Dual Attention GANs for Semantic Image Synthesis

Hao Tang1, Song Bai2, Nicu Sebe13
1DISI, University of Trento
2Department of Engineering Science, University of Oxford
3Huawei Research Ireland
hao.tang@unitn.it, songbai.site@gmail.com, sebe@disi.unitn.it

ABSTRACT
In this paper, we focus on the semantic image synthesis task that aims at transferring semantic label maps to photo-realistic images. Existing methods lack effective semantic constraints to preserve the semantic information and ignore the structural correlations in both spatial and channel dimensions, leading to unsatisfactory blurry and artifact-prone results. To address these limitations, we propose a novel Dual Attention GAN (DAGAN) to synthesize photo-realistic and semantically-consistent images with fine details from the input layouts without imposing extra training overhead or modifying the network architectures of existing methods. We also propose two novel modules, i.e., position-wise Spatial Attention Module (SAM) and scale-wise Channel Attention Module (CAM), to capture semantic structure attention in spatial and channel dimensions, respectively. Specifically, SAM selectively correlates the pixels at each position by a spatial attention map, leading to pixels with the same semantic label being related to each other regardless of their spatial distances. Meanwhile, CAM selectively emphasizes the scale-wise features at each channel by a channel attention map, which integrates associated features among all channel maps regardless of their scales. We finally sum the outputs of SAM and CAM to further improve feature representation. Extensive experiments on four challenging datasets show that DAGAN achieves remarkably better results than state-of-the-art methods, while using fewer model parameters. The source code and trained models are available at https://github.com/Ha0Tang/DAGAN.

CCS CONCEPTS
• Computing methodologies → Computer vision; Machine learning; Computer graphics.

KEYWORDS
Generative Adversarial Networks (GANs); Semantic Image Synthesis; Spatial Attention; Channel Attention

ACM Reference Format:
Hao Tang, Song Bai, Nicu Sebe. 2020. Dual Attention GANs for Semantic Image Synthesis. In Proceedings of the 28th ACM International Conference...

1 INTRODUCTION
In this paper, we aim to address the semantic image synthesis task that generates realistic images conditioned on input layouts. This has been widely investigated in the recent years [5, 16, 26, 31, 33, 44]. Most existing methods typically use Generative Adversarial Networks (GANs) [13] to learn the translation mapping from semantic layouts to realistic images. For instance, Wang et al. [44] adopt multi-scale generators and discriminators to generate high-resolution images. Park et al. [31] propose a novel spatially-adaptive normalization for generating realistic images. Despite the interesting exploration of these methods, we can still see blurriness and artifacts in the generated images (see Fig. 1), which is mainly due to two reasons. First, these methods lack effective semantic constraints to maintain the semantic information of the input semantic label. Second, these methods ignore the semantic correlations between the spatial pixels and channel features which cause intra-class semantic inconsistencies such as the roads, buses, lands and waters generated by GauGAN in Fig. 1.

To solve these limitations, we propose a novel Dual Attention GAN (DAGAN) and two novel modules, i.e., Position-Wise Spatial Attention Module (SAM) and Scale-Wise Channel Attention Module (CAM). Spatial and channel selections are two crucial factors for translating the input layout to a realistic image. Thus both SAM and...
CAM aim to explore the semantic attention in spatial and channel dimensions for generating high-quality and semantically-consistent images. In particular, SAM selectively correlates the pixels at each position by a spatial attention map, leading pixels with the same label to relate to each other regardless of their spatial distances. Simultaneously, CAM selectively emphasizes scale-wise features at each channel by a channel attention map integrating associated features among all channel maps regardless of their scales.

Differently from us, the spatial attention module proposed in [3] is a non-local model, which requires computing the correlation between every two pixels in the feature map, leading to expensive computational costs and thus limiting its applicability. At the same time, the channel attention module proposed in [3] did not use multi-scale features, which are quite important for generating small-objects in the semantic image synthesis tasks.

Finally, we sum the outputs of both SAM and CAM to further improve the feature representation. Notably, both SAM and CAM can be readily applied to existing GAN frameworks without imposing training overheads or modifying network architectures.

We perform comprehensive experiments on four challenging datasets with different image resolutions, i.e., ADE20K [58] (256×256), Cityscapes [9] (512×256), CelebAMask-HQ [21] (512×512) and Facades [41] (1024×1024). Both qualitative and quantitative results show that the proposed DAGAN is able to produce remarkably better results than existing models including CRN [5], SIMS [33], Pix2pixHD [44], GauGAN [31] and CC-FPSE [26], regarding both the visual fidelity and the alignment with the input layouts.

Overall, the contributions of our paper are:

- We propose a novel Dual Attention GAN (DAGAN) for the challenging task of semantic image synthesis, which can effectively model the semantic attention in both spatial and channel dimensions for improving the ability of feature representations.
- We design two novel modules, i.e., position-wise Spatial Attention Module (SAM) and scale-wise Channel Attention Module (CAM), to learn the spatial and channel attention of local features, respectively. Both significantly improve the generation results by modeling intra-class correlations. Moreover, both modules are lightweight and general modules, and can be seamlessly integrated into any existing GAN-based architectures to strengthen the feature representation with negligible overheads.
- We extensively evaluate the proposed DAGAN to confirm that it achieves new state-of-the-art performance on different datasets with different image resolutions, i.e., ADE20K [58] (256×256), Cityscapes [9] (512×256), CelebAMask-HQ [21] (512×512) and Facades [41] (1024×1024), while using significantly fewer model parameters compared with CC-FPSE [26]. Thus it presents new strong baselines for the research community.

2 RELATED WORK

Generative Adversarial Networks (GANs) [13] are widely used techniques to learn a complex and high-dimensional data distribution for generating new images [2, 17, 18, 34, 53]. A vanilla GAN consists of a generator and a discriminator. The generator aims to produce realistic images to fool the discriminator while the discriminator aims to accurately tell whether an image is real or generated. Based on GANs, Mirza and Osindero proposed Conditional GANs (CGANs) [29] by incorporating conditional guidance information to generate user-specific images. Conditional guidance information can be category labels [8, 27, 47, 56], text descriptions [22, 51, 54], human pose [1, 4, 30, 35, 37], segmentation maps [14, 25, 31, 38, 40, 43, 44, 55] and attention maps [7, 19, 28, 32, 39].

Semantic Image Synthesis aims to turn semantic label maps into photo-realistic images [5, 26, 31, 33, 44]. For instance, Park et al. [31] propose a novel spatially-adaptive normalization to preserve semantic information of input labels for generating realistic images. Although GauGAN [31] has achieved promising results, we still observe unsatisfactory aspects mainly in the generated scene details and layouts (see Fig. 1), which we believe are mainly due to the
problem of missing of spatial and channel semantic information associated with deep network operations.

To tackle this limitation, we propose two novel modules, i.e., position-wise Spatial Attention Module (SAM) and scale-wise Channel Attention Module (CAM), which try to enhance features in both spatial and channel dimensions for generating semantically-consistent results. To the best of our knowledge, this idea has not been considered by any existing semantic image generation method.

**Semantic Attention Modeling** aims to model semantic dependencies of distant regions and has widely been applied in many tasks such as semantic segmentation [10, 12, 23], depth estimation [49], sentiment classification [24], machine translation [42], action classification [45], image classification [53], cross-modal translation [11], text-to-image synthesis [3], and crowd counting [6, 52]. For instance, Wang et al. [45] explore the non-local operation in the space-time dimension for video and image processing. Zhang et al. [53] introduce a self-attention mechanism in the generator for image generation. However, these methods require computing the correlation between every two points in the feature map, leading to expensive computation cost and thus limiting its applications.

Different from the above-mentioned methods, we propose a novel DAGAN for the semantic image synthesis task, and carefully design two modules (i.e., SAM and CAM) to capture semantic attention in both spatial and channel dimensions for improving feature representations. Extensive experiments validate the effectiveness of the proposed method.

### 3 DUAL ATTENTION GANS

In this section, we first introduce an overview of our method and then present the two attention modules. Finally, we introduce the optimization objective and training details of the proposed whole framework.

**Overview.** We start by presenting the details of the proposed Dual Attention GANs (DAGAN). An illustration of the overall framework is shown in Fig. 2, which consists of a generator G and discriminator D. The generator G mainly consists of three parts, i.e., a backbone network B extracting deep multi-scale features from the input layout, a position-wise Spatial Attention Module (SAM) modeling the pixel attention in the spatial dimension, and a scale-wise Channel Attention Module (CAM) capturing the feature attention in the channel dimension.

![Figure 3: Overview of the proposed position-wise Spatial Attention Module (SAM), which aims to learn the position-wise attention between spatial pixels with the same label regardless of their spatial distances. The symbols ⊗, © and ⋆ denote element-wise multiplication, channel-wise concatenation, and Sigmoid() operation, respectively.](image)

Intuitively, stuff and objects in the input semantic layout are diverse on scales, lighting, and views. The features corresponding to the pixels with the same semantic label may have some differences due to traditional convolution operations that would lead to a local receptive view, resulting in intra-class semantic inconsistency and affect the generation performance (see Fig. 1). To address this issue, we explore long-range semantic correlations by building attention among spatial pixels and channel features, thus improving feature representation for image generation. As illustrated in Fig. 2, we design two types of attention modules to improve feature representations.

**Multi-Scale Feature Extraction.** We follow previous works [26, 31, 44] and employ the semantic layout S as the input of our backbone network B, as shown in Fig. 2. The network B aims to extract deep multi-scale features of S, which can be formulated as,

$$F^i_b = B(S), \quad \text{for} \quad i = 1, 2, \cdots, I,$$

where $F^i_b$ denotes the feature map extracted from the $i^{th}$ layer of the backbone network $B$, as shown in Fig. 2. By doing so, we obtain a multi-scale feature representation of $S$ for further processing.

**Position-Wise Spatial Attention Modeling.** Existing methods such as [26, 31, 44] always use local features generated by convolutional operations, leading different generation results of the same label. To model pixel correlations over local features, we propose a position-wise Spatial Attention Module (SAM), which encodes spatial pixel correlations into local features, enhancing their representation capability.

The framework of the proposed SAM is elaborated in Fig. 3. Specifically, given the local feature $F^l_b \in \mathbb{R}^{C \times H \times W}$ extracted from the last layer of the backbone network $B$, we first feed it into average and max pooling operations to produce two new feature maps $P_a \in \mathbb{R}^{1 \times H \times W}$ and $P_m \in \mathbb{R}^{1 \times H \times W}$,

$$P_a = \text{AvePool}(F^l_b),$$

$$P_m = \text{MaxPool}(F^l_b),$$

where AvePool() and MaxPool() represent average and max pooling, respectively. Although [3] and our method both use average and max pooling, the way of using them is different. Specifically, we use both average and max pooling in the spatial dimension since we need to model the correlations between the regions with the same semantic label, while [3] uses average and max pooling in the channel dimension to enhance the features in the channel dimension.
We then concatenate both $P_a$ and $P_m$ to form a new feature $P_{am} \in \mathbb{R}^{2 \times H \times W}$. After that, we perform a convolutional operation on $P_{am}$, and apply a Sigmoid($\cdot$) activation function to calculate the spatial attention map $A_s \in \mathbb{R}^{1 \times H \times W}$. Mathematically,

$$A_s = \sigma(\text{Conv}(\text{Concat}(P_a, P_m)),$$

where $\sigma(\cdot)$, $\text{Conv}(\cdot)$ and $\text{Concat}(\cdot)$ denote Sigmoid function, convolution operation and channel-wise concatenation, respectively. Next, we perform a matrix multiplication between $A_s$ and the original feature $F^{l}_{b}$ to obtain the updated feature $F^{l}_{b} \in \mathbb{R}^{C \times H \times W}$. The computation process is summarized as follow,

$$F_s = A_s \otimes F^{l}_{b},$$

(4) where $\otimes$ denote element-wise multiplication. Therefore, the spatial attention guided feature $F_s$ has a global contextual view in the spatial dimension. By doing so, the pixels with the same semantic label achieve mutual gains, thus improving intra-class semantic consistency (see Fig. 1).

**Scale-Wise Channel Attention Modeling.** Each channel map of features can be regarded as a scale-specific response, and the same object/stuff with different semantic responses should be correlated and associated with each other. To exploit the correlations between channel maps for enhancing the consistency, we propose a scale-wise Channel Attention Module (CAM) to explicitly reason the scale-wise correlations between channels.

The structure of the proposed CAM is illustrated in Fig. 4. Different from SAM, we first reshape $(F^{l}_{b})^{i\rightarrow l}$ to the same size of $F^{l}_{b}$ and then feed them to convolution layers. Next, we concatenate all of them and feed the result into a new convolution layer to obtain a new feature $F^{l-1}_{b} \in \mathbb{R}^{C \times H \times W}$. This process can expressed as,

$$F^{l-1}_{b} = \text{Conv} \left( \text{Concat}(\text{Conv}(F^{l}_{b}), \text{Conv}(P^{2}_{b}), \cdots, \text{Conv}(F^{l-1}_{b})) \right),$$

(5) where $\text{Conv}(\cdot)$ and $\text{Concat}(\cdot)$ denote convolution operation and channel-wise concatenation, respectively. After that, we concatenate $F^{l-1}_{b}$ and $F^{l}_{b}$, and feed the result to an average pooling layer to obtain a scale vector $\sigma \in \mathbb{R}^{2C \times 1 \times 1}$. Mathematically,

$$\sigma = \text{AvePool}(\text{Concat}(F^{l-1}_{b}, F^{l}_{b})),$$

(6) where $\text{AvePool}(\cdot)$ denote the average pooling. To reduce the number of the channel of $\sigma$, we feed it to two successively convolution layers to obtain a new scale vector $\gamma \in \mathbb{R}^{C \times 1 \times 1}$, indicating the weights of different channels are equal.

However, the features in different scales have different degrees of discrimination, which leads to different consistency of generation. To obtain the intra-class consistent generation, we extract the discriminative features within the same label and inhibit the indiscriminative features between different labels. Specifically, we apply a Sigmoid($\cdot$) activation function to obtain the channel attention weight $\delta \in \mathbb{R}^{C \times 1 \times 1}$. This process can be formulated as,

$$\delta = \sigma(\gamma).$$

(7) By doing so, each item in the channel attention weight $\delta$ measures the importance of the corresponding channel. Finally, we introduce two ways to calculate the updated feature $F_{c} \in \mathbb{R}^{C \times H \times W}$, which represents the feature selection with CAM. The first one is using the channel weight $\delta$ to multiply $F^{l-1}_{b}$ and perform an element-wise sum with $F^{l}_{b}$,

$$F_{c} = \delta \otimes F^{l-1}_{b} + F^{l}_{b},$$

(8) where $\otimes$ denotes element-wise multiplication. The second one is to use the channel weight $\delta$ to multiply $F^{l}_{b}$ and perform an element-wise sum with $F^{l-1}_{b}$,

$$F_{c} = \delta \otimes F^{l-1}_{b} + F^{l}_{b}.$$  

(9) In this way, the new feature $F_{c}$ spotlights attention within the same category regardless of their scales and channel distances, boosting feature discriminability.

Note that the proposed CAM is designed to change the weights of the features on each scale to enhance the scale consistency. With this design, we can make the generator to obtain scale-wise discriminative features, making the generated image to be intra-class consistent. Oppositely, the channel attention module proposed in [3] did not consider the multi-scale features causing as such the generated image to be intra-class inconsistent.

**Attention Modeling with GANs.** To take full advantage of pixel and feature attention in both spatial and channel dimensions, we sum the outputs from the two attention modules to obtain better feature representations for image generation. At last, we adopt a convolution layer to generate the final result $I$, as shown in Fig. 2. Notably, the proposed attention modules are simple and
can be directly inserted in the existing GAN frameworks without introducing too many parameters computational costs.

**Optimization Objective.** We follow [26, 31] and employ three different losses as our optimization objective.

\[
L = \lambda_{\text{cgan}} L_{\text{cgan}} + \lambda_f L_f + \lambda_p L_p.
\]

(10)

where \(L_{\text{cgan}}, L_f\) and \(L_p\) denote the conditional adversarial loss, the discriminator feature matching loss and the perceptual loss, respectively. We set \(\lambda_{\text{cgan}} = 1, \lambda_f = 10\) and \(\lambda_p = 10\) in our experiments.

**Training Details.** We employ the multi-scale discriminator used in [31, 44] as our discriminator \(D\). We follow the training procedures of GANs and alternatively train the generator \(G\) and discriminator \(D\), i.e., one gradient descent step on discriminator and generator alternately. We use the Adam solver [20] and set \(\beta_1=0, \beta_2=0.999\).

We conduct the experiments on NVIDIA DGX1 with 8 32GB V100 GPUs.

## 4 EXPERIMENTS

**Datasets.** We conduct extensive experiments on four public datasets to validate the proposed DAGAN, i.e., Cityscapes [9], ADE20K [58], CelebAMask-HQ [21] and Facades [41]. Notably, we follow the same train/test split used in their papers. Moreover, to verify the robustness of the proposed DAGAN on different image resolutions, we resize the images to 256x256, 512x256, 512x512, and 1024x1024 on ADE20K, Cityscapes, CelebAMask-HQ, and Facades, respectively.
**Evaluation Metrics.** We follow GauGAN [31] and use segmentation accuracy, i.e., mean Intersection-over-Union (mIoU) and pixel accuracy (Acc), as our evaluation metrics. Moreover, we use Fréchet Inception Distance (FID) [15], Fréchet ResNet Distance (FRD) [36] and Learned Perceptual Image Patch Similarity (LPIPS) [57] to evaluate the feature distance between generated images and real samples.

**4.1 Comparisons with State-of-the-Art**

**Qualitative Comparisons.** We compare the proposed DAGAN with several leading methods, i.e., GauGAN [31], Pix2pixHD [44] and CC-FPSE [26]. Comparison results are shown in Fig. 5, 6, 7 and 8. We observe that the proposed method generates more clear and visually plausible images than the existing baselines, validating the effectiveness of the proposed DAGAN. Note that we cannot reproduce the results of CC-FPSE on both Facades and CelebAMask-HQ datasets because we cannot fit the CC-FPSE model to our GPUs on both datasets.

**User Study.** We follow the evaluation protocol of GauGAN [31] and perform a user study to measure the quality of generated images. Comparison results are shown in Table 1. We observe that users strongly favor the images generated by the proposed DAGAN than both GauGAN and CC-FPSE on all the four challenging datasets, further validating that the generated images by our method are more photo-realistic.

**Quantitative Comparisons.** We also provide quantitative results in Table 2. Clearly, the proposed DAGAN achieves the best results compared with the baselines except CC-FPSE [26] on ADE20K. However, we see that the proposed DAGAN generates more photo-realistic images with fewer artifacts than CC-FPSE in Fig. 8. Moreover, we provide the number of model parameters in Table 3. We see that the proposed DAGAN has remarkably fewer model parameters than CC-FPSE, which means DAGAN requires significantly less training time and GPU memory than CC-FPSE. Notably, for CC-FPSE, we cannot generate high-resolution images on both Facades (1024×1024) and CelebAMask-HQ (512×512) datasets since

![Figure 7: Qualitative comparison on CelebAMask-HQ. From left to right: Input, GauGAN [31], DAGAN (Ours) and GT.](image-url)
Figure 8: Qualitative comparison on ADE20K. From left to right: Input, GauGAN [31], CC-FPSE [26], DAGAN (Ours) and GT.

Table 2: Quantitative comparison of different methods. For all metrics except mIoU and Acc, lower is better.

| Method         | Cityscapes | ADE20K | Facades | CelebAMask-HQ |
|----------------|------------|--------|---------|---------------|
|                | mIoU ↑     | Acc ↑  | FID ↓   | mIoU ↑ | Acc ↑ | FID ↓ | LPiPS ↓ | FRD ↓ | mIoU ↑ | Acc ↑ | FID ↓ | LPiPS ↓ | FRD ↓ |
| CRN [5]        | 52.4       | 77.1   | 104.7   | 22.4   | 68.8 | 73.3  | -      | -     | -      | -     | -      | -      | -     |
| SIMS [33]      | 47.2       | 75.5   | 49.7    | -      | -    | -     | -      | -     | -      | -     | -      | -      | -     |
| Pix2pixHD [44] | 58.3       | 81.4   | 95.0    | 20.3   | 69.2 | 81.8  | 128.5  | 0.6466 | 3.7402 | -      | -     | -      | -     |
| GauGAN [31]    | 62.3       | 81.9   | 71.8    | 38.5   | 79.9 | 33.9  | 127.2  | 0.6268 | 3.5309 | 42.2   | 0.4870 | -      | -     |
| CC-FPSE [26]   | 65.5       | 82.3   | 54.3    | 43.7   | 82.9 | 31.7  | 116.6  | 0.6224 | 3.4929 | 23.9   | 0.4796 | -      | -     |
| DAGAN (Ours)   | **66.1**   | **82.6** | 60.3    | 40.5   | 81.6 | 31.9  | **116.6** | 0.6224 | **3.4929** | **23.9** | **0.4796** | **3.4523** | - |

Table 3: Quantitative comparison of model parameters. 'Gen.' and 'Dis.' denote Generator and Discriminator, respectively.

| Method         | Cityscapes | ADE20K | Facades | CelebAMask-HQ |
|----------------|------------|--------|---------|---------------|
|                | Gen.       | Dis.   | Total   | Gen.       | Dis.   | Total   | Gen.   | Dis.   | Total   |
| GauGAN [31]    | 93.0M      | 5.6M   | **98.6M** | 96.5M      | 5.8M   | **102.3M** | 92.4M  | 5.6M   | **98.0M** |
| CC-FPSE [26]   | 138.6M     | 5.2M   | 143.8M  | 151.2M     | 5.2M   | 156.4M  | 398.7M | 5.2M   | 403.9M  |
| DAGAN (Ours)   | 93.1M      | 5.6M   | 98.7M   | 96.6M      | 5.8M   | 102.4M  | 92.4M  | 5.6M   | **98.0M** |

CC-FPSE has many parameters that need to be learned on both datasets, resulting in GPU memory overflow.

Visualization of Learned Attention Maps. For spatial attention map, we randomly select two classes on each sample and display their corresponding spatial attention map in columns 2 and 3 in Fig. 9, respectively. We see that the spatial attention module captures global relationships within each semantic class. For instance, in the first row, the point '1+' is marked on a car and its spatial attention map (in column 2) highlights most the areas where the cars are. For channel attention map, we show the 7th, 25th and 30th channel attention map in column 4, 5 and 6 in Fig. 9, respectively.
We see that the difference of each channel is noticeable after going through the proposed channel attention module. For example, the 7th channel map is brighter than both 25th and 30th channel maps, which means the 7th channel map contains more information and is more important than both 25th and 30th channel maps. Moreover, we show the learned channel attention weight of all channels, which indicates the same brightness has the same importance regardless of object scales and channel distances.

**Visualization of Generated Semantic Maps.** We follow GauGAN [31] and apply pretrained semantic segmentation models on the generated images to produce semantic maps. Specifically, we employ DRN-D-105 [50] for Cityscapes and UperNet101 [48] for ADE20K. The generated semantic maps are shown in Fig. 1. We see that the proposed DAGAN generates more intra-class semantic consistency labels than GauGAN, confirming our initial motivation.

### 4.2 Ablation Study

We conduct extensive ablation studies on Cityscapes [9] to evaluate each component of the proposed DAGAN.

**Baselines of DAGAN.** The proposed DAGAN has six baselines (i.e., B1, B2, B3, B4, B5, B6) as shown in Table 4. (i) B1 is our baseline. (ii) B2 uses the proposed Spatial Attention Module (SAM) to model the position-wise pixel attention in the spatial dimension. (iii) B3 employs the proposed Channel Attention Module (CAM) to reason the scale-wise feature attention in channel dimension. Note that it uses Eq. (8) to calculate the new channel feature $F_c$. (iv) The difference between B4 and B3 is that B4 uses Eq. (9) to calculate the new channel feature $F_c$. (v) B5 combines both SAM and CAM-I to model the semantic attention in both spatial and channel dimensions. (vi) B6 is our final model and adopts the combination of SAM and CAM-II to reason both spatial and channel semantic attentions.

**Ablation Analysis.** The results of ablation study are shown in Table 4. By comparison B2 with B1, the proposed SAM improves mIoU, Acc and FID by 2.9, 0.7 and 7.3, respectively, which confirms the importance of modeling the position-wise spatial pixel attention. By using the proposed CAM-I, B3 outperforms B1 on mIoU, Acc and FID by 3.3, 0.9 and 8.3, respectively, confirming the effectiveness of the proposed CAM. B4 outperforms B3 showing CAM-II is more effective than CAM-I. B5 significantly outperforms both B2 and B3, demonstrating the effectiveness of modeling both spatial and channel semantic attentions for generating photo-realistic and semantically-consistent images. Finally, we observe that by combining both SAM and CAM-II, the overall performance is further boosted, demonstrating the advantage of our full model.

### 5 CONCLUSIONS

We propose a novel Dual Attention GAN (DAGAN) for the challenging semantic image synthesis task. Specifically, we present two new modules, i.e., SAM and CAM. SAM is used to model position-wise pixel attention in spatial dimension. CAM is used to reason scale-wise feature attention in channel dimension. The outputs of SAM and CAM are combined to further improve feature representation. Experiments on four datasets show that DAGAN achieves remarkably better results than existing methods. Moreover, both SAM and CAM are lightweight and general modules, and can be seamlessly integrated into any existing GAN-based architectures to strengthen feature representation with negligible overheads.

**ACKNOWLEDGMENTS**

This work has been partially supported by the Italy-China collaboration project TALENT.
This document provides additional experimental results on the semantic image synthesis task. First, we compare the proposed DAGAN with state-of-the-art methods, i.e., Pix2pixHD [44], GauGAN [31] and CC-FPSE [26] (Sec. 6). Additionally, we show some learned attention maps of the proposed DAGAN (Sec. 7). Finally, we provide the visualization results of the generated semantic maps (Sec. 8).

6 STATE-OF-THE-ART COMPARISON

In this section, we show more generation results of the proposed DAGAN compared with those from the leading semantic image synthesis models, i.e., Pix2pixHD [44], GauGAN [31] and CC-FPSE [26]. The results of Cityscapes [9], Facades [41], CelebAMask-HQ [21], and ADE20K [58] are shown in Fig. 10, 11, 12, 13, 14, and 15. We observe that the proposed DAGAN achieves visually better results than the competing methods on all the four datasets.

7 VISUALIZATION OF LEARNED ATTENTION MAPS

In Fig. 16 we present the learned spatial and channel attention maps. We observe that the spatial attention module captures global relationships within each semantic class. For instance, in the first row, the point ‘2+’ is marked on a tree and its spatial attention map (in column 3) highlights most of the areas where the trees are. In the fourth row, the point ‘2+’ is marked on a person and its spatial attention map (in column 3) highlights most of the areas where the people are.

Moreover, we see that the difference of each channel is noticeable after going through the proposed channel attention module. For example, the $7^{th}$ channel map is brighter than both $25^{th}$ and $30^{th}$ channel maps, which means the $7^{th}$ channel map contains more information and it is more important than both $25^{th}$ and $30^{th}$ channel maps. Both visualization results confirm the design motivation of the proposed DAGAN.

8 VISUALIZATION OF GENERATED SEMANTIC MAPS

We follow GauGAN [31] and use the state-of-the-art segmentation networks on the generated images to produce semantic maps: DRN-D-105 [50] for Cityscapes and UperNet101 [48] for ADE20K. The generated semantic maps of the proposed DAGAN, GauGAN, and the ground truth on Cityscapes [9] and ADE20K [58] datasets are shown in Fig. 17, 18 and 19, respectively. We observe that the proposed DAGAN generates more semantically-consistent results than GauGAN, further validating our motivation.
Figure 10: Qualitative comparison on Cityscapes. From left to right: Input, GauGAN [31], CC-FPSE [26], DAGAN (Ours) and GT. These samples were randomly selected without cherry-picking for visualization purposes.
Figure 11: Qualitative comparison on Facades. From left to right: Input, Pix2PixHD [44], GauGAN [31], DAGAN (Ours) and GT. These samples were randomly selected without cherry-picking for visualization purposes.
Figure 12: Qualitative comparison on CelebAMask-HQ. From left to right: Input, GauGAN [31], DAGAN (Ours) and GT. These samples were randomly selected without cherry-picking for visualization purposes.
Figure 13: Qualitative comparison on CelebAMask-HQ. From left to right: Input, GauGAN [31], DAGAN (Ours) and GT. These samples were randomly selected without cherry-picking for visualization purposes.
Figure 14: Qualitative comparison on ADE20K. From left to right: Input, GauGAN [31], CC-FPSE [26], DAGAN (Ours) and GT. These samples were randomly selected without cherry-picking for visualization purposes.
Figure 15: Qualitative comparison on ADE20K. From left to right: Input, GauGAN [31], CC-FPSE [26], DAGAN (Ours) and GT. These samples were randomly selected without cherry-picking for visualization purposes.
Figure 16: Visualization of learned spatial and channel attention maps on Cityscapes.

Figure 17: Visualization of generated semantic maps compared with those from GauGAN [31] on Cityscapes. These samples were randomly selected without cherry-picking for visualization purposes. Most improved regions are highlighted in the ground truths with white dash boxes.
Figure 18: Visualization of generated semantic maps compared with those from GauGAN [31] on ADE20K. These samples were randomly selected without cherry-picking for visualization purposes. Most improved regions are highlighted in the ground truths with white dash boxes.
Figure 19: Visualization of generated semantic maps compared with those from GauGAN [31] on ADE20K. These samples were randomly selected without cherry-picking for visualization purposes. Most improved regions are highlighted in the ground truths with white dash boxes.