Retrieval-based Spatially Adaptive Normalization for Semantic Image Synthesis

Yupeng Shi [1],   Xiao Liu [1], Yuxiang Wei [2], Zhongqin Wu [1], Wangmeng Zuo[2,3] (✉)

[1]Tomorrow Advancing Life, [2]Harbin Institute of Technology, [3]Peng Cheng Laboratory

{csypshi, ender.liux, yuxiang.wei.cs}@gmail.com wuzhongqin@tal.com wmzuo@hit.edu.cn

| (a) SPADE | (b) SEAN | (c) Ours |

Figure 1. Synthesized results of SPADE [22], SEAN [35] and our method. (a) With the class-level guidance, SPADE produces blurry synthesis results. (b) With the region-level style vector, SEAN generates better details, but still prefers spatially uniform synthesis result. (c) Benefited from pixel level fine-grained guidance, our RESAIL is effective in generating visually plausible image with clear details.

Abstract

Semantic image synthesis is a challenging task with many practical applications. Albeit remarkable progress has been made in semantic image synthesis with spatially-adaptive normalization, existing methods usually normalize the feature activations under the coarse-level guidance (e.g., semantic class). However, different parts of a semantic object (e.g., wheel and window of car) are quite different in structures and textures, making blurry synthesis results usually inevitable due to the missing of fine-grained guidance. In this paper, we propose a novel normalization module, termed as REtrieval-based Spatially AdaptIve normaLization (RESAIL), for introducing pixel level fine-grained guidance to the normalization architecture. Specifically, we first present a retrieval paradigm by finding a content patch of the same semantic class from training set with the most similar shape to each test semantic mask. Then, the retrieved patches are composited into retrieval-based guidance, which can be used by RESAIL for pixel level fine-grained modulation on feature activations, thereby greatly mitigating blurry synthesis results. Moreover, distorted ground-truth images are also utilized as alternatives of retrieval-based guidance for feature normalization, further benefiting model training and improving visual quality of generated images. Experiments on several challenging datasets show that our RESAIL performs favorably against state-of-the-arts in terms of quantitative metrics, visual quality, and subjective evaluation. The source code is available at https://github.com/Shi-Yupeng/RESAIL-For-SIS.

1. Introduction

Semantic image synthesis aims to generate photo-realistic image from the given semantic map. It is an im-portant problem in computer vision that can be adopted in a variety of downstream tasks such as virtual idol, special effect, robotics [13] and image manipulation [9].

Humans have a remarkable ability to produce new creation from past experiences as references. In their early ages, children can paint a picture including flowers, sky and buildings by referring to templates of representative objects and backgrounds. Thus, producing something from references is a natural way for image generation because editing the references and stitching them is relatively easy than creating the entire image out of thin air. Inspired by this spirit, early works have well studied reference-based image synthesis, where proper references are searched from external memories [4, 7, 11, 14, 16]. Nonetheless, the retrieval, editing and stitching are conducted in separated and handcrafted manners, which are optimized in a sub-optimal way. SIMS [23] leverages deep network for further improving the quality of reference-based synthesized results, but it simply takes the retrieved image as network input, which is limited in synthesizing complex real-world scenes.

With the recent advance of deep generative networks, some recent studies [21,22,25,26,35] tackle semantic image synthesis using a spatially-adaptive normalization architecture, achieving significant performance improvements. However, with the coarse-level guidance (*e.g.*, semantic class), these methods modulate the activations inside each semantic object in spatially uniform manner, regardless of the huge internal variation of the objects. This inevitably leads to blurry results, especially for large semantic object with complex parts. We take two representative spatially-adaptive normalization architectures as examples in Fig. 1. SPADE [22] leverages the semantic layout as input and learns the modulation parameters through several convo-

lution layers, being limited in generating high-quality object parts and leading to blurry synthesis results (Fig. 1(a)). SEAN [35] improves SPADE by extracting style codes from selected regions, leading to flexible style control. However, the style map is generated by broadcasting the style codes to the corresponding semantic regions, which also prefers spatially uniform synthesis result (Fig. 1(b)). Most recent methods, *e.g.*, CLADE [26] and OASIS [25], intrinsically are also based on coarse-level guidance.

In this paper, we tackle the above issues by presenting a novel feature normalization method, termed as REtrieval-based Spatially AdaptIve normaLization (RESAIL). Our intuition is two-fold. On the one hand, the object segment mask of the input semantic map can not only provide the semantic class but also the object shape. On the other hand, the training dataset contains rich shape and texture information of objects which cannot be entirely captured by the learned deep generative networks. Taking these intuitions into account, given a object segment mask, we present a retrieval paradigm for retrieving a segment image with the most similar shape from the training dataset. The retrieved segment images are then composited into a retrieval-based guidance, which naturally is spatially variant in pixel level. We further propose a retrieval-based spatially adaptive normalization, where retrieval-based guidance and semantic map collaborate to provide pixel level fine-grained modulation on feature activations. As shown in Fig. 1(c), benefited from pixel level fine-grained guidance, our RESAIL is effective in generating visually plausible image with clear details. In contrast to SIMS [23], our method leverages retrieval-based guidance for spatially adaptive normalization, which is more effective in synthesizing photo-realistic images. In comparison to SPADE [22] and SEAN [22], our RESAIL can effectively leverage pixel level fine-grained guidance for improving synthesized results.

When retrieval-based guidance is used for feature normalization, it is difficult to exploit perceptual supervision for training, due to that the ground-truth image corresponding to retrieval-based guidance is missing. On the contrary, the ground-truth image of a semantic map can be naturally treated as a retrieval-based guidance, while the ground-truth image itself can also be used to facilitate perceptual supervision. However, ground-truth image is quite different from real retrieval-based guidance, and using it as guidance cannot make the learned model generate better synthesis results in the testing stage. Instead, we introduce a data distortion mechanism on ground-truth images to mimic the quality of retrieval-based guidance. During training, the distorted ground-truth images are also used as alternatives of retrieval-based guidance, making it feasible to leverage perceptual supervision for improving model training and visual quality. Experiments on several challenging datasets show that our RESAIL performs favorably against state-of-the-

arts. The contributions of this work are summarized as:

- A novel retrieval-based synthesis model is proposed by leveraging the retrieval-based guidance as pixel level fine-grained modulation, *i.e.*, Retrieval-based Spatially Adaptive Normalization (RESAIL), for semantic image synthesis.
- During training, a data distortion mechanism on the ground-truth images is introduced to facilitate model training and improves visual quality of synthesized results.
- Extensive experiments show the effectiveness of our proposed method in synthesizing photo-realistic image from given semantic map.

2. Related Work

2.1. Semantic Image Synthesis

Many methods have been proposed to tackle semantic image synthesis. Here we focus on GAN-based methods, and also list other related methods [3, 17, 23].

Pix2pix [12] proposed a general framework for image-to-image translation, and Pix2pixHD [29] improved it for generating high-resolution images. In these methods, the semantic map is simply used as input to the network. SPADE [22] exploited the semantic maps to predict transformation parameters for modulating the activations in normalization layers. Auxiliary guidance (*e.g.*, style map [35] or 3D noise map [25]) are incorporated with semantic map for diverse synthesis and easier controlling (details of normalization layer are surveyed in Sec. 2.2). Instead of injecting semantic map into the network directly, CC-FPSE [19] and SC-GAN [30] leveraged semantic map to predict the external parameters (convolution kernels [19] or semantic vectors [30]), which are further used by another network to guide the image synthesis.

Elaborate networks have also been explored in semantic image synthesis. SPADE [22] employed a generator consisting of several residual blocks with upsampling layers and the PatchGAN discriminator. LGGAN [27] explored the local context information and introduce a local pathway in the generator for details synthesizing. CC-PFSE [19] and SC-GAN [30] employed two generators for coarse and fine image synthesis. Besides generator, CC-FPSE [19] proposed a feature-pyramid discriminator for semantically aligned image synthesis. SESAME [21] and OSAIS [25] improved the PatchGAN discriminator with a semantics-related mechanism. In addition, CollogeGAN [18] used the StyleGAN [15] as the generator to improve visual quality and explored the local context with class-specific models.

Among these methods, CC-FPSE and SC-GAN first synthesize a coarse image and use it to guide the fine image synthesis. While our method directly uses retrieval-based guidance to facilitate pixel level fine-grained modulation on activations.

(a) Retrieval-based Guidance

(b) Generator Network

(c) RESAIL ResBlk

Figure 2. Illustration of our method. (a) Given a semantic map M, we first retrieve a set of segments from the training dataset according to each semantic region of M and composite them into the retrieval-based guidance I^r. It provides a pixel-level fine-grained guidance for the semantic image synthesis. (b) The architecture of our generator. It takes the semantic map and guidance as input, and consists of several RESAIL ResBlocks following upsample layers. (c) Detailed architecture of the RESAIL ResBlock used in (b). It learns the pixel level fine-grained modulation parameters from the semantic map and guidance for modulating the normalized activations.

2.2. Conditional Normalization

Conditional normalization [6, 10, 22, 35] has been extensively studied in conditional image synthesis. Different from the earlier normalization techniques, conditional normalization layers require external data to learn the affine transformation parameters which are then used to modulate the normalized activations. For example, Conditional Instance Normalization (CIN) [6] modified the γ and β parameters of Instance Normalization (IN) from length C vectors to $N \times C$ matrices, and the external style s is used to index the row of γ and β. AdaIN [10] learned a neural network that mapping the given style vectors to the γ and β parameters of IN. CIN and AdaIN perform uniformly across spatial coordinates, which may not be beneficial for the spatially-varying synthesis tasks, such as semantic image synthesis. Instead, SPADE [22] proposed to learn a spatially-varying affine transformation in the semantic class level. SEAN [35] extended the SPADE with a style map which is composed of the style vectors for each region, and learned the transformation parameters from both semantic map and the style map in the region level. OASIS [25] introduced a 3D noise concatenating with the semantic map to perform the spatially-variant normalization, but the 3D noise provides limited semantic information for the synthe-

sis. CLADE [26] learned a parameter bank for each semantic class, which is used to generate the parameters for modulation, but still limited to coarse-level guidance.

In contrast, our RESAIL module takes the retrieved results to introduce pixel level fine-grained guidance for semantic image synthesis.

2.3. Retrieval-based Image Synthesis

In the early studies, many retrieval-based methods [4, 7, 11, 14, 16] have been proposed for conditional image synthesis. For example, Hays et al. [7] used a collection of images as retrieval database for image completion. In testing stage, similar images are retrieved via the descriptor matching and used to complete the missing regions. Lalonde et al. [16] retrieved object segments from a large image database and then interactively composited them into an image. Chen et al. [4] developed a system that retrieved and synthesized an image from a freehand sketch with associated text labels. Isola and Liu [11] presented an analysis-by-synthesis method that retrieved segments according to the given query image and combined these segments to form a "scene collage" that explains the query. Recently, SIMS [23] leveraged deep network for improving the quality of synthesized results. However, it simply takes the retrieved image as network input, failing in exploiting progress in conditional

| Ground-truth I^{gt} | Decomposed | Distortion | Distorted | Distorted GT $\tilde{I}^{gt}$ |

Figure 3. Illustration of data distortion on ground-truth image I^{gt}. Specifically, I^{gt} is first decomposed into several segments based on semantic map. Then each segment is distorted separately by modifying shape, color, and resolution. Finally, distorted segments are composited into a distorted ground-truth image $\tilde{I}^{gt}$.

normalization. In contrast, our method uses retrieval-based guidance for spatially adaptive normalization, which is beneficial for synthesizing photo-realistic images.

3. Proposed Method

Given a semantic map $M \in \{0,1\}^{H \times W \times C}$, semantic image synthesis aims to generate the corresponding images $\hat{I} \in \mathbb{R}^{H \times W \times 3}$. Here H, W, and C denote the height, width, and number of categories in semantic map, respectively. In this section, we first present a retrieval paradigm to produce a retrieval-based guidance I^r (Sec. 3.1). We also introduce the distorted ground-truth as the alternative of retrieval-based guidance, and introduce the perceptual supervision to facilitate model training, (Sec. 3.2). With the guidance, we propose a Retrieval-based Spatially Adaptive Normalization (RESAIL) to perform pixel level fine-grained modulation on activations (Sec. 3.3). Finally, we introduce several loss terms for training the model to generate the photo-realistic images (Sec. 3.4).

3.1. Retrieval-based Guidance

Given the semantic map M, we first present a retrieval paradigm to obtain the retrieval-based guidance from the training dataset which contains pixel level fine-grained information. As shown in Fig. 2(a), the semantic map M can be decomposed into several object segment masks $M = \{(M_i^s, y_i^c)\}$, where M_i^s denotes the cropped binary segment mask of one object and y_i^c is the corresponding category. Similarly, a training image can also be decomposed into segment images according the semantic map. We define these segments as the retrieval unit. In training or testing stage, the retrieval-based guidance is obtained by,

$$I^r = \Theta\left(\{\Gamma(\mathcal{D}^{tr}, M_i^s, y_i^c) \mid (M_i^s, y_i^c) \in M\}\right) \quad (1)$$

where $\Gamma(\mathcal{D}^{tr}, M_i^s, y_i^c)$ denotes the retrieval function defined on training dataset $\mathcal{D}^{tr}$. It finds a segment image with category y_i^c and the most similar shape with M_i^s. When there is no matching segment image in training dataset, we replace it with a black image. $\Theta(\cdot)$ function recomposes the retrieved segments to form the guidance. Note that, in the training stage, we ignore the original segment images corresponding to M and retrieve the other most compatible segment images based on the geometric consistency score [28]. More details are provided in the *Suppl*.

3.2. Distorted Ground-truth as Guidance

The retrieval-based guidance image I^r lacks of paired ground-truth, making it impossible to exploit perceptual supervision during training. Intuitively, the ground-truth image can be used as both the guidance and the ground-truth, resulting a paired training data. However, ground-truth image is quite different from real retrieval-based guidance (*e.g.*, color, shape and resolution distortion usually are inevitable in retrieval-based guidance, see Fig. 2(a)). Thus, directly using ground-truth as guidance in training benefits little to learn generator that works well for retrieval-based guidance. Instead, we introduce a data distortion mechanism on ground-truth images to mimic the quality of retrieval-based guidance. As illustrated in Fig. 3, the ground-truth is first decomposed into a set of separate segments. Then these segments are distorted by changing shape, color and resolution, respectively. Finally, the distorted segment images are recomposed into the distorted ground-truth $\tilde{I}^{gt}$, which can be utilized as alternative of retrieval-based guidance. Due to that the distorted ground-truth has the real paired image (*i.e.*, original ground-truth), we can introduce perceptual supervision on synthesis results to facilitate model training and improve visual quality.

3.3. Network Architecture

Retrieval-based Spatially Adaptive Normalization. With the guidance I^r (or $\tilde{I}^{gt}$) and semantic map M, we propose a REtrieval-based Spatially AdaptIve normaLization (RESAIL) to perform pixel level fine-grained modulation on feature activations. Specifically, we adopt the conditional normalization architecture with spatially adaptive modulation. As the guidance image contains pixel level information about the object class, we first use it to learn the *fine-grained modulation parameters* (*i.e.*, γ^r for scale and β^r for bias) by a four-layer convolutional network. Due to there are some semantic regions missing in the retrieval-based guidance image (no matching segment images or shape gaps), 3×3 kernel is used in the convolutional layer to complete the information in the missing region. Besides, we use the AdaIN incorporated with the semantic map in the intermediate two layers to further enrich the semantic information of the missing area. The detailed structure is shown in Fig. 2(c). Analogous to [22,35], we also learn the *coarse modulation parameters* (*i.e.*, γ^s and β^s) from the semantic map. Two sets of parameters are weighted summed to get the final pixel level fine-grained modulation parameters,

$$\begin{aligned} \gamma &= \alpha_\gamma \gamma^s + (1 - \alpha_\gamma)\gamma^r, \\ \beta &= \alpha_\beta \beta^s + (1 - \alpha_\beta)\beta^r, \end{aligned} \quad (2)$$

where α_γ and α_β are learnable weight parameters, and the input activations are finally modulated by,

$$RESAIL(\mathbf{h}, M, I^r) = \gamma_{c,y,x} \frac{\mathbf{h}_{n,c,y,x} - \mu_c}{\sigma_c} + \beta_{c,y,x}, \quad (3)$$

4

(a) Qualitative comparison on Cityscapes.

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

Figure 4. Qualitative comparison of our method with the competing methods on the (a) Cityscapes, (b) ADE20K and COCO-Stuff datasets. Our model generates images with better perceptual quality and finer details.

where $\mathbf{h}$ denotes the input activations with a batch of N samples, μ and σ denote the mean and standard deviation of the activations. $(n \in N, c \in C, y \in H, x \in W)$ sites the modulated activations value. More details about the RESAIL module are provided in the *Suppl.*

Generator. Fig. 2(b) illustrates the architecture of our generator G, which is built on the generator of SPADE [22]. Analogous to [22], we employ a generator consisting of several RESAIL residual blocks (RESAIL ResBlk) with up-sampling layers. The semantic map M and guidance (I^r or $\tilde{I}^{gt}$) are resized and fed to each RESAIL module to guide the image synthesis,

$$\hat{I} = G(M, I^r), \quad \hat{I}^{gt} = G(M, \tilde{I}^{gt}). \quad (4)$$

3.4. Loss Functions

As discussed above, we first introduce the perceptual loss $\mathcal{L}_{vgg}$ [29] and feature matching loss $\mathcal{L}_{FM}$ [29] between I^{gt} and the synthesized image $\hat{I}^{gt}$ to facilitate the model training. To encourage the generator to synthesize photo-realistic images, we also introduce the adversarial loss [21] on synthesized images (both $\hat{I}$ and $\hat{I}^{gt}$). Besides, to emphasize the synthesis of each semantic region, we incorporate a segmentation loss with the model training. Specifically, we

introduce a pretrained segmentation network S to classify the category of each entry on the generated image,

$$\mathcal{L}_{cls} = -\mathbb{E}_M \left[\sum_c \alpha_c \sum_{i,j} M_{i,j,c} \log S(\hat{I})_{i,j,c} \right], \quad (5)$$

where α_c denotes the class balancing weight [25], and S is pretrained on the training dataset. $\mathcal{L}_{cls}$ is introduced on both $\hat{I}$ and $\hat{I}^{gt}$. Finally, we combine all the above losses to give the overall learning objective,

$$\mathcal{L} = \lambda_{vgg}\mathcal{L}_{vgg} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls}, \quad (6)$$

where λ_* denotes tradeoff parameters for different losses.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our model on four common used datasets, Cityscapes [5], ADE20K [34], ADE20K-outdoor and COCO-Stuff [1]. The training set of Cityscapes consists of 3,000 images, including 35 semantic categories, while the validation set consists of 500 images. The ADE20K dataset contains over 20,000 images for training and 2,000 images for validation with 150 semantic classes in total. The ADE20K-outdoor dataset is a subset of ADE20K

Figure 5. Multi-modal synthesis capability of our method. Each column represents the synthesized results with the given semantic map (top row). During testing, we retrieve a set of different guidance images, resulting diverse synthesized images (*i.e.*, left 2 columns). We can also fix most semantic regions and change the retrieved segments of certain objects to achieve local editing results (*i.e.*, right 3 columns). The retrieval-based guidance images used for the image synthesis are given in the red rectangle.

Figure 6. Ablation study on the RESAIL module. Model+ denotes introducing the retrieval-based guidance to the model (see Sec. 4.4 and the *Suppl* for more details). With the proposed RESAIL module and the retrieval-based guidance, our method produces more photo-realistic details (red circle). Zoom for a better view.

only containing outdoor scenes. COCO-Stuff consists of 118,000 training images and 5,000 validation images.

Evaluation Metric. Pixel ACcuracy (AC) and mean Intersection-Over-Union (mIOU) are adopted, which measure the agreement between synthesized image and given input [3, 21, 22]. They both require a pretrained segmentation model to compute segmentation accuracy [2, 31, 32]. We also utilize Frechet Inception Distance (FID) [8] to evaluate the quality of synthesized images.

Implementation Details. We train our model on four Tesla v100 GPUs and adopt ADAM optimizer with $\beta_1 = 0$ and $\beta_2 = 0.999$ where the learning rates are set to 0.0001 for generator and 0.0004 for discriminator. Additionally, we

apply the spectral normalization [20] to each layer in both generator and discriminator, and use synchronized Batch-Norm [33] in RESAIL blocks.

4.2. Qualitative Results

We first qualitatively compare our model with the state-of-the-art methods [19, 22, 25] on the Cityscapes, ADE20K and COCO-Stuff datasets, and the results are illustrated in Fig. 4. For SPADE [22] and CC-FPSE [19], degenerated synthesis results on some objects can be observed, such as car and bed. Although OASIS [25] introduces semantic discriminator to improve the visual quality of synthesized image, it is still limited in avoiding unrealistic details and obvious artifacts. In contrast, benefited from the retrieval-

6

Table 1. Quantitative comparison on ADE20K [34], ADE20K-outdoor, Cityscapes [5] and COCO-Stuff [1]. For AC and mIOU, higher is better, and for FID, lower is better. Our method achieves very competitive results on the four datasets.

Method	ADE20K			ADE20K-outdoor			Cityscapes			COCO-Stuff		
	FID ($\downarrow$)	AC ($\uparrow$)	mIOU ($\uparrow$)	FID ($\downarrow$)	AC ($\uparrow$)	mIOU ($\uparrow$)	FID ($\downarrow$)	AC ($\uparrow$)	mIOU ($\uparrow$)	FID ($\downarrow$)	AC ($\uparrow$)	mIOU ($\uparrow$)
CRN [3]	73.3	68.8	22.4	99.0	68.6	16.5	104.7	77.1	52.4	70.4	40.4	23.7
Pix2pixHD [29]	81.8	69.2	20.3	97.8	71.6	17.4	95.0	81.4	58.3	111.5	45.7	14.6
SIMS [23]	n/a	n/a	n/a	67.7	74.7	13.1	49.7	75.5	47.2	n/a	n/a	n/a
SPADE [22]	33.9	79.9	38.5	63.3	82.9	30.8	71.8	81.9	62.3	22.6	67.9	37.4
CC-FPSE [19]	31.7	82.9	43.7	n/a	n/a	n/a	54.3	82.3	65.5	19.2	70.7	41.6
SESAME [21]	31.9	**85.5**	49.0	n/a	n/a	n/a	54.2	82.5	66.0	n/a	n/a	n/a
SC-GAN [30]	29.3	83.8	45.2	n/a	n/a	n/a	49.5	82.5	66.9	18.1	72.0	42.0
OASIS [25]	**28.3**	n/a	48.8	**48.6**	n/a	40.4	47.7	n/a	69.3	**17.0**	n/a	44.1
Ours	30.2	84.8	**49.3**	**48.6**	**86.5**	**41.1**	45.5	83.2	**69.7**	18.3	**73.1**	**44.7**

(a) original I^r, w/o GT   (b) distorted I^r, w/o GT   (c) w/o I^r, distorted GT   (d) original I^r, original GT   (e) original I^r, distorted GT

Figure 7. Ablation study on the data distortion method. (a)(b) When only retrieval-based guidance I^r is used in training, generator fails to synthesize certain objects details marked in red rectangle; (c) Synthesized images also suffer from poor details only with the guidance of *distorted GT*. (d) Using both I^r and *original GT* as guidance, inconsistent edge and illumination can still be observed. (e) Using both I^r and *distorted GT* $\tilde{I}^{gt}$ as guidance, our model synthesizes photo-realistic image with fine details. Please zoom for a better view.

based guidance, our model generates more photo-realistic images with finer details such as edges, textures, color, and less artifacts.

Moreover, we retrieve the segment image for each semantic region separately, which allows us to edit the synthesis images either globally or locally. As shown in Fig. 5, given the same semantic map, we can achieve globally diverse synthesis results by changing all the retrieved segments of the whole image (left two columns). Moreover, We can also fix most semantic regions and change the retrieved segments of the remaining objects to edit the results locally (right three columns). More qualitative results are provided in the *Suppl*.

4.3. Quantitative Results

We further quantitatively compare with the competing methods [3, 19, 21–23, 25, 29, 30] on four datasets, and Table 1 lists the results. From the table, our method performs favorably against the competing methods on Cityscapes [5] and ADE20K-outdoor datasets, and also is very competitive on ADE20K [34] and COCO-Stuff [1] datasets, demonstrating the effectiveness of our method. Note that, SIMS [23] also uses a retrieved image to guide semantic image synthesis but is inferior to our method, partially due to that it is more effective to use retrieval-based guidance for spatially adaptive normalization other than use it as network input.

User Study. Following the previous works [21, 22, 30], we

Table 2. User study on Cityscapes. The numbers indicate the percentage of volunteers who favor the results of our method over those of the competing methods or even the ground-truth.

Ours vs. SPADE	Ours vs. CC-FPSE	Ours vs. OASIS	Ours vs. GT
87.8	80.2	85.4	16.8

conduct user study on Cityscapes dataset. Participants have been informed their identities will not be recorded. Each volunteer is given a semantic map and two corresponding images containing one by our method and another one by a randomly selected competing method (*i.e.*, SPADE [22], CC-FPSE [19], OASIS [25] or even the ground-truth image), and is asked to vote for the image with better visual quality. The orders of the two images are random to avoid the effect caused by potential bias. There are totally 2,000 questions for 200 volunteers, and Table 2 lists the results. Volunteers strongly favor (more than 80%) our results in contrast to the competing methods. In comparison with the ground-truth images, our results still have a chance of about 17% to be recognized as the better one, further indicating our method is able to generate photo-realistic images.

4.4. Ablation Studies

We conduct ablation studies on Cityscapes to assess the effect of RESAIL module and data distortion mechanism.
Effectiveness of RESAIL Module. To demonstrate the effectiveness of our RESAIL module, we compare our

Table 3. Ablation study on RESAIL module. Model+ denotes using the retrieval-based guidance as input to the given module. With the proposed RESAIL module and the retrieval-based guidance, our method achieves better quantitative performance.

Variants	Guidance Inject	FID($\downarrow$)	mIOU($\uparrow$)	AC($\uparrow$)
SPADE	w/o	58.7	62.2	81.9
Pix2pixHD+	Conv Layer	47.8	66.7	81.9
SPADE+	SPADE Module	53.4	68.6	82.8
SEAN+	SEAN Module	66.6	69.4	82.1
Ours	RESAIL Module	**45.5**	**69.7**	**83.2**

Table 4. Effect of data distortion mechanism on ground-truth guidance. Among all variants, using both I^r and distorted ground-truth $\tilde{I}^{gt}$ as guidance achieves better performance.

I^r	Ground-truth	FID ($\downarrow$)	mIOU ($\uparrow$)	AC ($\uparrow$)
original	w/o	47.7	66.3	82.5
distorted	w/o	48.8	65.3	82.6
w/o	distorted	49.0	64.9	82.1
original	original	52.8	64.0	81.2
original	distorted	**45.5**	**69.7**	**83.2**

method with 4 variants which vary on whether the retrieval-based guidance used and how to use it: (i) *SPADE* denotes the original SPADE module without exploiting the guidance. (ii) *Pix2pixHD+* denotes concatenating the guidance into the conv layer of pix2pixHD model. (iii) *SPADE+* denotes using the guidance as input to the SPADE module. (iv) *SEAN+* denotes using the guidance as input to the SEAN module. (v) *Ours* denotes using the guidance as input to the RESAIL module. More details about the architecture of each variant can be found in the *Suppl*. For a fair comparison, we use the same backbone for all variants and only change the normalization layer. Thus for *Pix2pixHD+*, we use the decoder part as the generator.

Table 3 lists the quantitative comparison among the variants. From the table, directly incorporating the guidance into the SPADE or conv layer improves the performance, indicating that the retrieval-based guidance is beneficial to image synthesis. As for SEAN, regarding the style map is heavy in GPU memory-consuming, we reduce the dimension of style vector to 128 to conduct the experiments, which may cause potential performance degradation but not affect the fair comparison. With the RESAIL module, our method achieves the best performance, clearly demonstrating the effectiveness of our RESAIL module. As shown in Fig. 6, without pixel level guidance information, *SPADE* and *SEAN+* generate blurry details. In compared to *Pix2pixHD+*, our RESAIL generates more photo-realistic results with finer details and consistent illumination. The result shows that spatially adaptive normalization is a more effective way to use retrieval-based guidance than simply concatenating it with feature of conv layer.

Effectiveness of Distorted Ground-truth. We also conduct the ablation study to assess the effect of data distortion mechanism on ground-truth (GT) images. Specifically, we

consider five variants. (i) Only the retrieval-based guidance I^r is used as guidance during training. (ii) Only the distorted I^r is used as guidance during training. (iii) Only the distorted GT is used as guidance during training. (iv) Both I^r and the original GT can be used as guidance during training. (v) Ours: both I^r and distorted GT $\tilde{I}^{gt}$ can be used as guidance during training.

Table 4 lists the quantitative results on Cityscapes. From the table, performing data distortion on retrieval-based guidance brings little gain or even adverse effect on semantic image synthesis. This is because the retrieval-based guidance is already distorted and further distorting it may make it more unrealistic and is not beneficial to synthesis performance. Also, using the original GT as guidance cannot improve the quality of generated images, because there exists obvious gap between original GT and retrieval-based guidance. With the data distortion on the ground-truth, we can reduce the gap between them and thus benefits the model training. Fig. 7 shows the qualitative results. One can see that, using both retrieval-based guidance and distorted ground-truth as guidance during training, our method produces more photo-realistic details and consistent color.

Additional ablation study on the segmentation loss is provided in the *Suppl*, please check it for more details.

5. Discussion

In this paper, we proposed a novel feature normalization method, termed as REtrieval-based Spatially AdaptIve normaLization (RESAIL). With the retrieval-based guidance and distorted ground-truth, the model can be trained with perceptual supervision, and produces diverse and photo-realistic synthesized images. Experimental results demonstrate that our method performs favorably against the state-of-the-art methods on several challenging datasets both qualitatively and quantitatively.

Impact. This work presents a RESAIL module for semantic image synthesis. Although we have not conducted the experiments on human face synthesis tasks, it has the potential for being used to face synthesis and editing. From this viewpoint, our work may be improperly used for deepfake techniques which trigger potential negative social impacts.

Limitation. Albeit our method synthesizes photo-realistic images and outperforms existing methods, the inference speed is still a limitation. Retrieving operation in our method is time consuming, which makes it unable to perform realtime inference. In the future, we will explore feasible method to accelerate or avoid the retrieving process.

Acknowledgement

References

[1] 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. 5, 7, 13

[2] 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

[3] 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, 6, 7

[4] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Transactions on Graphics*, 28(5):1–10, 2009. 1, 3

[5] 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. 5, 7, 13

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 3

[7] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics*, 26(3):4–es, 2007. 1, 3

[8] 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

[9] Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *Advances in Neural Information Processing Systems*, 31:2708–2718, 2018. 1

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 3

[11] Phillip Isola and Ce Liu. Scene collaging: Analysis and synthesis of natural images with semantic layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3048–3055, 2013. 1, 3

[12] 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. 2

[13] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. 1

[14] Micah K Johnson, Kevin Dale, Shai Avidan, Hanspeter Pfister, William T Freeman, and Wojciech Matusik. Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Transactions on Visualization and Computer Graphics*, 17(9):1273–1285, 2010. 1, 3

[15] 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

[16] Jean-François Lalonde, Derek Hoiem, Alexei A Efros, Carsten Rother, John Winn, and Antonio Criminisi. Photo clip art. *ACM Transactions on Graphics*, 26(3):3–es, 2007. 1, 3

[17] 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

[18] 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. 2

[19] 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

[20] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 6

[21] 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. 1, 2, 5, 6, 7, 12

[22] 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, 3, 4, 5, 6, 7, 12

[23] 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, 3, 7, 12

[24] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 11

[25] 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, 3, 5, 6, 7, 13

[26] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 3

[27] 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

[28] Hao Wang, Qilong Wang, Hongzhi Zhang, Jian Yang, and Wangmeng Zuo. Constrained online cut-paste for object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 4, 11

[29] 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. 2, 5, 7, 12

[30] 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, 7

[31] 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

[32] 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

[33] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 6

[34] 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. 5, 7, 13

[35] 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. 1, 2, 3, 4, 12

Supplemental Materials

A. Additional Implementation Details

A.1. Retrieval-based Guidance Image

Given a semantic map M, we use it to retrieve and composite a guidance image I^r for image synthesis.

Preprocess of Dataset. The training dataset $\mathcal{D}^{tr}$ is firstly used to create a retrieval database consisting of a set of segments. Specifically, for each image $I \in \mathcal{D}^{tr}$ and its corresponding semantic map M, we use the available instance-level annotation to decompose I and M as a number of segments,

$$I, M = \{(M_i^s, y_i^c, I_i^s)\}, \tag{a}$$

where M_i^s, y_i^c and I_i^s denote the cropped binary mask of the i-th object, its category and its corresponding RGB segment image, respectively. Besides, for background region without instance-level annotation, we take the maximal connected component as a single background object. Decomposing all the images in training dataset, we create a retrieval database, which is used in both training and testing stage.

Retrieval Strategy. Given a semantic map M, we first decompose it into a number of segment masks $\{(M_i^s, y_i^c)\}$. Then, we retrieve the most compatible segment from the retrieval database for each segment mask. Specifically, for segment mask M_i^s with category y_i^c, we retrieve a segment (M_j^s, y_j^c, I_j^s) which has the same category ($y_j^c = y_i^c$) and similar shape with M_i^s. To measure the similarity between two segment masks (M_i^s and M_j^s), we adopt the geometric score [28] to measure both scale and shape consistency,

$$\sigma_{scale}\left(M_i^s, M_j^s\right) = \begin{cases} 0, & t \geq 0.5 \\ 1, & t < 0.5 \end{cases}, \tag{b}$$

$$\sigma_{shape}\left(M_i^s, M_j^s\right) = \frac{SSD\left(\hat{M}_i^s, \hat{M}_j^s\right)}{\max\left(\left\|\hat{M}_i^s\right\|_1, \left\|\hat{M}_j^s\right\|_1\right)}, \tag{c}$$

where $t = \frac{\min\left(\|M_i^s\|_1, \|M_j^s\|_1\right)}{\max\left(\|M_i^s\|_1, \|M_j^s\|_1\right)}$. $\hat{M}_i^s$ and $\hat{M}_j^s$ denote the resized versions (*i.e.*, 128×128) of M_i^s and M_j^s using nearest neighbor interpolation, respectively. $SSD\left(\cdot\right)$ denotes the sum square difference. The final consistency is calculated as,

$$\sigma\left(M_i^s, M_j^s\right) = \sigma_{scale}\left(M_i^s, M_j^s\right) + \gamma\sigma_{shape}\left(M_i^s, M_j^s\right). \tag{d}$$

where γ is the balance coefficient and we set $\gamma = 1$ in practice. Lower $\sigma\left(M_i^s, M_j^s\right)$ indicates more similarity between two segment masks.

Composition of Guidance Image. Finally, we recompose the retrieved segments as the guidance image. Let (M_r^s, y_r^c, I_r^s) denotes the retrieved segment for the given segment mask M_i^s. As illustrated in Fig. A, I_r^s and the corresponding mask M_r^s are first resized to the size of M_i^s. The resized mask and image are denoted as $\hat{M}_r^s$ and $\hat{I}_r^s$. Then, the resized image is pasted into the guidance image according to the original position of M_i^s. To maintain integrity of instance, we paste the segment image following the below rules:

- Pixels of $\hat{I}_r^s$ in both $\hat{M}_r^s$ and M_i^s are preserved.
- If y_r^c belongs to background things categories, pixels of $\hat{I}_r^s$ in $\hat{M}_r^s$ but not in M_i^s are zeroed out.
- If y_r^c belongs to foreground (*i.e.*, instance object) and pixels of $\hat{I}_r^s$ in $\hat{M}_r^s$ but not in M_i^s are located in the *background* categories in M, they are preserved.
- If y_r^c belongs to foreground and pixels of $\hat{I}_r^s$ in $\hat{M}_r^s$ but not in M_i^s are located in the *foreground* categories in M, they are zeroed out.

We finally obtain the retrieval-based guidance image I^r to guide the image synthesis.

A.2. Distortion of Ground-truth Image.

To distort the ground-truth image I^{gt}, we first decompose it into a set of segment images $I^{gt} = \{I_i^s\}$. Then we apply the distortion (*i.e.*, color, shape and resolution) on each segment image I_i^s.

Color. We employ the method proposed by [24] to transfer the color of segment image I_i^s to a random segment image I_t^s with the same category. Specifically, we first convert I_i^s and I_t^s from RGB space into $l\alpha\beta$ space. Then the color transferred image $\tilde{I}_i^s$ in each channel of $l\alpha\beta$ space is calculated by,

$$\tilde{l}_i = (l_i - \mu(l_i)) \cdot \frac{\sigma(l_t)}{\sigma(l_i)} + \mu(l_t)$$

$$\tilde{\alpha}_i = (\alpha_i - \mu(\alpha_i)) \cdot \frac{\sigma(\alpha_t)}{\sigma(\alpha_i)} + \mu(\alpha_t) \tag{e}$$

$$\tilde{\beta}_i = (\beta_i - \mu(\beta_i)) \cdot \frac{\sigma(\beta_t)}{\sigma(\beta_i)} + \mu(\beta_t)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation of corresponding channel. Finally, we convert $\tilde{I}_i^s$ from $l\alpha\beta$ into RGB space to obtain the color distorted image.

Shape. To distort the shape of a segment image, we first sample 10 points uniformly on the edge of the segment image as source points, and shift three of them randomly

Figure A. Process to paste a retrieved segment into the semantic map. We here take "Tree" labeled in cyan as an example.

to produce the target points. The source points and target points are used to produce a dense flow utilizing thin plate spline algorithm. Then we use the produced flow to warp the segment image to obtain the shape distorted image.

Resolution. To distort the resolution of a segment image, we downsample it with a random scale $\tau(0.5 < \tau < 1)$, and upsample it to the original size.

After distortion, distorted segment images from ground-truth I^{gt} recompose the distorted ground-truth $\tilde{I}^{gt}$ to facilitate model training. The distortion results are shown in Fig. B.

B. Additional Details of Training Architecture

Details of RESAIL module. The RESAIL module takes both the guidance image (*i.e.*, retrieval-based guidance I^r or distorted ground-truth $\tilde{I}^{gt}$) and the semantic map M as input and learns to modulate the activations. We here represent the input activations as $\mathbf{h}$ with a batch of N samples. H, W and C denote the height, width and the number of channels in $\mathbf{h}$, and the modulated activations at site ($n \in N, c \in C, y \in H, x \in W$) is represented as,

$$RESAIL(\mathbf{h}, I^r, M) = \gamma_{c,y,x}(I^r, M)\frac{\mathbf{h}_{n,c,y,x} - \mu_c}{\sigma_c} \quad \text{(f)}$$
$$+ \beta_{c,y,x}(I^r, M),$$

where μ_c and σ_c denote the mean and standard deviation of the activation in channel c,

$$\mu_c = \frac{1}{NHW}\sum_{n,y,x}\mathbf{h}_{n,y,x}$$
$$\sigma_c = \sqrt{\frac{1}{NHW}\left(\sum_{n,y,x}\mathbf{h}_{n,y,x}^2\right) - \mu_c^2}. \quad \text{(g)}$$

$\gamma(\cdot)$ and $\beta(\cdot)$ have the same architectures and learn the parameters for modulating the scales and biases, respectively. We here take $\gamma(\cdot)$ as an example, which consists of two separated convolutional neural networks to produce coarse and fine-grained guidance for modulation. The one network

$\gamma^s(\cdot)$ takes the semantic map M to learn the coarse modulation parameters. The other network $\gamma^r(\cdot)$ takes the retrieved image I^r to learn the pixel-level fine-grained modulation parameters, and we also take the semantic map M to modulate the intermediate features with AdaIN blocks.

$$\gamma_{c,y,x}(I^r, M) = \alpha_\gamma\gamma_{c,y,x}^s(M) + (1 - \alpha_\gamma)\gamma_{c,y,x}^r(I^r, M)$$
$$\beta_{c,y,x}(I^r, M) = \alpha_\beta\beta_{c,y,x}^s(M) + (1 - \alpha_\beta)\beta_{c,y,x}^r(I^r, M), \quad \text{(h)}$$

where the $0 < \alpha_\beta, \alpha_\gamma < 1$ are learnable scalars.

Discriminator. In practice, we adopt two multi-scale discriminators proposed by [21] to facilitate our model training. As shown in Fig. C, the discriminator consists of two pathways and processes the RGB image and the semantic labels respectively; then the final features are merged by element-wise addition and element-wise multiplication.

C. Additional Ablation Studies

Comparison with SIMS. Also introducing an image synthesis mechanism based on reference, SIMS [23] simply takes the retrieved image as network input, resulting in low mIOU and blurs shown as Fig. D and Table 1. While our method leverages the retrieved images to provide pixel level fine-grained guidance via spatially adaptive normalization, making it more effective in synthesizing photo-realistic images.

Variants of RESAIL. We compare our RESAIL module with 4 variants and in each comparison experiment we employ the same generator architecture while only replacing the RESAIL ResBlk with other variants. We show the different ResBlks in Fig. E. In *SPADE*, we just employ the module proposed by [22]. In *SPADE+*, semantic map concatenating with the guidance image is convolved to produce the modulation parameters β and γ. In *Pix2pixHD+*, we concatenate the feature with the semantic map and the guidance image following with convolution layer, and we discard the encoder part of Pix2pixHD [29]. In *SEAN+*, we extract per region style vectors from the guidance image with a style encoder network and input the style vector and semantic map into the SEAN [35] module. Limited by GPU memory, dimension of style vector is set to 128.

Effectiveness of $\mathcal{L}_{seg}$. To prompt the model to synthesize images aligning well with the semantic layout, we introduce

12

Figure B. Distortion of ground-truth images. The top row shows the produced retrieval-based guidance images; the middle row shows the distorted ground-truth and the bottom row shows the corresponding ground-truth images.

Figure C. Discriminator network.

Table A. Ablation study of $\mathcal{L}_{seg}$ in Cityscapes dataset. It shows that $\mathcal{L}_{seg}$ facilitates the model learning.

$\mathcal{L}_{seg}$	FID($\downarrow$)	mIOU($\uparrow$)	AC($\uparrow$)
✗	46.8	66.3	82.7
✓	**45.5**	**69.7**	**83.2**

a pretrained segmentation network C to classify each pixel of the generated image and optimize the segmentation loss $\mathcal{L}_{seg}$. The designed segmentation network C follows [25], which consists of 12 ResBlks based on a U-Net architecture as shown in Fig. F. We report the results of training our model with and without $\mathcal{L}_{seg}$ on Cityscapes [5] in Table A. From the table, we can see segmentation loss $\mathcal{L}_{seg}$ improves the learning process. Albeit $\mathcal{L}_{seg}$ helps segmentation based metrics, it may introduce inconsistent edge transitions among instances, occurring in [25] which introduces a discriminator based on a segmentation network shown as

Table B. FID w.r.t non-similarity threshold.

Threshold	0.15	0.25	0.35	0.45	0.55	0.58
FID	45.49	46.38	48.18	48.3	50.56	51.04

Fig. G. However, with other losses (*e.g.*, GAN loss and perceptual loss) prompting model training, this kind of artifacts are suppressed and no obvious transitions are found in our results with $\mathcal{L}_{seg}$.

Effect of Shape Non-similarity Threshold. Computed as Eq. d, non-similarity σ is adopted to measure the shape consistency between two segment masks. We have tested the FID results by adopting different non-similarity thresholds. From Table B, higher threshold (*i.e.*, using more non-similar guidance) leads to worse guidance, resulting in worse FID.

D. Additional Visual Results

To demonstrate the effectiveness of our method on synthesizing the photo-realistic images, we show more visual results in this section. Fig. H $\sim$ J show the comparisons on Cityscapes [5] and as shown in figures, our synthesized images are more photo-realistic with fine details. Fig. K and Fig. M show more results on ADE20K [34]. Comparisons on COCO-Stuff [1] can be found in Fig. L. The guidance image and its corresponding generated image are shown as Fig. N and Fig. O.

Figure D. Comparison with SIMS. SIMS suffers from low mIOU (marked in green rectangle) and blurs (marked in red rectangle) of some objects.

(a) SPADE ResBlk

(b) Pix2pixHD+ ResBlk

(c) SPADE+ ResBlk

(d) SEAN+ ResBlk

Figure E. Variants of RESAIL ResBlk. (a) *SPADE* employs the SPADE module; (b) *Pix2pixHD+* denotes concatenating the guidance into the conv layer of pix2pixHD model. (c) *SPADE+* denotes using the guidance as input to the SPADE module. (d) *SEAN+* denotes using the guidance as input to the SEAN module.

14

(a) Segmentation Network

(b) Residual Block

Figure F. Segmentation network. (a) The network is designed based on U-Net. (b) Each downsampling or upsampling operation employs a ResBlk.

Figure G. Effect of segmentation loss $\mathcal{L}_{seg}$. Red rectangles mark the affected instances. OASIS suffers from inconsistent edge transitions whose discriminator based on a segmentation network. With the help of other losses (*e.g.*, GAN loss and perceptual loss), no obvious edge transitions are found in our results with $\mathcal{L}_{seg}$.

Figure H. Comparison results on Cityscapes.

Figure I. Comparison results on Cityscapes.

Figure J. Comparison results on Cityscapes.

Semantic Map	Ground-truth	SPADE

CC-FPSE	OASIS	Ours

Semantic Map	Ground-truth	SPADE

CC-FPSE	OASIS	Ours

Figure K. Comparison results on ADE20K.

| Semantic Map | Ground-truth | SPADE | CC-FPSE | OASIS | Ours |

Figure L. Comparison results on COCO-Stuff.

Semantic Map	Ground-truth	SPADE

CC-FPSE	OASIS	Ours

Semantic Map	Ground-truth	SPADE

CC-FPSE	OASIS	Ours

Figure M. Comparison results on ADE20K.

Semantic Map	Ground-truth	Guidance Image	Synthesis

Figure N. Synthesis results on ADE20K.

Semantic Map	Ground-truth	Guidance Image	Synthesis

Figure O. Synthesis results on Cityscapes.

Semantic Map Ground-truth Synthesis Semantic Map Ground-truth Synthesis

Ground-truth Synthesis Ground-truth Synthesis

Figure P. Synthesis results on ADE20K(top) and Cityscapes(bottom).