pix2pix++, X-Fork++ and X-Seq++ using their public source code. Results are shown in Tables 1 and 2. We ob-

Figure 6: Results generated by different methods on the Cityscapes dataset. The proposed LGGAN produces realistic images while respecting the spatial semantic layout at the same time. These samples were randomly selected without cherry-picking for visualization purposes.

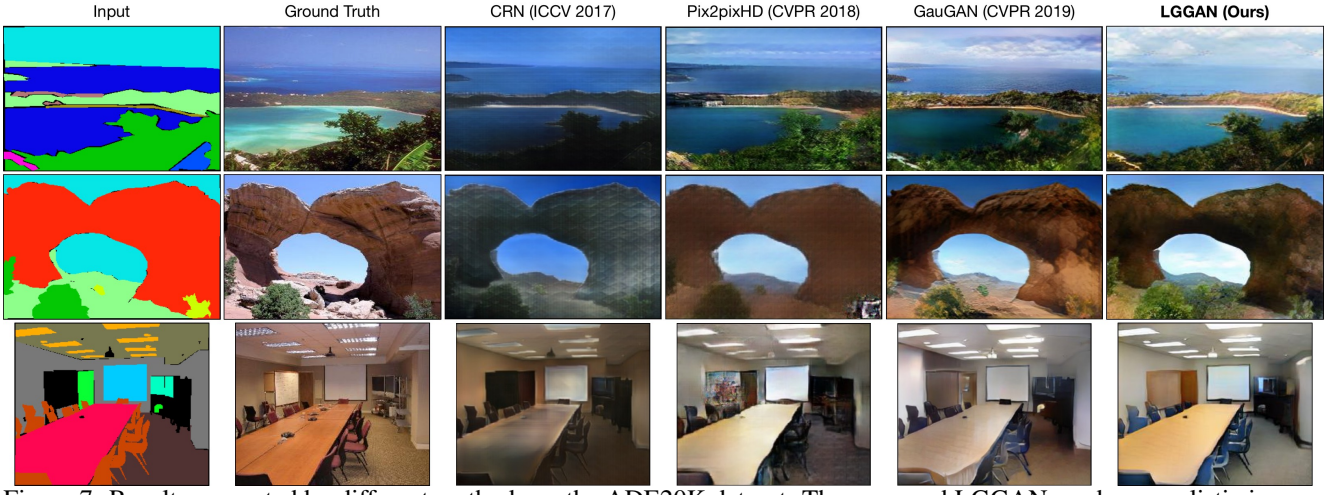

Figure 7: Results generated by different methods on the ADE20K dataset. The proposed LGGAN produces realistic images while respecting the spatial semantic layout at the same time. These samples were randomly selected without cherry-picking for visualization purposes.

serve that LGGAN achieves significantly better results than Pix2pix++, X-Fork++ and X-Seq++, confirming the advantage of the proposed LGGAN. A direct comparison with SelectionGAN is also shown in the tables providing better results on most metrics except pixel-level evaluation metrics, i.e., SSIM, PSNR and SD. SelectionGAN uses a two-stage generation strategy and an attention selection module, achieving slightly better results than ours on these three metrics. However, we generate much more photo-realistic results than SelectionGAN as shown in Fig. 4 and 5.

Qualitative Evaluation. The qualitative results are shown in Fig. 4 and 5. We observe that the generated results of LGGAN are visually significantly better than other approaches. It can be seen that our method generates more clear details on objects such as cars, buildings, road, trees than the other methods in the generated images.

4.2. Results on Semantic Image Synthesis

Datasets. We follow GauGAN [32] and conduct extensive experiments on both Cityscapes [11] and ADE20K [56] datasets. Cityscapes contains street scenes in German cities. The training and testing set sizes of Cityscapes are 2,975 and 500, respectively. To evaluate the proposed LGGAN on more challenging datasets, we conduct experiments on the ADE20K dataset [56]. This dataset contains challenging scenes with 150 semantic classes, and has 20,210 training and 2,000 validation images.

Evaluation Metric. We adopt the same evaluation metrics from previous work [8, 32, 48], and use the mean Intersection-over-Union (mIoU) and pixel accuracy (Acc) to measure the segmentation accuracy. Specifically, we follow GauGAN [32] and use the state-of-the-art segmentation networks on the generated images to produce se-

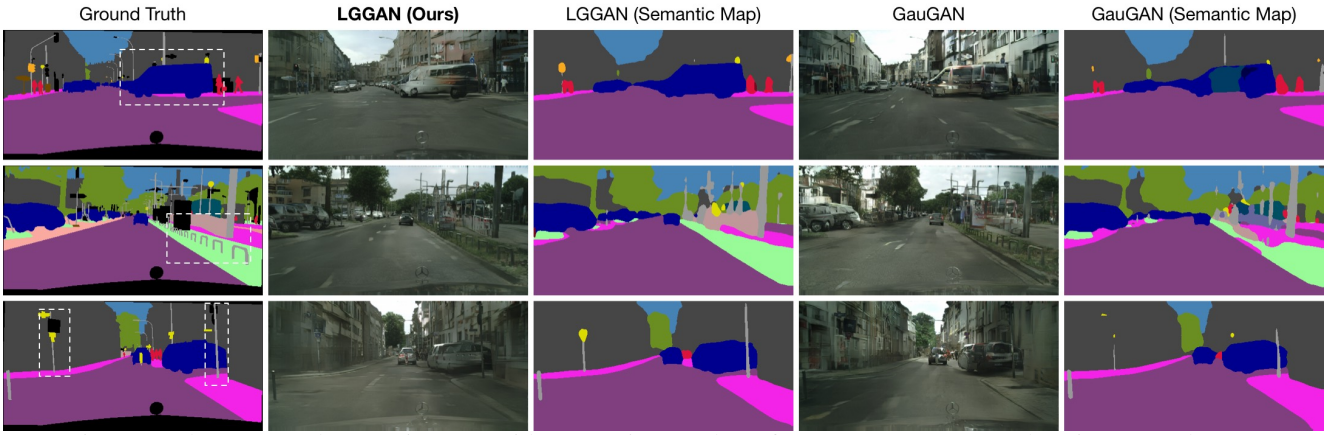

Figure 8: The generated semantic maps with comparison to those from GauGAN [32] on the Cityscapes dataset.

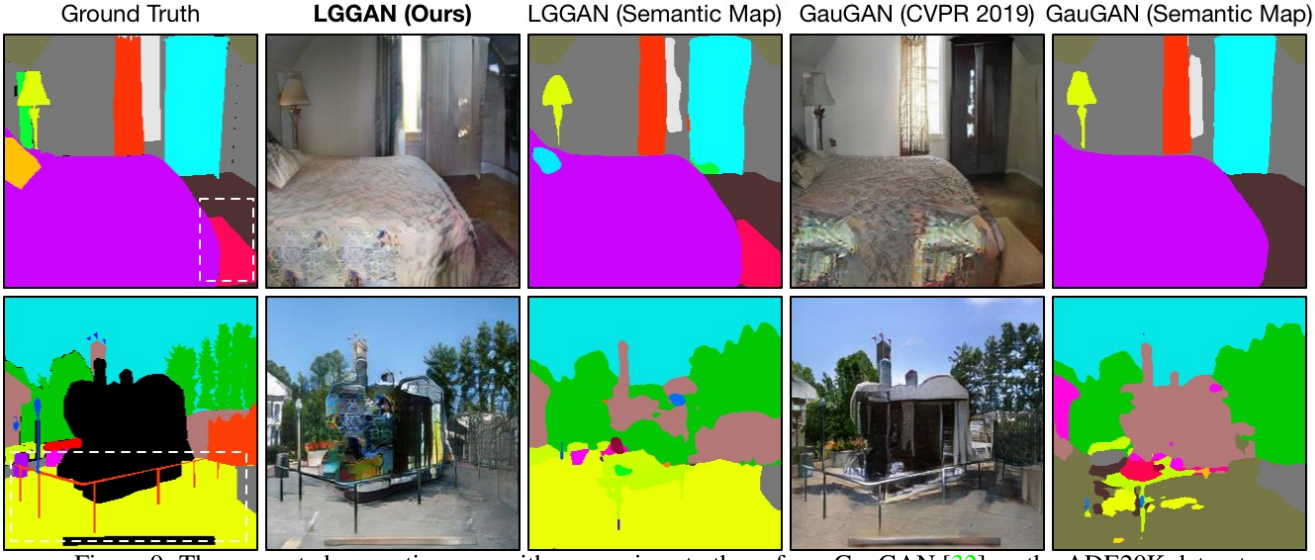

Figure 9: The generated semantic maps with comparison to those from GauGAN [32] on the ADE20K dataset.

Table 3: Our method significantly outperforms current leading methods in semantic segmentation scores (mIoU), pixel accuracy (Acc) and FID. For mIoU, higher is better. For FID, lower is better.

Method	Cityscapes			ADE20K		
	mIoU ↑	Acc ↑	FID ↓	mIoU ↑	Acc ↑	FID ↓
CRN [8]	52.4	77.1	104.7	22.4	68.8	73.3
SIMS [34]	47.2	75.5	**49.7**	N/A	N/A	N/A
Pix2pixHD [48]	58.3	81.4	95.0	20.3	69.2	81.8
GauGAN [32]	62.3	81.9	71.8	38.5	79.9	33.9
LGGAN (Ours)	**68.4**	**83.0**	57.7	**41.6**	**81.8**	**31.6**

mantic maps: DRN-D-105 [51] for Cityscapes and UperNet101 [50] for ADE20K. We also use the Fréchet Inception Distance (FID) [18] to measure the distance between the distribution of generated samples and the distribution of real samples. Finally, we follow [32] and employ Amazon Mechanical Turk (AMT) to measure the perceived visual fidelity of the generated images.

State-of-the-Art Comparisons. We compare the proposed

Table 4: User preference study. The numbers indicate the percentage of users who favor the results of the proposed method over the competing method.

AMT ↑	Cityscapes	ADE20K
Ours vs. CRN [8]	67.38	79.54
Ours vs. Pix2pixHD [48]	56.16	85.69
Ours vs. SIMS [34]	54.84	N/A
Ours vs. GauGAN [32]	53.19	57.31

LGGAN with several leading semantic image synthesis methods, i.e., Pix2pixHD [48], CRN [8], SIMS [34] and GauGAN [32]. Results of the mIoU, Acc and FID metrics are shown in Table 3. We find that the proposed LGGAN outperforms the existing competing methods by a large margin on both mIoU and Acc metrics. For FID, the proposed method is only worse than SIMS on Cityscapes. However, SIMS has poor segmentation performance. The reason is that SIMS produces an image by searching and copying image patches from the training dataset. The generated images are more realistic since the method uses the real image

Figure 10: Results and weight maps generated by the proposed LGGAN with different settings on the Cityscapes dataset.

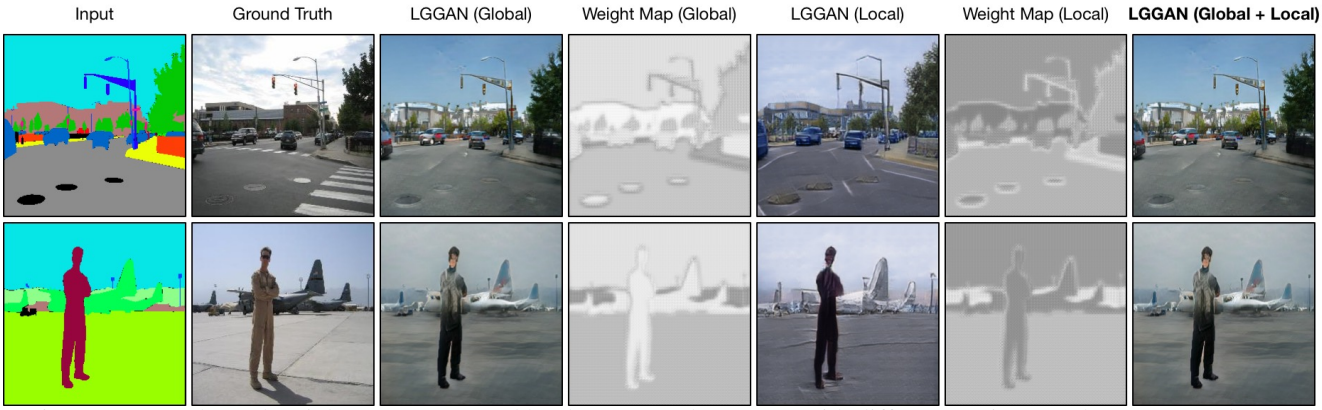

Figure 11: Results and weight maps generated by the proposed LGGAN with different settings on the ADE20K dataset.

patches. However, the approach always tends to copy objects with mismatched patches due to queries that cannot be guaranteed to have results in the dataset. Moreover, we follow the evaluation protocol of GauGAN and also provide AMT results, as shown in Table 4. We observe that users favor our synthesized results on both datasets compared with other competing methods including SIMS.

Qualitative Evaluation. The qualitative comparison results are shown in Fig. 6 and 7. We can see that the proposed method generates much better results with fewer visual artifacts while the spatial semantic layout of the generated images is also closer to the input semantic map.

Visualization of Generated Semantic Maps. We follow GauGAN [32] and apply pretrained segmentation networks on the generated images to produce semantic maps. The generated semantic maps of our LGGAN, GauGAN and the ground truths are shown in Fig. 8 and 9. We observe that the proposed LGGAN generates better semantic maps than GauGAN, especially on local texture ('car' in the first row and 'terrain' in the second row of Fig. 8) and small objects ('traffic sign' and 'pole' in the third row of Fig. 8), confirming our initial motivation.

4.3. Ablation Study

We conduct extensive ablation studies on the Cityscapes dataset to evaluate different components of our LGGAN.

Baseline Models. The proposed LGGAN has five baselines as shown in Table 5: (i) 'w/ Global' means only adopting the global generator; (ii) 'w/ Global + Local (Add.)' combines the global generator and the proposed local generator to produce the final results, in which the local results are produced by using an addition operation as proposed in Eq. (3). (iii) The difference between 'w/ Global + Local (Con.)' and the previous model is that it uses a convolutional layer to generate the local results as presented in Eq. (4). (iv) 'w/ Global + Local (Con.) + Class. Loss' employ the proposed classification-based discriminative feature learning module. (v) 'w/ Global + Local (Con.) + Class. Loss + Weight Map' is our full model and adopts the proposed weight map fusion strategy.

Table 5: Quantitative comparison of different variants of the proposed LGGAN on the semantic image synthesis tasks.

Setup of LGGAN	mIoU ↑	FID ↓
w/ Global	62.3	71.8
w/ Global + Local (Add.)	64.6	66.1
w/ Global + Local (Con.)	65.8	65.6
w/ Global + Local (Con.) + Class. Loss	67.0	61.3
w/ Global + Local (Con.) + Class. Loss + Weight Map	**68.4**	**57.7**

Effect of Local and Global Generation. The results of the ablation study are shown in Table 5. When using an addition operation to generate the local result, the local and global generation strategy improves mIoU and FID by 2.3 and 5.7, respectively. When adopting a convolutional operation to produce the local results, the performance boosts further, i.e, 3.5 and 6.2 gain on the mIoU and FID metrics, respectively. Both results confirm the effectiveness of the pro-

posed local and global generation framework. Moreover, we also provide qualitative results of the local and global generation in Fig. 1, 10 and 11. We observe that our full model, i.e., Global + Local, generates visually much better results than both the global and local method.

Effect of Classification-Based Feature Learning. 'w/ Global + Local (Con.) + Class. Loss' significantly outperforms the previous baseline with around 1.2 and 4.3 gain on the mIoU and FID metric, respectively. This means that the model indeed learns a more discriminative class-specific feature representation, confirming our design motivation.

Effect of Weight Map Fusion. By adding the proposed weight map fusion scheme, the overall performance is further boosted with 1.4 and 3.6 improvement on the mIoU and FID metric, respectively. This means the proposed LGGAN indeed learns complementary information from the local and the global generation branch. In Fig. 1, 10 and 11, we show some samples of the generated global and local weight maps. We observe that the generated global weight maps mainly focus on learning the global layout and structure, while the learned local weight maps focus on the local details, especially the connection between different classes.

5. Conclusion

We proposed Local class-specific and Global image-level Generative Adversarial Networks (LGGAN) for semantic-guided scene generation. The proposed LGGAN contains three generation branches, i.e., global image-level generation, local class-level generation and pixel-level fusion weight map generation, respectively. A new class-specific local generation network is designed to alleviate the influence of imbalanced training data and size difference of scene objects in joint learning. To learn more class-specific discriminative feature representations, a novel classification module is further proposed. To stabilize the model optimization, we further introduce a novel dual-discriminator, so that the synthesis results are not only visually appealing but also preserve the semantic layout. Experimental results demonstrate the superiority of the proposed approach and show new state-of-the-art results on both cross-view image translation and semantic image synthesis tasks.

References

[1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *ICCV*, 2019. 3

[2] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *ICCV*, 2019. 3

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 3

[4] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM TOG*, 38(4):59, 2019. 3

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 3

[6] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *ICCV*, 2019. 3

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 3

[8] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *CVPR*, 2017. 1, 3, 7, 8

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[10] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 3

[11] 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 *CVPR*, 2016. 3, 7

[12] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE TPAMI*, 39(4):692–705, 2017. 3

[13] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. 3

[14] Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *CVPR*, 2019. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 3, 5

[16] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *CVPR*, 2019. 3

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NIPS*, 2017. 3

[18] 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 *NIPS*, 2017. 8

[19] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017. 3

[20] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM TOG*, 36(4):107, 2017. 3

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 3, 5, 6

[22] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018. 3

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[25] Peipei Li, Yibo Hu, Qi Li, Ran He, and Zhenan Sun. Global and local consistent age generative adversarial networks. In *ICPR*, 2018. 3

[26] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *ICCV*, 2019. 3

[27] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 3

[28] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018. 3

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[30] Augustus Odena. Semi-supervised learning with generative adversarial networks. In *ICML Workshop*, 2016. 3

[31] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 3

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1, 3, 6, 7, 8, 9

[33] Guo-Jun Qi, Liheng Zhang, Hao Hu, Marzieh Edraki, Jingdong Wang, and Xian-Sheng Hua. Global versus localized generative adversarial nets. In *CVPR*, 2018. 3

[34] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *CVPP*, 2018. 1, 3, 8

[35] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, 2016. 3

[36] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018. 1, 3, 5, 6

[37] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Elsevier CVIU*, 187:102788, 2019. 1, 3, 5, 6

[38] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *ICCV*, 2019. 1

[39] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *ICCV*, 2019. 3

[40] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and remapping the dna of a natural image. In *ICCV*, 2019. 3

[41] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *arXiv preprint arXiv:1911.11897*, 2019. 3

[42] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. In *ACM MM*, 2018. 3

[43] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019. 1, 3, 5, 6

[44] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 3

[45] Mehmet Ozgur Turkoglu, William Thong, Luuk Spreeuwers, and Berkay Kicanaoglu. A layer-based sequential framework for scene generation with gans. In *AAAI*, 2019. 3

[46] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *ECCV*, 2016. 2, 6

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 3

[48] 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 *CVPR*, 2018. 1, 3, 7, 8

[49] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *ICCV*, 2015. 2, 6

[50] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 8

[51] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017. 8

[52] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *CVPR*, 2017. 1, 3, 6

[53] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 3

[54] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 3

[55] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, 2019. 3

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3, 7

[57] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 3

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3

[59] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 3