Inferring and Leveraging Parts from Object Shape for Improving Semantic Image Synthesis

Yuxiang Wei[1,2] Zhilong Ji[3] Xiaohe Wu[1(✉)] Jinfeng Bai[3] Lei Zhang[2] Wangmeng Zuo[1]
[1]Harbin Institute of Technology [2] The Hong Kong Polytechnic University [3]Tomorrow Advancing Life
yuxiang.wei.cs@gmail.com {wuxiaohe, wmzuo}@hit.edu.cn cslzhang@comp.polyu.edu.hk

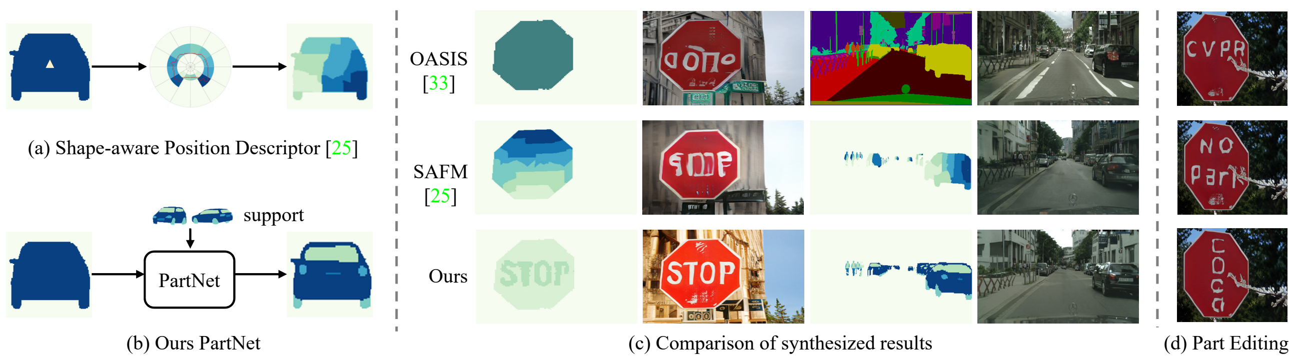


Figure 1. Illustration of our iPOSE for part map prediction and semantic image synthesis. (a) Calculation of SPD [25]. (b) Our PartNet for part map prediction. (c) Comparison of results with existing methods [25,33]. (d) Part editing results of our iPOSE.

Abstract

*Despite the progress in semantic image synthesis, it remains a challenging problem to generate photo-realistic parts from input semantic map. Integrating part segmentation map can undoubtedly benefit image synthesis, but is bothersome and inconvenient to be provided by users. To improve part synthesis, this paper presents to **i**nfer **P**arts from **O**bject **S**hap**E** (iPOSE) and leverage it for improving semantic image synthesis. However, albeit several part segmentation datasets are available, part annotations are still not provided for many object categories in semantic image synthesis. To circumvent it, we resort to few-shot regime to learn a PartNet for predicting the object part map with the guidance of pre-defined support part maps. PartNet can be readily generalized to handle a new object category when a small number (e.g., 3) of support part maps for this category are provided. Furthermore, part semantic modulation is presented to incorporate both inferred part map and semantic map for image synthesis. Experiments show that our iPOSE not only generates objects with rich part details, but also enables to control the image synthesis flexibly. And our iPOSE performs favorably against the state-of-the-art methods in terms of quantitative and qualitative evaluation. Our code will be publicly available at https://github.com/csyxwei/iPOSE.*

1. Introduction

Semantic image synthesis allows to generate an image with input semantic map, which provides significant flexibility for controllable image synthesis. Recently, it has attracted intensive attention due to its wide applications, *e.g.*, content generation and image editing [28,36], and extensive benefits for many vision tasks [33].

Albeit rapid progress has been made in semantic image synthesis [10, 25, 28, 30, 33, 36, 43–45], it is still challenging to generate photo-realistic parts from the semantic map. Most methods [28, 33] tackle semantic image synthesis with a spatially adaptive normalization architecture, while other frameworks are also explored, such as StyleGAN [19] and diffusion model [43, 45], *etc*. However, due to the lack of fine-grained guidance (*e.g.*, object part information), these methods only exploited the object-level semantic information for image synthesis, and usually failed to generate photo-realistic object parts (see the top row of Fig. 1(c)). SAFM [25] adopted the shape-aware position descriptor to exploit the pixel's position feature inside the object. However, as illustrated in Fig. 1(a), the obtained descriptor tends to be only region-aware instead of part-aware, leading to limited improvement in synthesized results (see the middle row of Fig. 1(c)).

To improve image parts synthesis, one straightforward solution is to ask the user to provide part segmentation map and integrate it into semantic image synthesis during training and inference. However, it is bothersome and inconvenient for users to provide it, especially during inference. Fortunately, with the existing part segmentation datasets [7], we propose a method to **i**nfer **P**arts from **O**bject **S**hap**E** (iPOSE), and leverage it for improving semantic image synthesis. Specifically, based on these datasets, we first construct an object part dataset that consists of paired (object shape, object part map) to train a part prediction network (PartNet). Besides, although the part dataset contains part annotations for several common object categories, many object categories in semantic image synthesis are still not covered. To address this issue, we introduce the few shot mechanism to our proposed PartNet. As shown in Fig. 1(b), for each category, a few annotated object part maps are selected as pre-defined supports to guide the part prediction. Cross attention block is adopted to aggregate the part information from support part maps. Benefited from the few shot setting, our PartNet can be readily generalized to handle a new object category not in the object part dataset. Particularly, we can manually label k object part maps for this category (*e.g.*, 3) as supports, and use them to infer the part map without fine-tuning the part prediction model.

By processing each object in the semantic map, we obtain the part map. With that, we further present a part semantic modulation (PSM) residual block to incorporate the part map with semantic map to improve the image synthesis. Specifically, the part map is first used to modulate the normalized activations spatially to inject the part structure information. Then, to inject the semantic texture information, the semantic map and a randomly sampled 3D noise are further used to modulate the features with the SPADE module [28]. We find that performing part modulation and semantic modulation sequentially disentangles the structure and texture for image synthesis, and is also beneficial for generating images with realistic parts. Additionally, to facilitate model training, a global adversarial loss and an object-level CLIP style loss [54] are further introduced to encourage model to generate photo-realistic images.

Experiments show that our iPOSE can generate more photo-realistic parts from the given semantic map, while having the flexibility to control the generated objects (as illustrated in Fig. 1 (c)(d)). Both quantitative and qualitative evaluations on different datasets further demonstrate that our method performs favorably against the state-of-the-art methods.

The contributions of this work can be summarized as:

- We propose a method iPOSE to infer parts from object shape and leverage them to improve semantic image synthesis. Particularly, a PartNet is proposed to predict the part map based on a few support part maps, which can be easily generalized to new object categories.
- A part semantic modulation Resblock is presented to incorporate the predicted part map and semantic map for image synthesis. And global adversarial and object-level CLIP style losses are further introduced to generate photo-realistic images.
- Experimental results show that our iPOSE performs favorably against state-of-the-art methods and can generate more photo-realistic results with rich part details.

2. Related Work

2.1. Semantic Image Synthesis

Semantic image synthesis predicts image from the given semantic map. With the development of generative models, many methods have been proposed to solve this problem [4, 10, 15, 17, 23, 25, 27, 28, 30, 33, 36, 37, 43–46, 56]. Pix2pix [10] first explored a conditional GAN for translating semantic map to a real image, Instead of taking semantic map as input directly, SPADE [28] proposed to use it to modulate the features layer-wisely by spatially adaptive denormalization. CC-FPSE [23] leveraged semantic map to predict the spatial variant convolutional kernels, which were used to generate the intermediate feature maps. While SC-GAN [46] transformed semantic map to semantic vector, and used it for semantic render generation. RESAIL [36] proposed to retrieve and compose a guidance image based on the given semantic map, and incorporated the image to guide the image synthesis. SAFM [25] calculated a shape-aware position descriptor for each object in semantic map, and proposed a semantic-shape adaptive feature modulation block to improve object synthesis.

In addition to semantic injection, different networks [19, 23, 27, 37, 44, 45] have also been explored in semantic image synthesis, such as multi-scale discriminator [44], feature-pyramid discriminator [23], and semantic-related discriminator [27, 33], *etc*. Besides, LGGAN [37] explored the local context information and introduced a local pathway in the generator for details synthesizing. CollogeGAN [19] used the StyleGAN [11, 21] as the generator to improve visual quality and also explored the local context with class-specific models. Recently, SDM [45] proposed a semantic diffusion model to generate images with an iterative denoising process conditioned on the semantic maps.

Most methods have only exploited the object-level semantic information for image synthesis, resulting in unrealistic parts synthesis. Albeit SAFM [25] exploited to use the object's pixel position feature for image synthesis, the obtained descriptor is only region-aware and also limited. In contrast, our iPOSE infers the part map from the given semantic map and leverage it to improve parts synthesis. It not only generates objects with realistic parts, but also enables to control the image synthesis flexibly.

2.2. Few Shot Segmentation

Few shot segmentation [1, 13, 24, 34, 35, 38, 39, 41, 47, 49, 51–53] aims to predict a dense segmentation for new class with only a few annotated support images. Shaban *et al*. [34] introduced the few shot segmentation task and proposed a two-branch networks to generate classifier weights from the support images for query image segmentation. Rakelly *et al*. [31] constructed the global conditioning prototype from the support set and concatenated it to the query representation. Following this prototype paradigm, MM-Net [47] introduced a set of learnable memory embeddings to store the meta-class information during training and transfer them to novel classes during the inference stage. Besides, several methods have also exploited pixel-level information for few shot segmentation. PGNet [51] used a graph attention unit to build pixel level dense similarity between the query and support images. PFENet [39] calculated the cosine similarity on high-level features without trainable parameters to create a prior mask and introduced a feature enrichment module to reduce the spatial inconsistency between the query and support samples. DCAMA [35] further proposed a cross attention weighted mask aggregation with multi-scale mechanism for few shot segmentation.

2.3. Part Segmentation

Part segmentation has also received considerable attention in semantic segmentation. Earlier methods usually treated part segmentation as a semantic segmentation problem, and most researches focused on part segmentation for humans [8, 14, 16, 22, 32]. For example, PGN [8] proposed a part grouping network to solve multi-person human parsing in a unified network, which contains two twinned grouping tasks, *i.e*., semantic part segmentation and instance aware edge detection. CE2P [32] introduced a simple yet efficient context embedding with edge perceiving framework by leveraging the useful properties to conduct human parsing. Besides human, several other categories have also been studied, such as faces [20] and animals [5, 42]. Recently, PPS [7] proposed two part annotated datasets, *i.e*., Cityscapes PPS and Pascal VOC PPS for panoptic part segmentation. Subsequently, Panoptic-PartFormer [18] proposed a unify model to learn thing, stuff, and part prediction tasks simultaneously in an end-to-end manner.

3. Proposed Method

Given a semantic map $M \in \{0,1\}^{H\times W\times C}$ with C categories, semantic image synthesis aims to generate the corresponding images $\hat{I} \in \mathbb{R}^{H\times W\times 3}$. One main challenge is to generate photo-realistic parts from the semantic map. To address this problem, we present to **i**nfer **P**arts from **O**bject **S**hap**E** (iPOSE), and leverage it for improving semantic image synthesis. Specifically, a PartNet is first proposed to infer the part map from the object shapes with the guidance of pre-defined support part maps (Sec. 3.1). Following [25, 44], we adopt the instance-level segmentation map to obtain the shape of each object in M. With the part map P, we further present a part semantic modulation ResBlock to incorporate it with the semantic map M and the 3D noise Z to modulate the generation process (Sec. 3.2). To facilitate model training, several loss terms are introduced to encourage the model to generate photo-realistic images (Sec. 3.3).

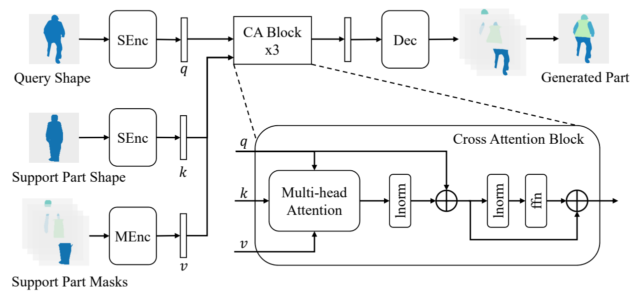


Figure 2. Illustration of our PartNet. The support part map is first decomposed into the support part shape and the support part masks as inputs. Cross attention block is adopted to aggregate the part information from the support features.

3.1. Inferring Parts from Object Shape

To improve the part synthesis, we first present to learn a PartNet based on existing part segmentation datasets [7], which infers the part map P from the given semantic map M. Specifically, the semantic map M can be decomposed into several object shapes $\{(O_i, y_i)\}$, where O_i denotes the cropped object shape and y_i is the corresponding category. For each object shape O_i, PartNet predicts the corresponding object part map P_i. To handle those objects with novel categories (*i.e*., category not in part training datasets [7]), we further introduce the few shot mechanism to learn the PartNet. And for each category y, a few object part maps $S_y = \{S_{y,0}, \cdots, S_{y,k-1}\}$ are selected as support set to guide the part prediction. For brevity, we take $k = 1$ as example, and our design can be easily extended to few shot, *e.g*., $k = 3$ in our implementation. The architecture of our PartNet is illustrated in Fig. 2. For each query object O_q, the corresponding support S_{y_q} is decomposed into the support part shape $O^S_{y_q}$ and the support part masks $O^M_{y_q}$ as inputs. Each mask denotes an object part, *e.g*., human's head, arm and leg, *etc*.

Architecture of PartNet. Following few shot segmentation, our PartNet aims to predict each part of O_q based on its similarity with $O^S_{y_q}$ and the part prior $O^M_{y_q}$. However, there is a lack of texture information for the object shape, which is unsuitable for the pre-trained image encoder. Instead, we utilize a shape encoder to extract the shape and

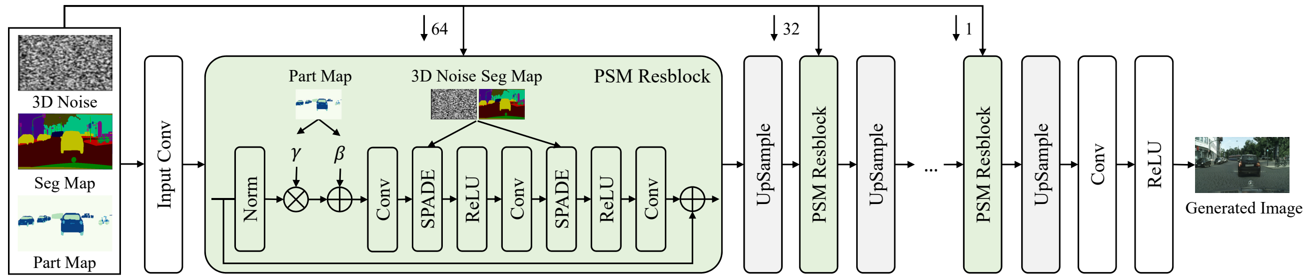


Figure 3. Architecture of our generator. It takes part map, semantic map and 3D noise as the input, while performing part semantic modulation for image synthesis.

position information from the object shapes O_q and $O^S_{y_q}$. A position embedding $\mathbf{p}$ [40] is further concatenated with O_q and $O^S_{y_q}$ as input. Besides, to exploit the part information, a mask encoder is further introduced to encode $O^M_{y_q}$,

$$\mathbf{q} = \mathrm{SEnc}(O_q),\ \mathbf{k} = \mathrm{SEnc}(O^S_{y_q}),\ \mathbf{v} = \mathrm{MEnc}(O^M_{y_q}), \quad (1)$$

where SEnc and MEnc denote the shape and mask encoder, respectively. Additionally, to perceive pixels' relative position of the whole object shape, the multi-scale mechanism is also adopted by the encoders. Features from different scales are upsampled and concatenated as encoder output. Furthermore, to aggregate the part information for part prediction, we introduce a cross attention (CA) block. In particular, a shape position similarity matrix $\mathbf{qk}^T$ is calculated between $\mathbf{q}$ and $\mathbf{k}$. The same part region in two object shapes tends to have a large similarity (*e.g.*, head part in O_q and $O^S_{y_q}$). Then, it multiplies with $\mathbf{v}$ to aggregate the part information. Layer norm and FFN are also adopted in CA block to improve the performance. Finally, the obtained feature is passed through the decoder to generate object part map P_q.

Generalizing to novel categories. We pre-train the PartNet on the basis categories of constructed part object dataset (see Sec. 4.1) and fix it during the following training. For those objects in M with novel categories (not in basis categories, *e.g.*, stop sign in COCO), our PartNet can be easily generalized to handle these objects. For example, to predict the part map for stop sign category, we just need to select k stop sign shapes from the COCO dataset, and manually annotate their part maps as supports (*e.g.*, the word STOP and the background). Then, given a new stop sign shape, our PartNet can leverage the supports to infer the part map without fine-tuning (as shown in Fig. 1). More details about the pre-training and novel categories can be found in *Suppl.*

3.2. Leveraging Parts for Semantic Image Synthesis

With the part map P, we then take it with the semantic map M to generate the image. Following [33], a random sampled 3D noise Z is also incorporated. Intuitively, the part map represents the structure information of the generated image, yet the semantic map and 3D noise provide the semantic texture information. To disentangle the structure with texture for image synthesis, we further present a part semantic modulation Resblock to modulate the activations.

Part Semantic Modulation Resblock. As illustrated in Fig. 3, to achieve a disentangled synthesis, we inject the part and semantic sequentially. Firstly, to inject the structure information, the part map P is used to modulate the activations. To encourage it to guide the structure synthesis independently, the modulation is performed spatially, *i.e.*, $\gamma \in \mathbb{R}^{H\times W\times 1}$ and $\beta \in \mathbb{R}^{H\times W\times 1}$. Besides, for those regions in the part map without structure information, we introduce additional noise to increase the diversity. Then, SPADE [28] is adopted to take semantic map M and 3D noise Z together to modulate the activations to inject the semantic texture information. By the separate modulation, our iPOSE can disentangle structure and texture successfully, and we also found that performing part modulation and semantic modulation sequentially is beneficial for generating images with photo-realistic parts.

Generator. Following [33], we stack the part semantic modulation (PSM) Resblocks and upsampling layers to constitute our generator (as shown in Fig. 3). The semantic map M, part map P and 3D noise Z are resized and fed to each PSM Resblock to guide the image synthesis.

$$\hat{I} = G(M, P, Z). \quad (2)$$

3.3. Learning Objective

We incorporate several loss terms to encourage the model to generate photo-realistic images. Following OASIS [33], we first introduce the $(N+1)$-class adversarial loss $\mathcal{L}^{N+1}_G$, $\mathcal{L}^{N+1}_D$, and the LabelMix loss $\mathcal{L}_{label}$ to train our model. In addition, to improve the quality of synthesized objects, we suggest a global adversarial loss [28] and an object-based CLIP style loss [54].

Global Adversarial Loss. Although the $(N+1)$-class adversarial loss can improve the semantic alignment of each pixel, it lacks the global constraint on images. To further improve the generator to synthesize realistic images, we suggest a global adversarial loss,

$$\mathcal{L}^{global}_D = \mathbb{E}_I[\log D(I)] + \mathbb{E}_{M,Z}[\log(1 - D(G(M,P,Z))], \quad (3)$$

$$\mathcal{L}^{global}_G = \mathbb{E}_{M,Z}[\log(1 - D(G(M,P,Z))], \quad (4)$$

(a) Qualitative comparison on Cityscapes

(b) Qualitative comparison on ADE20K (top two rows) and COCO-Stuff (bottom two rows)

Figure 4. Visual comparisons on the Cityscapes (1st $\sim$ 2nd rows), ADE20K (3rd $\sim$ 4th row) and COCO-stuff (5th $\sim$ 6th rows) datasets.

where Z is the 3D noise, and I, M denote the real image and its corresponding semantic map, respectively.

Object-level CLIP Style Loss. To facilitate the learning of object synthesis, we also introduce an object-level CLIP style loss [54]. Specifically, CLIP image encoder is adopted as the feature extractor, and extracts the intermediate tokens of images (I and $\hat{I}$) from the l-th layer. Then we align each token of generated image $\boldsymbol{F}_{\hat{I}}$ with the closest token of real image $\boldsymbol{F}_I$, where $\boldsymbol{F}_{\hat{I}} = \{\boldsymbol{F}_{\hat{I}}^1, \dots, \boldsymbol{F}_{\hat{I}}^n\}$ and $\boldsymbol{F}_I = \{\boldsymbol{F}_I^1, \dots, \boldsymbol{F}_I^n\}$ are the extracted tokens,

$$\mathcal{L}_{style} = \max\left(\frac{1}{n}\sum_i \min_j \boldsymbol{C}_{i,j}, \frac{1}{m}\sum_j \min_i \boldsymbol{C}_{i,j}\right), \quad (5)$$

where $\boldsymbol{C}$ is the cost matrix to measure the token-wise distances from F_I to $F_{\hat{I}}$, and each element of $\boldsymbol{C}$ is given by,

$$\boldsymbol{C}_{i,j} = 1 - \frac{\boldsymbol{F}_{\hat{I}}^i \cdot \boldsymbol{F}_I^j}{|\boldsymbol{F}_{\hat{I}}^i||\boldsymbol{F}_I^j|}. \quad (6)$$

It is worthy noting that, to emphasize object synthesis, we calculate the $\mathcal{L}_{style}$ with the objects cropped from I and $\hat{I}$.

The overall learning objective can be summarized as,

$$\mathcal{L}_G = \mathcal{L}_G^{N+1} + \lambda_{global}\mathcal{L}_G^{global} + \lambda_{style}\mathcal{L}_{style}, \quad (7)$$

$$\mathcal{L}_D = \mathcal{L}_D^{N+1} + \lambda_{global}\mathcal{L}_D^{global} + \lambda_{label}\mathcal{L}_{label}, \quad (8)$$

where λ_{global}, λ_{style} and λ_{label} denote the trade-off parameters for different loss terms.

4. Experiments

4.1. Experimental settings

Dataset for Part Seg. We first construct an object part dataset based on the Cityscapes PPS and Pascal VOC PPS datasets [7]. Specifically, each object part is cropped and resized to 64×64 based on its bounding box, and formed as paired (object shape, object part map). There is a total of 21 categories in the dataset. Following [34], we use 20 basis categories for PartNet training (*e.g.*, human, car, bus, and sheep, *etc.*) and use the remaining category as novel category for openset testing (*i.e.*, cat). To obtain the support object part maps for each category, we use k-means to cluster the training shapes into k clusters based on the shape similarity metric [36]. Besides, for those novel categories in image synthesis datasets, we have annotated k support part maps manually to perform part prediction, including the washing machine, van, zebra, cat, and stop sign. More details can be found in *Suppl*.

Figure 5. Multi-modal synthesis results of our iPOSE. By injecting different 3D noises into the assigned region, our iPOSE can achieve global-level (first 2 rows), object-level (3rd row), and also part-level (last row) image editing.

Table 1. Quantitative comparison with existing methods on different datasets. ↑ (↓) denotes the higher (lower) is better.

Method	Cityscapes			ADE20K			COCO-Stuff		
	FID (↓)	AC (↑)	mIOU (↑)	FID (↓)	AC (↑)	mIOU (↑)	FID (↓)	AC (↑)	mIOU (↑)
SIMS [30]	49.7	75.5	47.2	n/a	n/a	n/a	n/a	n/a	n/a
SPADE [28]	71.8	81.9	62.3	33.9	79.9	38.5	22.6	67.9	37.4
CC-FPSE [23]	54.3	82.3	65.5	31.7	82.9	43.7	19.2	70.7	41.6
SC-GAN [46]	49.5	82.5	66.9	29.3	83.8	45.2	18.1	72.0	42.0
OASIS [33]	47.7	n/a	69.3	28.3	n/a	48.8	17.0	n/a	44.1
RESAIL [36]	45.5	**83.2**	69.7	30.2	84.8	49.3	18.3	73.1	44.7
SAFM [25]	49.5	83.1	70.4	32.8	86.6	50.1	24.6	73.4	43.3
SDM [45]	42.1	n/a	**77.5**	27.5	n/a	39.2	n/a	n/a	n/a
Ours	**41.3**	82.2	70.6	**26.9**	**87.1**	**53.8**	**15.7**	**74.8**	**45.1**

Figure 6. Part editing results of our iPOSE. Our method allows editing parts of generated objects through the support part maps. For example, we can remove the light, wheel, or license of the car.

Datasets for Semantic Image Synthesis. Following [25, 28, 33], we conduct the experiments on Cityscapes [6], ADE20K [55], and COCO-Stuff [2]. Cityscapes includes 35 semantic categories, and consists of 3,000 training images and 500 validation images. ADE20K contains over 20,000 training images and 2,000 validation images with 150 semantic classes. COCO-Stuff consists of 118,000 training images and 5,000 validation images. In our experiments, the images in ADE20K and COCO-Stuff are resized and cropped to 256 × 256, while those in Cityscapes are processed to 256 × 512.

Evaluation Metrics. To evaluate our method, we adopt Pixel ACcuracy (AC), mean Intersection-Over-Union (mIOU), and Frechet Inception Distance (FID) [9] as the metrics. AC and mIOU measure the semantic consistency between the generated image and given input [25, 28, 33], and pretrained segmentation models are adopted to compute segmentation accuracy [3, 48, 50]. Furthermore, FID evaluates the quality and diversity of generated images.

Implementation Details. We implement our model with Pytorch [29] and train it with 4 Tesla V100 GPUs. For PartNet, we set the number of support part maps to be $k = 3$. For synthesis model, Adam [12] is adopted with $\beta_1 = 0$ and $\beta_2 = 0.999$ and the learning rates are set to 0.0001 for gener-

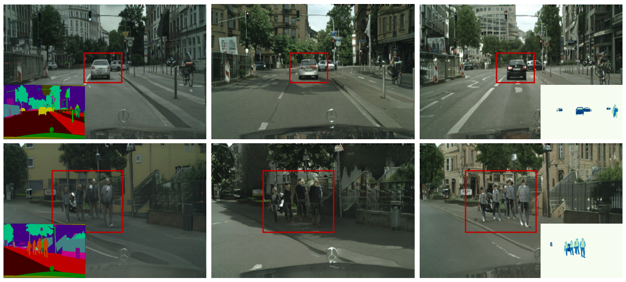


Figure 7. Visual comparisons of different part injection methods.

ator and 0.0004 for discriminator. The trade-off parameters λ_{global}, λ_{style} and λ_{label} are set to 1, 10 and 10, respectively. Following [28], we apply the spectral normalization [26] to each layer in both generator and discriminator.

4.2. Qualitative Results

Fig. 4 gives the qualitative comparisons with the SPADE [28], CC-PFSE [23], OASIS [33], and SAFM [25] on the three datasets. For illustration, the inferred part maps are also shown on the right. From the figure we can see that, our iPOSE predicts the plausible part map based on the given semantic map. For the novel categories not in object part training set (*e.g.*, stop sign), our PartNet still works by providing 3 annotated stop sign part maps as supports, demonstrating its generalization ability. Furthermore, our iPOSE generates images that are semantically consistent with the part maps, and also with high quality and fine details (*e.g.*, light of car, head of sheep, and stop sign). In contrast, only exploiting the object information, SPADE, CC-FPSE, and OASIS generate images with noticeable artifacts (*e.g.*, car, sheep and stop sign). Although SAFM explores the shape-aware position descriptor, it is region-aware and lacks part prior, resulting in limited improvement (*e.g.* sheep and stop sign). In comparison with the competing methods, our iPOSE can generate more photo-realistic images, clearly demonstrating its superiority.

In addition, our iPOSE learns to disentangle the structure and texture for image synthesis, and allows to edit the texture and structure independently. Fig. 5 illustrates the multi-modal synthesis results by injecting different sampled noises into assigned region. As shown in the figure, benefited from the predicted part map, our iPOSE can achieve global-level, object-level, and part-level image editing. Moreover, we can edit the structure of generated objects. As shown in Fig. 6, we can remove the license, light, and wheel of generated images, thus enabling the users to control the image synthesis more flexibly. Similar results in Fig. 1(d). More qualitative results are shown in the *Suppl*.

4.3. Quantitative Results

Table 1 lists the quantitative comparison between our iPOSE with the state-of-the-art methods [23, 25, 28, 30, 33, 36, 45, 46]. From the table, our method performs favorably against the competing methods on most datasets in terms of the three metrics. Benefited from the introduced part map and losses, our iPOSE can generate images with higher quality and finer details, especially for object regions, resulting in a notable improvement of FID metric (+0.8, +0.6, +1.3 on three datasets, respectively). Besides, the images generated by our iPOSE are better semantically aligned with the input semantic map on ADE20K and COCO-stuff, further demonstrating its effectiveness.

Table 2. User study on different datasets. The numbers indicate the percentage (%) of volunteers who favor the results of our method over those of the competing methods.

Dataset	Ours vs. SPADE	Ours vs. CC-FPSE	Ours vs. OASIS	Ours vs. SAFM
Cityscapes [6]	75.6	64.8	67.3	58.6
ADE20K [55]	72.3	62.4	60.5	57.6
COCO-Stuff [2]	67.7	57.2	57.4	61.9

Table 3. Ablation studies on the losses, part map, and different part injection methods. With the introduced $\mathcal{L}^g_{G/D}$, $\mathcal{L}_{style}$ and the proposed part semantic modulation, our method achieves better quantitative performance.

Part Inject	$\mathcal{L}^g_{G/D}$	$\mathcal{L}_{style}$	FID(↓)	mIOU(↑)	AC(↑)	obj FID (↓)
			47.7	66.9	81.5	44.1
	✓		43.6	66.7	81.9	39.2
	✓	✓	42.8	70.5	82.1	37.5
Ours PSM	✓	✓	**41.3**	**70.6**	**82.2**	**30.4**
SPADE	✓	✓	42.9	69.9	81.9	31.2
Concat	✓	✓	42.7	70.6	82.0	31.5

Table 4. Ablation study on the number of support part maps. Basis AC and Novel AC denote the testing accuracy on basis and novel categories of object part dataset, repectively. We also list the corresponding FID score on COCO dataset.

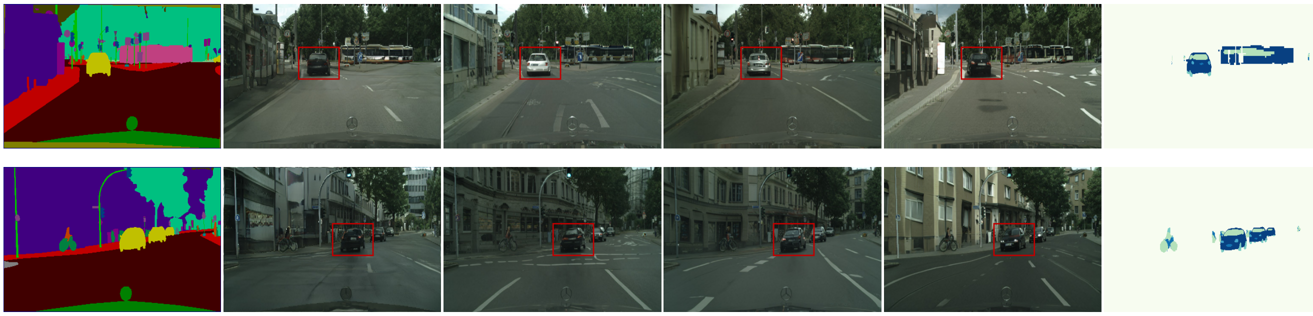
Methods	1-shot	2-shot	3-shot	4-shot
Basis AC (↑)	94.30	94.37	94.38	94.37
Novel AC (↑)	81.41	84.48	85.00	85.16
COCO FID (↓)	15.76	15.73	15.72	15.72

User Study. Following [28], we conduct the user study to compare with four competing methods [23, 25, 28, 33]. Specifically, the participants[1] were given an input segmentation mask and generated results from two different methods, and required to select the image that has better performance in semantic alignment and photo-realistic appearance. For each comparison, we randomly generate 500 questions for each dataset, and each question is answered by 10 different workers. From Table 2, users tend to favor our results on all the datasets, especially on Cityscapes.

4.4. Ablation Studies

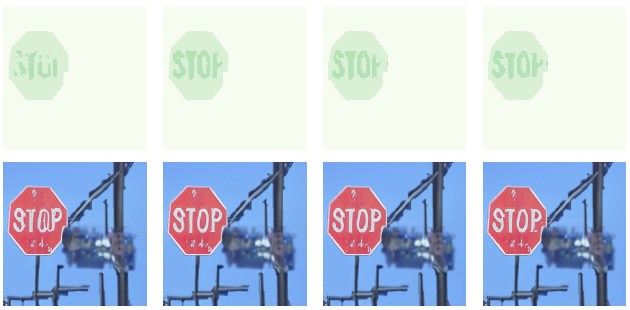
We conduct the ablation studies on Cityscapes to verify the effectiveness of our introduced part map and losses.

[1] The identities will not be recorded.



Semantic Map | Baseline | $+\mathcal{L}^{g}_{G/D}$ | $+\mathcal{L}^{g}_{G/D} + \mathcal{L}_{style}$ | Ours | Part Map

Figure 8. Visual comparisons of different variants. Our full model with the introduced losses and part semantic modulation achieves better visual quality, especially for object details.



$k = 1$ | $k = 2$ | $k = 3$ | $k = 4$

Figure 9. Visual comparisons on the effect of different number of support part maps.

Furthermore, the effect of the number of support part maps is also analyzed.

Effectiveness of part map and losses. To demonstrate the effectiveness of our proposed part map and losses, we conduct a comparison with the following variants. (i) Baseline. We adopt the OASIS [33] as our baseline. (ii) We introduce the global adversarial loss $\mathcal{L}^{g}_{G/D}$ to the baseline model. (iii) We further introduce the object-level CLIP style loss $\mathcal{L}_{style}$. (iv) Ours. We finally incorporate the part map and the part semantic modulation resblock to improve image synthesis. The results are listed in Table 3 and Fig. 8. From the table, with introduction of $\mathcal{L}^{g}_{G/D}$, $\mathcal{L}_{style}$, and the part map with PSM, the quality of generated images can be improved gradually. As shown in Fig. 8, $\mathcal{L}^{g}_{G/D}$ improves the layout of generated images, and $\mathcal{L}_{style}$ can further enhance the texture synthesis. Finally, with the introduced part map and PSM Resblock, our method can generate images with high-quality and more photo-realistic parts. Furthermore, we have cropped and resized each object to 128×128 to calculate the object-level FID. From Table 3, our Part&PSM brings biggest improvement to object-level FID, demonstrating its effectiveness.

Effectiveness of Part Injection. We further analyze the effect of the injection method for the part map, and compare it with two variants. (i) SPADE. We concatenate the part map with semantic map and 3D noise, and use SPADE to modulate the activations. (ii) Concat. We use part as input, and concatenate it with the input feature of each layer. From the Table 3 and Fig. 7, injecting the part map through SPADE and Concat is limited in leveraging part map, and the results degenerated to some extent. In contrast, our methods use the part map and semantic map to modulate the activations sequentially, and achieves better synthesis quality.

Effectiveness of the number of support part maps. We have also conducted the ablations on the number of support part maps. Specifically, we train the model with mixed numbers of supports ($k = 1 \sim 4$), and test it with different numbers of supports, respectively. From Table 4 and Fig. 9, with the increasing support number, the testing accuracy on both basis and novel categories increases, resulting in more plausible part maps and photo-realistic images. Similar trends with the FID on COCO dataset. Moreover, our generator has the ability to fix the incorrect part map to generate plausible results (*e.g.*, $k = 1$). To balance the efficiency and the performance, we select $k = 3$ in default as our final model. More ablations can be found in the *Suppl*.

5. Conclusion

In this paper, we propose a novel method, termed iPOSE, to exploit the part-level information for improving semantic image synthesis. A PartNet is first proposed to infer the part map from object shapes based on pre-defined support part maps. A part semantic modulation Resblock is then presented to leverage the inferred part map with semantic map to perform structure-texture disentangled image synthesis. With the further introduced global adversarial and object-level CLIP style losses, our iPOSE can generate photo-realistic images, especially for the object parts. Experiments show that our iPOSE performs favorably against the state-of-the-art methods on the three challenging datasets both qualitatively and quantitatively.

Acknowledgement. This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0104500, and by the National Natural Science Foundation of China (NSFC) under Grant No.s U19A2073 and 62006064.

References

[1] Wei Ao, Shunyi Zheng, and Yan Meng. Few-shot semantic segmentation via mask aggregation. *arXiv preprint arXiv:2202.07231*, 2022. 3

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 6, 7

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 6

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017. 2

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6, 7

[7] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 2, 3, 5

[8] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision*, pages 770–785, 2018. 3

[9] 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. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1, 2

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[13] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021. 3

[14] Jianshu Li, Jian Zhao, Yunpeng Chen, Sujoy Roy, Shuicheng Yan, Jiashi Feng, and Terence Sim. Multi-human parsing machines. In *Proceedings of the ACM International Conference on Multimedia*, pages 45–53, 2018. 3

[15] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4220–4229, 2019. 2

[16] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. *arXiv preprint arXiv:1709.03612*, 2017. 3

[17] Shijie Li, Ming-Ming Cheng, and Juergen Gall. Dual pyramid generative adversarial networks for semantic image synthesis. *arXiv preprint arXiv:2210.04085*, 2022. 2

[18] Xiangtai Li, Shilin Xu, Yibo Yang Cheng, Yunhai Tong, Dacheng Tao, et al. Panoptic-partformer: Learning a unified model for panoptic part segmentation. *arXiv preprint arXiv:2204.04655*, 2022. 3

[19] Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, and Krishna Kumar Singh. Collaging class-specific gans for semantic image synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 14418–14427, 2021. 1, 2

[20] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5654–5663, 2019. 3

[21] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. A survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *arXiv preprint arXiv:2207.10309*, 2022. 2

[22] Si Liu, Yao Sun, Defa Zhu, Guanghui Ren, Yu Chen, Jiashi Feng, and Jizhong Han. Cross-domain human parsing via adversarial feature and label adaptation. In *Proceedings of the AAAI Conference On Artificial Intelligence*, volume 32, 2018. 3

[23] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, 32:570–580, 2019. 2, 6, 7

[24] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11573–11582, 2022. 3

[25] Zhengyao Lv, Xiaoming Li, Zhenxing Niu, Bing Cao, and Wangmeng Zuo. Semantic-shape adaptive feature modulation for semantic image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11214–11223, 2022. 1, 2, 3, 6, 7

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 7

[27] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: semantic editing of scenes by adding, manipulating or erasing objects. In *Proceedings of the European Conference on Computer Vision*, pages 394–411, 2020. 2

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1, 2, 4, 6, 7

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[30] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 1, 2, 6, 7

[31] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018. 3

[32] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference On Artificial Intelligence*, volume 33, pages 4814–4821, 2019. 3

[33] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 6, 7, 8

[34] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 3, 5

[35] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. *arXiv preprint arXiv:2207.08549*, 2022. 3

[36] Yupeng Shi, Xiao Liu, Yuxiang Wei, Zhongqin Wu, and Wangmeng Zuo. Retrieval-based spatially adaptive normalization for semantic image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11224–11233, 2022. 1, 2, 5, 6, 7

[37] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020. 2

[38] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2022. 3

[39] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[41] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9197–9206, 2019. 3

[42] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1573–1581, 2015. 3

[43] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 1, 2

[44] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 1, 2, 3

[45] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 1, 2, 6, 7

[46] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13749–13758, 2021. 2, 6, 7

[47] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 517–526, 2021. 3

[48] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 418–434, 2018. 6

[49] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5475–5484, 2021. 3

[50] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 472–480, 2017. 6

[51] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9587–9595, 2019. 3

[52] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5217–5226, 2019. 3

[53] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020. 3

[54] Yabo Zhang, Mingshuai Yao, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Towards diverse and faithful one-shot adaption of generative adversarial networks. *arXiv preprint arXiv:2207.08736*, 2022. 2, 4, 5

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 6, 7

[56] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2