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$^{1,4}$
$^1$University of Trento $^2$University of Oxford $^3$Texas State University $^4$Huawei Research Ireland

Abstract

In this paper, we address the task of semantic-guided scene generation. One open challenge widely observed in global image-level generation methods is the difficulty of generating small objects and detailed local texture. 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 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. 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

Semantic-guided scene generation is a hot research topic covering several main-stream research directions, including cross-view image translation [21, 55, 38, 39, 46, 40] and semantic image synthesis [51, 8, 36, 34]. The cross-view image translation task proposed in [38] 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 [46] 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 [51, 8, 36, 34], as shown in Fig. 1(top). Recently, Park et al. [34] 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.

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 the number of training samples of different scene classes is imbalanced. For instance, for the Dayton dataset [49], 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 doing so, 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 [49] and CVUSA [52] datasets, and semantic image synthesis on the Cityscapes [11] and ADE20K [59] 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, 56, 7, 24, 17, 13, 42, 29, 41]. 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) [31] has been proposed. A CGAN combines a vanilla GAN and an external information, such as class labels [32, 33, 9], text descriptions [26, 57, 25], object keypoint [37, 45], human body/hand skeleton [1, 44, 3, 62], conditional images [61, 21], semantic maps [51, 46, 34, 50], scene graphs [22, 58, 2] and attention maps [56, 30, 43].

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, 28, 27, 35, 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, 48, 14, 4, 5]. In this paper, we mainly focus on two scene generation tasks, i.e., cross-view image translation [55, 38, 39, 46] and semantic image synthesis [51, 8, 36, 34]. Most existing works on cross-view image translation have been conducted to synthesize novel views of the same objects [12, 60, 47, 10]. Moreover, several works deal with image translation problems with drastically different views and generate a novel scene from a given different scene [55, 38, 39, 46]. For instance, Tang et al. [46] 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 [51, 8, 36, 34]. For example, Park et al. propose GauGAN [34], 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 by these global image-level generation methods are often unsatisfactory, especially on detailed local texture. In contrast, our proposed approach focuses on generating a 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 [34] 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 [46] and concatenate the input image \( I_l \) 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 to generate 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 being thus able to largely avoid the interference from the large object classes in the joint optimization. Each sub-generation branch has independent network parameters and concentrates on a specific class, being therefore capable of effectively producing similar gen-
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 $L_1$ loss is applied at the end for class-level reconstruction. The symbols $\otimes$ and $\oplus$ denote element-wise multiplication and channel-wise concatenation, respectively.

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:

$$F_i = M_i \ast f', \quad i = 1, 2, ..., c,$$

where $c$ is the number of semantic classes. Then the filtered feature map $F_i$ is fed into several convolutional layers for the corresponding $i^{th}$ class and generate an output image $I_{lg}^i$. For better learning each class, we utilize a semantic-mask guided pixel-wise $L_1$ reconstruction loss, which can be expressed as follows:

$$L_{L_1}^{local} = \sum_{i=1}^{c} E_{I_{lg}, I_{lg}^i} [||I_g \ast M_i - I_{lg}^i||_1].$$

The final output $I_g^L$ 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:

$$I_g^L = I_{lg}^1 \oplus I_{lg}^2 \oplus \cdots \oplus I_{lg}^c.$$  

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

$$I_g^L = \text{Conv}(\text{Concat}(I_{lg}^1, I_{lg}^2, \cdots, I_{lg}^c)),$$

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 $F_i$ 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., $\{F_1, ..., F_c\}$. First, the packed feature map $F_p \in \mathbb{R}^{c \times n \times h \times w}$ (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 $c \times 1 \times 1 \times 1$. Then the pooled feature map is connected with a fully connected layer to predict classification probability of the $c$ object classes of the scene. The output after the FC layer is $Y' \in \mathbb{R}^{c \times c}$, where $c$ is the number of semantic classes, as for each filtered feature map $F_i (i=1, ..., c)$, we predict a $c \times 1$ one-hot vector for the probabilities of the $c$ classes.

Since some object classes may not exist in the input se-
mantic mask sample, the features from the local branches corresponding to the void classes should not contribute to the classification loss. Therefore, 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:

$$L_{CE} = - \sum_{m=1}^{c} H_m \sum_{i=1}^{c} 1\{Y(i) = i\} \log(f(F_i)),$$

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 $F_i$. $Y$ 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 targeted 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 generators can also be used together 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 $I_l^G$. In our implementation, $G_w$ 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 $3 \times 3$ with stride 2 except for the last layer which has a kernel size of $1 \times 1$ with stride 1 for dense prediction. We predict a two-channel weight map $W_f$ using the following calculation:

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

where Softmax(·) 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:

$$I_g^C = I_g^G \otimes W_g + I_l^G \otimes W_l,$$

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 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:

$$L_{CGAN}(G, D_s) = \mathbb{E}_{S_g, I_g} \left[ \log D_s(S_g, I_g) \right] + \mathbb{E}_{S_g, I_g^C} \left[ \log(1 - D_s(S_g, I_g^C)) \right],$$

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 $D_l$, which takes the conditional image $I_l$ and the final generated image $I_g^C$ (or the ground-truth image $I_g$) as input:

$$L_{CGAN}(G, D_l) = \mathbb{E}_{I_l, I_g^C} \left[ \log D_l(I_l, I_g^C) \right] + \mathbb{E}_{I_l, I_g} \left[ \log(1 - D_l(I_l, I_g)) \right].$$

In this case, the total loss of our Dual-Discriminator $D$ is $L_{CGAN}$=$L_{CGAN}(G, D_l)$+$L_{CGAN}(G, D_s)$.

4. Experiments

The proposed LGGAN can be applied to different generative tasks such as the cross-view image translation [46] and the semantic image synthesis [34]. In this section, we present experimental results and analysis on both tasks.

4.1. Results on Cross-View Image Translation

**Datasets and Evaluation Metric.** We follow [46] and perform cross-view image translation experiments on both Dayton [49] and CVUSA datasets [52]. Similarly to [38, 46], we employ Inception Score (IS), Accuracy (Acc.), KL Divergence Score (KL), Structural-Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Sharpness Difference (SD) to evaluate the proposed model.

**State-of-the-Art Comparisons.** We compare our LGGAN with several recently proposed state-of-the-art methods, i.e., Zhai et al. [55], Pix2pix [21], X-SO [39], X-Fork [38] and X-Seq [38]. 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++ [38], X-Seq++ [38] and SelectionGAN [46]. We implement Pix2pix++, X-Fork++ and X-Seq++ using their public source code. Results are shown in Tables 1 and 2. We observe that LGGAN achieves significantly better results than
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Inception Score*</th>
<th>SSIM</th>
<th>PSNR</th>
<th>SD</th>
<th>KL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1 Top-5</td>
<td>Top-1 Top-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix2pix [21]</td>
<td>6.80 9.15</td>
<td>23.55 27.00</td>
<td>2.8515</td>
<td>1.9432</td>
<td>2.9683</td>
<td>0.4180</td>
</tr>
<tr>
<td>X-SO [39]</td>
<td>27.56 41.15</td>
<td>57.96 73.20</td>
<td>2.9459</td>
<td>2.0963</td>
<td>2.9980</td>
<td>0.4772</td>
</tr>
<tr>
<td>X-Fork [38]</td>
<td>30.00 48.68</td>
<td>61.57 78.84</td>
<td>3.0720</td>
<td>2.2402</td>
<td>3.0932</td>
<td>0.4963</td>
</tr>
<tr>
<td>X-Seq [38]</td>
<td>30.16 49.85</td>
<td>62.59 80.70</td>
<td>2.7384</td>
<td>2.1304</td>
<td>2.7674</td>
<td>0.5031</td>
</tr>
<tr>
<td>Pix2pix++ [21]</td>
<td>32.06 54.70</td>
<td>63.19 81.01</td>
<td>3.1709</td>
<td>2.1200</td>
<td>3.2001</td>
<td>0.4871</td>
</tr>
<tr>
<td>X-Fork++ [38]</td>
<td>34.67 59.14</td>
<td>66.37 84.70</td>
<td>3.0737</td>
<td>2.1508</td>
<td>3.0983</td>
<td>0.4982</td>
</tr>
<tr>
<td>X-Seq++ [33]</td>
<td>31.58 51.67</td>
<td>65.21 82.48</td>
<td>3.1703</td>
<td>2.2185</td>
<td>3.2444</td>
<td>0.4912</td>
</tr>
<tr>
<td>SelectionGAN</td>
<td>42.11 68.12</td>
<td>77.74 92.89</td>
<td>3.0613</td>
<td>2.2707</td>
<td>3.1336</td>
<td>0.5938</td>
</tr>
<tr>
<td>LGGAN (Ours)</td>
<td>48.17 79.35</td>
<td>81.14 94.91</td>
<td>3.3994</td>
<td>2.3478</td>
<td>3.4261</td>
<td>0.5457</td>
</tr>
</tbody>
</table>

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 3.9943 for all, top-1 and top-5 setups, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Inception Score*</th>
<th>SSIM</th>
<th>PSNR</th>
<th>SD</th>
<th>KL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1 Top-5</td>
<td>Top-1 Top-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhai et al. [55]</td>
<td>13.97 14.03</td>
<td>42.09 52.29</td>
<td>1.8434</td>
<td>1.5171</td>
<td>1.8666</td>
<td>0.4147</td>
</tr>
<tr>
<td>Pix2pix [21]</td>
<td>7.33 9.25</td>
<td>25.81 32.67</td>
<td>3.2771</td>
<td>2.2219</td>
<td>3.4312</td>
<td>0.3923</td>
</tr>
<tr>
<td>X-SO [39]</td>
<td>0.29 0.21</td>
<td>6.14 9.08</td>
<td>1.7575</td>
<td>1.4145</td>
<td>1.7791</td>
<td>0.3451</td>
</tr>
<tr>
<td>X-Fork [33]</td>
<td>20.58 31.24</td>
<td>50.51 63.66</td>
<td>3.4432</td>
<td>2.5447</td>
<td>3.5567</td>
<td>0.4356</td>
</tr>
<tr>
<td>X-Seq [38]</td>
<td>15.98 24.14</td>
<td>42.91 54.41</td>
<td>3.8151</td>
<td>2.6738</td>
<td>4.0077</td>
<td>0.4231</td>
</tr>
<tr>
<td>Pix2pix++ [21]</td>
<td>26.45 41.87</td>
<td>57.26 72.87</td>
<td>3.2592</td>
<td>2.4175</td>
<td>3.5078</td>
<td>0.4617</td>
</tr>
<tr>
<td>X-Fork++ [38]</td>
<td>31.03 49.65</td>
<td>64.47 81.16</td>
<td>3.3758</td>
<td>2.5375</td>
<td>3.5711</td>
<td>0.4769</td>
</tr>
<tr>
<td>X-Seq++ [33]</td>
<td>34.69 54.61</td>
<td>67.12 83.46</td>
<td>3.3919</td>
<td>2.5474</td>
<td>3.4858</td>
<td>0.4740</td>
</tr>
<tr>
<td>SelectionGAN</td>
<td>41.52 65.51</td>
<td>74.32 89.66</td>
<td>3.8074</td>
<td>2.7181</td>
<td>3.9197</td>
<td>0.5323</td>
</tr>
<tr>
<td>LGGAN (Ours)</td>
<td>44.75 70.68</td>
<td>78.76 93.40</td>
<td>3.9180</td>
<td>2.8383</td>
<td>3.9878</td>
<td>0.5238</td>
</tr>
</tbody>
</table>

Figure 4: Qualitative comparison in a2g direction on Dayton (top two rows) and CVUSA (bottom two rows).

4.2. Results on Semantic Image Synthesis

Datasets and Evaluation Metric. We follow [34] and conduct extensive experiments on both Cityscapes [11] and ADE20K [59] datasets. We use the mean Intersection-over-Union (mIoU), pixel accuracy (Acc) and Fréchet Inception Distance (FID) [18] as the evaluation metrics.

State-of-the-Art Comparisons. We compare the proposed LGGAN with several leading semantic image synthesis methods, i.e., Pix2pixHD [51], CRN [8], SIMS [36] and GauGAN [34]. Results of the mIoU, Acc and FID metrics are shown in Table 3(left). 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 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 provide AMT results, as shown in Table 3 (middle). We observe that users favor our synthesized results on both datasets compared with other competing methods including SIMS.

**Qualitative Evaluation.** The qualitative results compared with the leading method GauGAN [34] are shown in Fig. 5.

We can see that the proposed LGGAN generates much better results with fewer visual artifacts than GauGAN.

**Visualization of Learned Feature Maps.** In Fig. 6, we randomly show some channels from the learned class-specific feature maps (30th to 32th, and 50th to 52th) on Cityscapes to see if they clearly highlight particular semantic classes. We show the visualization results on 3 different classes, i.e., road, vegetation and car. We can easily observe that...
each local sub-generator learns well the class-level deep representations, further confirming our motivations.

**Visualization of Generated Semantic Maps.** We follow GauGAN [34] and apply pretrained segmentation networks on the generated images to produce semantic maps, i.e., DRN-D-105 [54] for Cityscapes and UperNet101 [53] for ADE20K. The generated semantic maps of the proposed LGGAN, GauGAN and the ground truths are shown in Fig. 7. We observe that our LGGAN generates better semantic maps than GauGAN, especially on local texture (‘car’ in the first row) and small objects (‘traffic sign’ and ‘pole’ in the second row), 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 5 baselines (i.e., S1, S2, S3, S4, S5) as shown in Table 3(right): (i) S1 means only adopting the global generator. (ii) S2 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 S3 and S2 is that S3 uses a convolutional layer to generate the local results, as presented in Eq. (4). (iv) S4 employ the proposed classification-based discriminative feature learning module. (v) S5 is our full model and adopts the proposed weight map fusion strategy.

**Effect of Local and Global Generation.** The results of the ablation study are shown in Table 3(right). 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 proposed local and global generation framework. We also provide qualitative results of the local and global generation in Fig. 1. We observe that our full model, i.e., Global + Local, generates visually better results than both the global and local method, which further confirms our motivations.

**Effect of Classification-Based Feature Learning.** S4 significantly outperforms S3 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 in S5, 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, we show some samples of the generated weight maps.

### 5. Conclusion

We propose 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. 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.

**Acknowledgments.** This work was supported by the Italy-China collaboration project TALENT: CN19GR09.
References


Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In CVPP, 2018. 1, 3, 6, 7


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


Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. INGAN: Capturing and remapping the “dna” of a natural image. In ICCV, 2019. 3


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

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


Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In ECCV, 2016. 1, 2, 5


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, 6, 7


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


Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In CVPR, 2017. 1, 3, 5, 6


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


