

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

Hao Tang, 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 aiming at synthesizing new images from an input image and several external semantic guidance, as shown in Fig. 1. This task received 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, etc., most existing methods 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 a novel 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 propose a bi-directional feature transformation to

better utilize the constraints of the semantic guidance. Although this approach performs 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 [1], [8], and cross-view image translation [9], [10]. 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 we propose a novel Multi-Channel Attention Selection Generative Adversarial Network (SelectionGAN), which contains two generation stages. The overall framework of 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 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

- Hao Tang is with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland. E-mail: [hao.tang@vision.ee.ethz.ch](mailto:hao.tang@vision.ee.ethz.ch)
- Philip H.S. Torr is with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, United Kingdom.
- Nicu Sebe is with the Department of Information Engineering and Computer Science (DISI), University of Trento, Trento 38123, Italy.

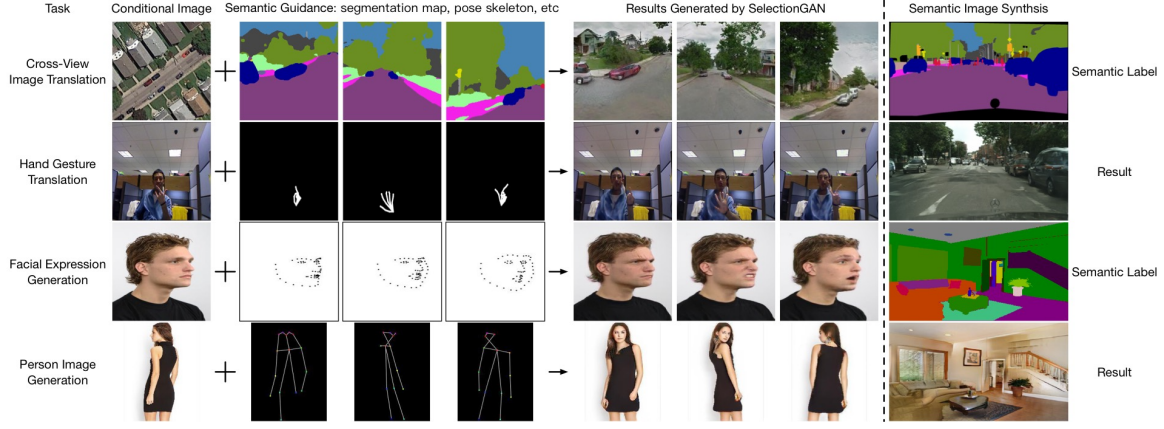


Fig. 1: SelectionGAN’s capabilities: (left) Guided image-to-image translation (including cross-view image translation, hand gesture translation, facial expression generation, and person image generation): synthesizing images from a single input image as well as semantic guidance (e.g., segmentation map,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

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 (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 effective-

ness 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 [9]. The current paper extends [9] 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 channel selection, to automatically enhance the multi-scale feature representation in the feature channel dimension. Equipped with this new module, our SelectionGAN proposed in [9] 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)** [11] have shown the capability of generating high-quality images [12]. A vanilla GAN model [11] 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  $D$  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) [13] have been proposed. One specific application of CGANs is image-to-image translation [14].

**Image-to-Image Translation** frameworks learn a parametric mapping between inputs and outputs. For example, Isola et al. [14] propose Pix2pix, which is a supervised model and uses a CGAN to learn a translation function from input to



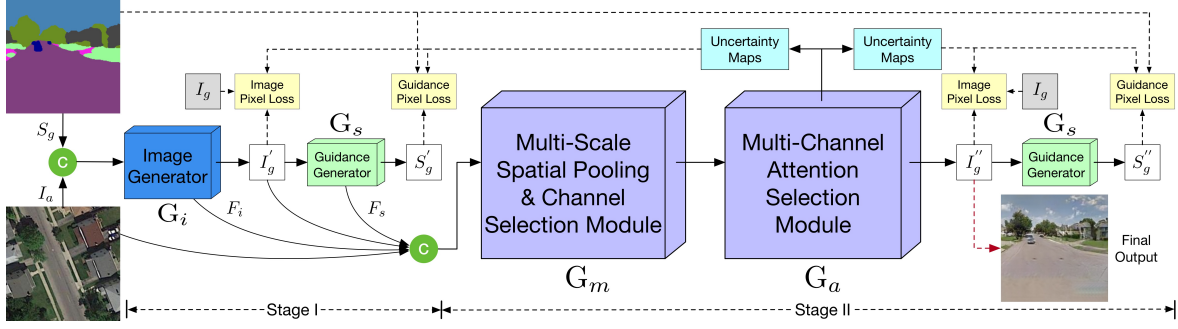


Fig. 2: Overview of the proposed SelectionGAN. Stage I presents a cycled semantic-guided generation sub-network which accepts both the input image  $I_a$  and the conditional semantic guidance  $S_g$ , and simultaneously synthesizes the target image  $I'_g$  and reconstructs the semantic guidance  $S'_g$ . Stage II takes the input image  $I_a$ , the coarse prediction  $I'_g$ , and the learned deep features ( $F_i$  and  $F_s$ ) 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.  $\odot$  denotes channel-wise concatenation.

output image domains. Based on Pix2pix, Wang et al. [15] 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 GAN-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 [16], [17], [18], text descriptions [19], [20], [21], human keypoint/skeleton [1], [2], [8], [22], [23], [24], segmentation maps [3], [9], [10], [25], [26], [27], [28], [29], [30], [31], [32], [33], and reference images [4], [34]. 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 (see Fig. 1).

**Attention Learning in Image-to-Image Translation.** Attention learning has been extensively exploited in computer

vision and natural language processing, e.g., [35], [36], [37], [38], [39], [40]. To improve the image generation performance, the attention mechanism has also been recently investigated in GANs such as [41], [42], [43], [44]. For example, Zhang et al. propose SAGAN [43], which introduces a self-attention mechanism into convolutional GANs to help with modeling long range, multi-level dependencies across image regions.

Unlike existing attention methods, we aim at a more effective network design and propose a novel SelectionGAN, 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 tasks.

### 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 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 addi-

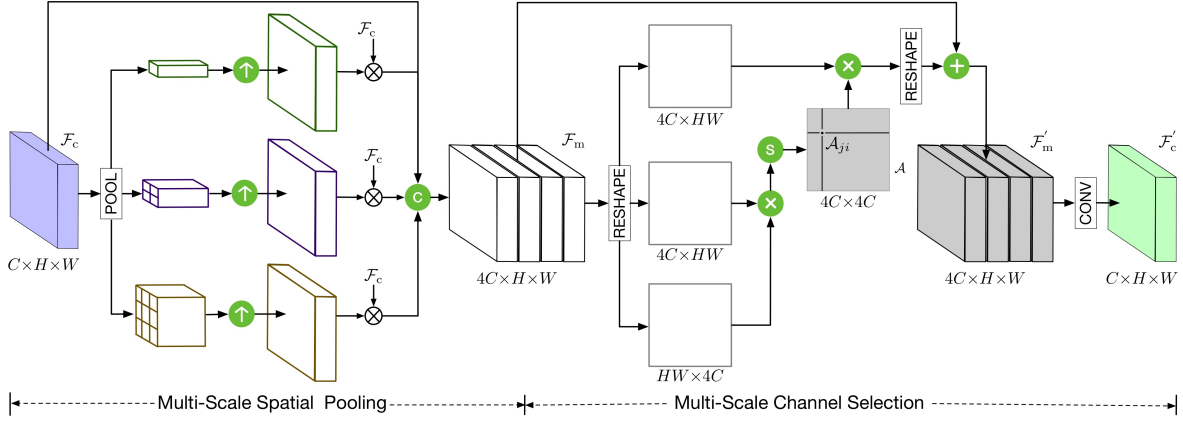


Fig. 3: Overview of the proposed multi-scale spatial pooling & channel selection module. The multi-scale spatial pooling pools features from different receptive fields in order to have a better generation of image details. The multi-scale channel selection aims at automatically emphasizing interdependent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

tional semantic guidance into the image-to-image translation model [4], [8] 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], [45] 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 [1], [8], [24], 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 [46] and semantic segmentation [47], 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 and the target domains, 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  $W$

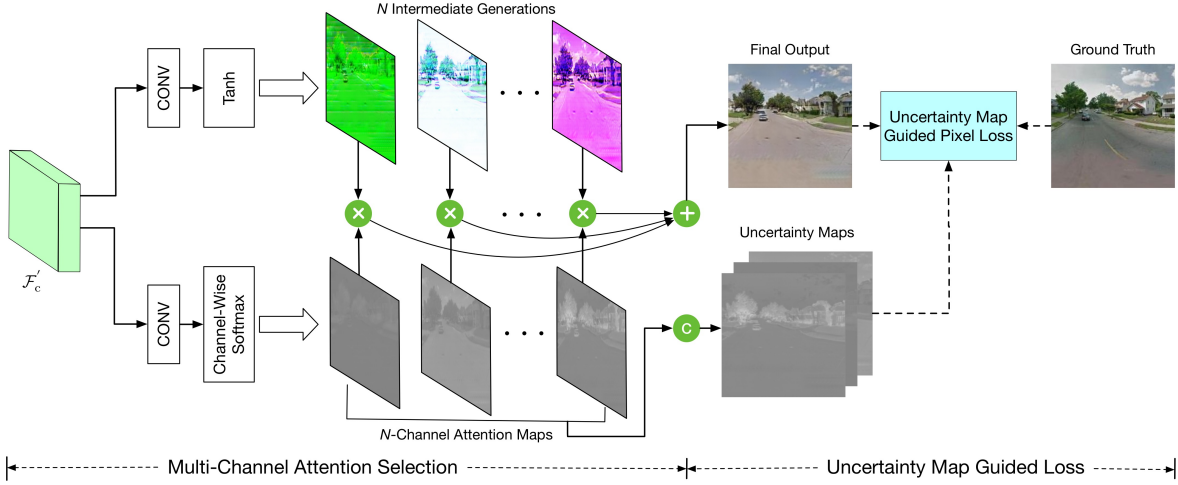


Fig. 4: Overview of the proposed multi-channel attention selection module. This module aims to automatically select from a set of intermediate diverse generations in a larger generation space to improve the generation quality. Meanwhile, the module also effectively learns uncertainty maps to guide the 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.

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 [48] 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:

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

which produces new multi-scale features  $\mathcal{F}_m \in \mathbb{R}^{4C \times H \times W}$  (we set  $M=3$  in our experiments.) 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.

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  $i^{\text{th}}$  channel’s impact on the  $j^{\text{th}}$  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  $\alpha$  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  $\mathcal{F}'_m$  is fed into a convolutional layer to obtain  $\mathcal{F}'_c \in \mathbb{R}^{C \times H \times W}$ , 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 [49] 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 loss. Let  $\mathcal{L}_p^i$  denote a pixel-level loss map and  $U_i$  denote the  $i$ -th uncertainty map, we have:

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

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 [14] 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:

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

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:

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

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 four images/guidances (i.e.,  $I_g'$ ,  $S_g'$ ,  $I_g''$ , and  $S_g''$ ) and the corresponding real images/guidances.  $\mathcal{L}_{tv}$  is the total variation regularization [50] on the final synthesized image  $I_g''$ .  $\lambda_i$  and  $\lambda_{tv}$  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 [14] 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  $G_s$ . 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  $3 \times 3$  and  $1 \times 1$ , respectively. We adopt PatchGAN [14] 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 [10] and use RefineNet [6] and [7] to generate segmentation maps on Dayton, SVA, and Ego2Top datasets as training data, respectively. For facial expression generation, we follow [51] 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 [52] 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 [11] to optimize the proposed SelectionGAN, i.e., one gradient descent step

TABLE 1: Ablations study of the proposed SelectionGAN.

Baseline	Setups of SelectionGAN	SSIM $\uparrow$	PSNR $\uparrow$	SD $\uparrow$	FID $\downarrow$	Inception Score $\uparrow$		
						All	Top-1	Top-5
A	$I_a \xrightarrow{G_i} I'_g$	0.4555	19.6574	18.8870	91.47	3.2359	2.1903	3.3110
B	$S_g \xrightarrow{G_i} I'_g$	0.5223	22.4961	19.2648	87.51	3.4849	2.2544	3.4217
C	$[I_a, S_g] \xrightarrow{G_i} I'_g$	0.5374	22.8345	19.2075	84.10	3.4478	2.2616	3.4668
D	$[I_a, S_g] \xrightarrow{G_i} I'_g \xrightarrow{G_s} S'_g$	0.5438	22.9773	19.4568	82.81	3.1655	2.2561	3.2401
E	D + Uncertainty-Guided Pixel Loss	0.5522	23.0317	19.5127	79.84	3.2741	2.2687	3.3063
F	E + Multi-Channel Attention Selection	0.5989	23.7562	20.0000	75.57	3.3365	<b>2.2749</b>	3.4664
G	F + Total Variation Regularization	0.6047	23.7956	20.0830	74.11	3.3172	2.1397	3.3509
H	G + Multi-Scale Spatial Pooling	<b>0.6167</b>	<b>23.9310</b>	<b>20.1214</b>	<b>72.23</b>	<b>3.4978</b>	2.1880	<b>3.4786</b>

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 [53] 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.

## 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 [9], [10], [54] and perform experiments on four public cross-view image translation datasets: 1) The Dayton dataset [55] contains 76,048 images and the train/test split is 55,000/21,048. The images in the original dataset have  $354 \times 354$  resolution. We resize them to  $256 \times 256$ . 2) The CVUSA dataset [56] consists of 35,532/8,884 image pairs in train/test split. Following [10], [57], the aerial images are center-cropped to  $224 \times 224$  and resized to  $256 \times 256$ . For the ground level images and corresponding segmentation maps, we take the first quarter of both and resize them to  $256 \times 256$ . 3) The Surround Vehicle Awareness (SVA) dataset [58] is a synthetic dataset collected from Grand Theft Auto V (GTAV) video game. Following [54], 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. 4) The Ego2Top dataset [59] 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 [9] 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 [10], [14]. All images are scaled to  $256 \times 256$ , and we enabled image flipping and random crops for data augmentation. Similar to [10], the experiments for Dayton are trained for 35 epochs with a batch size of 4. For CVUSA, we follow the same setup as in [10], [57], and train our network for 30 epochs with batch size of 4. For SVA, all models are trained with 20 epoch using batch size 4.

**Evaluation Metrics.** Similar to [9], [10], we employ Inception Score [60], top-k prediction accuracy, KL score, and Fréchet Inception Distance (FID) [61] 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) [62], 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 SelectionGAN. 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 [14] 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  $\mathcal{F}_c$  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. Note that Baseline B is better than A since  $S_g$  contains more structural information than  $I_a$ . When comparing Baselines A and C, the semantic-guided generation improves SSIM, PSNR and SD by 8.19, 3.1771 and 0.3205, respectively, confirming the importance of the conditional semantic guidance information. By using the proposed cycled

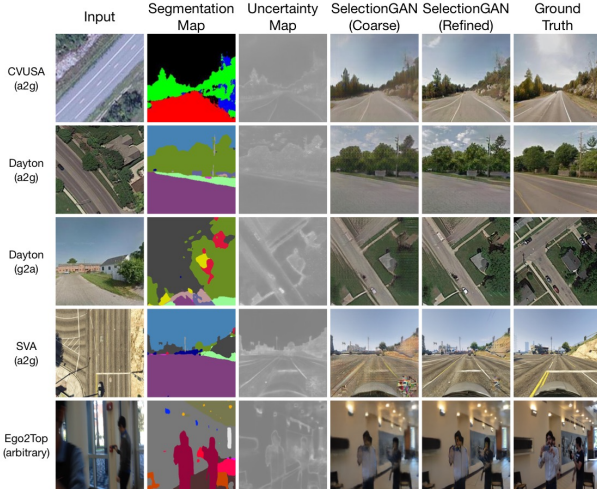


Fig. 5: Results of cross-view image translation generated by the proposed SelectionGAN on different datasets.

TABLE 2: Quantitative results of coarse-to-fine generation.

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

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. The results 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.

**Comparisons with SENet [63].** The proposed multi-scale spatial pooling shares a similar intuition with SENet [63] which amplifies the channels via attention based on pooling. Unlike SENet that employs positive attention via the Sigmoid function, the proposed multi-scale spatial pooling selects each pooled feature via an element-wise multiplication with the original feature. Since in our task the input features are from different sources, highly correlated features would preserve more useful information for the generation. We also conduct experiments to compare the proposed method with SENet on Dayton. Specifically, we use the SE layer proposed in [63] to replace our multi-scale spatial pooling

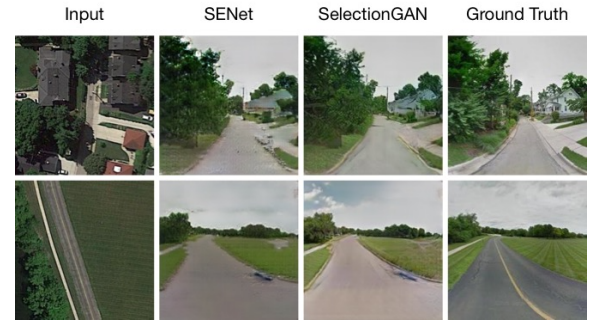


Fig. 6: Comparison results of SENet and the proposed SelectionGAN on Dayton.

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	<b>20.0033</b>
10	<b>0.5986</b>	23.7336	19.9993
32	0.5950	<b>23.8265</b>	19.9086

module, obtaining the following results in terms of SSIM, PSNR, and SD: 0.5912, 23.3857, and 19.8061, respectively. We can see that our method (see the Baseline H in Table 1) still significantly outperforms [63]. Moreover, we provide the visualization results in Fig. 6 (note that our method achieves better results than SENet).

**Influence of the Number of Attention Channels.** We investigate the influence of the number of attention channel  $N$  in Eq. (4). The 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 set  $N=10$  in all our experiments.

**State-of-the-Art Comparisons.** We compare our SelectionGAN with several recently proposed state-of-the-art methods, which are Pix2pix [14], Zhai et al. [57], X-Fork [10], X-Seq [10] and X-SO [54]. Moreover, to study the effectiveness of SelectionGAN, we introduce five strong baselines which use both segmentation map and RGB image as inputs, including Pix2pix++ [14], X-Fork++ [10], X-Seq++ [10], Pix2pixHD [15], and GauGAN [64]. The comparison results are shown in Table 4, 5, and 6. We can observe that SelectionGAN consistently outperforms existing methods on most metrics. Qualitative results compared with the leading baselines are shown in Fig. 7 and 8. We can see that our method generates more clear details on objects/scenes such as road, trees, clouds, car than the other comparison methods. Moreover, the results generated by our method are closer to the ground truths in layout and structure.

**Visualization of Learned Uncertainty Maps.** In Fig. 5 and 9, 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 methods can reconstruct the semantic guidance (here, the segmentation maps), we also compare the generated semantic guidance with X-Fork [10] and X-Seq [10] on Dayton. Fol-



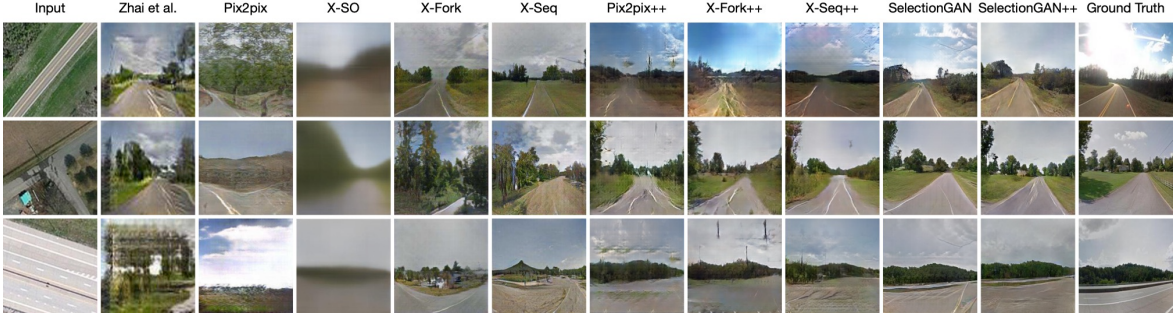


Fig. 7: Results of cross-view image translation on CVUSA.

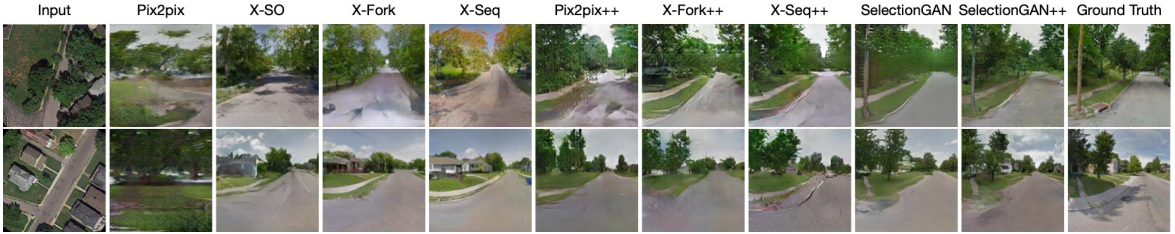


Fig. 8: Results of cross-view image translation on Dayton in a2g direction.

TABLE 4: Quantitative results of cross-view image translation on CVUSA. (\*) 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	Accuracy (%) $\uparrow$				Inception Score* $\uparrow$			SSIM $\uparrow$	PSNR $\uparrow$	SD $\uparrow$	KL $\downarrow$
	Top-1	Top-5	All	Top-1	Top-5	All	Top-1	Top-5			
Zhai et al. [57]	13.97	14.03	42.09	52.29	1.8434	1.5171	1.8666	0.4147	17.4886	16.6184	$27.43 \pm 1.63$
Pix2pix [14]	7.33	9.25	25.81	32.67	3.2771	2.2219	3.4312	0.3923	17.6578	18.5239	$59.81 \pm 2.12$
X-SO [54]	0.29	0.21	6.14	9.08	1.7575	1.4145	1.7791	0.3451	17.6201	16.9919	$414.25 \pm 2.37$
X-Fork [10]	20.58	31.24	50.51	63.66	3.4432	2.5447	3.5567	0.4356	19.0509	18.6706	$11.71 \pm 1.55$
X-Seq [10]	15.98	24.14	42.91	54.41	3.8151	2.6738	4.0077	0.4231	18.8067	18.4378	$15.52 \pm 1.73$
Pix2pix++ [14]	26.45	41.87	57.26	72.87	3.2592	2.4175	3.5078	0.4617	21.5739	18.9044	$9.47 \pm 1.69$
X-Fork++ [10]	31.03	49.65	64.47	81.16	3.3758	2.5375	3.5711	0.4769	21.6504	18.9856	$7.18 \pm 1.56$
X-Seq++ [10]	34.69	54.61	67.12	83.46	3.3919	2.5474	3.4858	0.4740	21.6733	18.9907	$5.19 \pm 1.31$
SelectionGAN	41.52	65.51	74.32	89.66	3.8074	2.7181	3.9197	0.5323	<b>23.1466</b>	19.6100	$2.96 \pm 0.97$
SelectionGAN++	<b>43.27</b>	<b>68.36</b>	<b>77.15</b>	<b>91.74</b>	<b>3.8296</b>	<b>2.8977</b>	<b>4.0238</b>	<b>0.5355</b>	22.8532	<b>19.7672</b>	<b>2.76 <math>\pm</math> 0.96</b>

TABLE 5: Quantitative results of cross-view image translation on Dayton in a2g direction. (\*) 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	Accuracy (%) $\uparrow$				Inception Score* $\uparrow$			SSIM $\uparrow$	PSNR $\uparrow$	SD $\uparrow$	KL $\downarrow$
	Top-1	Top-5	All	Top-1	Top-5	All	Top-1	Top-5			
Pix2pix [14]	6.80	9.15	23.55	27.00	2.8515	1.9342	2.9083	0.4180	17.6291	19.2821	$38.26 \pm 1.88$
X-SO [54]	27.56	41.15	57.96	73.20	2.9459	2.0963	2.9980	0.4772	19.6203	19.2939	$7.20 \pm 1.37$
X-Fork [10]	30.00	48.68	61.57	78.84	3.0720	2.2402	3.0932	0.4963	19.8928	19.4533	$6.00 \pm 1.28$
X-Seq [10]	30.16	49.85	62.59	80.70	2.7384	2.1304	2.7674	0.5031	20.2803	19.5258	$5.93 \pm 1.32$
Pix2pix++ [14]	32.06	54.70	63.19	81.01	3.1709	2.1200	3.2001	0.4871	21.6675	18.8504	$5.49 \pm 1.25$
X-Fork++ [10]	34.67	59.14	66.37	84.70	3.0737	2.1508	3.0893	0.4982	21.7260	18.9402	$4.59 \pm 1.16$
X-Seq++ [10]	31.58	51.67	65.21	82.48	3.1703	2.2185	3.2444	0.4912	21.7659	18.9265	$4.94 \pm 1.18$
SelectionGAN	42.11	68.12	77.74	92.89	3.0613	2.2707	3.1336	<b>0.5938</b>	<b>23.8874</b>	<b>20.0174</b>	$2.74 \pm 0.86$
SelectionGAN++	<b>47.01</b>	<b>73.54</b>	<b>80.19</b>	<b>94.97</b>	<b>3.2315</b>	<b>2.3367</b>	<b>3.3245</b>	0.5786	23.5385	19.8729	<b>2.45 <math>\pm</math> 0.83</b>

lowing [10], we compute the per-class accuracy and mean IOU for the most common classes in this dataset (see Table 7). We see that our SelectionGAN and SelectionGAN++ achieve better results than X-Fork [10] and X-Seq [10] on both metrics.

**Controllable Cross-View Image Translation.** We further adopt Ego2Top to conduct the controllable cross-view image translation experiments. The results are shown in Fig. 9. 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.

**SelectionGAN vs. SelectionGAN++.** We also provide comparison results of SelectionGAN [9] and SelectionGAN++ in Table 4, 5, and 6. SelectionGAN++ achieves better results on most metrics, meaning that the proposed multi-scale channel selection module indeed enhances the feature representation, and thus is improving the generation performance. Note that SelectionGAN++ generates sharper and more realistic images than SelectionGAN, but SelectionGAN has higher pixel-wise similarity scores (i.e., SSIM, PSNR, and SD). This is also observed in other image generation [1], super-resolution [50], and human perceptual judgment [65] tasks. From the visualization results in Fig. 7, 8, and 10,

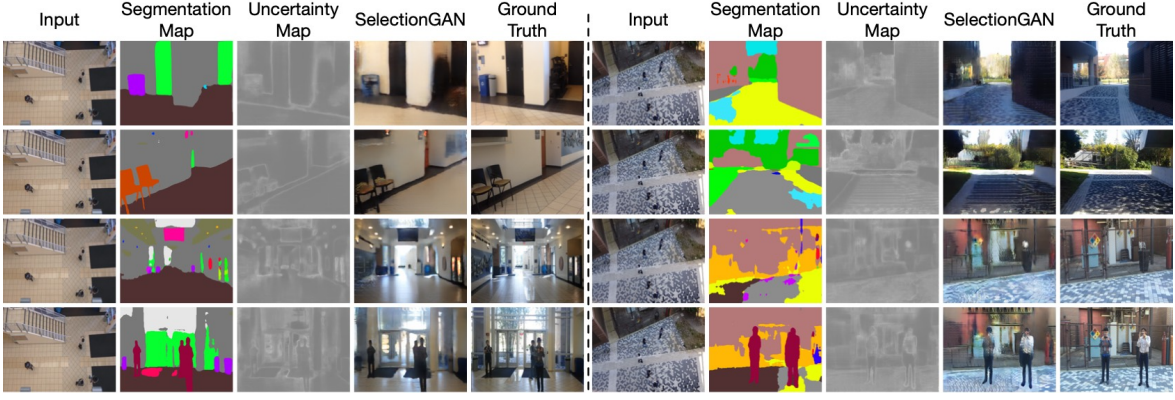
Fig. 9: Results of controllable cross-view image translation for both indoor (*left*) and outdoor (*right*) scenes.

TABLE 6: Quantitative results of cross-view image translation on SVA. (\*) 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	Accuracy (%) $\uparrow$				Inception Score* $\uparrow$			SSIM $\uparrow$	PSNR $\uparrow$	SD $\uparrow$	KL $\downarrow$	FID $\downarrow$
	Top-1	Top-5	Top-1	Top-5	All	Top-1	Top-5					
X-Pix2pix [14]	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 [54]	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 [10]	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 [10]	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 [54]	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 [54]	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 [54]	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 [54]	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 [54]	15.4803	48.0767	21.8225	56.8994	2.6328	2.0732	2.8347	0.4044	20.9848	17.6858	6.0638	88.78
Pix2pix++ [14]	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++ [10]	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++ [10]	11.2580	36.8018	11.9838	36.9231	2.6849	2.1325	2.9397	0.3617	17.4893	18.4122	11.8560	154.80
Pix2pixHD [15]	35.0018	72.9430	52.2181	85.6375	2.5820	2.1436	2.8730	0.5437	23.1823	18.9723	2.6322	32.79
GauGAN [64]	34.6740	71.4061	50.1152	81.4900	2.6462	<b>2.2112</b>	2.9550	0.5195	22.0174	18.7762	2.6714	27.93
SelectionGAN	33.9055	71.8779	50.8878	85.0019	2.6576	2.1279	2.9267	<b>0.5752</b>	<b>24.7136</b>	<b>19.7302</b>	2.6183	<b>26.09</b>
SelectionGAN++	<b>35.9008</b>	<b>73.3249</b>	<b>52.5346</b>	<b>86.9432</b>	<b>2.7370</b>	2.1914	<b>3.0271</b>	0.5481	24.2886	19.2001	<b>2.5788</b>	37.17

TABLE 7: Per-class accuracy and mean IOU for the generated segmentation maps on Dayton.

Method	Per-class Acc. $\uparrow$	mIOU $\uparrow$
X-Fork [10]	0.6262	0.4163
X-Seq [10]	0.4783	0.3187
SelectionGAN	0.6415	0.5455
SelectionGAN++	<b>0.6619</b>	<b>0.5741</b>

we see that SelectionGAN++ generates more photo-realistic images with fewer visual artifacts than SelectionGAN on both tasks. For example, SelectionGAN generates road lines in the first and second rows of Fig. 10, but there are no road lines in the corresponding ground truths.

## 4.2 Results on Facial Expression Generation

**Datasets.** We follow C2GAN [51] and conduct facial expression generation experiments on the Radboud Faces dataset [66]. This dataset contains over 8,000 face images with eight different emotional expressions. We follow C2GAN and all the images are resized to  $256 \times 256$  without any pre-processing. Then, we adopt OpenFace [5] to extract facial landmarks as the semantic guidance. Consequently, we collect 5,628 training image pairs and 1,407 testing pairs.

**Parameter Settings.** Following C2GAN [51], the experiments on Radboud Faces are trained for 200 epochs with batch size of 4.

**Evaluation Metrics.** We follow C2GAN [51] and employ Structural Similarity (SSIM) [62] and Peak Signal-to-Noise



Fig. 10: Comparison results of SelectionGAN and SelectionGAN++ on SVA.

Ratio (PSNR) to evaluate the quantitative quality of generated images. 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 image they thought was real. The same exact images are presented to the workers for all baselines for fair comparisons. Finally, we also use a neural network based metric LPIPS [65] to evaluate the proposed method.

**State-of-the-Art Comparisons.** We compare the proposed SelectionGAN with several state-of-the-art methods, i.e., StarGAN [16], Pix2pix [14], GPGAN [67], PG2 [1], Pix2pixHD [15], GauGAN [64], and C2GAN [51]. Quanti-



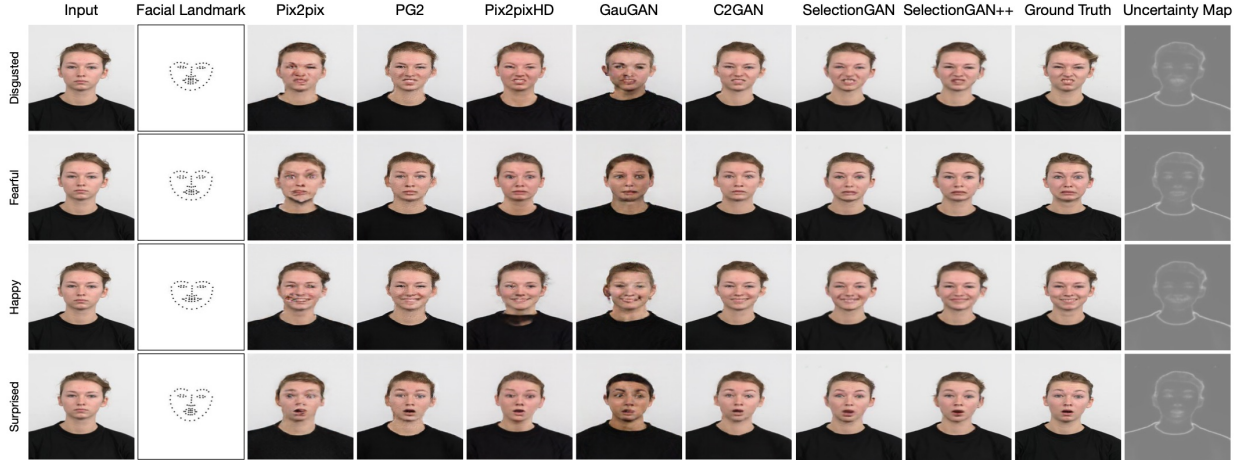


Fig. 11: Results of facial expression generation on Radboud Faces.

Fig. 12: Results of hand gesture-to-gesture translation on NTU Hand Digit (*top two rows*) and Senz3D (*bottom two rows*).

TABLE 8: Quantitative results of facial expression generation on Radboud Faces.

Method	AMT $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$
StarGAN [16]	24.7	0.8345	19.6451	-
Pix2pix [14]	13.4	0.8217	19.9971	0.1334
GPAN [67]	0.3	0.8185	18.7211	0.2531
PG2 [1]	28.4	0.8462	20.1462	0.1130
Pix2pixHD [15]	20.5	0.8269	24.5621	0.1228
GauGAN [64]	10.7	0.7528	20.8430	0.2170
C2GAN [51]	34.2	0.8618	21.9192	0.0934
SelectionGAN	37.5	0.8760	<b>27.5671</b>	0.0917
SelectionGAN++	<b>39.1</b>	<b>0.8761</b>	27.5158	<b>0.0905</b>

tative results of the SSIM, PSNR, LPIPS, and AMT metrics are shown in Table 8. We can see that the proposed SelectionGAN and SelectionGAN++ achieve better results than the existing methods on all metrics, validating the effectiveness of our methods. Note that GauGAN achieves unsatisfactory results in this task since it is proposed to use segmentation maps as input. However, this task uses facial landmarks as guidances, which is quite different from segmentation maps. On the contrary, our methods achieve good results in this task, which further proves the generalizability of our proposed methods. Qualitative results are shown in Fig. 11. Clearly, the image generated by our SelectionGAN and SelectionGAN++ are more sharper and contains more image details compared to 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.

**Efficiency.** We also compared the proposed methods with existing methods on facial expression generation. Our proposed SelectionGAN and SelectionGAN++ takes about 24 and 27 hours to finish the training on a single NVIDIA DGX1 V100 GPU, while C2GAN, GauGAN, Pix2pixHD, and PG2 takes around 27, 32, 36, and 30 hours, respectively. This also validates that the proposed methods are efficient.

**SelectionGAN vs. SelectionGAN++.** We also provide comparison results of SelectionGAN [9] and SelectionGAN++ in Table 8. SelectionGAN++ achieves better results than SelectionGAN on most metrics, i.e., AMT, SSIM, and LPIPS. Meanwhile, SelectionGAN++ generates more realistic details (e.g., eyes and mouth) than SelectionGAN (see Fig. 11).

### 4.3 Results on Hand Gesture Translation

**Datasets.** We follow GestureGAN [2] and conduct experiments on both NTU Hand Digit [70] and Senz3D [71] 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  $256 \times 256$ , and we enabled image flipping and random



TABLE 9: Quantitative results of hand gesture-to-gesture translation on NTU Hand Digit and Senz3D.

Method	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]	28.2403	2.4152	3.5	24.2093	2.6319	26.5138	3.3699	2.8	31.7333	3.0933
SAMG [68]	28.0185	2.4919	2.6	31.2841	2.7453	26.9545	3.3285	2.3	38.1758	3.1006
DPIG [69]	30.6487	2.4547	7.1	<b>6.7661</b>	2.6184	26.9451	3.3874	6.9	26.2713	3.0846
PoseGAN [45]	29.5471	2.4017	9.3	9.6725	2.5846	27.3014	3.2147	8.6	24.6712	3.0467
Pix2pixHD [15]	<b>38.1295</b>	2.2358	21.3	8.4003	<b>1.1475</b>	-	-	-	-	-
GauGAN [64]	32.2218	<b>2.6210</b>	13.2	18.4373	1.8229	-	-	-	-	-
GestureGAN [2]	32.6091	2.5532	<b>26.1</b>	7.5860	2.5223	27.9749	<b>3.4107</b>	<b>22.6</b>	<b>18.4595</b>	2.9836
SelectionGAN	30.6465	2.4472	15.8	16.2159	2.1560	30.4036	2.4595	14.1	30.9775	2.7014
SelectionGAN++	31.4580	2.5197	20.9	12.4843	2.0221	<b>31.1875</b>	2.8194	18.7	23.6390	<b>2.6711</b>

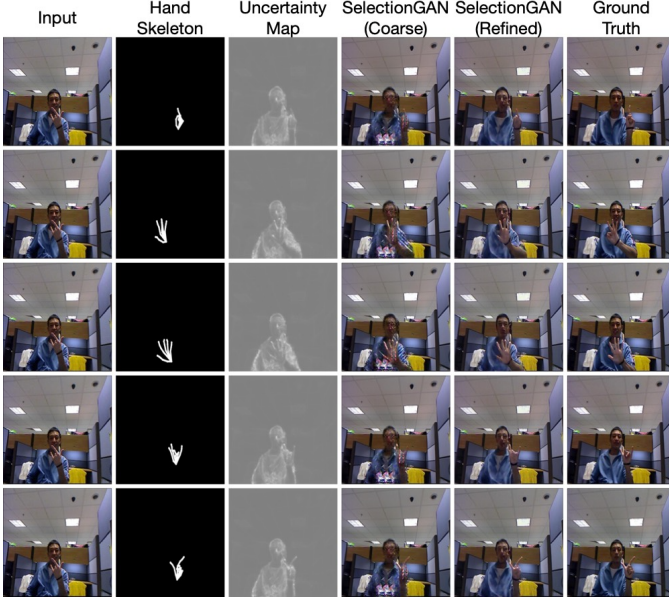


Fig. 13: Results of controllable hand gesture translation.

crops for data augmentation. Following GestureGAN [2], the experiments on both datasets are trained for 20 epochs with batch size of 4.

**Evaluation Metrics.** We follow [2] and employ Peak Signal-to-Noise Ratio (PSNR), Inception score (IS) [60], Fréchet Inception Distance (FID) [61], and Fréchet ResNet Distance (FRD) [2] to evaluate the generated images. Moreover, we follow the same settings as in [2], [14] to conduct the Amazon Mechanical Turk (AMT) perceptual studies.

**State-of-the-Art Comparisons.** We compare the proposed methods with the leading hand gesture translation methods, i.e., PG2 [1], SAMG [68], DPIG [69], PoseGAN [45], Pix2pixHD [15], GauGAN [64], and GestureGAN [2]. Comparison results are shown in Table 9. We can see that our SelectionGAN and SelectionGAN++ achieve competitive results on both datasets. Note that GestureGAN achieves better results than the proposed methods. The reason is that GestureGAN is carefully tailored and designed for this task, meaning that GestureGAN is fine-tuned to this task with the network structure, loss objective, and hyper-parameter selection. However, the proposed methods are novel and unified GAN models, which can be used to handle various settings of guided image-to-image translation without modifying the network structure, the loss objective, and hyper-parameters. Qualitative results compared with existing methods are shown in Fig. 12. We can see that our

SelectionGAN and SelectionGAN++ also generate photo-realistic results on this challenging task. Moreover, we show the learned uncertainty maps in Fig. 13.

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

**SelectionGAN vs. SelectionGAN++.** We also provide comparison results of SelectionGAN [9] and SelectionGAN++. The results of hand gesture translation are shown in Table 9. We can see that SelectionGAN++ achieves better results than SelectionGAN on all metrics. Meanwhile, SelectionGAN++ generates more photo-realistic results than SelectionGAN, as shown in Fig. 13.

#### 4.4 Results on Person Image Generation

**Datasets.** We follow PATN [73] and conduct person image generation experiments on both Market-1501 [74] and DeepFashion [75] datasets. Following [73], 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 PATN [73], images are re-scaled to  $128 \times 64$  and  $256 \times 256$  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.** We follow previous works [1], [45], [45], [51], [73] and adopt Structure Similarity (SSIM) [62], Inception score (IS) [60] and their corresponding masked versions, i.e., M-SSIM and M-IS, as our evaluation metrics. We also recruit 30 volunteers to conduct a user study. Specifically, given six results (four generated by existing methods, two generated by our proposed SelectionGAN and SelectionGAN++), each participant needs to answer two questions: ‘Q1: Which generated image is more realistic regardless of the target image?’ and ‘Q2: Which generated image matches the conditioning image better (e.g., clothes)?’.

**State-of-the-Art Comparisons.** We compare the proposed SelectionGAN and SelectionGAN++ with several leading person image generation methods, i.e., PG2 [1], DPIG [69], PoseGAN [45], VUNet [72], C2GAN [51], BTF [4], Pix2pixHD [15], GauGAN [64], and PATN [73]. Quantitative results of the SSIM, IS, M-SSIM, and M-IS metrics

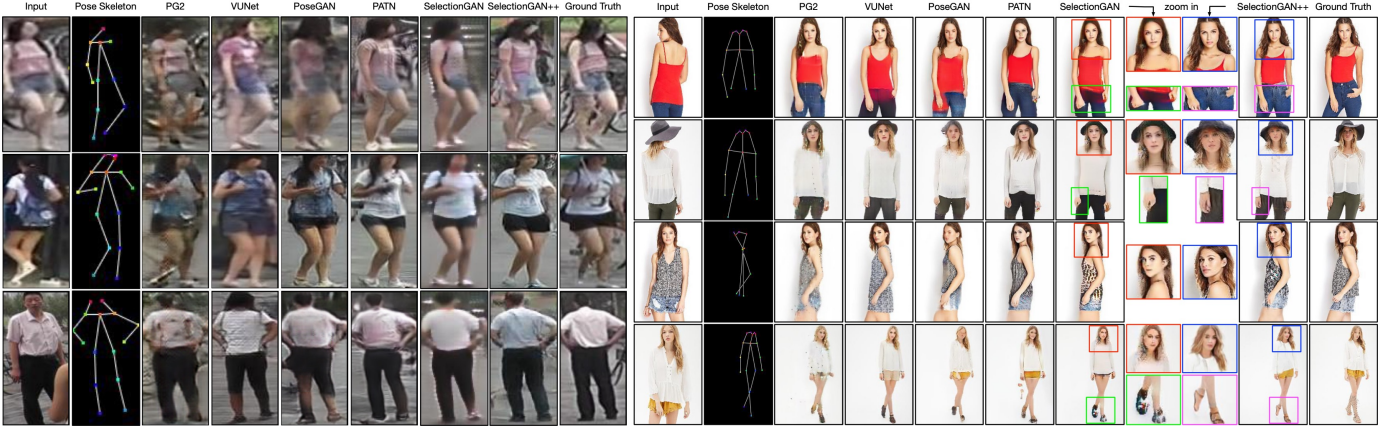


Fig. 14: Results of person image generation on Market-1501 (left) and DeepFashion (right). Key differences in DeepFashion are highlighted by colored boxes.

TABLE 10: (left) Quantitative results of person image generation on Market-1501 and DeepFashion. (\*) denotes the results tested on our test set. (right) User study (%) of person image generation. For each comparison, the participant is asked to answer two questions, i.e., ‘Q1: Which generated image is more realistic regardless of the target image?’, and ‘Q2: Which generated image matches the conditioning image better (e.g., clothes)?’.

Method	Market-1501				DeepFashion	
	SSIM $\uparrow$	IS $\uparrow$	M-SSIM $\uparrow$	M-IS $\uparrow$	SSIM $\uparrow$	IS $\uparrow$
PG2 [1]	0.253	3.460	0.792	3.435	0.762	3.090
DPIG [69]	0.099	3.483	0.614	3.491	0.614	3.228
PoseGAN [45]	0.290	3.185	0.805	3.502	0.756	3.439
C2GAN [51]	0.282	3.349	0.811	3.510	-	-
BTF [4]	-	-	-	-	0.767	3.220
PG2* [1]	0.261	3.495	0.782	3.367	0.773	3.163
PoseGAN* [45]	0.291	3.230	0.807	3.502	0.760	3.362
Pix2pixHD* [15]	-	-	-	-	0.762	3.224
GauGAN* [64]	-	-	-	-	0.754	3.165
VUNet* [72]	0.266	2.965	0.793	3.549	0.763	3.440
PATN* [73]	0.311	3.323	0.811	<b>3.773</b>	0.773	3.209
SelectionGAN	0.331	3.449	0.816	3.376	0.776	3.341
SelectionGAN++	<b>0.333</b>	<b>3.512</b>	<b>0.818</b>	3.651	<b>0.778</b>	<b>3.445</b>
Real Data	1.000	3.890	1.000	3.706	1.000	4.053

Method	Market-1501		DeepFashion	
	Q1 $\uparrow$	Q2 $\uparrow$	Q1 $\uparrow$	Q2 $\uparrow$
PG2 [1]	4.2	3.1	6.3	6.5
PoseGAN [45]	8.3	6.7	10.5	8.2
C2GAN [51]	16.1	17.6	-	-
PATN [73]	20.3	19.9	22.9	23.1
SelectionGAN	23.7	24.2	28.1	29.3
SelectionGAN++	<b>27.4</b>	<b>28.5</b>	<b>32.2</b>	<b>32.9</b>

are shown in Table 10(left). We see that the proposed SelectionGAN and SelectionGAN++ achieve competitive performance compared with the carefully designed methods on this task such as PATN [73] and PoseGAN [45]. Moreover, we show the user study results in Table 10(right). We observe that our methods achieve better results over [1], [45], [51], [73] in terms of both image realism and style consistency, further validating that our generated images are more photo-realistic. Qualitative results are shown in Fig. 14. The images generated by our SelectionGAN and SelectionGAN++ are more realistic and sharp compared with other leading methods. Moreover, the person layouts of the generated images by our methods are closer to the target skeletons.

**SelectionGAN vs. SelectionGAN++.** We also provide comparison results of SelectionGAN [9] and SelectionGAN++ in Table 10. We see that SelectionGAN++ achieves better results than SelectionGAN on all metrics. Moreover, the results of the user study indicate that SelectionGAN++ generates much better results than SelectionGAN. Meanwhile, SelectionGAN++ generates more photo-realistic results (especially in DeepFashion) than SelectionGAN (see Fig. 14). In order to better prove that SelectionGAN++ produces more realistic images than SelectionGAN, we provide the com-

parison in a zoomed-in manner on the DeepFashion dataset in Fig. 14. For example, in the last row, SelectionGAN++ generates better hair, face, and feet than SelectionGAN.

#### 4.5 Results on Semantic Image Synthesis

To explore the generality of SelectionGAN and SelectionGAN++ on other generation tasks, we also conduct experiments on the semantic image synthesis task. Specifically, we adopt GauGAN [64] as our backbone network in this task and we combine it with the proposed multi-channel attention selection module to form our final model.

**Datasets.** We follow GauGAN [64] and conduct semantic image synthesis experiments on two challenging datasets, i.e., Cityscapes [78] 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 re-scaled to  $512 \times 256$  and  $256 \times 256$  on Cityscapes and ADE20K datasets, respectively. Following GauGAN [64], the experiments on both datasets are trained for 200 epochs with batch size of 32.

**Evaluation Metrics.** We follow [64] and employ the mean Intersection-over-Union (mIoU) and pixel accuracy (Acc) to measure the segmentation accuracy. Specifically, we adopt





Fig. 15: Results of semantic image synthesis on Cityscapes (*top two rows*) and ADE20K (*bottom three rows*).

TABLE 11: (*left*) Quantitative results of semantic image synthesis on Cityscapes and ADE20K. (*right*) 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 SelectionGAN++ over the competing method.

Method	Cityscapes			ADE20K			AMT $\uparrow$	Cityscapes	ADE20K
	mIoU $\uparrow$	Acc $\uparrow$	FID $\downarrow$	mIoU $\uparrow$	Acc $\uparrow$	FID $\downarrow$			
CRN [76]	52.4	77.1	104.7	22.4	68.8	73.3	Ours vs. CRN [76]	65.80	72.15
SIMS [77]	47.2	75.5	<b>49.7</b>	-	-	-	Ours vs. Pix2pixHD [15]	60.93	80.61
Pix2pixHD [15]	58.3	81.4	95.0	20.3	69.2	81.8	Ours vs. SIMS [77]	56.78	-
GauGAN [64]	62.3	81.9	71.8	38.5	79.9	33.9	Ours vs. GauGAN [64]	55.22	57.54
SelectionGAN	63.8	82.4	65.2	40.1	81.2	33.1	Ours vs. SelectionGAN	53.17	55.75
SelectionGAN++	<b>64.5</b>	<b>82.7</b>	63.4	<b>41.7</b>	<b>81.5</b>	<b>32.2</b>			

the state-of-the-art segmentation networks to evaluate the generated images, i.e., DRN-D-105 [79] for Cityscapes and UperNet101 [80] for ADE20K. We also employ the Fréchet Inception Distance (FID) [61] 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 [15], CRN [76], SIMS [77], and GauGAN [64]. The results of mIoU, Acc, and FID are show in Table 11(*left*). We note that our SelectionGAN and SelectionGAN++ achieve better results than the existing competing methods on both mIoU and Acc metrics. For FID, SelectionGAN and SelectionGAN++ are only worse than SIMS on Cityscapes. However, SIMS has poor segmentation results. Moreover, we follow GauGAN and provide AMT results in Table 11(*right*). We observe that users favor our translated images on both datasets compared with existing leading methods. Qualitative results compared with the exiting methods are shown in Fig. 15. We observe that SelectionGAN and SelectionGAN++ produce 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. The results compared with the state-of-the-art method GauGAN are shown in

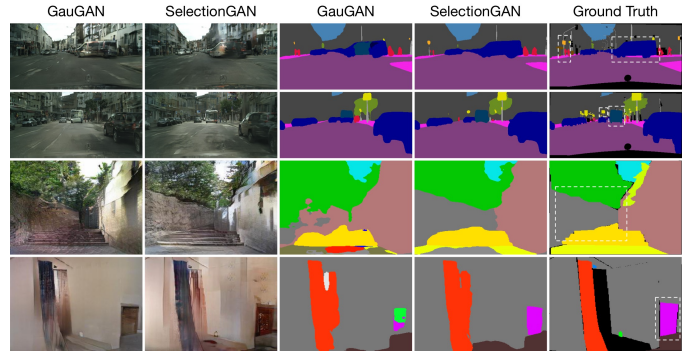


Fig. 16: Generated segmentation maps on Cityscapes (*top two rows*) and ADE20K (*bottom two rows*).

Fig. 16. We observe that the proposed SelectionGAN generates better semantic maps than GauGAN on both datasets.

**SelectionGAN vs. SelectionGAN++.** We also provide comparison results of SelectionGAN [9] and SelectionGAN++ in Table 11. SelectionGAN++ achieves better results than SelectionGAN on all metrics, i.e., mIoU, Acc, and FID. Moreover, the results of the user study indicate that SelectionGAN++ generates more photo-realistic results than SelectionGAN. We also note that SelectionGAN++ generates better results than SelectionGAN on both datasets (see Fig. 15).

## 5 CONCLUSION

We propose SelectionGAN to address a novel image synthesis task by conditioning on an 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 focuses 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 semantic image synthesis tasks with 11 public datasets show that our method obtains much better results than the state-of-the-art models.

**Acknowledgments.** This work has been supported by the Italy-China collaboration project TALENT, by the EU H2020 project AI4Media (No. 951911) and by the PRIN project PREVUE (Prot. 2017N2RK7K).

## 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, 6, 9, 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, 6, 11, 12
- [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, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, no. 2, 2016. 1, 6, 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, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *ECCV*, 2020. 1, 3, 4
- [9] 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. 1, 2, 3, 7, 9, 11, 12, 13, 14
- [10] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *CVPR*, 2018. 1, 3, 6, 7, 8, 9, 10
- [11] 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, 6
- [12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019. 2
- [13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint:1411.1784*, 2014. 2
- [14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017. 2, 6, 7, 8, 9, 10, 11, 12
- [15] 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, 8, 10, 11, 12, 13, 14
- [16] 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
- [17] 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
- [18] H. Tang, W. Wang, S. Wu, X. Chen, D. Xu, N. Sebe, and Y. Yan, "Expression conditional gan for facial expression-to-expression translation," in *ICIP*, 2019. 3
- [19] Z. Xu, T. Lin, H. Tang, F. Li, D. He, N. Sebe, R. Timofte, L. Van Gool, and E. Ding, "Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model," in *CVPR*, 2022. 3
- [20] X. Liu, Z. Lin, J. Zhang, H. Zhao, Q. Tran, X. Wang, and H. Li, "Open-edit: Open-domain image manipulation with open-vocabulary instructions," in *ECCV*, 2020. 3
- [21] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *CVPR*, 2022. 3
- [22] H. Tang, H. Liu, and N. Sebe, "Unified generative adversarial networks for controllable image-to-image translation," *IEEE TIP*, vol. 29, pp. 8916–8929, 2020. 3
- [23] H. Tang and N. Sebe, "Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes," *IEEE TMM*, 2021. 3
- [24] H. Tang, S. Bai, P. H. Torr, and N. Sebe, "Bipartite graph reasoning gans for person image generation," in *BMVC*, 2020. 3, 4
- [25] H. Tang, S. Bai, and N. Sebe, "Dual attention gans for semantic image synthesis," in *ACM MM*, 2020. 3
- [26] G. Liu, H. Tang, H. Latapie, and Y. Yan, "Exocentric to egocentric image generation via parallel generative adversarial network," in *ICASSP*, 2020. 3
- [27] 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," in *CVPR*, 2020. 3
- [28] S. Wu, H. Tang, X.-Y. Jing, J. Qian, N. Sebe, Y. Yan, and Q. Zhang, "Cross-view panorama image synthesis with progressive attention gans," *Elsevier PR*, 2022. 3
- [29] S. Wu, H. Tang, X.-Y. Jing, H. Zhao, J. Qian, N. Sebe, and Y. Yan, "Cross-view panorama image synthesis," *IEEE TMM*, 2022. 3
- [30] H. Tang, L. Shao, P. H. Torr, and N. Sebe, "Local and global gans with semantic-aware upsampling for image generation," *IEEE TPAMI*, 2022. 3
- [31] B. Ren, H. Tang, and N. Sebe, "Cascaded cross mlp-mixer gans for cross-view image translation," in *BMVC*, 2021. 3
- [32] H. Tang and N. Sebe, "Layout-to-image translation with double pooling generative adversarial networks," *IEEE TIP*, 2021. 3
- [33] G. Liu, H. Tang, H. M. Latapie, J. J. Corso, and Y. Yan, "Cross-view exocentric to egocentric video synthesis," in *ACM MM*, 2021. 3
- [34] 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
- [35] 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
- [36] 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
- [37] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE TGRS*, 2020. 3
- [38] B. Duan, W. Wang, H. Tang, H. Latapie, and Y. Yan, "Cascade attention guided residue learning gan for cross-modal translation," in *ICPR*, 2020. 3
- [39] D. Shi, X. Diao, L. Shi, H. Tang, Y. Chi, C. Li, and H. Xu, "Charformer: A glyph fusion based attentive framework for high-precision character image denoising," in *ACM MM*, 2022. 3
- [40] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, T. Hao, X. Alamedd-Pineda, and E. Ricci, "Continual attentive fusion for incremental learning in semantic segmentation," *IEEE TMM*, 2022. 3
- [41] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE TNNLS*, 2021. 3
- [42] 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
- [43] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *ICML*, 2019. 3
- [44] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *IJCNN*, 2019. 3
- [45] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018. 4, 12, 13
- [46] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *ECCV*, 2014. 4
- [47] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *CVPR*, 2016. 4
- [48] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017. 5

- [49] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018. **6**
- [50] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016. **6, 9**
- [51] 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. **6, 10, 11, 12, 13**
- [52] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017. **6**
- [53] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. **7**
- [54] K. Regmi and A. Borji, "Cross-view image synthesis using geometry-guided conditional gans," *Elsevier CVIU*, vol. 187, p. 102788, 2019. **7, 8, 9, 10**
- [55] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *ECCV*, 2016. **7**
- [56] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *ICCV*, 2015. **7**
- [57] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *CVPR*, 2017. **7, 8, 9**
- [58] A. Palazzi, G. Borghi, D. Abati, S. Calderara, and R. Cucchiara, "Learning to map vehicles into bird's eye view," in *ICIAP*, 2017. **7**
- [59] S. Ardeshtir and A. Borji, "Ego2top: Matching viewers in egocentric and top-view videos," in *ECCV*, 2016. **7**
- [60] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016. **7, 12**
- [61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017. **7, 12, 14**
- [62] 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, 12**
- [63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018. **8**
- [64] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019. **8, 10, 11, 12, 13, 14**
- [65] 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. **9, 10**
- [66] 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**
- [67] X. Di, V. A. Sindagi, and V. M. Patel, "Gp-gan: Gender preserving gan for synthesizing faces from landmarks," in *ICPR*, 2018. **10, 11**
- [68] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *ACM MM*, 2017. **12**
- [69] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *CVPR*, 2018. **12, 13**
- [70] 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**
- [71] 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**
- [72] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018. **12, 13**
- [73] 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**
- [74] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *CVPR*, 2015. **12**
- [75] 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**
- [76] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *ICCV*, 2017. **14**
- [77] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," in *CVPR*, 2018. **14**
- [78] 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**
- [79] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *CVPR*, 2017. **14**
- [80] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *ECCV*, 2018. **14**



**Hao Tang** is currently a Postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He received the master's degree from the School of Electronics and Computer Engineering, Peking University, China and the Ph.D. degree from the Multimedia and Human Understanding Group, University of Trento, Italy. He was a visiting scholar in the Department of Engineering Science at the University of Oxford. His research interests are deep learning, machine learning, and their applications to computer vision.



**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 in the University of Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding. He was the General Co-Chair of the IEEE FG 2008 and ACM Multimedia 2013. He was a program chair of ACM Multimedia 2011 and 2007, ECCV 2016, ICCV 2017 and ICPR 2020. He is a general chair of ACM Multimedia 2022 and a program chair of ECCV 2024. He is a fellow of IAPR.