# Multi-Channel Attention Selection GAN with Cascaded Semantic Guidance for Cross-View Image Translation

Hao Tang[1,2*] Dan Xu[3*] Nicu Sebe[1,4] Yanzhi Wang[5] Jason J. Corso[6] Yan Yan[2]

[1]DISI, University of Trento, Trento, Italy [2]Texas State University, San Marcos, USA
[3]University of Oxford, Oxford, UK [4]Huawei Technologies Ireland, Dublin, Ireland
[5]Northeastern University, Boston, USA [6]University of Michigan, Ann Arbor, USA

## Abstract

*Cross-view image translation is challenging because it involves images with drastically different views and severe deformation. In this paper, we propose a novel approach named Multi-Channel Attention SelectionGAN (SelectionGAN) that makes it possible to generate images of natural scenes in arbitrary viewpoints, based on an image of the scene and a novel semantic map. The proposed SelectionGAN explicitly utilizes the semantic information and consists of two stages. In the first stage, the condition image and the target semantic map 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 a multi-channel attention selection mechanism. Moreover, uncertainty maps automatically learned from attentions are used to guide the pixel loss for better network optimization. Extensive experiments on Dayton [41], CVUSA [43] and Ego2Top [1] datasets show that our model is able to generate significantly better results than the state-of-the-art methods. The source code, data and trained models are available at* https://github.com/Ha0Tang/SelectionGAN.

## 1. Introduction

Cross-view image translation is a task that aims at synthesizing new images from one viewpoint to another. It has been gaining a lot interest especially from computer vision and virtual reality communities, and has been widely investigated in recent years [40, 20, 54, 34, 47, 15, 31, 52, 45]. Earlier works studied this problem using encoder-decoder Convolutional Neural Networks (CNNs) by involving viewpoint codes in the bottle-neck representations for city scene synthesis [52] and 3D object translation [45]. There also exist some works exploring Generative Adversarial Networks (GAN) for similar tasks [31]. However, these existing works consider an application scenario in which the objects and the scenes have a large degree of overlapping in appearances and views.

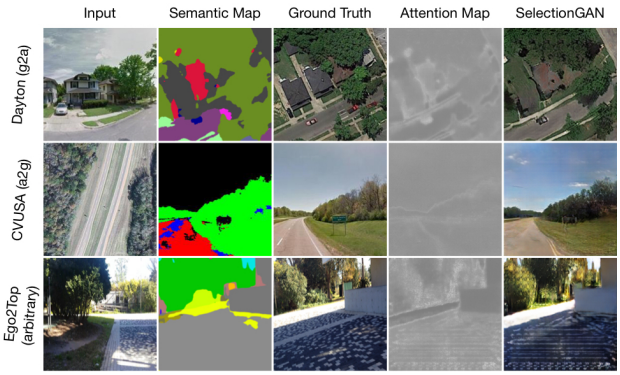


Figure 1: Examples of our cross-view translation results on two public benchmarks *i.e.* Dayton [41] and CVUSA [43], and on our self-created large-scale benchmark based on Ego2Top [1].

Different from previous works, in this paper, we focus on a more challenging setting in which fields of views have little or even no overlap, leading to significantly distinct structures and appearance distributions for the input source and the output target views, as illustrated in Fig. 1. To tackle this challenging problem, Regmi and Borji [34] recently proposed a conditional GAN model which jointly learns the generation in both the image domain and the corresponding semantic domain, and the semantic predictions are further utilized to supervise the image generation. Although this approach performed an interesting exploration, we observe unsatisfactory aspects mainly in the generated scene structure and details, which are due to different reasons. First, since it is always costly to obtain manually annotated semantic labels, the label maps are usually produced from pretrained semantic models from other large-scale segmentation datasets, leading to insufficiently accurate predictions for all the pixels, and thus misguiding the image generation. Second, we argue that the translation with a single

*Equal contribution.

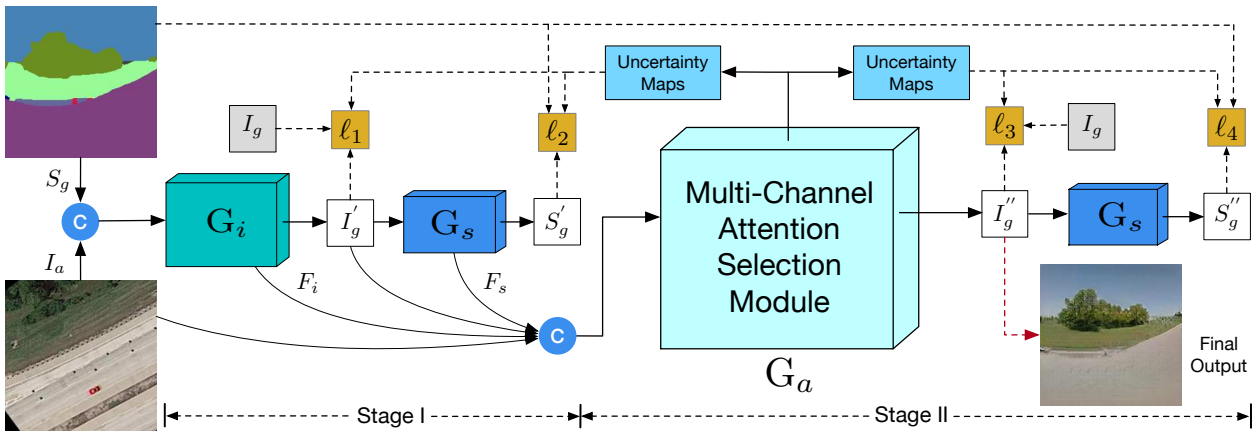


Figure 2: Overview of the proposed SelectionGAN. Stage I presents a cycled semantic-guided generation sub-network which accepts images from one view and conditional semantic maps and simultaneously synthesizes images and semantic maps in another view. Stage II takes the coarse predictions and the learned deep semantic features from stage I, and performs a fine-grained generation using the proposed multi-channel attention selection module.

phase generation network is not able to capture the complex scene structural relationships between the two views. 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 this first stage, we learn a cycled image-semantic generation sub-network, which accepts a pair consisting of an image and the target semantic map, and generates images for the other view, which further fed into a semantic generation network to reconstruct the input semantic maps. This cycled generation adds more strong supervision between the image and semantic 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. 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 label 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 baselines such as Pix2pix [16], Zhai *et al.* [47], X-Fork [34] and X-Seq [34]. Moreover, we establish state-of-the-art results on three different datasets for the arbitrary cross-view image synthesis task.

Overall, the contributions of this paper are as follows:

- A novel multi-channel attention selection GAN framework (SelectionGAN) for the cross-view 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-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 two public benchmarks, *i.e.* Dayton [41] and CVUSA [43]. Meanwhile, we also create a larger-scale cross-view synthesis benchmark using the data from Ego2Top [1], and present results of multiple baseline models for the research community.

## 2. Related Work

**Generative Adversarial Networks (GANs)** [11] have shown the capability of generating better high-quality images [42, 18, 12], compared to existing methods such as Restricted Boltzmann Machines [13, 35] and Deep Boltzmann Machines [14]. A vanilla GAN model [11] has two important components, *i.e.* 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 [18, 48, 32], there are still some challenges, *i.e.* how to generate images in a controlled setting. To generate domain-specific images, Conditional GAN (CGAN) [27]

has been proposed. CGAN usually combines a vanilla GAN and some external information, such as class labels or tags [29, 30, 4, 39, 36], text descriptions [33, 49], human pose [8, 37, 28, 22] and reference images [25, 16].

**Image-to-Image Translation** frameworks adopt input-output data to learn a parametric mapping between inputs and outputs. For example, Isola *et al.* [16] propose Pix2pix, which is a supervised model and uses a CGAN to learn a translation function from input to output image domains. Zhu *et al.* [53] introduce CycleGAN, which targets unpaired image translation using the cycle-consistency loss. To further improve the generation performance, the attention mechanism has been recently investigated in image translation, such as [3, 44, 38, 24, 26]. However, 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 task.

**Learning Viewpoint Transformations.** Most existing works on viewpoint transformation have been conducted to synthesize novel views of the same object, such as cars, chairs and tables [9, 40, 5]. Another group of works explore the cross-view scene image generation, such as [46, 52]. However, these works focus on the scenario in which the objects and the scenes have a large degree of overlapping in both appearances and views. Recently, several works started investigating image translation problems with drastically different views and generating a novel scene from a given arbitrary one. This is a more challenging task since different views have little or no overlap. To tackle this problem, Zhai *et al.* [47] try to generate panoramic ground-level images from aerial images of the same location by using a convolutional neural network. Krishna and Ali [34] propose a X-Fork and a X-Seq GAN-based structure to address the aerial to street view image translation task using an extra semantic segmentation map. However, these methods are not able to generate satisfactory results due to the drastic difference between source and target views and their model design. To overcome these issues, 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 improves the generation quality.

# 3. Multi-Channel Attention Selection GAN

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 cascade semantic-guided generation sub-network, which utilizes the images from one view and conditional semantic maps from another view as inputs, and reconstruct images in another view. These images are further input into a semantic generator to recover the input semantic map 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-channel attention selection module, which aims at producing more fine-grained synthesis from a larger generation space and also at generating uncertainty maps to guide multiple optimization losses.

## 3.1. Cascade Semantic-guided Generation

**Semantic-guided Generation.** Cross-view synthesis is a challenging task, especially when the two views have little overlapping as in our study case, which apparently leads to ambiguity issues in the generation process. To alleviate this problem, we use semantic maps as conditional guidance. Since it is always costly to obtain annotated semantic maps, following [34] we generate the maps using segmentation deep models pretrained from large-scale scene parsing datasets such as Cityscapes [6]. However, [34] uses semantic maps only in the reconstruction loss to guide the generation of semantics, which actually provides a weak guidance. Different from theirs, we apply the semantic maps not only in the output loss but also as part of the network's input. Specifically, as shown in Fig. 2, we concatenate the input image $I_a$ from the source view and the semantic map $S_g$ from a target view, and input them into the image generator $G_i$ and synthesize the target view image $I_g^{'}$ as $I_g^{'}{=}G_i(I_a, S_g)$. In this way, the ground-truth semantic maps provide stronger supervision to guide the cross-view translation in the deep network.

**Semantic-guided Cycle.** Regmi and Borji [34] observed that the simultaneous generation of both the images and the semantic maps improves the generation performance. Along the same line, we propose a cycled semantic generation network to benefit more the semantic information in learning. The conditional semantic map $S_g$ 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 generator $G_s$ which reconstructs a new semantic map $S_g^{'}$. 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 generation cycle, *i.e.* $[I_a, S_g] \overset{G_i}{\rightarrow} I_g^{'} \overset{G_s}{\rightarrow} S_g^{'}{\approx}S_g$. The two generators are explicitly connected by the ground-truth semantic maps, which in this way provide extra constraints on the generators to learn better the semantic structure consistency.

**Cascade Generation.** Due to the complexity of the task, after the first stage, we observe that the image generator $G_i$ outputs a coarse synthesis, which yields blurred scene details and high pixel-level dis-similarity with the target-view images. This inspires us to explore a coarse-to-fine generation strategy in order to boost the synthesis performance based on the coarse predictions. Cascade models

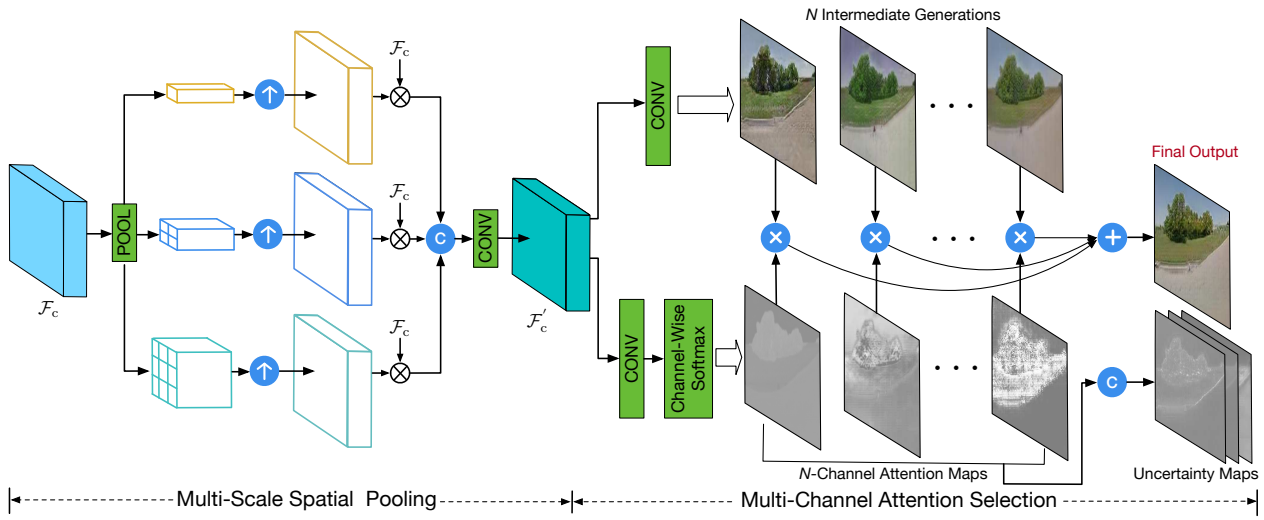


Figure 3: Illustration of the proposed multi-channel attention selection module. The multi-scale spatial pooling pools features in different receptive fields in order to have better generation of scene details; the multi-channel attention selection aims at automatically select from a set of intermediate diverse generations in a larger generation space to improve the generation quality. The symbols $\oplus$, $\otimes$, ⓒ and ⓣ denote element-wise addition, element-wise multiplication, concatenation, and up-sampling operation, respectively.

have been used in several other computer vision tasks such as object detection [2] and semantic segmentation [7], and have shown great effectiveness. In this paper, we introduce the cascade strategy to deal with the complex cross-view translation problem. In both stages we have a basic cycled semantic guided generation sub-network, while in the second stage, we propose a novel multi-channel attention selection module to better utilize the coarse outputs from the first stage and produce fine-grained final outputs. We observed significant improvement by using the proposed cascade strategy, illustrated in the experimental part.

### 3.2. Multi-Channel Attention Selection

An overview of the proposed multi-channel attention selection module $G_a$ is shown in Fig. 3. The module consists of a multi-scale spatial pooling and a multi-channel attention selection component.

**Multi-Scale Spatial Pooling.** Since there exists a large object/scene deformation between the source view and the target view, 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 size and stride 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 a different spatial context. More specifically, given the coarse inputs and the deep semantic features produced from the stage I, we first concatenate all of them as new features denoted as $\mathcal{F}_\mathrm{c}$ for the stage II as:

$$\mathcal{F}_\mathrm{c} = \mathrm{concat}(I_a, I_g^{'}, F_i, F_s) \tag{1}$$

where $\mathrm{concat}(\cdot)$ 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. 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 [50] 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 $\mathrm{pl\_up_s}(\cdot)$ 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:

$$\mathcal{F}_\mathrm{c} \leftarrow \mathrm{concat}\big(\mathcal{F}_\mathrm{c} \otimes \mathrm{pl\_up}_1(\mathcal{F}_\mathrm{c}), \ldots, \mathcal{F}_\mathrm{c} \otimes \mathrm{pl\_up}_M(\mathcal{F}_\mathrm{c})\big) \tag{2}$$

Then the features $\mathcal{F}_\mathrm{c}$ are fed into a convolutional layer, which produces new multi-scale features $\mathcal{F}_\mathrm{c}^{'}$ for the use in the multi-channel selection module.

**Multi-Channel Attention Selection.** In previous cross-view image synthesis works, the image is 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 multi-scale feature volume $\mathcal{F}_\mathrm{c}^{'} \in \mathbb{R}^{h \times w \times c}$, where $h$ and $w$ are width and height of the features, and $c$ is the number of channels, we consider two directions. One is for the generation of multiple intermediate image syn-

thesis, 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:

$$\begin{aligned} I_G^i &= \tanh(\mathcal{F}_c^{'} W_G^i + b_G^i), \qquad \text{for } i = 1, \ldots, N \\ I_A^i &= \mathrm{Softmax}(\mathcal{F}_c^{'} W_A^i + b_A^i), \qquad \text{for } i = 1, \ldots, N \end{aligned} \tag{3}$$

where $\mathrm{Softmax}(\cdot)$ 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:

$$I_g^{''} = (I_A^1 \otimes I_G^1) \oplus \cdots \oplus (I_A^N \otimes I_G^N) \tag{4}$$

where $I_g^{''}$ represents the final synthesized generation selected from the multiple diverse results, and the symbol $\oplus$ denotes the element-wise addition. We also generate a final semantic map 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 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 maps obtained from the pretrained model are not accurate for all the pixels, which leads to a wrong guidance during training. To tackle this issue, we propose the generated attention maps to learn uncertainty maps to control the optimization loss. The uncertainty learning has been investigated in [19] for multi-task learning, and here we introduce it for solving the noisy semantic label 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 of 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:

$$\begin{aligned} U_i &= \sigma\big(W_u^i(\mathrm{concat}(I_A^1, \ldots, I_A^N) + b_u^i\big) \\ \mathcal{L}_p^i &\leftarrow \frac{\mathcal{L}_p^i}{U_i} + \log U_i, \quad \text{for } i = 1, \ldots, K \end{aligned} \tag{5}$$

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 [16] 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 input. This pairwise input encourages $D$ to discriminate the diversity of image structure and capture the local-aware information.

### 3.3. 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,

$$\begin{aligned} \mathcal{L}_{cGAN}(I_a, I_g^{'}) =& \mathbb{E}_{I_a, I_g}\left[\log D(I_a, I_g)\right] + \\ & \mathbb{E}_{I_a, I_g^{'}}\left[\log(1 - D(I_a, I_g^{'}))\right]. \end{aligned} \tag{6}$$

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:

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

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. (6) and (7),

$$\mathcal{L}_{cGAN} = \mathcal{L}_{cGAN}(I_a, I_g^{'}) + \lambda \mathcal{L}_{cGAN}(I_a, I_g^{''}). \tag{8}$$

**Overall Loss.** The total optimization loss is a weighted sum of the above losses. Generators $G_i$, $G_s$, attention selection network $G_a$ and discriminator $D$ are trained in an end-to-end fashion optimizing the following min-max function,

$$\min_{\{G_i, G_s, G_a\}} \max_{\{D\}} \mathcal{L} = \sum_{i=1}^{4} \lambda_i \mathcal{L}_p^i + \mathcal{L}_{cGAN} + \lambda_{tv} \mathcal{L}_{tv}. \tag{9}$$

where $\mathcal{L}_p^i$ uses the L1 reconstruction to separately calculate the pixel loss between the generated images $I_g^{'}$, $S_g^{'}$, $I_g^{''}$ and $S_g^{''}$ and the corresponding real images. $\mathcal{L}_{tv}$ is the total variation regularization [17] 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.

### 3.4. Implementation Details

**Network Architecture.** For a fair comparison, we employ U-Net [16] 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 cross-view 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

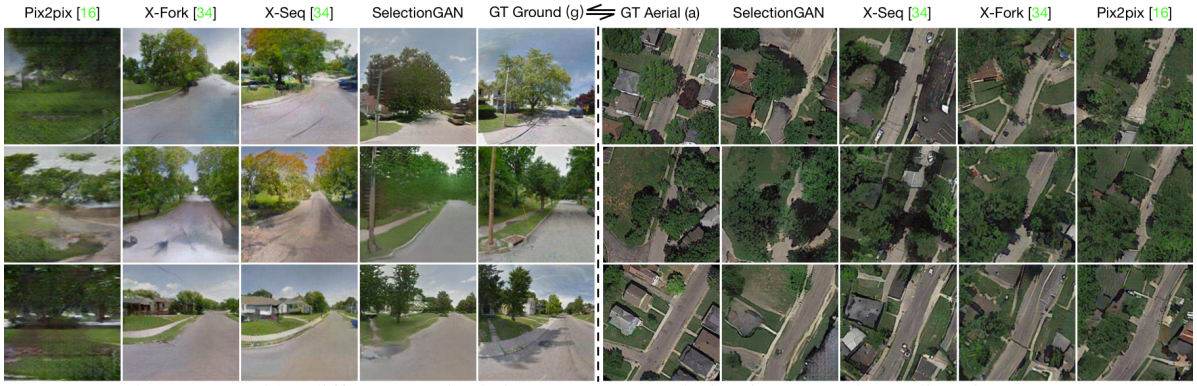


Figure 4: Results generated by different methods in 256×256 resolution in a2g and g2a directions on Dayton dataset.

Table 1: SSIM, PSNR, Sharpness Difference (SD) and KL score (KL) of different methods. For these metrics except KL score, higher is better. (*) These results are reported in [34].

| Direction | Method | Dayton (64×64) | | | | Dayton (256×256) | | | | CVUSA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ⇆ | | SSIM | PSNR | SD | KL | SSIM | PSNR | SD | KL | SSIM | PSNR | SD | KL |
| a2g | Zhai *et al.* [47] | - | - | - | - | - | - | - | - | 0.4147* | 17.4886* | 16.6184* | 27.43 ± 1.63* |
| | Pix2pix [16] | 0.4808* | 19.4919* | 16.4489* | 6.29 ± 0.80* | 0.4180* | 17.6291* | 19.2821* | 38.26 ± 1.88* | 0.3923* | 17.6578* | 18.5239* | 59.81 ± 2.12* |
| | X-Fork [34] | 0.4921* | 19.6273* | 16.4928* | 3.42 ± 0.72* | 0.4963* | 19.8928* | 19.4533* | 6.00 ± 1.28* | 0.4356* | 19.0509* | 18.6706* | 11.71 ± 1.55* |
| | X-Seq [34] | 0.5171* | 20.1049* | 16.6836* | 6.22 ± 0.87* | 0.5031* | 20.2803* | 19.5258* | 5.93 ± 1.32* | 0.4231* | 18.8067* | 18.4378* | 15.52 ± 1.73* |
| | SelectionGAN (Ours) | **0.6865** | **24.6143** | **18.2374** | **1.70 ± 0.45** | **0.5938** | **23.8874** | **20.0174** | **2.74 ± 0.86** | **0.5323** | **23.1466** | **19.6100** | **2.96 ± 0.97** |
| g2a | Pix2pix [16] | 0.3675* | 20.5135* | 14.7813* | 6.39 ± 0.90* | 0.2693* | 20.2177* | 16.9477* | 7.88 ± 1.24* | - | - | - | - |
| | X-Fork [34] | 0.3682* | 20.6933* | 14.7984* | 4.45 ± 0.84* | 0.2763* | 20.5978* | 16.9962* | 6.92 ± 1.15* | - | - | - | - |
| | X-Seq [34] | 0.3663* | 20.4239* | 14.7657* | 7.20 ± 0.92* | 0.2725* | 20.2925* | 16.9285* | 7.07 ± 1.19* | - | - | - | - |
| | SelectionGAN (Ours) | **0.5118** | **23.2657** | **16.2894** | **2.25 ± 0.56** | **0.3284** | **21.8066** | **17.3817** | **3.55 ± 0.87** | - | - | - | - |

$G_s$. 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 3×3 and 1×1, respectively. We adopt PatchGAN [16] for the discriminator $D$.

**Training Details.** Following [34], we use RefineNet [23] and [51] to generate segmentation maps on Dayton and Ego2Top datasets as training data, respectively. We follow the optimization method in [11] to optimize the proposed SelectionGAN, *i.e.* one gradient descent step on discriminator and generators alternately. We first train $G_i$, $G_s$, $G_a$ with $D$ fixed, and then train $D$ with $G_i$, $G_s$, $G_a$ fixed. The proposed SelectionGAN is trained and optimized in an end-to-end fashion. We employ Adam [21] with momentum terms $\beta_1$=0.5 and $\beta_2$=0.999 as our solver. The initial learning rate for Adam is 0.0002. The network initialization strategy is Xavier [10], weights are initialized from a Gaussian distribution with standard deviation 0.2 and mean 0.

# 4. Experiments

## 4.1. Experimental Setting

**Datasets.** We perform the experiments on three different datasets: (i) For the Dayton dataset [41], following the same setting of [34], we select 76,048 images and create a train/test split of 55,000/21,048 pairs. The images in the original dataset have 354×354 resolution. We resize them to 256×256; (ii) The CVUSA dataset [43] consists of 35,532/8,884 image pairs in train/test split. Following [47, 34], 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 Ego2Top dataset [1] 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 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 [16, 34]. All images are scaled to 256×256, and we enabled image flipping and random crops for data augmentation. Similar to [34], the low resolution (64×64) experiments on Dayton dataset are carried out for 100 epochs with batch size of 16, whereas the high resolution (256×256) experiments for this dataset are trained for 35 epochs with batch size of 4. For the CVUSA dataset, we follow the same setup as in [47, 34], and train our network for 30 epochs with batch size of 4. For the Ego2Top dataset, all models are trained with 10 epochs using batch size 8. In our experiment, we set $\lambda_{tv}$=1$e$−6, $\lambda_1$=100, $\lambda_2$=1, $\lambda_3$=200 and $\lambda_4$=2 in Eq. (9), and $\lambda$=4 in Eq. (8). The number of attention channels $N$ in Eq. (5) is set to 10. The proposed SelectionGAN is implemented in PyTorch. We perform our experiments on Nvidia GeForce GTX 1080 Ti GPU with 11GB memory to accelerate both training and inference.

**Evaluation Protocol.** Similar to [34], we employ Inception Score, top-k prediction accuracy and KL score for the quantitative analysis. These metrics evaluate the generated

Table 2: Accuracies of different methods. For this metric, higher is better. (*) These results are reported in [34].

| Dir. ⇆ | Method | Dayton (64×64) | | | | Dayton (256×256) | | | | CVUSA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 Accuracy (%) | | Top-5 Accuracy (%) | | Top-1 Accuracy (%) | | Top-5 Accuracy (%) | | Top-1 Accuracy (%) | | Top-5 Accuracy (%) | |
| a2g | Zhai *et al.* [47] | - | - | - | - | - | - | - | - | 13.97* | 14.03* | 42.09* | 52.29* |
| | Pix2pix [16] | 7.90* | 15.33* | 27.61* | 39.07* | 6.80* | 9.15* | 23.55* | 27.00* | 7.33* | 9.25* | 25.81* | 32.67* |
| | X-Fork [34] | 16.63* | 34.73* | 46.35* | 70.01* | 30.00* | 48.68* | 61.57* | 78.84* | 20.58* | 31.24* | 50.51* | 63.66* |
| | X-Seq [34] | 4.83* | 5.56* | 19.55* | 24.96* | 30.16* | 49.85* | 62.59* | 80.70* | 15.98* | 24.14* | 42.91* | 54.41* |
| | SelectionGAN (Ours) | **45.37** | **79.00** | **83.48** | **97.74** | **42.11** | **68.12** | **77.74** | **92.89** | **41.52** | **65.51** | **74.32** | **89.66** |
| g2a | Pix2pix [16] | 1.65* | 2.24* | 7.49* | 12.68* | 10.23* | 16.02* | 30.90* | 40.49* | - | - | - | - |
| | X-Fork [34] | 4.00* | 16.41* | 15.42* | 35.82* | 10.54* | 15.29* | 30.76* | 37.32* | - | - | - | |
| | X-Seq [34] | 1.55* | 2.99* | 6.27* | 8.96* | 12.30* | 19.62* | 35.95* | 45.94* | - | - | - | - |
| | SelectionGAN (Ours) | **14.12** | **51.81** | **39.45** | **74.70** | **20.66** | **33.70** | **51.01** | **63.03** | - | - | - | - |

Table 3: Inception Score of different methods. For this metric, higher is better. (*) These results are reported in [34].

| Dir. ⇆ | Method | Dayton (64×64) | | | Dayton (256×256) | | | CVUSA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | all classes | Top-1 class | Top-5 classes | all classes | Top-1 class | Top-5 classes | all classes | Top-1 class | Top-5 classes |
| a2g | Zhai *et al.* [47] | - | - | - | - | - | - | 1.8434* | 1.5171* | 1.8666* |
| | Pix2pix [16] | 1.8029* | 1.5014* | 1.9300* | 2.8515* | 1.9342* | 2.9083* | 3.2771* | 2.2219* | 3.4312* |
| | X-Fork [34] | 1.9600* | 1.5908* | 2.0348* | **3.0720*** | 2.2402* | 3.0932* | 3.4432* | 2.5447* | 3.5567* |
| | X-Seq [34] | 1.8503* | 1.4850* | 1.9623* | 2.7384* | 2.1304* | 2.7674* | **3.8151*** | 2.6738* | **4.0077*** |
| | SelectionGAN (Ours) | **2.1606** | **1.7213** | **2.1323** | 3.0613 | **2.2707** | **3.1336** | 3.8074 | **2.7181** | 3.9197 |
| | Real Data | 2.3534 | 1.8135 | 2.3250 | 3.8319 | 2.5753 | 3.9222 | 4.8741 | 3.2959 | 4.9943 |
| g2a | Pix2pix [16] | 1.7970* | 1.3029* | 1.6101* | 3.5676* | 2.0325* | 2.8141* | - | - | - |
| | X-Fork [34] | 1.8557* | 1.3162* | 1.6521* | 3.1342* | 1.8656* | 2.5599* | - | - | - |
| | X-Seq [34] | 1.7854* | 1.3189* | 1.6219* | **3.5849*** | 2.0489* | 2.8414* | - | - | - |
| | SelectionGAN (Ours) | **2.1571** | **1.4441** | **2.0828** | 3.2446 | **2.1331** | **3.4091** | - | - | - |
| | Real Data | 2.3015 | 1.5056 | 2.2095 | 3.7196 | 2.3626 | 3.8998 | - | - | - |

Table 4: Ablations study of the proposed SelectionGAN.

| Baseline | Setup | SSIM | PSNR | SD |
|---|---|---|---|---|
| A | $I_a \xrightarrow{G_i} I_g'$ | 0.4555 | 19.6574 | 18.8870 |
| B | $S_g \xrightarrow{G_i} I_g'$ | 0.5223 | 22.4961 | 19.2648 |
| C | $[I_a, S_g] \xrightarrow{G_i} I_g'$ | 0.5374 | 22.8345 | 19.2075 |
| D | $[I_a, S_g] \xrightarrow{G_i} I_g' \xrightarrow{G_s} S_g'$ | 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** |

images from a high-level feature space. We also employ pixel-level similarity metrics to evaluate our method, *i.e.* Structural-Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Sharpness Difference (SD).

## 4.2. Experimental Results

**Baseline Models.** We conduct ablation study in a2g (aerial-to-ground) direction on Dayton dataset. 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 4. Baseline A uses a Pix2pix structure [16] 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 map $S_g$. Baseline C also uses the Pix2pix structure, and inputs the combination of a conditional image $I_a$ and the target semantic map $S_g$ to the generator $G_i$. Baseline D uses the proposed cycled semantic generation upon Baseline C. Baseline E represents the


Figure 5: Qualitative results of coarse-to-fine generation on CVUSA dataset.

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 a scene image with a new viewpoint. 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 ablation study are shown in Table 4. We observe that Baseline B is better than baseline A since $S_g$ contains more structural information

Table 5: Quantitative results on Ego2Top dataset. For all metrics except KL score, higher is better.

| Method | SSIM | PSNR | SD | Inception Score | | | Accuracy | | | | KL Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | all classes | Top-1 class | Top-5 classes | Top-1 | | Top-5 | | |
| Pix2pix [16] | 0.2213 | 15.7197 | 16.5949 | 2.5418 | 1.6797 | 2.4947 | 1.22 | 1.57 | 5.33 | 6.86 | $120.46 \pm 1.94$ |
| X-Fork [34] | 0.2740 | 16.3709 | 17.3509 | 4.6447 | 2.1386 | 3.8417 | 5.91 | 10.22 | 20.98 | 30.29 | $22.12 \pm 1.65$ |
| X-Seq [34] | 0.2738 | 16.3788 | 17.2624 | 4.5094 | 2.0276 | 3.6756 | 4.78 | 8.96 | 17.04 | 24.40 | $25.19 \pm 1.73$ |
| SelectionGAN (Ours) | **0.6024** | **26.6565** | **19.7755** | **5.6200** | **2.5328** | **4.7648** | **28.31** | **54.56** | **62.97** | **76.30** | $\mathbf{3.05 \pm 0.91}$ |
| Real Data | - | - | - | 6.4523 | 2.8507 | 5.4662 | - | - | - | - | - |

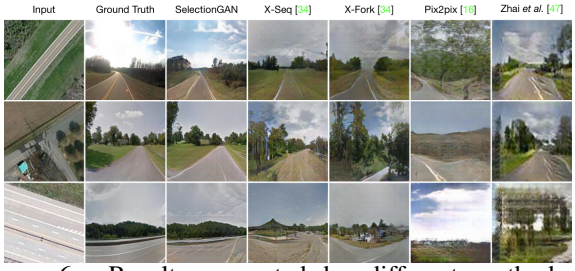


Figure 6: Results generated by different methods in $256\times256$ resolution in a2g direction on CVUSA dataset.

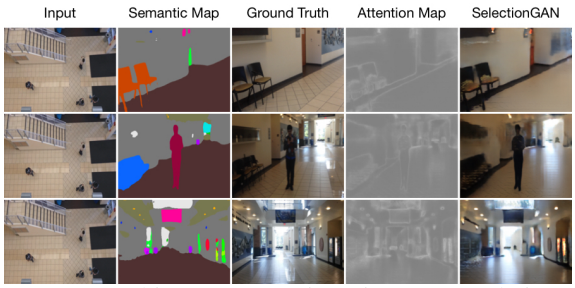


Figure 7: Arbitrary cross-view image translation on Ego2Top dataset.

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 information; By using the proposed cycled semantic generation, Baseline D further improves over C, meaning that the proposed semantic cycle structure indeed utilizes the semantic 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 labels produced from the pretrained segmentation model; 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 4 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. 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.

**State-of-the-art Comparisons.** We compare our SelectionGAN with four recently proposed state-of-the-art methods, which are Pix2pix [16], Zhai *et al.* [47], X-Fork [34] and X-Seq [34]. The comparison results are shown in Tables 1, 2, 3, and 5. We can observe the significant improvement of SelectionGAN in these tables. SelectionGAN consistently outperforms Pix2pix, Zhai *et al.*, X-Fork and X-Seq on all the metrics except for Inception Score. In some cases in Table 3 we achieve a slightly lower performance as compared with X-Seq. However, we generate much more photo-realistic results than X-Seq as shown in Fig. 4 and 6.

**Qualitative Evaluation.** The qualitative results in higher resolution on Dayton and CVUSA datasets are shown in Fig. 4 and 6. 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, 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 truths in layout and structure, such as the results in a2g direction in Fig. 4 and 6.

**Arbitrary Cross-View Image Translation.** Since Dayton and CVUSA datasets only contain two views in one scene, *i.e.* aerial and ground views. We further use the Ego2Top dataset to conduct the arbitrary cross-view image translation experiments. The quantitative and qualitative results are shown in Table 5 and Fig. 7, respectively. Given an image and some novel semantic maps, SelectionGAN is able to generate the same scene but with different viewpoints.

## 5. Conclusion

We propose the Multi-Channel Attention Selection GAN (SelectionGAN) to address a novel image synthesizing task by conditioning on a reference image and a target semantic map. 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 scene and Stage II focus on more appearance details via the proposed multi-channel attention selection module. We also propose an uncertainty map-guided pixel loss to solve the inaccurate semantic labels issue for better optimization. Extensive experimental results on three public datasets demonstrate that our method obtains much better results than the state-of-the-art.

**Acknowledgements:** This research was partially supported by National Institute of Standards and Technology Grant 60NANB17D191 (YY, JC), Army Research Office W911NF-15-1-0354 (JC) and gift donation from Cisco *Inc* (YY).

## References


[1] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV*, 2016. 1, 2, 6

[2] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *ECCV*, 2014. 4

[3] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. In *ECCV*, 2018. 3

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[5] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[7] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 4

[8] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*, 2018. 3

[9] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE TPAMI*, 39(4):692–705, 2017. 3

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *ICAIS*, 2010. 6

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 6

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NIPS*, 2017. 2

[13] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *MIT Press Neural computation*, 18(7):1527–1554, 2006. 2

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov. A better way to pretrain deep boltzmann machines. In *NIPS*, 2012. 2

[15] Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017. 1

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 3, 5, 6, 7, 8

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 5

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 1

[21] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[22] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro. Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In *ECCV*, 2018. 3

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 6

[24] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *CVPR*, 2018. 3

[25] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016. 3

[26] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018. 3

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[28] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018. 3

[29] Augustus Odena. Semi-supervised learning with generative adversarial networks. In *ICML Workshop*, 2016. 3

[30] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 3

[31] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017. 1

[32] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

[33] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, 2016. 3

[34] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[35] David E Rumelhart and James L McClelland. Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations. 1986. 2

[36] Hao Tang, Xinya Chen, Wei Wang, Dan Xu, Jason J. Corso, Nicu Sebe, and Yan Yan. Attribute-guided sketch generation. In *FG*, 2019. 3

[37] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. In *ACM MM*, 2018. 3

[38] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*, 2019. 3

[39] Hao Tang, Dan Xu, Wei Wang, Yan Yan, and Nicu Sebe. Dual generator generative adversarial networks for multi-domain image-to-image translation. In *ACCV*, 2018. 3

[40] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 1, 3

[41] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *ECCV*, 2016. 1, 2, 6

[42] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 2

[43] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *ICCV*, 2015. 1, 2, 6

[44] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 3

[45] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, 2015. 1

[46] Xiaochuan Yin, Henglai Wei, Xiangwei Wang, Qijun Chen, et al. Novel view synthesis for large-scale scene using adversarial loss. *arXiv preprint arXiv:1802.07064*, 2018. 3

[47] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *CVPR*, 2017. 1, 2, 3, 6, 7, 8

[48] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 2

[49] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 3

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 4

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 6

[52] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 1, 3

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3

[54] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative adversarial frontal view to bird view synthesis. In *3DV*, 2018. 1