

Multi-Channel Attention Selection GANs for Guided Image-to-Image Translation

Hao Tang, Dan Xu, Yan Yan, Jason J. Corso, Philip H.S. Torr, Nicu Sebe

Abstract—We propose a novel model named Multi-Channel Attention Selection Generative Adversarial Network (SelectionGAN) for guided image-to-image translation, where we translate an input image into another while respecting an external semantic guidance. The proposed SelectionGAN explicitly utilizes the semantic guidance information and consists of two stages. In the first stage, the input image and the conditional semantic guidance are fed into a cycled semantic-guided generation network to produce initial coarse results. In the second stage, we refine the initial results by using the proposed multi-scale spatial pooling & channel selection module and the multi-channel attention selection module. Moreover, uncertainty maps automatically learned from attention maps are used to guide the pixel loss for better network optimization. Exhaustive experiments on four challenging guided image-to-image translation tasks (face, hand, body and street view) demonstrate that our SelectionGAN is able to generate significantly better results than the state-of-the-art methods. Meanwhile, the proposed framework and modules are unified solutions and can be applied to solve other generation tasks, such as semantic image synthesis. The code is available at <https://github.com/Ha0Tang/SelectionGAN>.

Index Terms—GANs, Deep Attention Selection, Cascade Generation, Guided Image-to-Image Translation.

1 INTRODUCTION

GUIDED image-to-image translation is a task aiming at synthesizing new images from an input image and several external semantic guidance, as shown in Fig. 1. This task has been gaining a lot interest especially from the computer vision community, and has been widely investigated in recent years. Due to different forms of semantic guidance, e.g., segmentation maps, hand skeletons, facial landmarks and pose skeleton, most of the existing methods for this class of tasks are tailored toward specific applications, i.e., they need to specifically design the network architectures and training objectives according to different generation tasks. For example, Ma et al. propose PG2 [1], which is a two-stage framework and uses the pose mask loss for generating person images based on an image of that person and human pose keypoints. Tang et al. propose GestureGAN [2], which is a forward-backward consistency architecture and adopt the proposed color loss to generate novel hand gesture images based on the input image and conditional hand skeletons. Wang et al. propose the few-shot Vid2Vid framework [3], which uses the carefully designed weight generation module to synthesize videos that realistically reflect the style of the input image and the layout of conditional segmentation maps.

Different from previous works in guided image-to-image translation, in this paper, we focus on developing a framework that is application-independent. This makes our

framework and modules more widely applicable to many generation tasks with different forms of semantic guidance. To tackle this challenging problem, AlBahar and Huang [4] recently proposed a bi-directional feature transformation to better utilize the constraints of the semantic guidance. Although this approach performed an interesting exploration, we observe unsatisfactory aspects mainly in the generated image layout and content details, which are due to three different reasons. First, since it is always costly to obtain manually annotated semantic guidance, the semantic guidance is usually produced from pre-trained models trained on other large-scale datasets, e.g., pose skeletons are extracted using OpenPose [5] and segmentation maps are extracted using [6], [7], leading to insufficiently accurate predictions for all the pixels, and thus misguiding the image generation process. Second, we argue that the translation with a single phase generation network is not able to capture the complex image structural relationships between the source and target domains, especially when source and target domains only have little or even no overlap, e.g, person image generation and cross-view image translation. Third, a three-channel generation space may not be suitable enough for learning a good mapping for this complex synthesis problem. Given these problems, could we enlarge the generation space and learn an automatic selection mechanism to synthesize more fine-grained generation results?

Based on these observations, in this paper, we propose a novel Multi-Channel Attention Selection Generative Adversarial Network (SelectionGAN), which contains two generation stages. The overall framework of the proposed SelectionGAN is shown in Fig. 2. In the first stage, we learn a cycled image-guidance generation sub-network, which accepts a pair consisting of an image and the conditional semantic guidance, and generates target images, which are further fed into a semantic guidance generation network to reconstruct the input semantic guidance. This cycled

- Hao Tang and Nicu Sebe are with the Department of Information Engineering and Computer Science (DISI), University of Trento, Trento 38123, Italy. E-mail: hao.tang@unitn.it
- Dan Xu and Philip H.S. Torr are with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, United Kingdom.
- Yan Yan is with the Department of Computer Science, Texas State University, San Marcos 78666, USA.
- Jason J. Corso is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor 48109, USA.

Manuscript revised on Feb 03, 2020.

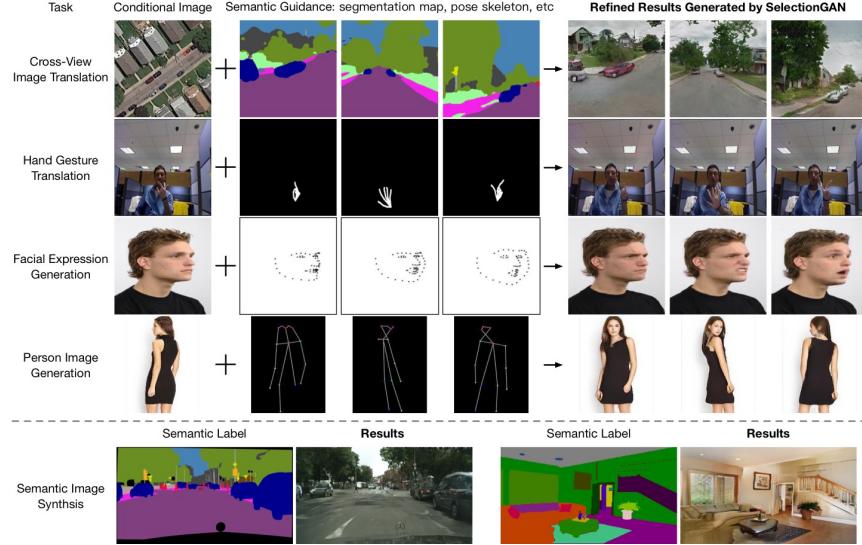


Fig. 1: SelectionGAN’s capabilities: (Top) Guided image-to-image translation (including cross-view image translation, hand gesture translation, facial expression generation and person image generation): synthesizing images from a single input image as well as semantic guidance (e.g., segmentation map, hand skeleton, facial landmark and human pose skeleton). (Bottom) Semantic image synthesis: SelectionGAN simultaneously produces realistic images while respecting the spatial semantic layout for both outdoor and indoor scenes.

guidance generation adds stronger supervision between the image and guidance domains, facilitating the optimization of the network.

The coarse outputs from the first generation network, including the input image, together with the deep feature maps from the last layer, are input into the second stage networks. We first employ the proposed multi-scale spatial pooling & channel selection module to enhance the multi-scale features in both spatial and channel dimensions. Next, several intermediate outputs are produced, and simultaneously we learn a set of multi-channel attention maps with the same number as the intermediate generations. These attention maps are used to spatially select from the intermediate generations, and are combined to synthesize a final output. Finally, to overcome the inaccurate semantic guidance issue, the multi-channel attention maps are further used to generate uncertainty maps to guide the reconstruction loss. Through extensive experimental evaluations, we demonstrate that SelectionGAN produces remarkably better results than the existing baselines on four different guided image-to-image translation tasks, i.e., segmentation map guided cross-view image translation, hand skeleton guided gesture-to-gesture translation, facial landmark guided expression-to-expression translation and pose guided person image generation. Moreover, the proposed framework and modules can be applied to other generation tasks such as semantic image synthesis.

Overall, the contributions of this paper are as follows:

- A novel multi-channel attention selection GAN framework (SelectionGAN) for guided image-to-image translation task is presented. It explores cascaded semantic guidance with a coarse-to-fine inference, and aims at producing a more detailed synthesis from richer and more diverse multiple intermediate generations.
- A novel multi-scale spatial pooling & channel selection module is proposed, which is utilized to automatically enhance the multi-scale feature representation in both

spatial and channel dimensions.

- A novel multi-channel attention selection module is proposed, which is utilized to attentively select interested intermediate generations and is able to significantly boost the quality of the final output. The multi-channel attention module also effectively learns uncertainty maps to guide the pixel loss for more robust optimization.
- Extensive experiments clearly demonstrate the effectiveness of the proposed SelectionGAN, and show state-of-the-art results on four guided image-to-image translation (including face, hand, body and street view) tasks. Moreover, we show the proposed SelectionGAN is effective on other generation tasks such as semantic image synthesis.

Part of the material presented here appeared in [8]. The current paper extends [8] in several ways. (1) We present a more detailed analysis of related works by including recently published works dealing with guided image-to-image translation. (2) We propose a novel module, i.e., multi-scale spatial pooling & channel selection, to automatically enhance the multi-scale feature representation in both spatial and channel dimensions. Equipped with this new module, our SelectionGAN proposed in [8] is upgraded to SelectionGAN++. (3) We extend the proposed framework to a more robust and general framework for handling different guided image-to-image translation tasks. (4) We extend the quantitative and qualitative experiments by comparing our SelectionGAN and SelectionGAN++ with the very recent works on four guided image-to-image translation tasks and one semantic image synthesis task with 11 public datasets.

2 RELATED WORK

Generative Adversarial Networks (GANs) [9] have shown the capability of generating high-quality images [10]. A vanilla GAN model [9] has two important components: a generator G and a discriminator D . The goal of G is to generate photo-realistic images from a noise vector, while

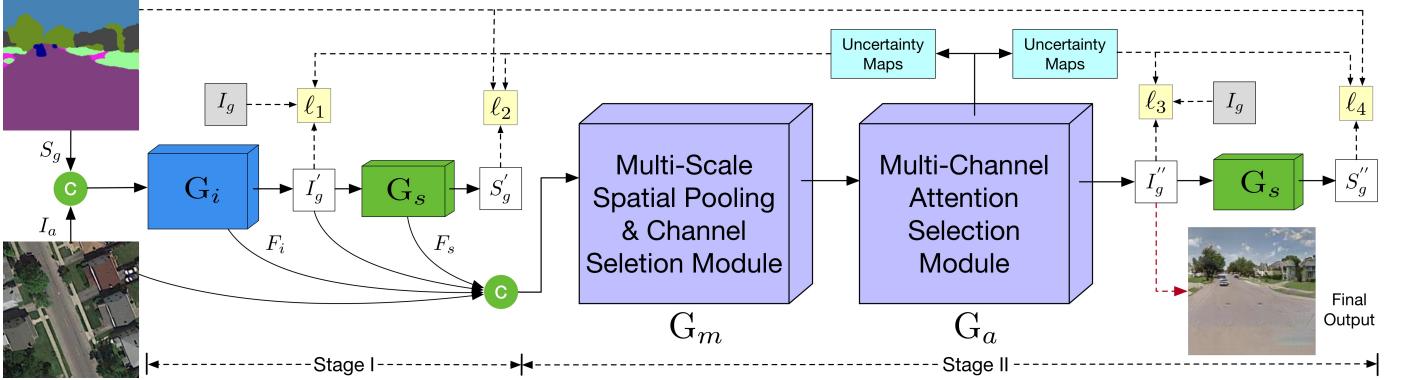


Fig. 2: Overview of the proposed SelectionGAN. Stage I presents a cycled semantic-guided generation sub-network which accepts both the input image and the conditional semantic guidance, and simultaneously synthesizes the target images and reconstructs the semantic guidance. Stage II takes the coarse predictions and the learned deep features from stage I, and performs a fine-grained generation using the proposed multi-scale spatial pooling & channel selection and the multi-channel attention selection modules.

D is trying to distinguish between a real image and the image generated by G . Although it is successfully used in generating images of high visual fidelity, there are still some challenges, i.e., how to generate images in a conditional setting. To generate domain-specific images, Conditional GANs (CGANs) [11] have been proposed. One specific application of CGANs is image-to-image translation [12].

Image-to-Image Translation frameworks learn a parametric mapping between inputs and outputs. For example, Isola et al. [12] propose Pix2pix, which is a supervised model and uses a CGAN to learn a translation function from input to output image domains. Based on Pix2pix, Wang et al. [13] propose Pix2pixHD, which can turn semantic maps into photo-realistic images.

Our work builds upon the recent advances in image-to-image translation, i.e., Pix2pix, and aims to extend it to a broader set of guided image-to-image translation problem, which provides users with more input. Moreover, the proposed multi-scale spatial pooling & channel selection and the multi-channel attention selection modules are network-agnostic and can be plugged into any existing CNN-based generation architectures.

Guided Image-to-Image Translation is a variant of image-to-image translation problem aimed at translating an input image to a target image while respecting certain constraints specified by some external guidance, such as class labels [14], [15], text descriptions [16], [17], human key-point/skeleton [1], [2], [18], segmentation maps [3], [8], [19], [20] and reference images [4], [21]. Given that different generation tasks need different guidance information, existing works are tailored to a specific application, i.e., with specifically designed network architectures and training objectives. For example, Ma et al. propose PG2 [1], which is a two-stage framework and uses the pose mask loss for generating person images based on an image of that person and human pose keypoints. Tang et al. propose GestureGAN [2], which is a forward-backward consistency architecture and adopt the proposed color loss to generate novel hand gesture images based on the input image and conditional hand skeletons. Wang et al. propose the few-shot Vid2Vid framework [3], which uses a carefully designed weight generation module to synthesize videos

that realistically reflect the style of the input image and the layout of conditional segmentation maps.

Compared to existing works in guided image-to-image translation, we develop a unified and robust framework that is application-independent. In this way, the proposed framework can be widely applied to many generation tasks with different forms of guidance, such as scene segmentation maps, hand skeletons, facial landmarks and human body skeleton, as shown in Fig. 1.

Attention Learning in Image-to-Image Translation. Attention learning have been extensively exploited in computer vision and natural language processing, e.g., [22], [23]. To improve the image generation performance, the attention mechanism has also been recently investigated in the image-to-image translation tasks [24], [25], [26].

Unlike existing attention methods, we aim at a more effective network design and propose a novel multi-channel attention selection GAN, which allows to automatically select from multiple diverse and rich intermediate generations, and thus significantly improving the generation quality. To the best of our knowledge, our model is the first attempt to incorporate a multi-channel attention selection module within a GAN framework for image-to-image translation.

3 SELECTIONGAN

In this section we present the details of the proposed multi-channel attention selection GAN. An illustration of the overall network structure is depicted in Fig. 2. In the first stage, we present a cascaded semantic-guided generation sub-network, which utilizes the input image and the conditional semantic guidance as inputs, and generate the target images while respecting the semantic guidance.

These generated images are further input into a semantic guidance generator to recover the input semantic guidance forming a generation cycle. In the second stage, the coarse synthesis and the deep features from the first stage are combined, and then are passed to the proposed multi-scale spatial pooling & channel selection module to model the long-range multi-scale dependencies between each channel of feature representations. Thus the enhanced feature maps

are fed to the proposed multi-channel attention selection module, which aims at producing more fine-grained synthesis from a larger generation space and also at generating uncertainty maps to jointly guide multiple optimization losses.

3.1 Cascade Semantic-Guided Generation

Semantic-Guided Generation. We target to translate an input image to another while respecting the semantic guidance. There are many strategies to incorporate the additional semantic guidance into the image-to-image translation model [4] and the most straight forward scheme is input concatenation. Specifically, as shown in Fig. 2, we concatenate the input image I_a and the semantic guidance S_g , and feed them into the image generator G_i and synthesize the target image I'_g as $I'_g = G_i(I_a, S_g)$. In this way, the semantic guidance provides stronger supervision to guide the image-to-image translation in the deep network.

Semantic-Guided Cycle. Existing guided image-to-image translation methods [1], [4], [27] only use semantic guidance as input to guide the image generation process, which actually provide a weak guidance. Different from theirs, we apply the semantic guidance not only as input but also as part of the network’s output. Specifically, as shown in Fig. 2, we propose a cycled semantic guidance generation network to benefit more the semantic guidance information in learning jointly. The conditional semantic guidance S_g together with the input image I_a are input into the image generator G_i , and produce the synthesized image I'_g . Then I'_g is further fed into the semantic guidance generator G_s , which reconstructs a new semantic guidance S'_g . We can formalize the process as $S'_g = G_s(I'_g) = G_s(G_i(I_a, S_g))$. Then the optimization objective is to make S'_g as close as possible to S_g , which naturally forms a semantic guidance generation cycle, i.e., $[I_a, S_g] \xrightarrow{G_i} I'_g \xrightarrow{G_s} S'_g \approx S_g$. The two generators are explicitly connected by the ground-truth semantic guidance, which in this way provides extra constraints on the generators to better learn the semantic structure consistency. We observe that the simultaneous generation of both the images and the semantic guidance improves the generation performance in our experiments section.

Cascade Generation. Due to the complexity of the tasks such as in pose guided person image generation, input and output domains usually have little overlap, which apparently leads to ambiguity issues in the generation process. Moreover, we observe that the image generator G_i outputs a coarse synthesis after the first stage, which yields blurred image details and high pixel-level dissimilarity with the target images. Both inspire us to explore a coarse-to-fine generation strategy in order to boost the synthesis performance based on the coarse predictions. Cascade models have been used in several other computer vision tasks such as object detection [28] and semantic segmentation [29], and have shown great effectiveness. In this paper, we introduce the cascade strategy to deal with the guided image-to-image translation problems. In both stages we have a basic cycled semantic guidance generation sub-network, while in the second stage, we propose two novel multi-scale spatial pooling & channel selection and multi-channel attention

selection modules to better utilize the coarse outputs from the first stage and to produce fine-grained final outputs. We observed significant improvement by using the proposed cascade strategy, illustrated in the experimental part.

3.2 Multi-Scale Spatial Pooling & Channel Selection

An overview of the proposed multi-scale spatial pooling & channel selection module is shown in Fig. 3. The module consists of a multi-scale spatial pooling and a multi-scale channel selection components. In this way, the proposed module can learn multi-scale deep feature interdependencies in both spatial and channel dimensions.

Multi-Scale Spatial Pooling. Since there exists a large object/scene deformation between the source domain and the target domain, a single-scale feature may not be able to capture all the necessary spatial information for a fine-grained generation. Thus, we propose a multi-scale spatial pooling scheme, which uses a set of different kernel sizes and strides to perform a global average pooling on the same input features. By so doing, we obtain multi-scale features with different receptive fields to perceive different spatial contexts. More specifically, given the coarse inputs and the deep features produced from the stage I, we first concatenate all of them as new features denoted as $\mathcal{F}_c \in \mathbb{R}^{C \times H \times W}$ for the stage II as:

$$\mathcal{F}_c = \text{concat}(I_a, I'_g, F_i, F_s), \quad (1)$$

where $\text{concat}(\cdot)$ is a function for channel-wise concatenation operation; F_i and F_s are features from the last convolution layers of the generators G_i and G_s , respectively. H and W are width and height of the features, and C is the number of channels. We apply a set of M spatial scales $\{s_i\}_{i=1}^M$ in pooling, resulting in pooled features with different spatial resolution. Different from the pooling scheme used in [30] which directly combines all the features after pooling, we first select each pooled feature via an element-wise multiplication with the input feature. Since in our task the input features are from different sources, highly correlated features would preserve more useful information for the generation. Let us denote $\text{pl_up}_s(\cdot)$ as pooling at a scale s followed by an up-sampling operation to rescale the pooled feature at the same resolution, and \otimes as element-wise multiplication, we can formalize the whole process as follows:

$$\mathcal{F}_m \leftarrow \text{concat}(\mathcal{F}_c, \mathcal{F}_c \otimes \text{pl_up}_1(\mathcal{F}_c), \dots, \mathcal{F}_c \otimes \text{pl_up}_M(\mathcal{F}_c)), \quad (2)$$

which produces new multi-scale features $\mathcal{F}_m \in \mathbb{R}^{4C \times H \times W}$ (in our experiments, we set $M=3$) for the use in the next multi-scale channel selection module. By doing so, the “level” of features can be enriched by combining multiple scale feature maps.

Multi-Scale Channel Selection. Each channel map of \mathcal{F}_m can be now regarded as a scale-specific response, and different scale feature maps should be associated with each other. To exploit the interdependencies between each scale features of \mathcal{F}_m , we propose a multi-scale channel selection module to explicitly model interdependencies between channels of multi-scale feature \mathcal{F}_m . The structure of multi-scale channel selection module is illustrated in Fig. 3.

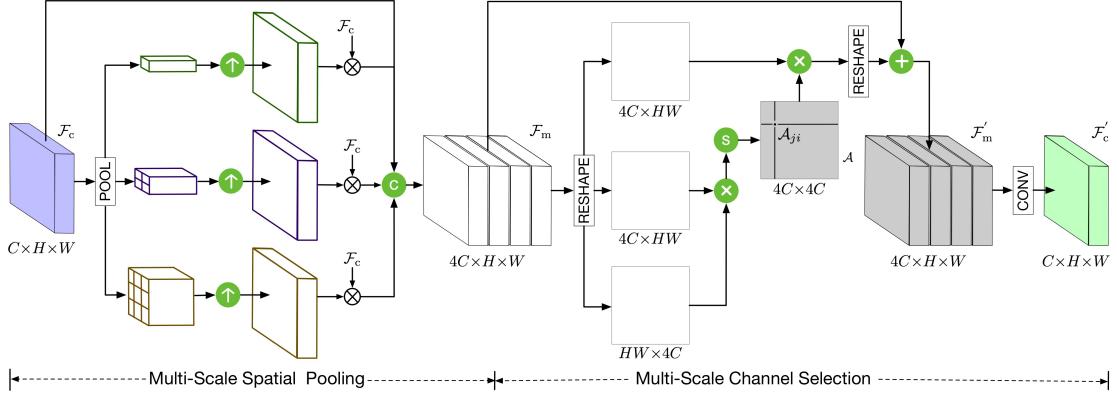


Fig. 3: Proposed multi-scale spatial pooling & channel selection module. The multi-scale spatial pooling pools features from different receptive fields in order to have better generation of image details; the multi-scale channel selection aims at automatically emphasizing interdependent channel maps by integrating associated features among all multi-scale channel maps to improve deep feature representation. \oplus , \otimes , \odot , \odot and \circledast denote element-wise addition, element-wise multiplication, channel-wise concatenation, softmax and up-sampling operation, respectively.

The channel attention map \mathcal{A} can be obtained from the multi-scale feature \mathcal{F}_m . More specific, \mathcal{F}_m is first reshaped to $\mathbb{R}^{4C \times HW}$, and then a matrix multiplication is preformed between \mathcal{F}_m and the transpose of \mathcal{F}_m . Next, we employ a Softmax activation function to obtain the channel attention map $\mathcal{A} \in \mathbb{R}^{4C \times 4C}$. Each pixel \mathcal{A}_{ji} in \mathcal{A} measures the i^{th} channel's impact on the j^{th} channel. In this way, the correlation can be built between features from different scales. Moreover, to reshape back to $\mathbb{R}^{4C \times H \times W}$, we perform a matrix multiplication between \mathcal{A} and the transpose of \mathcal{F}_m . Then, the result is multiplied by a parameter α and added to the original feature \mathcal{F}_m to obtain the channel-wise enhanced feature $\mathcal{F}'_m \in \mathbb{R}^{4C \times H \times W}$,

$$\mathcal{F}'_m = \alpha \sum_{i=1}^{4C} (\mathcal{A}_{ji} \mathcal{F}_{mi}) + \mathcal{F}_{mj}. \quad (3)$$

By doing so, each channel in the final feature \mathcal{F}'_m is a weighted sum of all channels and it models the long-range dependencies between multi-scale feature maps. Finally, the enhanced feature $\mathcal{F}'_m \in \mathbb{R}^{C \times H \times W}$, which has the same size as the original one \mathcal{F}_c . This design ensures that the proposed multi-scale spatial pooling & channel selection module can be plugged into existing computer vision architectures.

3.3 Multi-Channel Attention Selection

In previous image-to-image translation works, the image was generated only in a three-channel RGB space. We argue that this is not enough for the complex translation problem we are dealing with, and thus we explore using a larger generation space to have a richer synthesis via constructing multiple intermediate generations. Accordingly, we design a multi-channel attention mechanism to automatically perform spatial and temporal selection from the generations to synthesize a fine-grained final output.

Given the enhanced multi-scale feature volume $\mathcal{F}'_c \in \mathbb{R}^{C \times H \times W}$, where H and W are width and height of the features, and C is the number of channels, we consider two directions as shown in Fig. 4. One is for the generation of multiple intermediate image synthesis and the other is for

the generation of multi-channel attention maps. To produce N different intermediate generations $I_G = \{I_G^i\}_{i=1}^N$, a convolution operation is performed with N convolutional filters $\{W_G^i, b_G^i\}_{i=1}^N$ followed by a $\tanh(\cdot)$ non-linear activation operation. For the generation of corresponding N attention maps, the other group of filters $\{W_A^i, b_A^i\}_{i=1}^N$ is applied. Then the intermediate generations and the attention maps are calculated as follows:

$$\begin{aligned} I_G^i &= \tanh(\mathcal{F}'_c W_G^i + b_G^i), & \text{for } i = 1, \dots, N \\ I_A^i &= \text{Softmax}(\mathcal{F}'_c W_A^i + b_A^i), & \text{for } i = 1, \dots, N \end{aligned} \quad (4)$$

where $\text{Softmax}(\cdot)$ is a channel-wise softmax function used for the normalization. Finally, the learned attention maps are utilized to perform channel-wise selection from each intermediate generation as follows:

$$I_g'' = (I_A^1 \otimes I_G^1) \oplus \dots \oplus (I_A^N \otimes I_G^N) \quad (5)$$

where I_g'' represents the final synthesized generation selected from the multiple diverse results, and \oplus denotes the element-wise addition. We also generate a final semantic guidance in the second stage as in the first stage, i.e., $S_g'' = G_s(I_g'')$. Due to the same purpose of the two semantic guidance generators, we use a single G_s twice by sharing the parameters in both stages to reduce the network capacity.

Uncertainty-Guided Pixel Loss. As we discussed in the introduction, the semantic guidance obtained from the pre-trained model is not accurate for all the pixels, leading to a wrong guidance during training. To tackle this issue, we propose to learn uncertainty maps to control the optimization loss as shown in Fig. 4. The uncertainty learning has been investigated in [31] for multi-task learning, and here we introduce it for solving the noisy semantic guidance problem. Assume that we have K different loss maps which need a guidance. The multiple generated attention maps are first concatenated and passed to a convolution layer with K filters $\{W_u^i\}_{i=1}^K$ to produce a set of K uncertainty maps. The reason for using the attention maps to generate uncertainty maps is that the attention maps directly affect the final generation leading to a close connection with the

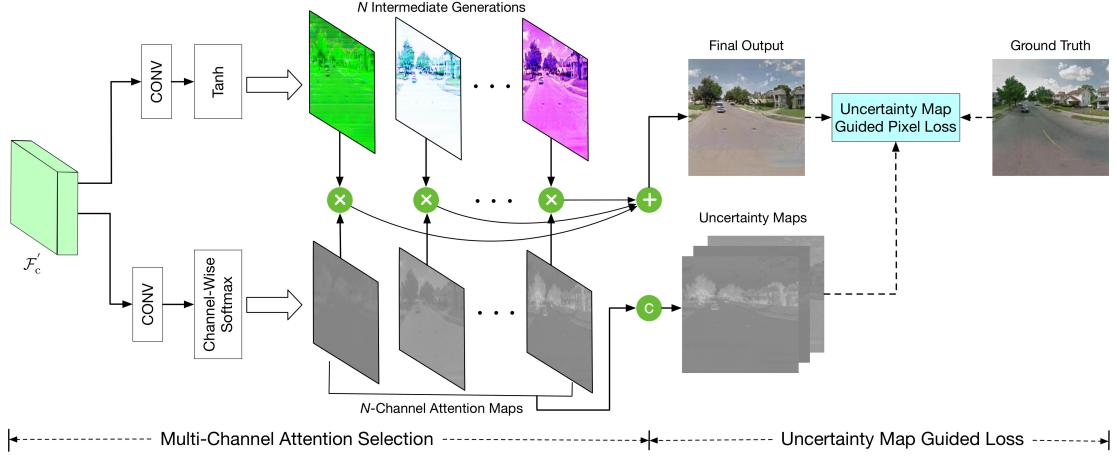


Fig. 4: Proposed multi-channel attention selection module. The multi-channel attention selection aims at automatically select from a set of intermediate diverse generations in a larger generation space to improve the generation quality; the multi-channel attention module also effectively learns uncertainty maps to guide the pixel loss for robust joint images and guidances optimization. \oplus , \otimes and \odot denote element-wise addition, element-wise multiplication and channel-wise concatenation, respectively.

loss. Let \mathcal{L}_p^i denote a pixel-level loss map and U_i denote the i -th uncertainty map, we have:

$$\begin{aligned} U_i &= \sigma(W_u^i(\text{concat}(I_A^1, \dots, I_A^N) + b_u^i)) \\ \mathcal{L}_p^i &\leftarrow \frac{\mathcal{L}_p^i}{U_i} + \log U_i, \quad \text{for } i = 1, \dots, K \end{aligned} \quad (6)$$

where $\sigma(\cdot)$ is a Sigmoid function for pixel-level normalization. The uncertainty map is automatically learned and acts as a weighting scheme to control the optimization loss.

Parameter-Sharing Discriminator. We extend the vanilla discriminator in [12] to a parameter-sharing structure. In the first stage, this structure takes the real image I_a and the generated image I_g' or the ground-truth image I_g as input. The discriminator D learns to tell whether a pair of images from different domains is associated with each other or not. In the second stage, it accepts the real image I_a and the generated image I_g'' or the real image I_g as inputs. This pairwise input encourages D to discriminate the diversity of image structure and to capture the local-aware information.

3.4 Overall Optimization Objective

Adversarial Loss. In the first stage, the adversarial loss of D for distinguishing synthesized image pairs $[I_a, I_g']$ from real image pairs $[I_a, I_g]$ is formulated as follows:

$$\mathcal{L}_{cGAN}(I_a, I_g') = \mathbb{E}_{I_a, I_g} [\log D(I_a, I_g')] + \mathbb{E}_{I_a, I_g'} [\log(1 - D(I_a, I_g'))]. \quad (7)$$

In the second stage, the adversarial loss of D for distinguishing synthesized image pairs $[I_a, I_g'']$ from real image pairs $[I_a, I_g]$ is formulated as follows:

$$\mathcal{L}_{cGAN}(I_a, I_g'') = \mathbb{E}_{I_a, I_g} [\log D(I_a, I_g'')] + \mathbb{E}_{I_a, I_g''} [\log(1 - D(I_a, I_g''))]. \quad (8)$$

Both losses aim to preserve the local structure information and produce visually pleasing synthesized images. Thus, the adversarial loss of the proposed SelectionGAN is the sum of Eq. (7) and (8),

$$\mathcal{L}_{cGAN} = \mathcal{L}_{cGAN}(I_a, I_g') + \lambda \mathcal{L}_{cGAN}(I_a, I_g''). \quad (9)$$

Overall Loss. The total optimization loss is a weighted sum of the above losses. Generators G_i , G_s , multi-scale spatial pooling & channel selection module G_m , multi-channel attention selection network G_a and discriminator D are trained in an end-to-end fashion optimizing the following min-max function:

$$\min_{\{G_i, G_s, G_m, G_a\}} \max_{\{D\}} \mathcal{L} = \sum_{i=1}^4 \lambda_i \mathcal{L}_p^i + \mathcal{L}_{cGAN} + \lambda_{tv} \mathcal{L}_{tv}. \quad (10)$$

where \mathcal{L}_p^i uses the L1 reconstruction to separately calculate the pixel loss between the generated 4 images (i.e., I_g' , S_g' , I_g'' and S_g'') and the corresponding real images. \mathcal{L}_{tv} is the total variation regularization [32] on the final synthesized image I_g'' . λ_i and λ_{tv} are the trade-off parameters to control the relative importance of different objectives. The training is performed by solving the min-max optimization problem.

3.5 Implementation Details

Network Architecture. For a fair comparison, we employ U-Net [12] as our generator architectures G_i and G_s . U-Net is a network with skip connections between a down-sampling encoder and an up-sampling decoder. Such architecture comprehensively retains contextual and textural information, which is crucial for removing artifacts and padding textures. Since our focus is on the image generation task, G_i is more important than G_s . Thus we use a deeper network for G_i and a shallow network for G_s , such asymmetric architecture design can also be observed in other generation papers [33]. Specifically, the filters in first convolutional layer of G_i and G_s are 64 and 4, respectively. For the network G_a , the kernel size of convolutions for generating the intermediate images and attention maps are 3×3 and 1×1 , respectively. We adopt PatchGAN [12] for the discriminator D .

Training Details. We mainly focus on four guided image-to-image translation tasks in this paper. For cross-view image translation, we follow [19] and use RefineNet [6] and [7] to generate segmentation maps on Dayton, SVA, Ego2Top

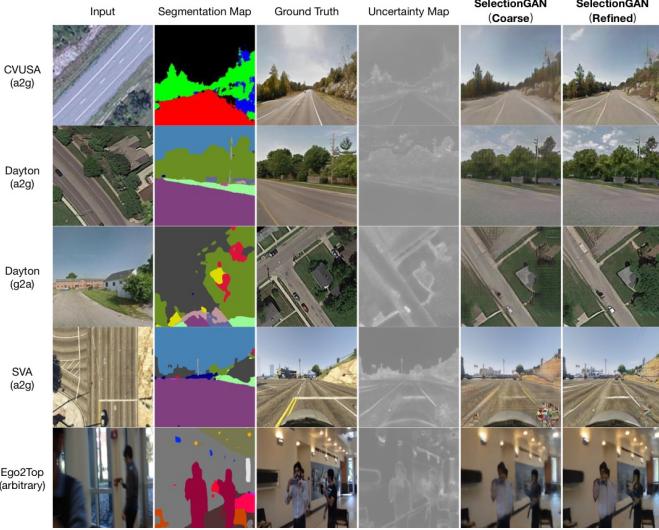


Fig. 5: Results of cross-view image translation generated by the proposed SelectionGAN on different datasets. From left to right: input image, segmentation map, ground truth, uncertainty map, coarse result and refined result.

datasets as training data, respectively. For facial expression generation, we follow [34] and use OpenFace [5] to extract facial landmarks on Radboud Faces dataset as training data. For both hand gesture generation and human pose generation tasks, we follow [1], [2] and employ OpenPose [35] as pose joints detector and filter out images where no human hand and body are detected in the associated datasets.

We follow the optimization method in [9] to optimize the proposed SelectionGAN, i.e., one gradient descent step on discriminator and generators alternately. We first train G_i, G_s, G_m, G_a with D fixed, and then train D with G_i, G_s, G_m, G_a fixed. The proposed SelectionGAN is trained and optimized in an end-to-end fashion. We employ Adam [36] with momentum terms $\beta_1=0.5$ and $\beta_2=0.999$ as our solver. In our experiments, we set $\lambda_{tv}=1e-6$, $\lambda_1=100$, $\lambda_2=1$, $\lambda_3=200$ and $\lambda_4=2$ in Eq. (10), and $\lambda=4$ in Eq. (9). The number of attention channels N in Eq. (4) is set to 10. The proposed SelectionGAN is implemented in PyTorch.

4 EXPERIMENTS

We conduct extensive experiments on a variety of guided image-to-image translation tasks such as segmentation map guided cross-view image translation, facial landmark guided expression-to-expression translation, hand skeleton guided gesture-to-gesture translation and pose skeleton guided person image generation. Moreover, to explore the generality of the proposed SelectionGAN on other generation tasks, we conduct experiments on the challenging semantic image synthesis task.

4.1 Results on Cross-View Image Translation

Datasets. We follow [8], [19], [37] and perform experiments on four public cross-view image translation datasets: (i) The Dayton dataset [38], which contains 76,048 images and the train/test split is 55,000/21,048. The images in the original dataset have 354×354 resolution. We resize them to 256×256 . (ii) The CVUSA dataset [39] consists of

TABLE 1: Ablations study of the proposed SelectionGAN.

	Setup of SelectionGAN	SSIM	PSNR	SD
A	$I_a \xrightarrow{G_a} I_g'$	0.4555	19.6574	18.8870
B	$S_g \xrightarrow{G_a} I_g'$	0.5223	22.4961	19.2648
C	$[I_a, S_g] \xrightarrow{G_a} I_g'$	0.5374	22.8345	19.2075
D	$[I_a, S_g] \xrightarrow{G_a} I_g' \xrightarrow{G_g} S_g'$	0.5438	22.9773	19.4568
E	D + Uncertainty-Guided Pixel Loss	0.5522	23.0317	19.5127
F	E + Multi-Channel Attention Selection	0.5989	23.7562	20.0000
G	F + Total Variation Regularization	0.6047	23.7956	20.0830
H	G + Multi-Scale Spatial Pooling	0.6167	23.9310	20.1214

TABLE 2: Quantitative results of coarse-to-fine generation on cross-view image translation task.

Baseline	Stage I	Stage II	SSIM	PSNR	SD
F	✓		0.5551	23.1919	19.6311
F		✓	0.5989	23.7562	20.0000
G	✓		0.5680	23.2574	19.7371
G		✓	0.6047	23.7956	20.0830
H	✓		0.5567	23.1545	19.6034
H		✓	0.6167	23.9310	20.1214

TABLE 3: Influence of the number of attention channels N .

N	SSIM	PSNR	SD
0	0.5438	22.9773	19.4568
1	0.5522	23.0317	19.5127
5	0.5901	23.8068	20.0033
10	0.5986	23.7336	19.9993
32	0.5950	23.8265	19.9086

35,532/8,884 image pairs in train/test split. Following [19], [40], the aerial images are center-cropped to 224×224 and resized to 256×256 . For the ground level images and corresponding segmentation maps, we take the first quarter of both and resize them to 256×256 . (iii) The Surround Vehicle Awareness (SVA) dataset [41] is a synthetic dataset collected from Grand Theft Auto V (GTAV) video game. Following [37], we select every tenth frame to remove redundancy in this dataset since the consecutive frames in each set are very similar to each other. Thus, we collect 46,030/22,254 image pairs for training and testing, respectively. (iv) The Ego2Top dataset [42] is more challenging and contains different indoor and outdoor conditions. Each case contains one top-view video and several egocentric videos captured by the people visible in the top-view camera. This dataset has more than 230,000 frames. For training data, we follow [8] and randomly select 386,357 pairs and each pair is composed of two images of the same scene but different viewpoints. We randomly select 25,600 pairs for evaluation.

Parameter Settings. For a fair comparison, we adopt the same training setup as in [12], [19]. All images are scaled to 256×256 , and we enabled image flipping and random crops for data augmentation. Similar to [19], the experiments for Dayton are trained for 35 epochs with batch size of 4. For CVUSA, we follow the same setup as in [19], [40], and train our network for 30 epochs with batch size of 4. For Ego2Top, all models are trained with 10 epochs using batch size 8. For SVA, all models are trained with 20 epoch using batch size 4.

Evaluation Metrics. Similar to [8], [19], we employ Inception Score [43], top-k prediction accuracy, KL score and Fréchet Inception Distance (FID) [44] for the quantitative analysis. These metrics evaluate the generated images from a high-level feature space. We also employ pixel-level similarity metrics to evaluate our method, i.e., Structural-Similarity (SSIM) [45], Peak Signal-to-Noise Ratio (PSNR) and Sharpness Difference (SD).

Baseline Models. We first conduct an ablation study on Dayton to evaluate the components of the proposed Se-

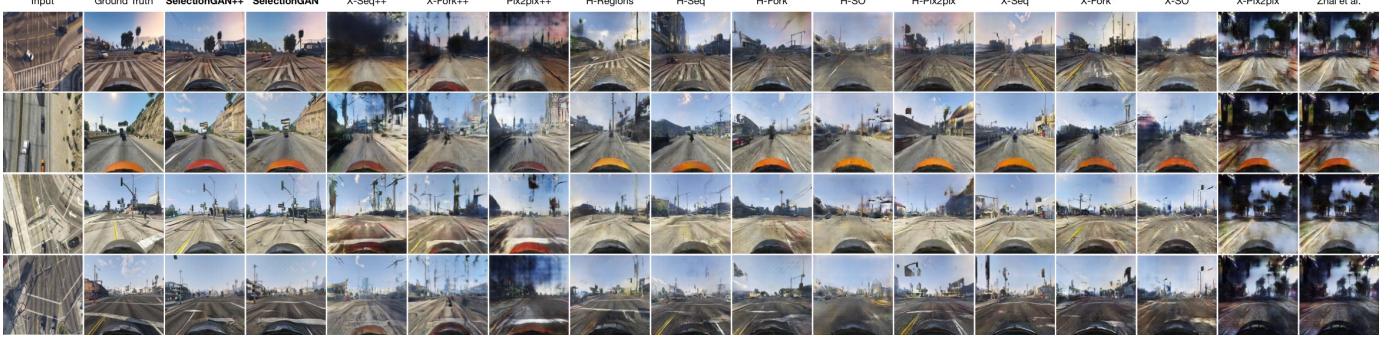


Fig. 6: Results of cross-view image translation on SVA. From left to right: input image, ground truth, SelectionGAN++ (Ours), SelectionGAN (Our), X-Seq++, X-Fork++, Pix2pix++, H-Regions, H-Seq, H-Fork, H-SO, H-Pix2pix, X-Seq, X-Fork, X-SO, X-Pix2pix and Zhai et al.

TABLE 4: Quantitative results of cross-view image translation on SVA. For all metrics except KL and FID, higher is better. (*) Inception Score for real (ground truth) data is 3.1282, 2.4932 and 3.4646 for all, top-1 and top-5 setups, respectively.

Method	Publish	Accuracy (%)			Inception Score*			SSIM	PSNR	SD	KL	FID	
		Top-1		Top-5	All	Top-1	Top-5						
		Top-1	Top-5										
X-Pix2pix [12]	CVPR 2017	8.5961	30.3288	9.0260	29.9102	2.0131	1.7221	2.2370	0.3206	17.9944	17.0254	19.5533	859.66
X-SO [37]	CVIU 2019	7.5146	30.9507	10.3905	38.9822	2.4951	1.8940	2.6634	0.4552	21.5312	17.5285	12.0906	443.79
X-Fork [19]	CVPR 2018	17.3794	53.4725	23.8315	63.5045	2.1888	1.9776	2.3664	0.4235	21.2400	16.9371	4.1925	129.16
X-Seq [19]	CVPR 2018	19.5056	57.1010	25.8807	65.3005	2.2232	1.9842	2.4344	0.4638	22.3411	17.4138	3.7585	118.70
H-Pix2pix [37]	CVIU 2019	18.0706	54.8068	23.4400	62.3072	2.1906	1.9507	2.4069	0.4327	21.6860	16.9468	4.2894	117.13
H-SO [37]	CVIU 2019	5.2444	26.4697	5.2544	31.9527	2.3202	1.9410	2.7340	0.4457	21.7709	17.3876	12.8761	1452.88
H-Fork [37]	CVIU 2019	18.0182	51.0756	26.6747	62.8166	2.3202	1.9525	2.3918	0.4240	21.6327	16.8653	4.7246	109.43
H-Seq [37]	CVIU 2019	20.7391	57.5378	28.5517	67.4649	2.2394	1.9892	2.4385	0.4249	21.4770	17.5616	4.4260	95.12
H-Regions [37]	CVIU 2019	15.4803	48.0677	21.8225	56.8994	2.6328	2.0732	2.8347	0.4044	20.9848	17.6858	6.0638	88.78
Pix2pix++ [12]	CVPR 2017	8.8687	34.5434	9.2713	35.7490	2.5625	2.0879	2.7961	0.3664	17.6549	18.4015	13.1153	220.23
X-Fork++ [19]	CVPR 2018	10.2658	37.8405	11.4138	38.7976	2.4280	2.0387	2.7630	0.3406	17.3937	18.2153	10.1403	166.33
X-Seq++ [19]	CVPR 2018	11.2580	36.8018	11.9838	36.9231	2.6849	2.1325	2.9397	0.3617	17.4893	18.4122	11.8560	154.80
SelectionGAN	Ours	33.9055	71.8779	50.8878	85.0019	2.6576	2.1279	2.9267	0.5752	24.7136	19.7302	2.6183	26.09
SelectionGAN++	Ours	35.9008	73.3249	52.5346	86.9432	2.7370	2.1914	3.0271	0.5481	24.2886	19.2001	2.5788	37.17

TABLE 5: Quantitative results of cross-view image translation on CVUSA. For all metrics except KL, higher is better. (*) Inception Score for real (ground truth) data is 4.8741, 3.2959 and 4.9943 for all, top-1 and top-5 setups, respectively.

Method	Publish	Accuracy (%)			Inception Score*			SSIM	PSNR	SD	KL	
		Top-1		Top-5	All	Top-1	Top-5					
		Top-1	Top-5									
Zhai et al. [40]	CVPR 2017	13.97	14.03	42.09	52.29	1.8434	1.5171	1.8666	0.4147	17.4886	16.6184	27.43 ± 1.63
Pix2pix [12]	CVPR 2017	7.33	9.25	25.81	32.67	3.2771	2.2219	3.4312	0.3923	17.6578	18.5239	59.81 ± 2.12
X-SO [37]	CVIU 2019	0.29	0.21	6.14	9.08	1.7575	1.4145	1.7791	0.3451	17.6201	16.9919	414.25 ± 2.37
X-Fork [19]	CVPR 2018	20.58	31.24	50.51	63.66	3.4432	2.5447	3.5567	0.4356	19.0509	18.6706	11.71 ± 1.55
X-Seq [19]	CVPR 2018	15.98	24.14	42.91	54.41	3.8151	2.6738	4.0077	0.4231	18.8067	18.4378	15.52 ± 1.73
Pix2pix++ [12]	CVPR 2017	26.45	41.87	57.26	72.87	3.2592	2.4175	3.5078	0.4617	21.5739	18.9044	9.47 ± 1.69
X-Fork++ [19]	CVPR 2018	31.03	49.65	64.47	81.16	3.3758	2.5375	3.5711	0.4769	21.6504	18.9856	7.18 ± 1.56
X-Seq++ [19]	CVPR 2018	34.69	54.61	67.12	83.46	3.3919	2.5474	3.4858	0.4740	21.6733	18.9907	5.19 ± 1.31
SelectionGAN [8]	Ours	41.52	65.51	74.32	89.66	3.8074	2.7181	3.9197	0.5323	23.1466	19.6100	2.96 ± 0.97

TABLE 6: Quantitative evaluation of cross-view image translation on Dayton in a2g direction. For all metrics except KL, higher is better. (*) Inception Score for real (ground truth) data is 3.8319, 2.5753 and 3.9222 for all, top-1 and top-5 setups, respectively.

Method	Publish	Accuracy (%)			Inception Score*			SSIM	PSNR	SD	KL	
		Top-1		Top-5	All	Top-1	Top-5					
		Top-1	Top-5									
Pix2pix [12]	CVPR 2017	6.80	9.15	23.55	27.00	2.8515	1.9342	2.9083	0.4180	17.6291	19.2821	38.26 ± 1.88
X-SO [37]	CVIU 2019	27.56	41.15	57.96	73.20	2.9459	2.0963	2.9980	0.4772	19.6203	19.2939	7.20 ± 1.37
X-Fork [19]	CVPR 2018	30.00	48.68	61.57	78.84	3.0720	2.2402	3.0932	0.4963	19.8928	19.4533	6.00 ± 1.28
X-Seq [19]	CVPR 2018	30.16	49.85	62.59	80.70	2.7384	2.1304	2.7674	0.5031	20.2803	19.5258	5.93 ± 1.32
Pix2pix++ [12]	CVPR 2017	32.06	54.70	63.19	81.01	3.1709	2.1200	3.2001	0.4871	21.6675	18.8504	5.49 ± 1.25
X-Fork++ [19]	CVPR 2018	34.67	59.14	66.37	84.70	3.0737	2.1508	3.0893	0.4982	21.7260	18.9402	4.59 ± 1.16
X-Seq++ [19]	CVPR 2018	31.58	51.67	65.21	82.48	3.1703	2.2185	3.2444	0.4912	21.7659	18.9265	4.94 ± 1.18
SelectionGAN [8]	Ours	42.11	68.12	77.74	92.89	3.0613	2.2707	3.1336	0.5938	23.8874	20.0174	2.74 ± 0.86

lectionGAN. To reduce the training time, we randomly select 1/3 samples from the whole 55,000/21,048 samples, i.e., around 18,334 samples for training and 7,017 samples for testing. The proposed SelectionGAN considers eight baselines (A, B, C, D, E, F, G, H) as shown in Table 1. Baseline A uses a Pix2pix structure [12] and generates I'_g using a single image I_a . Baseline B uses the same Pix2pix model and generates I'_g using the corresponding semantic guidance S_g . Baseline C also uses the Pix2pix structure, and inputs the combination of a conditional image I_a and the semantic guidance S_g to the generator G_i . Baseline D uses the proposed cycled semantic guidance generation upon Baseline C. Baseline E represents the pixel loss guided by

the learned uncertainty maps. Baseline F employs the proposed multi-channel attention selection module to generate multiple intermediate generations, and to make the neural network attentively select which part is more important for generating the target image. Baseline G adds the total variation regularization on the final result I'_g . Baseline H employs the proposed multi-scale spatial pooling module to refine the features F_c from stage I. All the baseline models are trained and tested on the same data using the configuration.

Ablation Analysis. The results of the ablation study are shown in Table 1. We observe that Baseline B is better than baseline A since S_g contains more structural information

TABLE 7: Quantitative results of cross-view image translation on Ego2Top. For all metrics except KL, higher is better. (*) Inception Score for real (ground truth) data is 6.4523, 2.8507 and 5.4662 for all, top-1 and top-5 setups, respectively.

Method	Publish	SSIM	PSNR	SD	Inception Score*			Accuracy (%)		KL Score		
					All	Top-1	Top-5	Top-1	Top-5			
Pix2pix [12]	CVPR 2017	0.2213	15.7197	16.5949	2.5418	1.6797	2.4947	1.22	1.57	5.33	6.86	120.46 ± 1.94
X-Fork [19]	CVPR 2018	0.2740	16.3709	17.3509	4.6447	2.1386	3.8417	5.91	10.22	20.98	30.29	22.12 ± 1.65
X-Seq [19]	CVPR 2018	0.2738	16.3788	17.2624	4.5094	2.0276	3.6756	4.78	8.96	17.04	24.40	25.19 ± 1.73
Pix2pix++ [12]	CVPR 2017	0.3779	21.1346	17.8056	5.0833	2.4096	4.4595	19.53	33.19	40.89	48.34	10.93 ± 1.87
X-Fork++ [19]	CVPR 2018	0.3560	20.5788	17.6183	5.2266	2.4100	4.5591	13.92	22.38	34.20	42.42	17.34 ± 1.98
X-Seq++ [19]	CVPR 2018	0.3878	21.2327	17.9469	4.9890	2.3519	4.2881	19.41	36.11	40.46	50.41	9.33 ± 1.64
SelectionGAN	Ours	0.6024	26.6565	19.7755	5.6200	2.5328	4.7648	28.31	54.56	62.97	76.30	3.05 ± 0.91



Fig. 7: Results of cross-view image translation on CVUSA. From left to right: input image, ground truth, SelectionGAN (Our), X-Seq++, X-Fork++, Pix2pix++, X-Seq, X-Fork, X-SO, Pix2pix and Zhai et al.



Fig. 9: Results of cross-view image translation on Ego2Top. From left to right: input image, segmentation map, ground truth, Pix2pix++, X-Fork++, X-Seq++, SelectionGAN (Ours) and uncertainty maps generated by SelectionGAN.

TABLE 8: Per-class accuracy and mean IOU for the generated segmentation maps on Dayton. For both metric, higher is better.

Method	Publish	Per-Class Acc.	mIOU
X-Fork [19]	CVPR 2018	0.6262	0.4163
X-Seq [19]	CVPR 2018	0.4783	0.3187
SelectionGAN	Ours	0.6415	0.5455



Fig. 8: Results of cross-view image translation on Dayton. From left to right: input image, ground truth, SelectionGAN (Our), X-Seq++, X-Fork++, Pix2pix++, X-Seq, X-Fork, X-SO and Pix2pix.

than I_a . By comparison Baseline A with Baseline C, the semantic-guided generation improves SSIM, PSNR and SD by 8.19, 3.1771 and 0.3205, respectively, which confirms the importance of the conditional semantic guidance information. By using the proposed cycled semantic guidance generation, Baseline D further improves over C, meaning that the proposed semantic guidance cycle structure indeed utilizes the semantic guidance information in a more effective way, confirming our design motivation. Baseline E outperforms D showing the importance of using the uncertainty maps to guide the pixel loss map which contains an inaccurate reconstruction loss due to the wrong semantic guidance produced from the pre-trained models. Baseline F significantly outperforms E with around 4.67 points gain on the SSIM metric, clearly demonstrating the effectiveness of the proposed multi-channel attention selection scheme. We can also observe from Table 1 that, by adding the proposed multi-scale spatial pool scheme and the TV regularization, the overall performance is further boosted. Finally, we demonstrate the advantage of the proposed two-stage strategy over the one-stage method. Several examples are shown in Fig. 5, 13 and Table 2. It is obvious that the coarse-to-fine generation model is able to generate sharper results and contains more details than the one-stage model, which further confirms our motivations.

Influence of the Number of Attention Channels. We investigate the influence of the number of attention channel N in Eq. (4). Results are shown in Table 3. We observe that the performance tends to be stable after $N=10$. Thus, taking both performance and training speed into consideration, we have set $N=10$ in all our experiments.

SelectionGAN vs. SelectionGAN++. We also provide comparison results of SelectionGAN and SelectionGAN++ on both SVA and Radboud Faces datasets. SelectionGAN is proposed in our conference paper [8] and SelectionGAN++ is proposed in this paper. Results of cross-view image translation are shown in Table 4 and Fig. 6. Results of facial expression generation are shown in Table 9 and Fig. 11. We can see that SelectionGAN++ achieves better results in both figures and both tables (on most metrics), meaning that the proposed multi-scale pooling & channel selection module indeed enhances the feature representation, confirming our design motivation.

State-of-the-art Comparison. We compare our SelectionGAN with several recently proposed state-of-the-art methods, which are Pix2pix [12], Zhai et al. [40], X-Fork [19], X-Seq [19] and X-SO [37]. Moreover, to study the effectiveness of SelectionGAN, we introduce three strong baselines which use both segmentation maps and RGB images as inputs, including Pix2pix++ [12], X-Fork++ [19], and X-Seq++ [19]. We implement Pix2pix++, X-Fork++ and X-Seq++ using their public source code. The comparison results are shown in Table 4, 5, 6 and 7. We can observe that SelectionGAN

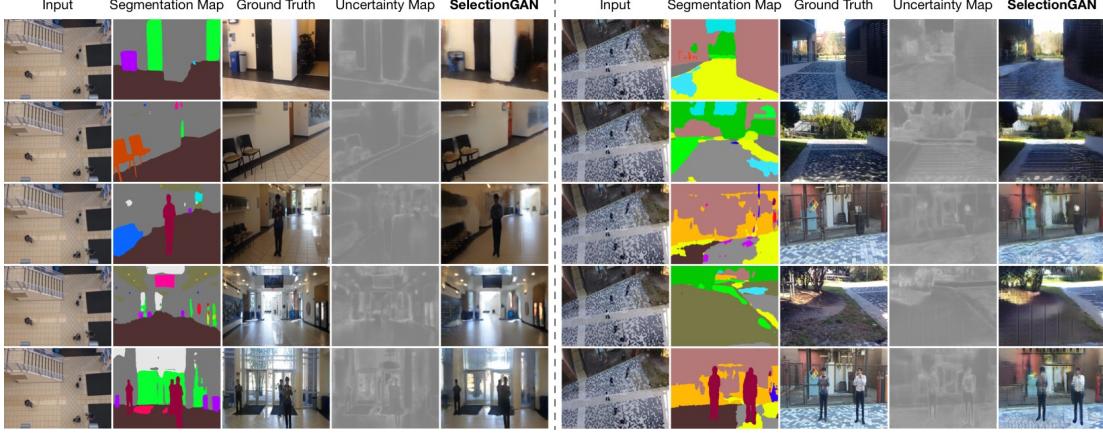


Fig. 10: Results of controllable cross-view image translation for both indoor and outdoor scenes. From left to right: input image, segmentation map, ground truth, uncertainty maps generated by SelectionGAN and SelectionGAN (Ours).

TABLE 9: Quantitative results of facial expression generation on Radboud Faces. For all metrics except LPIPS, higher is better.

Model	Publish	AMT \uparrow	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow
StarGAN [14]	CVPR 2018	24.7	0.8345	19.6451	N/A
Pix2pix [12]	CVPR 2017	13.4	0.8217	19.9971	0.1334
GPGAN [46]	ICPR 2018	0.3	0.8185	18.7211	0.2531
PG2 [1]	NeurIPS 2017	28.4	0.8462	20.1462	0.1130
C2GAN [34]	ACM MM 2019	34.2	0.8618	21.9192	0.0934
SelectionGAN	Ours	37.5	0.8760	27.5671	0.0917
SelectionGAN++	Ours	39.1	0.8761	27.5158	0.0905

consistently outperforms existing methods on most metrics.

Qualitative Evaluation. Qualitative results are shown in Fig. 6, 7, 8 and 9. It can be seen that our method generates more clear details on objects/scenes such as road, tress, clouds, car than the other comparison methods in the generated ground level images. For the generated aerial images in Fig. 8, we can observe that grass, trees and house roofs are well rendered compared to others. Moreover, the results generated by our method are closer to the ground truth in layout and structure.

Visualization of Learned Uncertainty Maps. In Fig. 5, 9 and 10, we show some samples of the generated uncertainty maps. We can see that the generated uncertainty maps learn the layout and structure of the target images. Note that most textured regions are similar in our generation images, while the junction/edge of different regions is uncertain, and thus the model learns to highlight these parts.

Generated Semantic Guidances. Since the proposed SelectionGAN can reconstruct the semantic guidance (here, the segmentation maps), we also compare the generated semantic guidance with X-Fork [19] and X-Seq [19] on Dayton. Following [19], we compute the per-class accuracy and mean IOU for the most common classes in this dataset (see Table 8). We see that our SelectionGAN achieves better results than X-Fork [19] and X-Seq [19] on both metrics.

Controllable Cross-View Image Translation. We further adopt Ego2Top to conduct the controllable cross-view image translation experiments. The quantitative and qualitative results are shown in Table 7 and Fig. 10, respectively. As shown in Fig. 10, given a single input image and some novel segmentation maps, SelectionGAN is able to generate the same scene but with different viewpoints in both indoor and outdoor environments. Moreover, we observe that the proposed SelectionGAN achieves significantly better results



Fig. 11: Results of facial expression generation on Radboud Faces. From left to right: input image, facial landmark, ground truth, StarGAN, Pix2pix, GPGAN, PG2, C2GAN, SelectionGAN (ours), SelectionGAN++ (ours) and uncertainty map generated by SelectionGAN.

than existing methods in Table 7 and Fig. 9 on this challenging task.

4.2 Results on Facial Expression Generation

Datasets. We follow C2GAN [34] and conduct facial expression generation experiments on the Radboud Faces dataset [49]. This dataset contains over 8,000 face images with eight different emotional expressions. We follow C2GAN and all the images are resized to 256×256 without any pre-processing. Then, we adopt OpenFace [5] to extract facial landmarks as the ground truths. Consequently, we collect 5,628 training image pairs and 1,407 testing pairs.

Parameter Settings. Following C2GAN [34], the experiments on Radboud Faces are trained for 200 epochs with batch size of 4.

Evaluation Metrics. Following C2GAN [34], we first employ Structural Similarity (SSIM) [45] and Peak Signal-to-Noise Ratio (PSNR) to evaluate the quantitative quality of generated images by different methods. Moreover, we adopt Amazon Mechanical Turk (AMT) perceptual studies to evaluate the quality of the generated images. Specifically, participants were shown a sequence of pairs of images, one a real image and one fake image, and asked to click on the

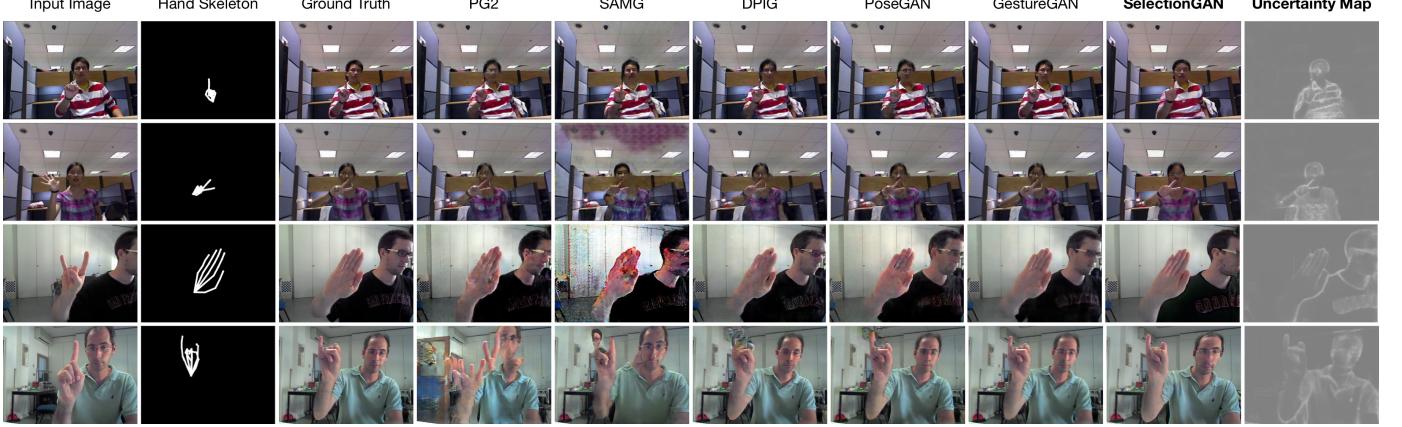


Fig. 12: Results of hand gesture-to-gesture translation on NTU Hand Digit (top) and the Senz3D (bottom) datasets. From left to right: input image, hand skeleton, ground truth, PG2, SAMG, DPIG, PoseGAN, GestureGAN, SelectionGAN (ours) and uncertainty map generated by SelectionGAN.

TABLE 10: Quantitative results of hand gesture-to-gesture translation on NTU Hand Digit and Senz3D datasets. For all metrics except FID and FRD, higher is better.

Method	Publish	NTU Hand Digit					Senz3D				
		PSNR \uparrow	IS \uparrow	AMT \uparrow	FID \downarrow	FRD \downarrow	PSNR \uparrow	IS \uparrow	AMT \uparrow	FID \downarrow	FRD \downarrow
PG2 [1]	NeurIPS 2017	28.2403	2.4152	3.5	24.2093	2.6319	26.5138	3.3699	2.8	31.7333	3.0933
SAMG [47]	ACM MM 2017	28.0185	2.4919	2.6	31.2841	2.7453	26.9545	3.3285	2.3	38.1758	3.1006
DPIG [48]	CVPR 2018	30.6487	2.4547	7.1	6.7661	2.6184	26.9451	3.3874	6.9	26.2713	3.0846
PoseGAN [27]	CVPR 2018	29.5471	2.4017	9.3	9.6725	2.5846	27.3014	3.2147	8.6	24.6712	3.0467
GestureGAN [2]	ACM MM 2018	32.6091	2.5532	26.1	7.5860	2.5223	27.9749	3.4107	22.6	18.4595	2.9836
SelectionGAN	Ours	30.6465	2.4472	15.8	16.2159	2.1560	30.4036	2.4595	14.1	30.9775	2.7014

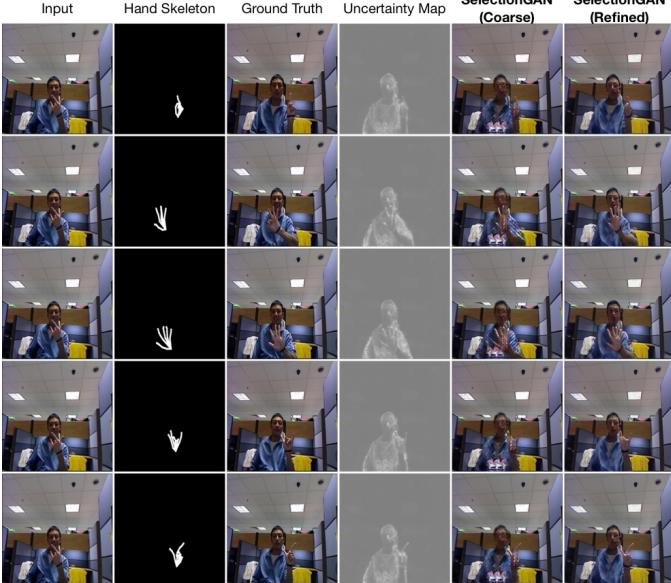


Fig. 13: Results of controllable hand gesture-to-gesture translation. From left to right: input image, hand skeleton, ground truth, uncertainty map , coarse result and refined result.

image they thought was real. Finally, we also use a neural network based metric LPIPS [50] to evaluate the proposed method.

State-of-the-Art Comparison. We compare the proposed SelectionGAN with several state-of-the-art methods, i.e., StarGAN [14], Pix2pix [12], GPGAN [51], PG2 [1] and C2GAN [34]. Quantitative results of the SSIM, PSNR, LPIPS and AMT metrics are shown in Table 9. We can see that

the proposed SelectionGAN achieves the best results on all metrics.

Qualitative Evaluation. Qualitative results are shown in Fig. 11. Clearly, the image generated by our SelectionGAN are more sharper and contains more image details compared with other leading methods.

Visualization of Learned Uncertainty Maps. We also show the learned uncertainty maps in Fig. 11. We observe that the proposed SelectionGAN can generate different uncertainty maps according to different facial expressions, which means the proposed model can learn the difference between different expression domains.

4.3 Results on Hand Gesture Translation

Datasets. We follow GestureGAN [2] and conduct experiments on both NTU Hand Digit [54] and Senz3D [55] datasets. NTU Hand Digit dataset contains 75,036 and 9,600 image pairs for training and testing sets, each of which is comprised of two images of the same person but different gestures. For Senz3D, which contains 135,504 pairs and 12,800 pairs for training and testing.

Parameter Settings. Images on both datasets are resized to 256×256 , and we enabled image flipping and random crops for data augmentation. Following GestureGAN [2], the experiments on both datasets are trained for 20 epochs with batch size of 4.

Evaluation Metrics. Following [2], we employ Peak Signal-to-Noise Ratio (PSNR), Inception score (IS) [43], Fréchet Inception Distance (FID) [44] and Fréchet ResNet Distance (FRD) [2] to evaluate the generated images. Moreover, we follow the same settings as in [2], [12] to conduct the Amazon Mechanical Turk (AMT) perceptual studies.

TABLE 11: Quantitative results of person image generation on Market-1501 and DeepFashion. For all metrics, higher is better. (*) denotes the results tested on our test set.

Method	Publish	Market-1501				DeepFashion	
		SSIM \uparrow	IS \uparrow	Mask-SSIM \uparrow	Mask-IS \uparrow	SSIM \uparrow	IS \uparrow
PG2 [1]	NeurIPS 2017	0.253	3.460	0.792	3.435	0.762	3.090
DPIG [48]	CVPR 2018	0.099	3.483	0.614	3.491	0.614	3.228
PoseGAN [27]	CVPR 2018	0.290	3.185	0.805	3.502	0.756	3.439
C2GAN [34]	ACM MM 2019	0.282	3.349	0.811	3.510	N/A	N/A
BTF [4]	ICCV 2019	N/A	N/A	N/A	N/A	0.767	3.220
PG2* [1]	NeurIPS 2017	0.261	3.495	0.782	3.367	0.773	3.163
PoseGAN* [27]	CVPR 2018	0.291	3.230	0.807	3.502	0.760	3.362
VUNet* [52]	CVPR 2018	0.266	2.965	0.793	3.549	0.763	3.440
Pose-Transfer* [53]	CVPR 2019	0.311	3.323	0.811	3.773	0.773	3.209
SelectionGAN	Ours	0.331	3.449	0.816	3.376	0.776	3.341
Real Data	-	1.000	3.890	1.000	3.706	1.000	4.053

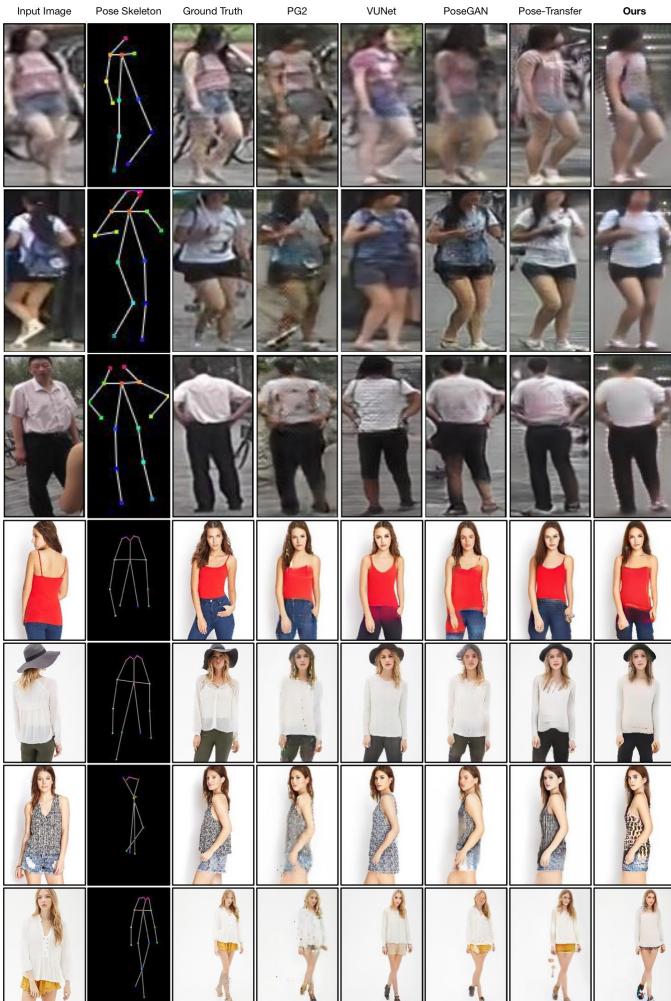


Fig. 14: Results of person image generation on Market-1501 (top) and DeepFashion (bottom). From left to right: input image, pose skeleton, ground truth, PG2, VUNet, PoseGAN, Pose-Transfer and SelectionGAN (Ours).

State-of-the-Art Comparison. We compare the proposed SelectionGAN with the leading hand gesture translation methods, i.e., PG2 [1], SAMG [47], DPIG [48], PoseGAN [27] and GestureGAN [2]. Comparison results are shown in Table 10. We can see that our SelectionGAN achieves competitive results on both datasets compared with existing methods except GestureGAN. GestureGAN is a model carefully designed for this task, thus it obtain slightly better results than

TABLE 12: User study of person image generation (%). R2G means the percentage of real images rated as generated w.r.t. all real images. G2R means the percentage of generated images rated as real w.r.t. all generated images. The results of other methods are drawn from their papers.

Method	Publish	Market-1501		DeepFashion	
		R2G	G2R	R2G	G2R
PG2 [1]	NeurIPS 2017	11.2	5.5	9.2	14.9
PoseGAN [27]	CVPR 2018	22.67	50.24	12.42	24.61
C2GAN [34]	ACM MM 2019	23.20	46.70	N/A	N/A
Pose-Transfer [53]	CVPR 2019	32.23	63.47	19.14	31.78
SelectionGAN	Ours	34.64	64.75	20.57	33.54

TABLE 13: Quantitative results of semantic image synthesis on Cityscapes and ADE20K. For mIoU and Acc, higher is better. For FID, lower is better.

Method	Publish	Cityscapes			ADE20K		
		mIoU \uparrow	Acc \uparrow	FID \downarrow	mIoU \uparrow	Acc \uparrow	FID \downarrow
CRN [56]	ICCV 2017	52.4	77.1	104.7	22.4	68.8	73.3
SIMS [57]	CVPR 2018	47.2	75.5	49.7	N/A	N/A	N/A
Pix2pixHD [13]	CVPR 2018	58.3	81.4	95.0	20.3	69.2	81.8
GauGAN [58]	CVPR 2019	62.3	81.9	71.8	38.5	79.9	33.9
SelectionGAN	Ours	63.8	82.4	65.2	40.1	81.2	33.1

TABLE 14: User preference study of semantic image synthesis on Cityscapes and ADE20K. The numbers indicate the percentage of users who favor the results of the proposed method over the competing method. For this metric, higher is better.

AMT \uparrow	Published	Cityscapes	ADE20K
Ours vs. CRN [56]	ICCV 2017	63.86	69.43
Ours vs. Pix2pixHD [13]	CVPR 2018	54.04	78.62
Ours vs. SIMS [57]	CVPR 2018	53.57	N/A
Ours vs. GauGAN [58]	CVPR 2019	52.89	55.15

ours. We also provide results of user study in Table 10. Note that the proposed SelectionGAN achieves the second best results compared with other strong baselines.

Qualitative Evaluation. Qualitative results compared with existing methods are shown in Fig. 12. We can see that the proposed SelectionGAN achieves competitive results compared with the leading approaches. Moreover, we show the learned uncertainty maps in Fig. 12 and 13.

Controllable Hand Gesture Translation. In Fig. 13, we provide results of controllable hand gesture translation. We can see that the proposed SelectionGAN can translates a single input image into several output images while each one respecting the constraints specified in the provided hand skeleton.

4.4 Results on Person Image Generation

Datasets. We follow Pose-Transfer [53] and conduct person image generation experiments on both Market-1501 [59]



Fig. 15: Results of semantic image synthesis on Cityscapes. From left to right: input semantic label, ground truth, CRN, SIMS, Pix2pixHD, GauGAN and SelectionGAN (ours).

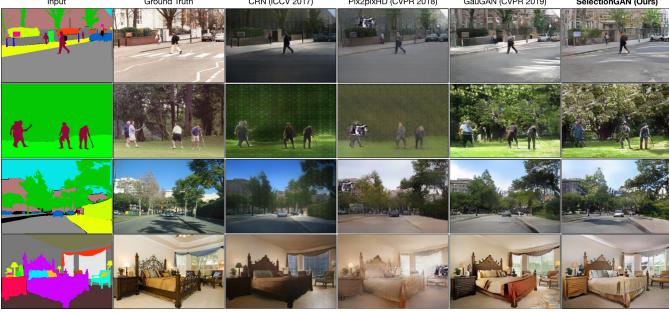


Fig. 16: Results of semantic image synthesis task on ADE20K. From left to right: input semantic label, ground truth, CRN, Pix2pixHD, GauGAN and SelectionGAN (ours).

and DeepFashion [60] datasets. Following [53], we collect 263,632 and 12,000 pairs for training and testing on Market-1501. For DeepFashion, 101,966 and 8,570 pairs are randomly selected for training and testing.

Parameter Settings. Following Pose-Transfer [53], images are rescaled to 128×64 and 256×256 on Market-1501 and DeepFashion datasets, respectively. Moreover, the experiments on both datasets are trained for around 90k iteration with batch size of 32 and 12 on Market-1501 and DeepFashion, respectively.

Evaluation Metrics. Following previous works [1], [27], [27], [34], we adopt Structure Similarity (SSIM) [45], Inception score (IS) [43] and their corresponding masked versions, i.e., Mask-SSIM and Mask-IS, as our evaluation metrics. Moreover, we follow Pose-Transfer [53] and recruit 30 volunteers to conduct a user study.

State-of-the-Art Comparison. We compare the proposed SelectionGAN with several leading person image generation methods, i.e., PG2 [1], DPG [48], PoseGAN [27], VUNet [52], C2GAN [34], BTF [4] and Pose-Transfer [53]. Quantitative results of the SSIM, IS, Mask-SSIM and Mask-IS metrics are show in Table 11. We can see that the proposed SelectionGAN achieves competitive performance compared with the carefully designed methods on this task such as Pose-Transfer [53] and PoseGAN [27]. Moreover, we show user study results in Table 12. We observe that our method achieve better results over [1], [27], [34], [53], further validating that our generated images are more photo-realistic.

Qualitative Evaluation. Qualitative results are shown in Fig. 14. The image generated by our SelectionGAN are more realistic and sharp compared with other leading methods. Moreover, the person layouts of generated images by our method are closer to the target skeletons.



Fig. 17: Generated segmentation maps on Cityscapes. From left to right: ground truth, result generated by SelectionGAN (ours), segmentation map generated on SelectionGAN’s result, result generated by GauGAN and segmentation map generated on GauGAN’s result.



Fig. 18: Generated segmentation maps on ADE20K. From left to right: ground truth, result generated by SelectionGAN (ours), segmentation map generated on SelectionGAN’s result, result generated by GauGAN and segmentation map generated on GauGAN’s result.

4.5 Results on Semantic Image Synthesis

To explore the generality of the proposed SelectionGAN on other generation tasks, we also conduct experiments on the challenging semantic image synthesis task.

Datasets. We follow GauGAN [58] and conduct semantic image synthesis experiments on two challenging datasets, i.e., Cityscapes [61] and ADE20K [7]. The training and testing set sizes of Cityscapes are 2,975 and 500, respectively. For ADE20K, which contains 150 semantic classes, and has 20,210 training and 2,000 validation images.

Parameter Settings. Images are rescaled to 512×256 and 256×256 on Cityscapes and ADE20K datasets, respectively. Following GauGAN [58], the experiments on both datasets are trained for 200 epochs with batch size of 32.

Evaluation Metrics. Following [58], we employ the mean Intersection-over-Union (mIoU) and pixel accuracy (Acc) to measure the segmentation accuracy. Specifically, we adopt the state-of-the-art segmentation networks to evaluate the generated images, i.e., DRN-D-105 [62] for Cityscapes and UperNet101 [63] for ADE20K. We also employ the Fréchet Inception Distance (FID) [44] to measure the distance between the distribution of generated samples and the distribution of real samples. Finally, we follow GauGAN and employ Amazon Mechanical Turk (AMT) to measure the perceived visual fidelity of the generated images.

State-of-the-Art Comparisons. We adopt several leading

semantic image synthesis methods as our baselines, i.e., Pix2pixHD [13], CRN [56], SIMS [57] and GauGAN [58]. Results of mIoU, Acc and FID are show in Table 13. We note that the proposed SelectionGAN achieves better results than the existing competing methods on both mIoU and Acc metrics. For FID, the proposed SelectionGAN is only worse than SIMS on Cityscapes. However, SIMS has poor segmentation results. Moreover, we follow GauGAN and provide AMT results in Table 14. We see that users favor our translated images on both datasets compared with existing leading methods.

Qualitative Evaluation. Qualitative results compared with exiting methods are shown in Fig. 15 and 16. We observe that the proposed SelectionGAN produces much better results with fewer visual artifacts than exiting methods.

Visualization of Generated Segmentation Maps. We follow GauGAN and apply pre-trained segmentation networks on the generated images to produce segmentation maps. The intuition behind this is that if the generated images are realistic, a well-trained semantic segmentation model should be able to predict the ground truth label. Results compared with the state-of-the-art methods, i.e., GauGAN, are shown in Fig. 17 and 18. We observe that the proposed SelectionGAN generates better semantic maps than GauGAN on both datasets.

5 CONCLUSION

We propose the Multi-Channel Attention Selection GAN (SelectionGAN) to address a novel image synthesizing task by conditioning on a input image and several conditional semantic guidances. In particular, we adopt a cascade strategy to divide the generation procedure into two stages. Stage I aims to capture the semantic structure of the target image and Stage II focus on more appearance details via the proposed multi-scale spatial pooling & channel selection and the multi-channel attention selection modules. We also propose an uncertainty map guided pixel loss to solve the inaccurate semantic guidance issue for better optimization. Extensive experimental results on four guided image-to-image translation and one semantic image synthesis tasks with 11 public datasets demonstrate that our method obtains much better results than the state-of-the-art approaches.

REFERENCES

- [1] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *NeurIPS*, 2017. [1](#), [3](#), [4](#), [7](#), [10](#), [11](#), [12](#), [13](#)
- [2] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Gesturegan for hand gesture-to-gesture translation in the wild," in *ACM MM*, 2018. [1](#), [3](#), [7](#), [11](#)
- [3] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," in *NeurIPS*, 2019. [1](#), [3](#)
- [4] B. AlBahar and J.-B. Huang, "Guided image-to-image translation with bi-directional feature transformation," in *ICCV*, 2019. [1](#), [3](#), [4](#), [12](#), [13](#)
- [5] B. Amos, B. Ludwiczuk, M. Satyanarayanan *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, 2016. [1](#), [7](#), [10](#)
- [6] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, 2017. [1](#), [6](#)
- [7] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017. [1](#), [6](#), [13](#)
- [8] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in *CVPR*, 2019. [2](#), [3](#), [7](#), [8](#), [9](#)
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014. [2](#), [7](#)
- [10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019. [2](#)
- [11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint:1411.1784*, 2014. [3](#)
- [12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017. [3](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)
- [13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018. [3](#), [12](#), [13](#)
- [14] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018. [3](#), [10](#), [11](#)
- [15] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," in *ACCV*, 2018. [3](#)
- [16] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *NeurIPS*, 2019. [3](#)
- [17] X. Yu, Y. Chen, S. Liu, T. Li, and G. Li, "Multi-mapping image-to-image translation via learning disentanglement," in *NeurIPS*, 2019. [3](#)
- [18] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis," in *NeurIPS*, 2018. [3](#)
- [19] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *CVPR*, 2018. [3](#), [6](#), [7](#), [8](#), [9](#), [10](#)
- [20] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," *arXiv preprint arXiv:1912.12215*, 2019. [3](#)
- [21] M. Wang, G.-Y. Yang, R. Li, R.-Z. Liang, S.-H. Zhang, P. Hall, S.-M. Hu *et al.*, "Example-guided style consistent image synthesis from semantic labeling," in *CVPR*, 2019. [3](#)
- [22] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018. [3](#)
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. [3](#)
- [24] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, "Attention-gan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *arXiv preprint arXiv:1911.11897*, 2019. [3](#)
- [25] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *ICLR*, 2020. [3](#)
- [26] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *ICML*, 2019. [3](#)
- [27] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018. [4](#), [11](#), [12](#), [13](#)
- [28] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *ECCV*, 2014. [4](#)
- [29] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *CVPR*, 2016. [4](#)
- [30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017. [4](#)
- [31] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018. [5](#)
- [32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016. [6](#)
- [33] H. Tang, D. Xu, H. Liu, and N. Sebe, "Asymmetric generative adversarial networks for image-to-image translation," *arXiv preprint arXiv:1912.06931*, 2019. [6](#)
- [34] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan, "Cycle in cycle generative adversarial networks for keypoint-guided image generation," in *ACM MM*, 2019. [7](#), [10](#), [11](#), [12](#), [13](#)
- [35] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017. [7](#)

- [36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. [7](#)
- [37] K. Regmi and A. Borji, "Cross-view image synthesis using geometry-guided conditional gans," *Elsevier CVIU*, vol. 187, p. 102788, 2019. [7, 8, 9](#)
- [38] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *ECCV*, 2016. [7](#)
- [39] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocation with aerial reference imagery," in *ICCV*, 2015. [7](#)
- [40] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *CVPR*, 2017. [7, 8, 9](#)
- [41] A. Palazzi, G. Borghi, D. Abati, S. Calderara, and R. Cucchiara, "Learning to map vehicles into birds eye view," in *ICIP*, 2017. [7](#)
- [42] S. Ardesir and A. Borji, "Ego2top: Matching viewers in egocentric and top-view videos," in *ECCV*, 2016. [7](#)
- [43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016. [7, 11, 13](#)
- [44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017. [7, 11, 13](#)
- [45] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004. [7, 10, 13](#)
- [46] X. Di, V. A. Sindagi, and V. M. Patel, "Gp-gan: Gender preserving gan for synthesizing faces from landmarks," in *ICPR*, 2018. [10](#)
- [47] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *ACM MM*, 2017. [11](#)
- [48] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *CVPR*, 2018. [11, 12, 13](#)
- [49] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Taylor & Francis Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010. [10](#)
- [50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. [11](#)
- [51] X. Di, V. A. Sindagi, and V. M. Patel, "Gp-gan: Gender preserving gan for synthesizing faces from landmarks," in *ICPR*, 2018. [11](#)
- [52] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018. [12, 13](#)
- [53] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *CVPR*, 2019. [12, 13](#)
- [54] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE TMM*, vol. 15, no. 5, pp. 1110–1120, 2013. [11](#)
- [55] A. Memo and P. Zanuttigh, "Head-mounted gesture controlled interface for human-computer interaction," *Springer MTA*, vol. 77, no. 1, pp. 27–53, 2018. [11](#)
- [56] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *ICCV*, 2017. [12, 13](#)
- [57] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," in *CVPR*, 2018. [12, 13](#)
- [58] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019. [12, 13](#)
- [59] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *CVPR*, 2015. [12](#)
- [60] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016. [12](#)
- [61] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016. [13](#)
- [62] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *CVPR*, 2017. [13](#)
- [63] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *ECCV*, 2018. [13](#)



Hao Tang is a Ph.D. candidate in the Department of Information Engineering and Computer Science, and a member of Multimedia and Human Understanding Group (MHUG) led by Prof. Nicu Sebe at the University of Trento. He received the Master degree in computer application technology in 2016 at the School of Electronics and Computer Engineering, Peking University, China. His research interests are machine learning, (deep) representation learning and their applications to computer vision.



Dan Xu is a Postdoc researcher in Visual Geometric Group at the University of Oxford. He received the Ph.D. Computer Science at the University of Trento. He was a research assistant in the Multimedia Laboratory in the Department of Electronic Engineering at the Chinese University of Hong Kong. His research focuses on computer vision, multimedia and machine learning. He received the Intel best scientific paper award at ICPR 2016.



Paper Award in ACM Multimedia 2015.

Yan Yan is currently an assistant professor in computer science at the Texas State University. He received the Ph.D. degree from University of Trento, Italy, in 2014. He was Research Fellow at the University of Michigan (2016-2017) and University of Trento (2014-2016). He was a Visiting Scholar with Carnegie Mellon University in 2013 and a Visiting Research Fellow with Advanced Digital Sciences Center (ADSC), UIUC, Singapore in 2015. Dr. Yan is the recipient of Best Student Paper Award in ICPR 2014 and Best



in video understanding such as video segmentation, activity recognition, and video-to-text.

Jason J. Corso is currently a Professor of Electrical Engineering and Computer Science at the University of Michigan. He received his Ph.D. in Computer Science at The Johns Hopkins University in 2005. He is a recipient of the NSF CAREER award (2009), ARO Young Investigator award (2010), Google Faculty Research Award (2015) and on the DARPA CSSG. His main research thrust is high-level computer vision and its relationship to human language, robotics and data science. He primarily focuses on problems



Philip H. S. Torr received the PhD degree from Oxford University. After working for another three years at Oxford, he worked for six years for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He is now a professor at Oxford University. He has won awards from top vision conferences, including ICCV, CVPR, ECCV, NIPS and BMVC. He is a senior member of the IEEE and a Royal Society Wolfson Research Merit Award holder.



Nicu Sebe is Professor with the University of Trento, Italy, leading the research in the areas of multimedia information retrieval and human behavior understanding. He was the General CoChair of the IEEE FG Conference 2008 and ACM Multimedia 2013, and the Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010, ACM Multimedia 2007 and 2011. He is the Program Chair of ICCV 2017 and ECCV 2016, and a General Chair of ACM ICMR 2017 and ICPR 2020. He is a fellow of the International Association for Pattern Recognition.