Abstract—In this paper, we address the task of semantic-guided image generation. One challenge common to most existing image-level generation methods is the difficulty in generating small objects and detailed local textures. To address this, in this work we consider generating images using local context. As such, we design a local class-specific generative network using semantic maps as guidance, which separately constructs and learns subgenerators for different classes, enabling it to capture finer details. To learn more discriminative class-specific feature representations for the local generation, we also propose a novel classification module. To combine the advantages of both global image-level and local class-specific generation, a joint generation network is designed with an attention fusion module and a dual-discriminator structure embedded. Lastly, we propose a novel semantic-aware upsampling method, which has a larger receptive field and can take far-away pixels that are semantically related for feature upsampling, enabling it to better preserve semantic consistency for instances with the same semantic labels. Extensive experiments on two image generation tasks show the superior performance of the proposed method. State-of-the-art results are established by large margins on both tasks and on nine challenging public benchmarks. The source code and trained models are available at https://github.com/Ha0Tang/LGGAN.

Index Terms—GANs, Local and Global, Feature Upsampling, Semantic-Guided, Image Generation.

1 INTRODUCTION

Semantic-guided image generation is a hot topic covering several mainstream research directions, including cross-view image translation [1], [2], [3], [4] and semantic image synthesis [5], [6], [7], [8], [9], [10]. The cross-view image translation task, proposed in [3], 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 [4], try to generate the target image based on another 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 photorealistic image from a semantic map (see Fig. 1 (top)). This has many real-world applications and has drawn much attention from the academic research community as well as industry [5], [6], [7], [8], [11], [12], [13]. With the use of semantic information, existing methods for both tasks have achieved promising performance in semantic-guided image generation.

However, there is still room for improvement, especially when it comes to generating local structures and details, as well as small-scale objects. We believe there are several reasons for this. First, the existing methods for both tasks are typically based on global image-level generation. In other words, they accept a semantic map containing several object classes and aim to generate the appearance of each one using the same network design or shared network parameters. In this case, all the classes are treated equally by the network. However, because different semantic classes have distinct properties, using specified network learning for each would intuitively facilitate the complex generation of multiple classes. Second, the number of training samples for different semantic classes is often imbalanced. For instance, for the Dayton dataset [14], cars and buses occupy less than 2% of all pixels in the training data. This naturally causes the model learning to be dominated by the classes with the largest number of training samples. Third, the sizes...
of objects in different semantic classes vary. As shown in the first row of Fig. 1, larger-scale object classes, such as roads and sky, usually occupy a bigger area of the image than smaller-scale classes, such as poles and traffic lights. Since convolutional networks usually share parameters at different convolutional positions, the larger-scale object classes would thus take advantage during the learning process, further increasing the difficulty in accurately generating the small-scale object classes.

To tackle these issues, a straightforward solution would be to model the generation of different image classes individually using 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 subgenerators for different classes with a shared encoded feature map. The input semantic map is utilized as the guidance to obtain feature maps corresponding spatially to each class, 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 advantages of both. It contains three main network branches (see Fig. 2). The first is the image-level global generator, which learns a global appearance distribution using the input. The second is the proposed class-specific local generator, which aims to generate different object 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 subnetworks in a weighted combination of their final generation results. The proposed LGGAN can be jointly trained in an end-to-end fashion, enabling the local and global generation to benefit from each other during optimization.

Moreover, existing methods, such as [8], [13], typically adopt nearest-neighbor interpolation to upsample feature maps and then generate final results, which leads to many visual artifacts and blurriness in the generated images. Feature upsampling is a key operation in the semantic-guided image synthesis task. Traditional upsampling methods, such as nearest-neighbor, bilinear, and bicubic interpolation only consider subpixel neighborhoods (indicated by white circles in Fig. 3(b)), failing to capture the complete semantic information, e.g., the head and body of the dog, and the front part of the car. Learnable upsampling methods, such as deconvolution [15] and pixel shuffle [16], are able to obtain the global information with a larger kernel size, but learn the same kernel (indicated by the white arrows in Fig. 3(c)) across the image, regardless of the semantic information. Other feature enhancement methods, such as spatial attention [17], can learn different kernels (indicated by different colored arrows in Fig. 3(d)), but they still inevitably capture a lot of redundant information, i.e., ‘grass’ and ‘soil’. Moreover, they are prohibitively expensive since they must consider all pixels.

To address these limitations, we further propose a novel semantic-aware upsampling (SAU) for this challenging task, as shown in Fig. 3(e). Our SAU dynamically upsamples a small subset of relevant pixels based on the semantic information, i.e., the green and tangerine circles represent the pixels within the dog and the car, respectively. In this way, SAU is more efficient than deconvolution, pixel shuffle, and spatial attention, and can capture more complete semantic information than traditional upsampling methods such as nearest-neighbor interpolation.

Overall, the contributions of this paper are as follows:

- We explore image generation from local context, which we
believe is beneficial for generating richer details compared with the existing global image-level generation methods. A new local class-specific generative structure is designed for this purpose. It can effectively handle the generation of small objects and 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.

- We introduce a novel semantic-aware upsampling (SAU) to dynamically upsample a small subset of relevant pixels based on the semantic information. SAU is more efficient than deconvolution, pixel shuffle, and spatial attention, and can capture more complete semantic information than traditional upsampling methods such as nearest-neighbor interpolation.

- Experiments for cross-view image translation on the Day- ton [14], CVUSA [18], and SVA [19] datasets, and semantic image synthesis on the Cityscapes [20], ADE20K [21], COCO-Stuff [22], DeepFashion [23], CelebAMask-HQ [24], and Facades [25] datasets demonstrate the effectiveness of the proposed method, and show significantly better results compared with state-of-the-art methods.

Part of the material presented here appeared in [26]. The current paper extends [26] in several ways. (1) We present a more detailed analysis of related works, including recently published works dealing with semantic-guided image generation and feature upsampling. (2) We propose a general and highly effective feature upsampling method, i.e., SAU. The proposed SAU has three advantages: i) Global view. Unlike traditional upsampling methods (e.g., nearest-neighbor) that only exploit local neighborhoods, SAU can aggregate semantic information in a global view. ii) Semantically adaptive. Instead of using a fixed kernel for all locations (e.g., deconvolution), SAU enables semantic class-specific upsampling by generating adaptive kernels for different locations. iii) Efficient. Unlike spatial attention, which uses a fully connected strategy to connect all pixels, SAU only considers the most relevant pixels, introducing little computational overhead. Equipped with this new upsampling method, our LGGAN proposed in [26] is upgraded to LGGAN++.

(3) We conduct extensive ablation studies to demonstrate the effectiveness of the proposed SAU against other feature upsampling and enhancement methods. (4) We extend the quantitative and qualitative experiments by comparing the proposed LGGAN and SAU with very recent works on two image synthesis tasks with diverse scenarios. We observe that the proposed methods achieve consistent and substantial gains on nine public datasets.

2 Related Work

Generative Adversarial Networks (GANs) [27] have been widely used for image generation [28], [29]. A vanilla GAN has two important components, i.e., a generator and a discriminator. The goal of the generator is to synthesize photorealistic images from a noise vector, while the discriminator tries to distinguish between the real and the generated images. To create user-specific images, the conditional GAN (CGAN) [30] was proposed. A CGAN combines a vanilla GAN and external information, such as class labels [31], [32], text descriptions [33], [34], object/face keypoints [35], [36], human body/hand skeletons [37], [38], [39], [40], [41], semantic maps [4], [5], or attention maps [42], [43].

Global and Local Generation in GANs. Modeling global and local information in GANs to generate better results has been explored in various generative tasks [44], [45], [46], [47], [48], [49]. For instance, Huang et al. [44] proposed TPGAN for frontal view synthesis by simultaneously perceiving global structures and local details. Gu et al. [49] proposed MaskGAN for face editing by separately learning every face component, e.g., mouth and eye. However, these methods have only been applied to face-related tasks, such as face rotation or face editing, where all the domains have large overlap and similarity. In contrast, we propose a new local and global image generation framework for the more challenging scene image generation task, where the local context modeling is based on semantic-guided class-specific generation, which has not been explored by any existing works.

Semantic-Guided Image Generation. Scene generation tasks are a hot topic as each image can be parsed into distinctive semantic objects. In this paper, we mainly focus on two image generation tasks, i.e., cross-view image translation [2], [3], [4], [50], [51] and semantic image synthesis [3], [6], [7], [8], [52]. Most existing works on cross-view image translation aim to synthesize novel views of the same objects [53], [54], [55]. For instance, Dosovitskiy et al. [53] used generative models to produce unseen views...
of cars, chairs, and tables. Moreover, several works deal with image translation problems with drastically different views and generate a novel scene from another given scene [2], [3], [4]. For instance, Zhai et al. [2] tried to generate panoramic ground-level images from aerial images of the same location. Tang et al. proposed SelectionGAN [4] 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 photorealistic image from a semantic map [5], [6], [8]. For example, Park et al. proposed GauGAN [8], which achieves the best results on this task.

With the semantic maps as guidance, existing approaches for both tasks achieve promising performance. However, the results produced by these global image-level generation methods are still often unsatisfactory, especially for detailed local textures. In contrast, our proposed approach focuses on generating a more realistic global structure/layout and local texture details. Both the local and global generation branches are jointly learned in an end-to-end fashion, enabling both to be improved through their mutually beneficial information.

**Feature Upsampling.** Traditional upsampling methods such as nearest-neighbor and bilinear interpolation use spatial distance and handcrafted kernels to capture the correlations between pixels. Recently, several deep learning methods, such as deconvolution [15] and pixel shuffle [16], have been proposed to upsample feature maps using learnable kernels. For instance, pixel shuffle [16] tries to reshape depth on the channel dimension into width and height on the spatial dimension. However, these methods either exploit semantic information in a small neighborhood or use a fixed kernel. Other works, including super-resolution [56], [57], inpainting [58], and denoising [59], also explore the use of learnable kernels.

Existing image generation methods, such as [8], [12], [13], typically adopt nearest-neighbor interpolation to upsample feature maps and then generate final results. However, this leads to unsatisfactory results, particularly in the generated content details and intra-object completions. To address this limitation, we propose a novel semantic-aware upsampling for this task. To the best of our knowledge, we are the first to investigate the influence of feature upsampling on this challenging task.

**3 THE PROPOSED SAU AND LGGAN**

We start by presenting the details of the proposed SAU and LGGAN. We first introduce the used backbone structure and then present the proposed semantic-aware upsampling, and finally introduce the design of the proposed local and global generation networks.

**3.1 Backbone Encoding Network Structure**

Semantic-Guided Generation. In this paper, we focus on two main tasks, i.e., semantic image synthesis and cross-view image translation. For the former, we follow GauGAN [8] 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 [4] 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 in learning the generation of another domain.

Parameter-Sharing Encoder. As we have three different branches for three different generators, the encoder $E$ shares parameters with all the branches to create a compact backbone network. The gradients from each branch all contribute 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. Thus, the encoded deep representation of the input $S_g$ can be represented as $f = E(S_g)$. This feature $f$ is then fed into the proposed SAU to obtain an upsampled feature map $f' = \text{SAU}(f)$, as shown in Fig. 2 and 4.

**3.2 Semantic-Aware Upsampling**

An illustration of the proposed SAU is shown in Fig. 4. It consists of two main branches, i.e., the semantically adaptive kernel generation (SAKG) branch, which predicts upsampled kernels according to the semantic information, and the semantically adaptive feature upsampling (SAFU) branch, which selectively performs the feature upsampling based on the kernels learned in SAKG. All components are trained in an end-to-end fashion.

Specifically, given a feature map $f \in \mathbb{R}^{C \times H \times W}$ and an upsample scale $s$, SAU aims to produce a new feature map $f' \in \mathbb{R}^{C \times H' \times W'}$. For any target location $l' = (i', j')$ in the output $f'$, there is a corresponding source location $l = (i, j)$ at

**Fig. 4: Overview of the proposed SAU, which consists of two branches, i.e., SAKG and SAFU. The SAKG branch aims to generate semantically adaptive kernels according to the input layout. The SAFU branch aims to selectively upsample the feature $f \in C \times H \times W$ to the target one $f' \in C \times H' \times W'$ based on the kernels learned in SAKG, where $s$ is the expected upsample scale.**
the input \( f \), where \( i = [i’, s] \), \( j = [j’, s] \). We denote \( Z(l, k) \) as the \( k \times k \) subregion of \( f \) centered at the location \( l \) in, e.g., the neighbor of the location \( l \). See Fig. 3 and 4 for illustration.

### 3.2.1 Semantically Adaptive Kernel Generation

The SAKG branch aims to generate a semantically adaptive kernel at each location, according to the semantic information. It consists of four modules, i.e., feature channel compression, semantic kernel generation, feature shuffle, and channel-wise normalization.

**Feature Channel Compression.** This module is used to reduce the network parameters and computational cost. Specifically, the input feature \( f \) is fed into a convolutional layer with a \( 1 \times 1 \) kernel to compress the input channel from \( C \) to \( C’ \), making SAU have fewer parameters.

**Semantic Kernel Generation.** This module receives the feature \( f_s \in \mathbb{R}^{C \times H \times W} \) as input (where \( H \) and \( W \) denote the height and width of the feature map) and tries to generate different semantically adaptive kernels, which can be represented as \( f_r \in \mathbb{R}^{k^2 \times H \times W} \). Here \( k \) is the semantically adaptive upsampling kernel size and \( s \) is the expected upsample scale. In our experiments, we set \( C’ = 64 \), \( k = 5 \), and \( s = 2 \), which achieve good results in most cases.

**Feature Shuffle.** We then feed the feature \( f_k \) through a feature shuffle layer to rearrange its elements, leading to a new feature map \( f_s \in \mathbb{R}^{k^2 \times H \times W} \), where \( k^2 \times k \) represents the learned semantic kernel. Note that the learned semantic kernels are quite different at different locations \( l’ \), as shown in Fig. 3 and 5.

**Channel-Wise Normalization.** Next, we apply a channel-wise softmax operation on each semantic kernel \( f_s \) to obtain the normalized kernel \( f_n \), i.e., the sum of the weight values in \( k^2 \) is equal to 1. In this way, we can guarantee that information from the combination will not explode. Moreover, the semantically adaptive kernels can determine which regions to emphasize or suppress according to the semantic information.

### 3.2.2 Semantically Adaptive Feature Upsampling

The SAFU branch aims to upsample the input feature \( f \) based on the kernel \( f_n \) learned in the SAKG branch, in a semantically adaptive way. It contains four modules, i.e., feature spatial expansion, sliding local block extraction, feature reshape, and upsampling feature selection.

**Feature Spatial Expansion.** The input feature \( f \) is fed into this module to expand the spatial size from \( H \times W \) to \( H’ \times W’ \).

**Sliding Local Block Extraction.** Then, the expanded feature \( f_c \in \mathbb{R}^{C \times H’ \times W’} \) is fed into this module to extract a sliding local block of each location in \( f_c \), leading to the new feature \( f_l \in \mathbb{R}^{C \times k \times H \times W} \).

**Feature Reshape.** We next reshape \( f_l \) by adding another dimension, resulting in a new feature \( f_r \in \mathbb{R}^{C \times k^2 \times H \times W} \). In this way, we can conduct a multiplication between the reshaped local block \( f_r \) and the learned kernel \( f_n \).

**Upsampling Feature Selection.** Finally, the feature map \( f_r \) and the kernel \( f_n \) learned in the SAKG branch are fed into the upsampling feature selection module to generate the final feature map \( f’ \in \mathbb{R}^{C \times H’ \times W’} \) in a weighted sum manner. The computational process at the location \( l = (i, j) \) can be expressed as follows:

\[
f’ = \sum_{p=-[k/2]}^{[k/2]} \sum_{q=-[k/2]}^{[k/2]} f_r(i + p, j + q) \times f_n(p, q). \tag{1}
\]

In this way, the pixels in the learned kernel \( f_n \) contribute to the upsampled pixel \( l’ \) differently, based on semantic information of features instead of the spatial distance between locations. The semantics of the upsampled feature map are stronger than those of the original one, since the information from relevant points in a local region can be more attended to, and the pixels with the same semantic label can achieve mutual gains, improving intra-object semantic consistency.

### 3.2.3 Why Does the SAU Work Better?

A short answer is that it can better preserve semantic information compared to other common upsampling methods. Specifically, while other methods, such as nearest-neighbor interpolation and deconvolution, are essential parts in almost all state-of-the-art image generation [60] and translation [8] models, they tend to ‘pollute’ semantic information when performing feature upsampling, since they inevitably incorporate contaminating information from irrelevant regions (see Fig. 3).

In contrast, the proposed SAU performs feature upsampling using itself, i.e., it uses the pixels belonging to the same semantic label to upsample the feature maps. Hence, the generator can better preserve semantic information. It enjoys the benefit of feature upsampling without losing the input semantic information. In Fig. 5, we show some examples of the learned semantically adaptive kernels. We can
Fig. 6: Overview of the proposed local class-specific generator $G_i$, which consists 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 classes filtered is applied to the feature representation of each class to learn more discriminative class-specific representations. A semantic-mask guided pixel-wise $L_1$ loss is applied at the end for class-level reconstruction. The symbols $\otimes$ and $\odot$ denote element-wise multiplication and channel-wise concatenation, respectively. Note that we assume the size of $f'$ is $C \times H \times W$ for simplicity, which is different from the one in Fig. 4 (i.e., $C \times H_s \times W_s$).

3.3 Local and Global GAN

An illustration of the overall framework is shown in Fig. 2. The generation module consists of three main parts, i.e., a semantic-guided class-specific generator modeling the local context, an image-level generator modeling the global layout, and a weight-map generator for fusing the local and the global generators.

Class-Specific Local Generation Network. As shown in Fig. 1 and discussed in the introduction, the training data imbalance between different classes and the size difference between semantic objects makes it extremely difficult to generate small object classes and fine details. To address this, we propose a novel local class-specific generation network. It constructs a separate generator for each semantic class, thus being able to largely avoid the interference from large object classes during the joint optimization. Each sub-generation branch has independent network parameters and concentrates on a specific class, therefore being capable of effectively producing similar generation quality for different classes and yielding richer local image details.

An overview of the local generation network $G_i$ is provided in Fig. 6. The upsampled 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:

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

where $N$ 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, which generates an output image $I_{g_i}$. To better learn each class, we utilize a semantic mask guided pixel-wise $L_1$ reconstruction loss, which can be expressed as follows:

$$L_{L_1}^{local} = \sum_{i=1}^{N} \mathbb{E}_{I_g, I_{g_i}} [||I_g \ast M_i - I_{g_i}||_1]. \quad (3)$$

The final output $I_{g}^L$ from the local generation network can be obtained in two ways. The first is by performing an element-wise addition of all the class-specific outputs:

$$I_{g}^L = I_{g_1}^l \oplus I_{g_2}^l \oplus ... \oplus I_{g_N}^l. \quad (4)$$

The second is by applying a convolutional operation on all the class-specific outputs, as shown in Fig. 6:

$$I_{g}^L = \text{Conv}((\text{Concat}(I_{g_1}, I_{g_2}, ..., I_{g_N}))), \quad (5)$$

where $\text{Concat}(\cdot)$ and $\text{Conv}(\cdot)$ denote a 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. To provide a 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. 6. One input sample of the module is a stack of feature maps produced by different local generation branches, i.e., $\{F_1, ..., F_N\}$. First, the stacked feature map $F_p \in \mathbb{R}^{N \times C \times H \times W}$ (with $C, H, W$ as the number of feature
map channels, height and width, respectively) is fed into a semantic-guided average pooling layer, and we obtain a pooled feature map with dimensions of \(N \times C \times 1 \times 1\). Then the pooled feature map is fed to a fully connected (FC) layer to predict the classification probabilities of the \(N\) object classes of the image. The output after the FC layer is \(Y \in \mathbb{R}^{N \times N}\), where \(N\) is the number of semantic classes. Since, for each filtered feature map \(F_i\) (\(i=1, \ldots, N\)), we predict an \(N \times 1\) one-hot vector for the probabilities of the \(N\) classes.

Since some object classes may not exist in the input semantic mask, 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 (CE) loss by multiplying it with a void class indicator for each input sample. The indicator is a one-hot vector \(H=\{H_i\}_{i=1}^N\), with \(H_i=1\) for a valid class and \(H_i=0\) for a void class. Then, the CE loss is defined as follows:

\[
\mathcal{L}_{CE} = - \sum_{m=1}^{N} H_m \sum_{i=1}^{N} 1\{Y(i)=i\} \log(f(F_i)),
\]

where \(1\{\cdot\}\) is an indicator function having a return of 1 if \(Y(i)=i\), otherwise 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 object classes.

**Image-Level Global Generation Network.** Similar to the local generation branch, the upsampled feature map \(f'\) is also fed into the global generation subnetwork \(G_g\) for image-level generation, as shown in Fig. 2. Global generation is capable of capturing the global structural information or layout of the target images. Thus, the global result \(I_g^G\) can be obtained through a feed-forward computation: \(I_g^G=\mathcal{G}_g(f')\). Besides the proposed \(G_g\), many existing global generators can also be used together with the local generator \(G_l\), making the proposed framework very flexible.

**Pixel-Level Fusion Weight-Map Generation Network.** To better combine the local and global generation subnetworks, we further propose a pixel-level weight map generator \(G_w\), which aims at predicting pixel-wise weights to fuse the global generation \(I_g^G\) and the local generation \(I_g^L\). In our implementation, \(G_w\) consists of two Transpose Convolution→InstanceNorm→ReLU blocks and one Convolution→InstanceNorm→ReLU block. The number of output channels for these three blocks are 128, 64, and 2, respectively. The kernel sizes are \(3\times3\) with stride 2, except for the last layer which has a kernel size of \(1\times1\) with stride 1 for dense prediction. We predict a two-channel weight map \(W_f\) using the following calculation:

\[
W_f = \text{Softmax}(G_w(f')),
\]

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 will 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 final fused generation result is calculated as follows:

\[
I_f = I_g^G \odot W_g + I_g^L \odot W_l,
\]

where \(\odot\) is an element-wise multiplication operation. This way, the pixel-level weights predicted by \(G_w\) directly operate on the output of \(G_g\) and \(G_l\). Moreover, the generators \(G_w\), \(G_g\) and \(G_l\) affect and contribute to each other during the model optimization.

**Dual-Discriminator.** To exploit prior domain knowledge, i.e., the semantic map, we extend the single-domain vanilla discriminator [27] to a cross-domain structure, which we refer to as the semantic-guided discriminator \(D_s\), as shown in Fig. 2. It takes the semantic map \(S_g\) and generated image \(I_g^C\) (or the real image \(I_a\)) as inputs:

\[
\begin{align*}
\mathcal{L}_{CGAN}(G, D_s) = & \mathbb{E}_{S_g, I_g} \log D_s(S_g, I_g) + \\
& \mathbb{E}_{S_g, I^C_g} \left[\log(1 - D_s(S_g, I^C_g))\right],
\end{align*}
\]

aiming to preserve image layout and capture the local 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 \(I_a\)) as input:

\[
\begin{align*}
\mathcal{L}_{CGAN}(G, D_i) = & \mathbb{E}_{I_a, I^C_g} \log D_i(I_a, I^C_g) + \\
& \mathbb{E}_{I_a, I_g} \left[\log(1 - D_i(I_a, I_g))\right].
\end{align*}
\]

In this case, the overall loss of our dual-discriminator \(D\) is \(\mathcal{L}_{CGAN}(G, D_i) + \mathcal{L}_{CGAN}(G, D_s)\).

### 4 Experiments

**4.1 Local and Global GAN**

The proposed LGGAN can be applied to different generative tasks, such as cross-view image translation [4] and semantic image synthesis [8].

![Fig. 7: Qualitative comparison of cross-view image translation in a2g direction on SVA.](image-url)
4.1.1 Results on Cross-View Image Translation

Datasets and Evaluation Metrics. We follow [4], [50] and perform experiments on the SVA [19], Dayton [14], and CVUSA [18] datasets. Similar to [4], [50], we employ inception score (IS), Fréchet inception distance (FID) [61], accuracy, 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. [2], Pix2pix [1], X-SO [50], X-Fork [3] and X-Seq [3]. The comparison results are shown in Tables 1, 2, and 3. LGGAN achieves significantly better results than Pix2pix++, X-Fork++, and X-Seq++, confirming its advantage. A direct comparison with SelectionGAN is also shown in the tables, where our model provides better results on all metrics except those for pixel-level evaluation, i.e., SSIM, PSNR, and SD. SelectionGAN achieves slightly better results in these three metrics because it uses a two-stage generation strategy and an attention selection module. However, we generate much more photorealistic results than SelectionGAN, as shown in Fig. 7, 8, and 9.

Qualitative Evaluation. The qualitative results of our model compared with the leading methods are shown in Fig. 7, 8, and 9. The results generated by the proposed LGGAN are visually better than those provided by the existing methods. Specifically, our method generates clearer details for objects such as cars, buildings, roads, and trees.
TABLE 1: Quantitative comparison of cross-view image translation on SVA in the a2g direction. (*) The inception scores for real (ground truth) data are 3.1282, 2.4932, and 3.4646 for the ‘all’, ‘top-1’, and ‘top-5’ setups, respectively.

| Method              | Top-1               | Top-5               | Inception Score ↑ | SSIM † | PSNR † | SD  † | KL  ↓ | FID  ↓ |
|---------------------|---------------------|---------------------|-------------------|--------|--------|------|------|-------|
| X-Pix2pix [1]       | 8.5961 30.3288      | 9.0260 29.9102      | 2.0131 1.7221     | 2.2370 | 0.3206 | 17.9944 | 15.0254 |
| X-So [50]           | 7.5146 30.9507      | 10.0365 38.9822     | 1.1888 1.9726     | 2.3664 | 0.4237 | 21.2400 | 16.9371 |
| X-Seq [3]           | 17.3794 53.4725     | 23.8315 63.5045     | 2.2232 1.9824     | 2.4344 | 0.4638 | 22.3411 | 17.4138 |
| H-Pix2pix [50]      | 18.0706 54.8068     | 23.4000 62.3072     | 2.1906 1.9507     | 2.4069 | 0.4327 | 21.6860 | 16.9468 |
| H-So [50]           | 5.2444 26.4697      | 5.2544 31.9527      | 2.3202 1.9410     | 2.7340 | 0.4457 | 21.7709 | 17.3876 |
| H-Seq [3]           | 18.0182 51.0756     | 26.6747 62.8166     | 2.3202 1.9525     | 2.3918 | 0.4240 | 21.6327 | 16.8563 |
| X-Regions [50]      | 15.4803 48.0767     | 21.8225 36.8992     | 2.6381 1.8940     | 2.6347 | 0.4552 | 20.9148 | 17.0858 |

Fig. 10: Examples of arbitrary cross-view image translation.

Arbitrary Cross-View Image Translation. We also follow SelectionGAN [4] and show some results on arbitrary cross-view image translation (Fig. 10). We observe that, given an aerial image and a few semantic maps, LGGAN is able to produce the same scene but with different viewpoints.

Network Parameter Comparisons. In Table 4, we compare the number of network parameters in LGGAN with several state-of-the-art models. As we can see, the proposed LGGAN achieves superior model capacity and better generation performance compared with existing methods.

4.1.2 Results on Semantic Image Synthesis

Datasets and Evaluation Metrics. We follow GauGAN [8] and conduct extensive experiments on Cityscapes [20], ADE20K [21], and COCO-Stuff [22]. We use mean intersection-over-union (mIoU), pixel accuracy (Acc), and Fréchet inception distance (FID) [61] as evaluation metrics.

State-of-the-Art Comparisons. We compare the proposed LGGAN with several leading semantic image synthesis methods, i.e., Pix2pixHD [5], CRN [6], SIMS [7], Gau-
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

(a) GauGAN vs. SMIS vs. LGGAN

(b) SMIS vs. LGGAN vs. LGGAN++ (i.e., LGGAN+SAU)

(c) SEAN vs. SEAN++ (i.e., SEAN+SAU) vs. LGGAN++

(d) MaskGAN vs. MaskGAN++ (i.e., MaskGAN+SAU) vs. LGGAN++

Fig. 11: Qualitative comparison of semantic image synthesis on Cityscapes (top) and ADE20K (bottom).

TABLE 5: Quantitative comparison of semantic image synthesis on Cityscapes, ADE20K, and COCO-Stuff.

| Method          | Cityscapes | ADE20K | COCO-Stuff |
|-----------------|------------|--------|------------|
|                 | mIoU ↑     | Acc ↑  | FID ↓      | mIoU ↑     | Acc ↑  | FID ↓      | mIoU ↑     | Acc ↑  | FID ↓      |
| CRN [6]         | 52.4       | 77.1   | 104.7      | 22.4       | 68.8   | 73.3       | 23.7       | 40.4   | 70.4       |
| SIMS [7]        | 47.2       | 75.5   | 49.7       | -          | -      | -          | -          | -      | -          |
| Pix2pixHD [5]   | 58.3       | 81.4   | 95.0       | 20.3       | 81.2   | 81.8       | 14.6       | 45.7   | 111.5      |
| SelectionGAN [4]| 63.8       | 82.4   | 65.2       | 40.1       | 81.2   | 33.1       | -          | -      | -          |
| PIS [62]        | 64.8       | 82.4   | 96.4       | -          | -      | -          | -          | -      | -          |
| TSIT [12]       | 65.9       | 82.7   | 59.2       | 38.6       | 80.8   | 31.6       | -          | -      | -          |
| DAGAN [13]      | 66.1       | 82.6   | 60.3       | 40.5       | 81.6   | 31.9       | -          | -      | -          |
| SMIS [9]        | 62.3       | 82.1   | 49.8       | 30.4       | 77.1   | 42.2       | -          | -      | -          |
| MaskGAN [24]    | 63.2       | 81.5   | 65.8       | 35.1       | 76.5   | 33.7       | -          | -      | -          |
| + SAU (Ours)    | 65.1 (+1.9)| 82.1 (+0.6)| 51.3 (-14.5)| 37.4 (+2.3)| 78.6 (+2.1)| 31.5 (-2.2)| -          | -      | -          |
| SEAN [10]       | 64.1       | 82.2   | 61.8       | 37.5       | 76.8   | 31.8       | -          | -      | -          |
| + SAU (Ours)    | 66.3 (+2.2)| 82.7 (+0.5)| 50.3 (-11.5)| 39.2 (+1.7)| 78.9 (+2.1)| 30.1 (-1.7)| -          | -      | -          |
| GauGAN [8]      | 62.3       | 81.9   | 71.8       | 38.5       | 79.7   | 33.9       | 37.4       | 67.9   | 22.6       |
| + SAU (Ours)    | 65.5 (+3.2)| 82.5 (+0.6)| 48.3 (-23.5)| 39.8 (+1.3)| 80.7 (+0.8)| 32.0 (-1.9)| 39.0 (+1.6)| 69.1 (+1.2)| 20.1 (-2.5)|
| LGGAN (Ours)    | 68.4       | 83.0   | 57.7       | 41.6       | 81.8   | 31.6       | -          | -      | -          |
| + SAU (Ours)    | 67.7 (-0.7)| 82.9 (-0.1)| 48.1 (-9.6)| 41.4 (-0.2)| 81.5 (-0.3)| 30.5 (-1.1)| -          | -      | -          |

TABLE 6: User study I. The numbers indicate the percentage of users who favor the results of the proposed LGGAN over the competing methods.

| AMT ↑           | Cityscapes | ADE20K |
|-----------------|------------|--------|
| LGGAN vs. CRN [6] | 67.38     | 79.54  |
| LGGAN vs. Pix2pixHD [5] | 56.16     | 85.69  |
| LGGAN vs. SIMS [7] | 54.84     | -      |
| LGGAN vs. GauGAN [8] | 53.19     | 57.31  |
| LGGAN vs. SMIS [9] | 56.83     | 62.45  |
| LGGAN vs. MaskGAN [24] | 54.61     | 60.93  |
| LGGAN vs. SEAN [10] | 55.18     | 60.76  |
| LGGAN vs. SelectionGAN [4] | 52.65     | 56.93  |

GAN [8], SelectionGAN [4], TSIT [12], PIS [62], DAGAN [13], SMIS [9], MaskGAN [24], and SEAN [10]. The results in terms of mIoU, Acc, and FID are shown in Table 5. The proposed LGGAN outperforms the existing competing methods by a large margin in both mIoU and Acc. For FID, LGGAN is only worse than SIMS and SMIS on Cityscapes. However, SIMS and SMIS have poor segmentation performance. This is because SIMS produces an image by searching and copying image patches from the training dataset. The generated images are more realistic since the method uses real image patches. However, SIMS tends to copy objects with mismatched patches due to the presence queries that cannot be guaranteed to have results.
User Study. We follow the evaluation protocol of GauGAN and provide Amazon Mechanical Turk (AMT) results, as shown in Tables 6 and 7. We see that users favor our synthesized results on both datasets compared with other competing methods. We also conduct another user study to evaluate the generated images. Participants are asked to pick the best realistic image among many choices, and we give them unlimited time to respond. Each trial consists of 30 comparisons and is performed by 30 distinct participants. The results are shown in Table 8. We observe that more participants consider our results to be more realistic than those provided by the other methods.

Qualitative Evaluation. The qualitative results compared with the leading methods e.g., GauGAN, SMIS, SEAN, and MaskGAN, are shown in Fig. 11. We can see that our LGGAN and LGGAN++ generate much better results with fewer visual artifacts than those methods.

Visualization of Learned Feature Maps. In Fig. 12, 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 for three different classes, i.e., road, vegetation, and cars. We can easily observe that...
each local subgenerator effectively learns the deep class-
level representations, further confirming our motivations.

**Visualization of Generated Semantic Maps.** We follow
GauGAN [8] and apply pretrained segmentation networks
on the generated images to produce semantic maps, i.e.,
DRN-D-105 [63] for Cityscapes and UperNet101 [64] for
ADE20K. The generated semantic maps of our LGGAN,
GauGAN, and ground truths are shown in Fig. 13. As
can be seen, LGGAN generates better semantic maps than
GauGAN, especially on local textures (‘car’ in the first row)
and small objects (‘traffic sign’ and ‘pole’ in the second row).

### 4.1.3 Ablation Study

**Baselines.** We conduct extensive ablation studies on
Cityscapes to evaluate different components of our LGGAN.

| No. | Setup of LGGAN                   | mIoU ↑ | FID ↓ |
|-----|----------------------------------|--------|-------|
| B1  | Ours w/ Global                   | 62.3   | 71.8  |
| B2  | B1 + Local (addition)            | 64.6   | 66.1  |
| B3  | B1 + Local (convolution)         | 65.8   | 65.6  |
| B4  | B3 + Class Discriminative Loss   | 67.0   | 61.3  |
| B5  | B4 + Weight Map                  | 68.4   | 57.7  |

The proposed LGGAN has five baselines (i.e., B1, B2, B3, B4,
B5), as shown in Table 9: (i) In B1, only the global generator
is adopted. (ii) B2 combines the global generator and the
proposed local generator to produce the final results, where
the local results are produced using an addition operation,
as proposed in Eq. (4). (iii) B3 is similar to B2 but uses a con-
Effect of Local and Global Generation. The results of the ablation study are shown in Table 9. When using an addition operation to generate the local result, the local and global generation strategy improves the mIoU and FID by 2.3 and 5.7, respectively. When adopting a convolutional operation to produce the local results, the performance is further improved, with a 3.5 and 6.2 gain in mIoU and FID, respectively. Both results confirm the effectiveness of the proposed local and global generation framework. We also provide qualitative results in Fig. 1. We observe that our full model, i.e., Global+Local, generates visually better results than both the individual global and local methods.

Effect of Classification-Based Feature Learning. B4 significantly outperforms B3, with gains of roughly 1.2 and 4.3 in mIoU and FID, respectively. This means that the model does indeed learn a more discriminative class-specific feature representation, confirming the superiority of our design.

Effect of Weight Map Fusion. By adding the proposed weight map fusion scheme in B5, the overall performance is further boosted, with improvements of 1.4 and 3.6 in mIoU and FID, respectively. This indicates that the proposed LGGAN can in fact learn complementary information from the local and global generation branches.
4.2 Semantic-Aware Upsampling

Datasets. We first follow GauGAN [8] and conduct experiments on Cityscapes [20], ADE20K [21], and COCO-Stuff [22]. Then we conduct experiments on three more datasets with diverse scenarios. 1) Facades [25] contains different city images with various architectural styles. The training and test sets have 378 and 228 images, respectively. We resize the images to $512 \times 512$ for high-resolution layout-to-image translation tasks. 2) CelebAMask-HQ [24] contains face images with 19 semantic facial attributes. The training and test sets are made up of 24,183 and 2,842 images, respectively. We also resize these images to $512 \times 512$. 3) DeepFashion [23] contains human body images. The number of images in the training and test sets are 30,000 and 2,247, respectively. We resize the images to $256 \times 256$, and use a well-trained model [65] to extract input semantic layouts for this dataset.

Evaluation Metrics. We follow GauGAN [8] and use mIoU, Acc, and FID [61] as the evaluation metrics on Cityscapes, ADE20K, and COCO-Stuff. For DeepFashion, CelebAMask-HQ, and Facades, we use FID and Learned Perceptual Image Patch Similarity (LPIPS) [66].

4.2.1 Comparisons with State-of-the-Art

Qualitative Comparisons. We first compare the proposed SEAN+SAU and MaskGAN+SAU with SEAN [10] and MaskGAN [24] on both Cityscapes and ADE20K datasets in Fig. 11(c) and 11(d). We observe that the models with our SAU generate more realistic results than the models without using it, validating the effectiveness of the proposed SAU. We then compare the proposed GauGAN+SAU with GauGAN [8] on DeepFashion, CelebAMask-HQ, and Facades. Specifically, we replace the feature upsampling layer in GauGAN with our SAU layer. Visualization results are shown in Fig. 14(b) and 15. We can see that the model with SAU generates more photorealistic results than the original GauGAN. Moreover, we compare GauGAN and the proposed method in a zoomed-in manner on CelebAMask-HQ in Fig. 17(b). As can be seen, the model with our SAU generates more vivid content than the original GauGAN model, further validating the effectiveness of SAU. Lastly, we compare the proposed method with GauGAN on Cityscapes, ADE20K, and COCO-Stuff. Comparison results are shown in Fig. 14(a) and 16. Our method produces clearer and more visually plausible results than both leading methods, further demonstrating the benefit of our design.

User Study. We follow the same evaluation protocol as GauGAN and perform a user study. The results compared with the original GauGAN are shown in Table 10. As can be seen, users strongly favor the results generated by our proposed method on all datasets, further validating that the images generated by our upsampling method are more photorealistic.

Quantitative Comparisons. Although the user study is most suitable for evaluating the quality of the generated images, we also follow GauGAN and use mIoU, Acc, FID, and LPIPS for quantitative evaluation. The results compared with several leading methods are shown in Tables 5 and 11. Firstly, we observe from Table 11 that the model with SAU achieves better results than GauGAN on DeepFashion, CelebAMask-HQ, and Facades. Moreover, from Table 5, we can see that our methods (i.e., GauGAN+SAU, SEAN+SAU, MaskGAN+SAU) achieve better results compared with the original models on Cityscapes, ADE20K, and COCO-Stuff.

Visualization of Generated Semantic Maps. We again follow GauGAN and apply the pretrained DRN-D-105 [63] to
the generated Cityscapes images to produce semantic maps. The results, compared with those produced by GauGAN, are shown in Fig. 17(a). We see that the method with our proposed SAU generates more semantically consistent results than the original GauGAN.

**LGGAN vs. LGGAN++.** Our SAU is general and can be seamlessly integrated into existing GANs. To demonstrate this generalization ability, we conduct more experiments on both Cityscapes and ADE20K. Specifically, we replace the upsampling layers in LGGAN [26] with the proposed SAU. The results are shown in Tables 5 and 7, and Fig. 11(b). We can see from Fig. 11(b) that LGGAN++ achieves more photorealistic results than LGGAN, validating the generalization ability of SAU. Moreover, LGGAN+SAU (i.e., LGGAN++) achieves a significantly better FID score on both datasets, as shown in Table 5. Finally, LGGAN++ achieves better user study results than LGGAN, as shown in Table 7.

### 4.2.2 Ablation Study

**Baselines.** We conduct an extensive ablation study on Cityscapes to evaluate the effectiveness of the proposed SAU. As shown in Table 12, B1, B2, and B3 are three traditional upsampling methods. B4 and B5 are two learnable upsampling methods. B6 is the spatial attention module proposed in [17]. B7 is the involution operation proposed in [67]. B8 is our proposed SAU.

**Ablation Analysis.** We first compare the proposed SAU with different upsampling strategies (i.e., B1, B2, B3, B4, B5). The results of the ablation study are shown in Table 12 and Fig. 18. We can see from Table 12 that the proposed SAU achieves significantly better results than other methods, indicating that the design of effective upsampling methods is critical for this challenging task. We also observe from Fig. 18 that the proposed SAU generates more photorealistic and semantically consistent results with fewer artifacts than other upsampling methods. Moreover, we add the spatial attention [17] and involution [67] module to GauGAN, obtaining 56.2 and 58.8 in FID, respectively. We can see that our method still significantly outperforms both spatial attention and involution methods.

**Model Parameter Comparisons.** We also compare the number of generator parameters with different baselines. The results are shown in Table 12. Traditional upsampling methods (B1, B2, and B3) have the same number of parameters. Moreover, we can see that the proposed SAU achieves superior model capacity compared to the learnable upsampling methods (i.e., B4 and B5) and spatial attention (i.e., B6).

### 5 Conclusion

We propose a local class-specific and global image-level generative adversarial network (LGGAN) for semantic-guided image 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 the size difference of objects for joint learning. To learn more discriminative and class-specific feature representations, a novel classification module is further proposed. Moreover, we introduce a novel semantic-aware upsampling method, which is able to aggregate semantic information in the input layout and adaptively conduct class-specific upsampling during the translation process. Experimental results demonstrate the superiority of the proposed approach, which achieves the new state-of-the-art on both cross-view image translation and semantic image synthesis, on nine public datasets.

**Acknowledgments.** This work was partially supported by the EU H2020 AH4Media Project under Grant 959111, the PRIN project CREATiVE Prot. 2020ZSL9F9, and the National Natural Science Foundation of China (No. 61929104).

### APPENDIX A

#### APPLICATIONS

**Semantic Manipulation.** Our model also supports semantic manipulation. For instance, we can replace a building with trees (Fig. 19(a)), insert a bed into a room (Fig. 19(b)), add a few zebras to some grass (Fig. 19(c)), or remove earrings and eyeglasses from a face (Fig. 19(d)). These applications provide users more controllability during the translation process.

**Multi-Modal Synthesis.** By using a random vector as the input of the generator, our model can perform multi-modal synthesis. The results are shown in Fig. 19(e). We can see that our model generates different outputs from the same input layout.

### REFERENCES

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in CVPR, 2017.

[2] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, “Predicting ground-level scene layout from aerial imagery,” in CVPR, 2017.

[3] K. Regmi and A. Borji, “Cross-view image synthesis using conditional gans,” in CVPR, 2018.

[4] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, “Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation,” in CVPR, 2019.
Fig. 19: Exemplar applications of the proposed method on different datasets.
[42] Y. A. Mejjati, C. Richards, J. Tompkin, D. Cosker, and K. I. Kim, “Unsupervised attention-guided image-to-image translation,” in NeurIPS, 2018.

[43] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, “Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks,” IEEE TNNSL, 2021.

[44] R. Huang, S. Zhang, T. Li, and R. He, “Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis,” in ICCV, 2017.

[45] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” ACM TOG, vol. 36, no. 4, p. 107, 2017.

[46] C. H. Lin, C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen, “Coco-gan: Generation by parts via conditional coordination,” in ICCV, 2019.

[47] P. Li, Y. Hu, Q. Li, R. He, and Z. Sun, “Global and local consistent age generative adversarial networks,” in ICPR, 2018.

[48] G.-J. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X.-S. Hua, “Global versus localized generative adversarial nets,” in CVPR, 2018.

[49] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, “Mask-guided portrait editing with conditional gans,” in CVPR, 2019.

[50] K. Regmi and A. Borji, “Cross-view image synthesis using geometry-guided conditional gans,” Elsevier CVIU, vol. 187, p. 102788, 2019.

[51] B. Ren, H. Tang, and N. Sebe, “Casceded cross mlp-mixer gans for cross-view image translation,” in BMVC, 2021.

[52] H. Tang and N. Sebe, “Layout-to-image translation with double poolig generative adversarial networks,” IEEE TIP, vol. 30, pp. 7900–7913, 2021.

[53] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox, “Learning to generate chairs, tables and cars with convolutional networks,” IEEE TIPAMI, vol. 39, no. 4, pp. 692–705, 2017.

[54] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, “View synthesis by appearance flow,” in ECCV, 2016.

[55] M. Tatarchenko, A. Dosovitskiy, and T. Brox, “Multi-view 3d models from single images with a convolutional network,” in ECCV, 2016.

[56] Y. Jo, S. Wug Ob, J. Kang, and S. Joo Kim, “Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation,” in CVPR, 2018.

[57] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, “Meta-sr: A magnification-arbitrary network for super-resolution,” in CVPR, 2019.

[58] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, “Carafe: Content-aware reassembly of features,” in ICCV, 2019.

[59] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carrell, “Burst denoising with kernel prediction networks,” in CVPR, 2018.

[60] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in ICLR, 2016.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in NeurIPS, 2017.

[62] A. Dundar, K. Sapra, G. Liu, A. Tao, and R. Catanzaro, “Panoptic-based image synthesis,” in CVPR, 2020.

[63] F. Yu, V. Kolotun, and T. Funkhouser, “Dilated residual networks,” in CVPR, 2017.

[64] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in ECCV, 2018.

[65] P. Li, Y. Xu, Y. Wei, and Y. Yang, “Self-correction for human parsing,” arXiv preprint arXiv:1910.09777, 2019.

[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in CVPR, 2018.

[67] D. Li, J. Hu, C. Wang, X. Li, Q. She, L. Zhu, T. Zhang, and Q. Chen, “Involution: Inverting the inherence of convolution for visual recognition,” in CVPR, 2021.