Attention-Based Spatial Guidance for Image-to-Image Translation

Yu Lin, Yigong Wang, Yifan Li, Yang Gao, Zhuoyi Wang, Latifur Khan
University of Texas at Dallas
800 W. Campbell Road, Richardson, Texas
yxl163430,yxw158830,yli,yxg122530,zxw151030,lkhan@utdallas.edu

Abstract

The aim of image-to-image translation algorithms is to tackle the challenges of learning a proper mapping function across different domains. Generative Adversarial Networks (GANs) have shown superior ability to handle this problem in both supervised and unsupervised ways. However, one critical problem of GAN in practice is that the discriminator is typically much stronger than the generator, which could lead to failures such as mode collapse, diminished gradient, etc. To address these shortcomings, we propose a novel framework, which incorporates a powerful spatial attention mechanism to guide the generator. Specifically, our designed discriminator estimates the probability of realness of a given image, and provides an attention map regarding this prediction. The generated attention map contains the informative regions to distinguish the real and fake images, from the perspective of the discriminator. Such information is particularly valuable for the translation because the generator is encouraged to focus on those areas and produce more realistic images. We conduct extensive experiments and evaluations, and show that our proposed method is both qualitatively and quantitatively better than other state-of-the-art image translation frameworks.

1. Introduction

Generative Adversarial Networks (GANs) [13] have drawn tremendous attention during the past few years, due to their proven ability to generate realistic and sharp looking images. Various computer vision problems are solved using this framework, such as colorization [5], super-resolution [23] and style transfer [44]. All these problems can be considered as an image-to-image translation problem: mapping an image from source domain to target domain. For instance, the super-resolution problem tries to convert a low-resolution image (source domain) to a corresponding high-resolution image (target domain). Existing literatures have show that variants of GAN achieve very impressive results under both supervised and unsupervised settings [7, 21, 24, 38, 46].

Even with such great success, most existing GAN-based approaches are suffering from the imbalance issue between the generator and discriminator [1]. In practice, the discriminator is ordinarily too powerful compared to the generator. As a consequence, the generator may obtain limited gradients from discriminator and is hard to converge. Most state-of-the-art solutions are trying to either find an alternative objective function [2, 15, 27, 32] or plugin some new regularization terms [1, 18, 43]. However, such paradigms ignore the rich information inside the discriminator, which may lead to blurry and artificial regions.

On the other hand, the attention mechanism has been widely adopted in image translation algorithms. Recently, *Mejjati et al.* [29] concatenates an attention network before the generator and mask out the background of the output image, so merely the objects are translated into the target domain. They achieved superior performance on the object-only translation while cannot be easily generalized to scene translation. InstaGAN [31] achieves object deformation (e.g. sheep $\rightarrow$ giraffe) on the image by using the attention mask from an auxiliary network. A contemporary work proposed by *Emami et al.* [11] utilizes the internal activation from the discriminator to guide the translation. However, this approach can only be applied to the unsupervised setting. In our paper, we propose that attention mechanism should not be restricted to object translation and can be further applied to both supervised and unsupervised settings.

Inspired by the close-loop feedback control systems [4], we propose that the high intensity regions in the attention map are more significant during the translation, so that the generator should allocate more resource on these particular areas. Our framework focuses on this key idea, which aims to compute an attention map based on the discriminator's internal activation, and then feed it back to the generator. Imagine that a student is learning how to draw an apple. The standard discriminator, as a painting master, merely grades the student's painting and hopes that can help the student improve his work. On the other hand, another master point

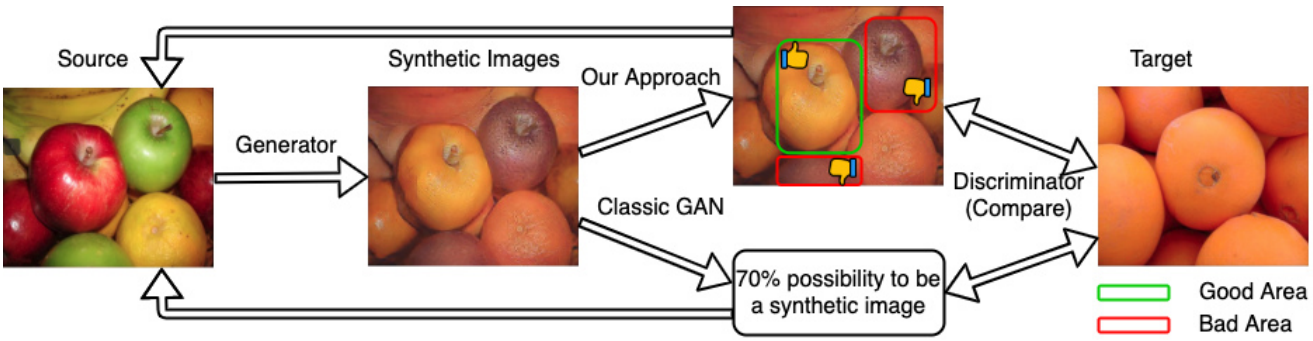

Figure 1. The discriminator distinguish real/fake image based on unrealistic regions. In this paper, we propose a novel framework that utilizes the internal information of the discriminator to enhance generator's capacity.

out areas for the student to improve for the next painting, such as incorrect regions (*e.g.* skin or stem). That is exactly our idea: we believe that the student (generator) would gain benefit from the second master (attention embedded discriminator), which provides better lead regarding spatial guidance. Our main contributions are threefold:

- **A flexible attention-augmented discriminator**: such discriminator provides not only the probability of realness, but also a valuable spatial attention map from its internal activation. We propose two types of attention mechanism in this paper.
- **A unified GAN framework with spatial attention feedback**: we propose two concatenation methods to combine the attention map with raw input 1) Adding an *alpha channel*; 2) compute the *Residual Hadamard Production* of the attention map and raw input. Noted that these methods naturally preserve the information of original input and amplify the signal of crucial regions.
- **Extensive validation on different benchmarks**: we provide extensive experimental validation of our proposed framework on different benchmarks. Both the qualitative results and quantitative comparisons against state-of-the-art methods demonstrate the effectiveness of our approach.

Different from previous approaches, our framework strengthens the communication and guidance between the generator and discriminator. At a high level, our work shed the light upon using auxiliary network attention information to improve the performance of image to image translation, which could be influential to other related research in the future as well.

2. Related Works

Generative Adversarial Network GANs have achieved impressive results in image translation tasks [10, 21, 22, 23, 33]. Typically, GAN consists of two components: a generator and a discriminator. The generator is trained to fool the discriminator, which in turn tries to distinguish between real and synthetic samples. Various improvements to GANs have been proposed regarding different aspects, for instance, improved objective functions [2, 27] and advanced training strategies [14, 32, 39]. A recently proposed framework, *FAL* [20], iteratively improves synthetic images with the signal returned by a well designed spatial discriminative decoder. However, they either don't collect enough information from the discriminator, or are computational expensive because of multiple forward passes.

Image Translation Image-to-image translation can be considered as a generative process conditioned on an input image. *pix2pix* [21] was the first unified framework for supervised image-to-image translation based on conditional GAN (cGAN) [30]. *TextureGAN* [41] solves the sketch-to-image problem using user defined texture patch, and *ContextualGAN* [25] addresses the same problem by learning a joint distribution of the sketch and its image. More recently, *Gonzalez et al.* [12] adopted disentanglement representation to improve the rendering process and *Tang et al.* [36] utilized the extra semantic information to guide the generation.

Despite the promising results they achieved, the above methods are generally not applicable in practice due to the lack of paired data. Several interesting frameworks have been proposed to solve the unsupervised image-to-image translation problem. Cycle consistency loss is first introduced in *CycleGAN* [46] and is then widely used by other unsupervised image translation frameworks. For example, *UNIT* [24] improves the translation with shared latent space assumption, and *MUNIT* [19] later uses it as backbone to handle multi-modal translation. In contrast, our flexible framework can be applied on both supervised and unsupervised settings.

Attention Mechanism Generally, the attention mechanism can be viewed as guidance to bias the allocation of available

processing resources towards the most informative components of an input.It's divided into two categories: post hoc network analysis and trainable attention module. The former scheme has been predominantly employed to access network reasoning for the visual object recognition task [6, 34, 35, 45]. Trainable attention models fall into two main sub-categories, hard (stochastic) that requires reinforcement training and soft (deterministic) that can be trained end-to-end [18, 37, 40].

The attention mechanism is quite useful to solve the image-to-image translation problem. *Ma et al.* [26] use a deep attention encoder to discover the instance level correspondences. *AGGAN* [29] utilizes an auxiliary trainable attention network to separate the instance and background. *InstaGAN* [31] further incorporates the instance information to improve the multi-instance transfiguraiton. Noted that any attention mechanism producing an attention map can be integrated into our framework. Without loss of generality, we implement one representative attention model each category in this paper.

3. Method

3.1. Overview

Consider images from two different domains, source domain $\mathcal{X}$ and target domain $\mathcal{Y}$. Data instances in source domain $x \in \mathcal{X}$ follow the distribution P_x, whereas instances in target domain $y \in \mathcal{Y}$ follow the distribution P_y. Our goal, in the problem setting of image-to-image translation, aims to learn mapping functions across these two different image domains, $G : x \rightarrow y$ and/or $F : y \rightarrow x$, such that the differences between P_x and $F \circ P_y$ and the difference between P_y and $G \circ P_x$ are minimized.

The main idea of our approach is to incorporate a spatial attention map generated by the discriminator, *i.e.*, augment a space of attention map M to the original input space $\mathcal{X}$, to improve the image-to-image translation task. Formally, our approach can be described as a joint-mapping learning from attention-augmented space $\mathcal{X} \oplus M_{\mathcal{X}}$ to $\mathcal{Y}$, and $\mathcal{Y} \oplus M_{\mathcal{Y}}$ to $\mathcal{X}$ if cycle consistency is applied, where $\oplus$ is the concatenate operation. Our method explicitly forces the generator to allocate more processing resources to the attended areas so it can conduct a sharp and clear translation. Generally, our method can be applied to any conditional GAN-based translation.

3.2. Architecture

Our framework, as illustrated in Figure 2, is built upon GAN and attention mechanism. For the supervised learning setting, it consists of three components, a generator G, a discriminator $D_{\mathcal{Y}}$ and an attention transfer block T. It can be extended to unsupervised setting by simply enforcing cycle consistency, which now has five components, including:

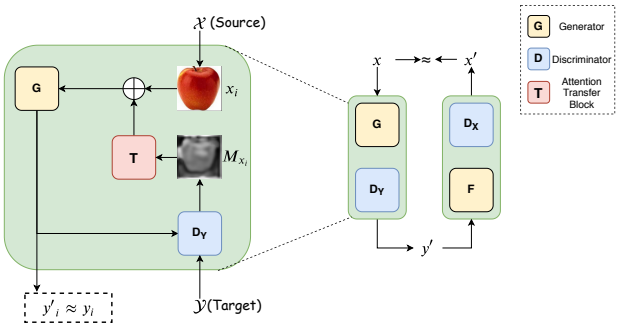

Figure 2. Overview of our framework. **Left**: standard GAN with an attention embedded discriminator. M_x is the attention map provided by the discriminator. The L1 loss between generated y'_i and corresponding ground truth y_i is computed. **Right**: the framework for unsupervised translation using cycle consistency. y_i is not available and the L1 loss between x and x' is calculated instead.

two generators G and F, two domain discriminators $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$, and one shared attention transfer component T.

The training is based on each generator-discriminator pair. Considering a standard GAN, the generator G translates an image x_i in $\mathcal{X}$ to an image in domain $\mathcal{Y}$, and the discriminator $D_{\mathcal{Y}}$ tries to distinguish whether its input is a real or fake image in domain $\mathcal{Y}$. Here, we denote $\hat{y_i} = G(x_i)$ as the output of generator, given x_i. Our attention embedded discriminator not only returns the probability of realness, $D_{\mathcal{Y}}(\hat{y_i}) \in [0, 1]$, but also an attention map A_{x_i} that highlights the attending areas of $D_{\mathcal{Y}}$. This attention map then will be transferred to a pixel-level weight map, M_{x_i} via the attention transfer block T and concatenated with the raw input. It's worth noting that the actual input of our generator G is the concatenation of x_i and M_{x_i}, formulated as $x'_i = x_i \oplus M_{x_i}$. At the start of the training, the attention map of each image is not available so we initialize it as an all-ones matrix $A_{x_i} \in \mathcal{R}^{m \times n}$, where $m \times n$ is the shape of the input image. Other initialization methods, like random noise, have also be examined but have limited impact on the final result. The translation process of generator G can be formulated as:

$$\hat{y}_i^{(k+1)} = G(x_i \oplus T(D(\hat{y}_i^{(k)})); \theta), k = 0, 1, 2, \dots \quad (1)$$

where k and $k + 1$ denote the index of iteration and θ is the parameter of G. Please note that we use the attention feedback from previous iteration for the same input, which is more efficient comparing to *FAL* [20] that requires multiple forward passes per instance. Assume we only provide the raw input to the generator, G may waste its processing resources on some peripheral locations thus $D_{\mathcal{Y}}$ can beat it easily. As a consequence, the loss of the discriminator quickly converges to zero and the generator can no longer efficiently update its parameter. Alternatively, by concate-

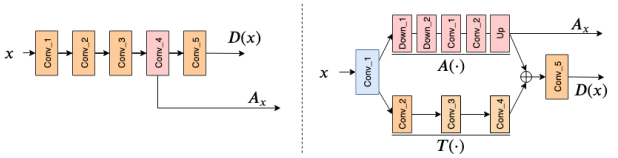

Figure 3. The architecture of different type of discriminator. **Left**: PHA that builds attention map from a specific layer. **Right**: TAM that builds attention map from an additional network branch.

nating the raw input with M_x, the generator knows exactly where the discriminator is noticing and can manage its resources appropriately. As illustrated in Figure 2, we can extend this framework to perform the unsupervised translation by adding another GAN component and enforcing cycle consistency.

3.3. Attention Map

Our discriminator provides an extra attention map A_{x_i} for each image generated from x_i. We consider both *post hoc attention* (PHA) that leaves the discriminator untouched, and *trainable attention module* (TAM) that leads to better distinguishing power.

Given input x, the PHA attention map can be constructed from the backward gradients, forward activation, or the mix of them [34]. We build our discriminator based on the classical PatchGAN [21]. The network is presented as $D = \{l_0, l_1, \ldots, l_m\}$ where l_i denotes i-th convolution layer in the network, and $Act_D = \{a_1, a_2, \ldots, a_m\}$ is the set of activation map of corresponding layer. The PHA attention is sensitive to layer selection, as different layer activation leads to different attention map [28]. Specifically, if l_t is the chosen layer, the attention map can be described as:

$$M = norm(\frac{1}{c}\sum_{i=1}^{c}|a_{t,i}|) \tag{2}$$

where c is the number of channels in t-th layer and $norm(\cdot)$ applies the min-max normalization. As suggested by [28], we chose the 4-th convolution layer in our experiment. This attention map only requires minor computation and works surprisingly well in most cases, but it may not achieve promising results facing complex images (e.g. scene images). On the contrary, a TAM is suitable for such complex input since it simultaneously increases the capacity of generator and discriminator.

Our TAM follows the same 2-branch architecture of the attention block in RAM [37]. Noted that the discriminator is ordinarily powerful than the generator, the enhancement over discriminator must be chosen wisely. Thus, we replaced the *Resblock* [16] by a simple convolution layer. As presented in the right part of Figure 3, first few layers of the discriminator extract the low-level information of the input, and passes it through following branches. Given the trunk branch output $T(x)$ with the input x, the attention branch learns an attention map $A(x)$ that softly weights the output of trunk branch. The output of such module is:

$$E_C = (A_C(x) + 1) \times T_C(x) \tag{3}$$

where C is the set of channels. Finally, a convolutional layer computes the probability of realness based on E and an attention map from the attention branch output, $M = \text{Avg}(A_C(x))$, will be returned.

3.4. Concatenation

In this section, we propose two methods to blend the attention map M_x with its corresponding input x. The first one is based on the aforementioned TAM. We compute the Residual Hadamard Production (RHP) of the attention map and original input. Such operation is superior comparing to dot production because dot production with the weight factor range from zero to one will degrade the pixel value and cause fractional pixel problem [29]. RHP can be formulated as:

$$x' = x \oplus M_x = (g(M_x; \theta) + 1) \times x \tag{4}$$

where $g(;\theta)$ is the attention transfer block T that transfer the attention map to corresponding pixel weight map. It's implemented as a small 3-layers convolution network.

Another intuitive concatenation is called *Alpha* concatenation and is inspired by RGBA and Depth image, which contains three-channel RGB color model supplemented with a 4-th channel that provides additional information, like the opaque level of each pixel. By using this method, the importance of each pixel is observed by the generator explicitly. Formally, it is described as:

$$x' = x \oplus M_x = \{x_r, x_g, x_b, g(M_x; \theta)\} \tag{5}$$

where $g(;\theta)$ is the same transfer function in RHP. Remeber that a gray scale image can be transformed into RGB image by repeating its intensity for each RGB channel. It's worth noting that these two concatenate methods only allow the attention map to amplify the pixel signal, and the generator can always receive the original input. It is crucial for the trick we used during the test since the generator won't completely rely on the attention map.

3.5. Training loss

Let's start with the supervised translation setting. The adversarial loss of a vanilla GAN consists of one generator G and one discriminator D can be expressed as:

$$\begin{aligned} L_{GAN}(G, D) = \; & \mathbb{E}_{y\sim\mathcal{Y}}[\log D(y)] + \\ & \mathbb{E}_{x\sim\mathcal{X}}[\log(1 - D(G(x')))] \end{aligned} \tag{6}$$

where x' is computed from Eq 4 or 5. This cost function is well known for its training difficulty [1]. We adopt the modified least-squares loss [27] to further stabilize the training process and improve the quality of generated images:

$$L_{GAN}(G, D) = \mathbb{E}_{y\sim\mathcal{Y}}[(D(y)-1)^2] + \mathbb{E}_{x\sim\mathcal{X}}[(G(x'))^2] \tag{7}$$

Noted that adversarial loss alone does not guarantee a sound translation. It is beneficial to mix traditional loss like L1 or L2 distance between synthesized image and ground truth. Based on the suggestion from *pix2pix* [21] that L1 loss encourages less blurry, L1 loss has be chosen as part of our supervised training objective:

$$L_{L1}(G) = \mathbb{E}_{x,y}[\|y - G(x')\|_1] \tag{8}$$

The final objective function in this setting is:

$$\arg\min_G \max_D L_{GAN}(G, D) + \lambda L_{L1}(G) \tag{9}$$

We can extend this framework to further conduct the unsupervised translation task by adding another pair of generator and discriminator, and enforcing cycle consistency. Assume the generator G simulates the map function $G : \mathcal{X} \to \mathcal{Y}$ and discriminator $D_{\mathcal{Y}}$ are trying to distinguish between $G(x)$ and y, the objective of this GAN component is $L_{GAN}(G, D_{\mathcal{Y}})$. The generator F and discriminator $D_{\mathcal{X}}$ is doing the same task in the opposite direction, its loss function is $L_{GAN}(F, D_{\mathcal{X}})$. Cycle consistency is employed in such unsupervised setting because it alleviate the shortness of paired data. It assumes that if a image x from domain $\mathcal{X}$ has be translated to a fake image in domain $\mathcal{Y}$, we should get the same image x by applying $F : \mathcal{Y} \to \mathcal{X}$. This behavior is formally presented as:

$$L_{cyc}(G, F) = \mathbb{E}_{x\sim\mathcal{X}}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y\sim\mathcal{Y}}[\|G(F(y)) - y\|_1] \tag{10}$$

The final objective in the unsupervised setting is:

$$\arg\min_{G,F} \max_{D_{\mathcal{X}},D_{\mathcal{Y}}} L_{GAN}(G, D_{\mathcal{Y}}) + L_{GAN}(F, D_{\mathcal{X}}) + \lambda L_{cyc}(G, F) \tag{11}$$

4. Experiment

To verify the effectiveness of our proposed framework, we evaluate it on both unsupervised setting and supervised setting in this paper. The source code is available at `https://github.com/voidstrike/ASGIT`

A crucial point of our framework is how can we perform the inference in test phase. The attention map of each image is not available beforehand, and some placeholders are required. Based on the training phase and the concatenation in Sec 3.4, an all-one attention is used as the placeholder as we assume the whole image is important by default.

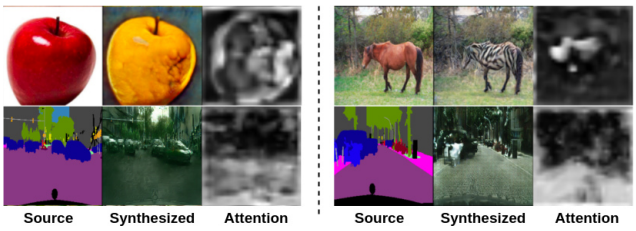

Figure 4. Examples of attention maps. **Left**: Attention maps generated by PHA; **Right**: Attention maps generated by TAM.

4.1. Settings

4.1.1 Datasets

We evaluate our method on four benchmarks for unsupervised translation. *orange2apple*, *horse2zebra* [9] are for object transfer; *summer2winter* [46] and *day2night* are two challenging scenery tasks. *day2night* contains 7870 daytime street images and 8592 night street images cropped from BDD110k [42]. Furthermore, we evaluate on *Cityscape* [8] for both supervised and unsupervised translations. All data are randomly split into train and test (80/20 split).

4.1.2 Baselines

For the unsupervised translation setting, we compare our framework to CycleGAN [46] that enforces cycle consistency, and UNIT [24] that leverages the latent space assumption between source/target images. Also, we compare with StarGAN [7], which is capable for multiple domains translation. Additionally, we include AGGAN algorithm [29] in the comparison, which separates the foreground and background via an attention network.

For the supervised translation setting, we compare to GAN [13] and cGAN [30]. The only difference between them is that cGAN is conditioned on the input. We also consider pix2pix [21] in the comparison, which extends cGAN by adding a reconstruction loss. Moreover, we compare with FAL [20], which iteratively modifies the hidden feature according to the discriminator's feedback.

4.1.3 Metrics

To be comparable with previous approaches [21, 29, 46], FCN score is computed to evaluate *Cityscape* tasks and Kernel Inception Distance (KID) [3] is for unsupervised translation. KID computes the squared MMD (Maximum Mean Discrepancy) between feature representations of real and

generated images. Different from the Fréchet Inception Distance [17], KID is more reliable because of the unbiased estimator. While KID is unbounded, the lower its value, the more shared visual similarities there are between real and generated images.

Two types of KID are reported based on the task. Target-KID measures the distance between generated images and target domain, while fused-KID denotes the distance between synthesized images and both domains. Generally, target-KID is suitable for object translation since we only care about the target object rather than the background. On the other hand, fused-KID is good for scenery task because both foreground and background matter [29].

4.1.4 implementation

To be comparable with previous methods [20, 21, 46], we use 256×256 images for the *unsupervised Cityscape translation* and all object and scenery tasks, and 128×128 images for the *supervised Cityscape translation*. In the preprocessing step, we resized the input image to 286×286 (143×143) then randomly cropping back to 256×256 (128×128). For all the unsupervised experiments, we set the weight factor of the GAN loss to 1, $\lambda_{GAN} = 1$, and the weight factor of cycle consistency to 10, $\lambda_{Cyc} = 10$. On the other hands, we set $\lambda_{GAN} = 1$ and $\lambda_{L1} = 100$. for the supervised setting.

We used Adam optimizer with batch size 1, training on a Quadro 8000 GPU. All networks were trained from scratch, with learning rate of 0.0002 for both the generator and discriminator, and $\beta_1 = 0.5$, $\beta_2 = 0.999$ for the optimizer. We kept learning rate for first 100 epochs and linearly decayed to 0 for next 100 epochs.

4.2. Attention and Concatenation

Recall that we implement two attention mechanisms and two concatenations for our experiment. The remaining problem is how to combine them properly. We first present qualitative results in Figure 5. As discussed in Section 3, TAM is not good at handling simple datasets, e.g. *apple2orange*, while the results are more attractive for more complex *summer2winter* dataset. By comparing alpha concatenation with RHP under post hoc attention, we find that the contrast ratio of the synthesized image is usually too high and leads to unrealistic images.

We also present attention map examples in Figure 4 and a quantitative evaluation for each combination in Table 1. Numerical results in the table justify our previous observations. Based on the overall performance across different tasks, most experiments use PHA and RHP in the following sections.

Figure 5. Different combination of attention and concatenation on *apple2orange* and *summer2winter*. First column is the real input. From second column to the right: PHA and alpha, PHA and RHP, TAM and alpha, TAM and RHP

4.3. Object and Scenery Translation

We present target-KID in Table 2 and fused-KID in Table 3. Our proposed framework outperforms all baselines in all tasks except *day→night*. Nevertheless, our result is very close to the winner. This observation is consistent with our qualitative evaluation in Figure 6, where our fake horse (zebra) is much more realistic than the counterparts produced by baselines. However, our method dramatically changes the background comparing to other methods, which means it is a better choice if the background doesn't play a important role in the translation.

Scenery translation results are presented in Figure 6. It's surprising to see the simplest CycleGAN model got first place in *day2night*, which is harder than two aforementioned object transfer datasets. Notwithstanding, CycleGAN got 2nd place on night→day, which is commonly considered easier. Another interesting observation is AGGAN does not perform any translation these cases. Based on the idea of AGGAN, it will decompose the image into foreground and background. But a proper 'foreground' cannot be found in scenery translation, thus no translation can be conduct. To sum up, our method produces more realistic scenery images comparing to baselines.

4.4. Cityscape translation

We evaluate our method on *Cityscape* [8] for both supervised and unsupervised settings. We train *photo→label* and *label→photo* on the *Cityscape*, and compare the output images with the ground truth.

As shown in Table 4, our method significantly outper-

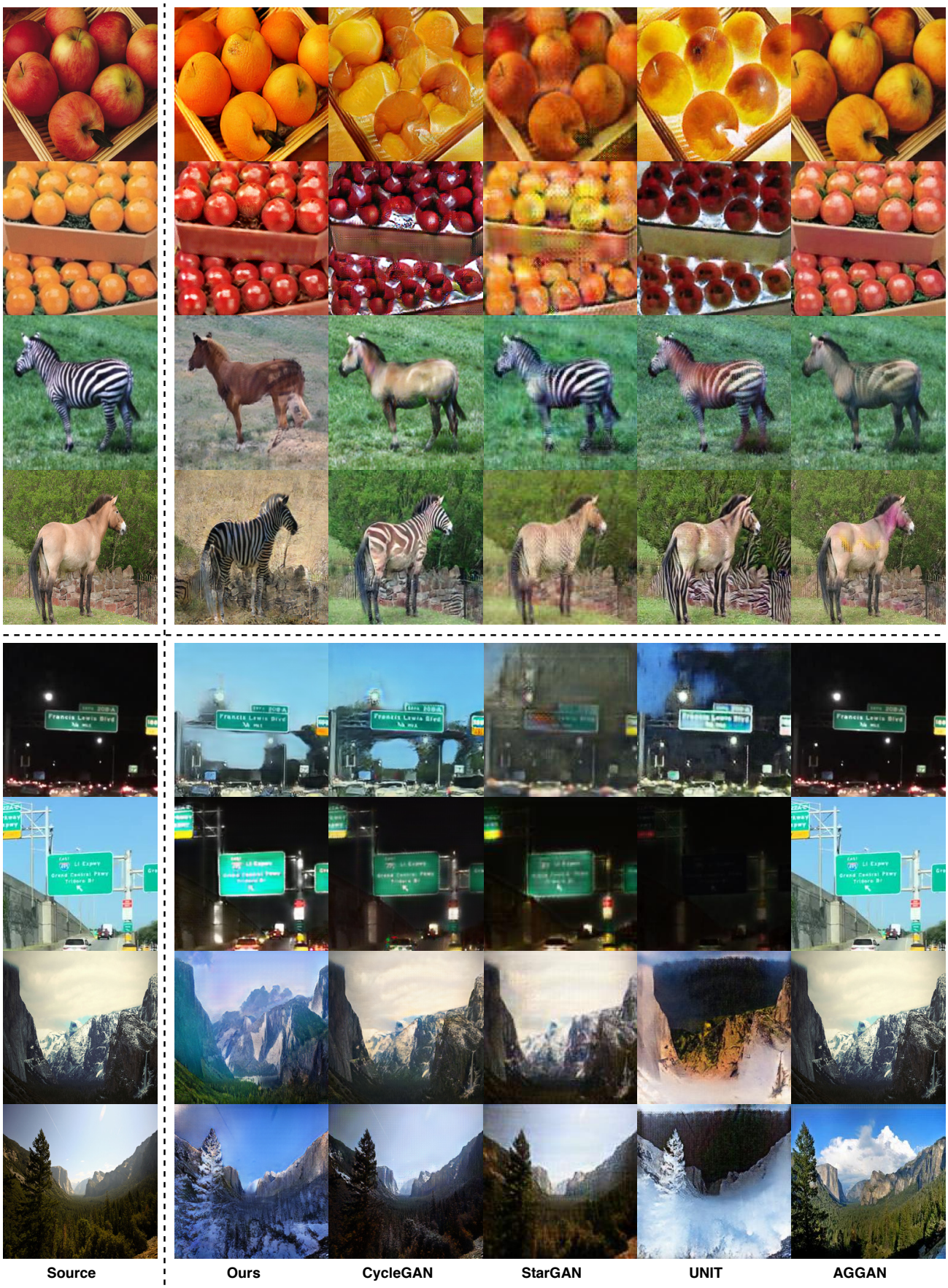

Figure 6. Image-to-Image translation results generated by different approaches on object translation and scenery translation. Every two rows from top: *apple↔orange*, *zebra↔horse*, *night↔day* and *winter↔summer*. More result is available in the supplementary

Method	(A)pple↔(O)range		(S)ummer↔(W)inter		(A)pple↔(O)range		(S)ummer↔(W)inter	
	A→O	O→A	S→W	W→S	A→O	O→A	S→W	W→S
PHA+Alpha	7.25 ± 0.83	3.69 ± 0.41	1.86 ± 0.24	**1.01 ± 0.23**	4.02 ± 0.37	**4.11 ± 0.31**	1.04 ± 0.12	**1.23 ± 0.12**
PHA+RHP	**6.31 ± 0.60**	**2.99 ± 0.38**	1.98 ± 0.33	1.03 ± 0.26	**3.69 ± 0.27**	4.30 ± 0.31	1.18 ± 0.16	1.55 ± 0.13
TAM+Alpha	10.80 ± 0.71	7.26 ± 0.47	2.37 ± 0.35	1.76 ± 0.37	5.93 ± 0.31	6.70 ± 0.36	1.45 ± 0.18	1.71 ± 0.15
TAM+RHP	10.06 ± 0.64	6.81 ± 0.45	**1.34 ± 0.29**	1.73 ± 0.30	5.54 ± 0.31	6.47 ± 0.37	**0.82 ± 0.14**	1.72 ± 0.15

Table 1. KID ± std. (scaled by 100) computed for different combination on *apple2orange* and *summer2winter*. Left 4 columns shown the target-KID and the rest 4 columns show the fused-KID (Lower the better).

Method	(A)pple↔(O)range		(H)orse↔(Z)ebra		(D)ay↔(N)ight		(S)ummer↔(W)inter	
	A→O	O→A	H→Z	Z→H	D→N	N→D	S→W	W→S
CycleGAN	8.48 ± 0.53	5.94 ± 0.65	3.94 ± 0.41	4.87 ± 0.52	**2.63 ± 0.20**	7.68 ± 0.35	2.78 ± 0.22	1.86 ± 0.26
StarGAN	13.32 ± 0.52	11.19 ± 0.51	12.42 ± 0.74	12.21 ± 0.89	5.37 ± 0.43	8.49 ± 0.34	8.05 ± 0.37	8.72 ± 0.47
AGGAN	10.61 ± 0.79	4.57 ± 0.30	4.12 ± 0.80	4.46 ± 0.40	8.09 ± 0.37	7.85 ± 0.29	3.45 ± 0.43	2.75 ± 0.20
UNIT	17.41 ± 1.13	7.26 ± 0.57	12.25 ± 0.74	12.37 ± 0.84	2.83 ± 0.30	11.00 ± 0.53	6.20 ± 0.25	5.99 ± 0.28
Ours (PHA+RHP)	**6.31 ± 0.60**	**2.99 ± 0.38**	**1.03 ± 0.35**	**3.42 ± 0.51**	2.76 ± 0.32	**6.96 ± 0.38**	**1.98 ± 0.33**	**1.03 ± 0.26**

Table 2. Target KID ± std. (scaled by 100) computed for different methods and on different datasets. Best results are bolded.

Method	(A)pple↔(O)range		(H)orse↔(Z)ebra		(D)ay↔(N)ight		(S)ummer↔(W)inter	
	A→O	O→A	H→Z	Z→H	D→N	N→D	S→W	W→S
CycleGAN	11.02 ± 0.60	9.82 ± 0.51	10.25 ± 0.25	11.44 ± 0.38	**1.95 ± 0.13**	**3.63 ± 0.20**	2.05 ± 0.12	3.34 ± 0.12
StarGAN	9.15 ± 0.43	8.31 ± 0.48	7.14 ± 0.48	4.50 ± 0.36	3.43 ± 0.20	5.18 ± 0.23	3.95 ± 0.17	4.14 ± 0.21
AGGAN	6.44 ± 0.69	5.32 ± 0.48	6.93 ± 0.27	6.71 ± 0.27	4.14 ± 0.14	4.97 ± 0.18	3.15 ± 0.19	2.45 ± 0.13
UNIT	11.68 ± 0.43	10.48 ± 0.67	**4.91 ± 0.36**	**4.39 ± 0.33**	2.48 ± 0.16	6.12 ± 0.29	3.51 ± 0.15	2.83 ± 0.12
Ours (PHA+RHP)	**3.69 ± 0.27**	**4.30 ± 0.31**	8.42 ± 0.47	8.46 ± 0.41	2.48 ± 0.15	4.58 ± 0.23	**1.18 ± 0.16**	**1.55 ± 0.13**

Table 3. Fused KID ± std. (scaled by 100) computed for different methods and on different datasets. Best results are bolded.

Method	Label→Photo			Photo→Label		
	Per-pixel acc.	Per-class acc.	IoU	Per-pixel acc.	Per-class acc.	IoU
CycleGAN	0.42	0.15	0.10	0.56	0.21	0.17
UNIT	0.48	0.17	0.11	0.58	0.18	0.14
AGGAN	0.37	0.11	0.09	0.49	0.14	0.10
StarGAN	0.47	0.16	0.11	**0.61**	0.21	0.17
Ours (PHA)	**0.52**	**0.20**	**0.12**	0.60	**0.24**	**0.19**
Ours (TAM)	0.49	0.19	0.10	0.59	0.23	**0.19**

Table 4. FCN-scores (Higher is better) for different methods, evaluated on *Cityscape* label↔photos in unsupervised setting.

Method	Label→Photo			Photo→Label		
	Per-pixel acc.	Per-class acc.	IoU	Per-pixel acc.	Per-class acc.	IoU
GAN	0.22	0.05	0.01	0.32	0.08	0.02
cGAN	0.57	0.20	0.14	0.71	0.26	0.21
FAL	0.57	0.18	0.13	0.77	0.25	0.21
pix2pix	0.61	0.22	**0.16**	0.80	**0.43**	**0.32**
Ours(PHA)	**0.63**	**0.23**	**0.16**	**0.81**	0.42	**0.32**
Ours(TAM)	**0.63**	0.22	**0.16**	0.75	0.40	0.30

Table 5. FCN-scores (Higher is better) for different methods, evaluated on *Cityscape* label↔photos in supervised setting.

forms the baselines in the unsupervised experiments. The compelling improvement in the pixel-level accuracy comes from the guidance of the attention map, which aligns with our expectations. However, the improvement of other metrics is somehow limited. We suggest that it's because only few domain specific classes are highlighted in the attention map, and the generator works too hard on these objects and ignores others. Another possible cause would be the number of classes per image, which is small in this task and we cannot increase the score for nonexistent classes. Since it's not our major contribution, we leave the justification in the supplementary.

Meanwhile, the improvement of the supervised translation is not as sharp as the unsupervised translation according to Table 5, yet it still shows that further improvement can be achieved with little extra computation. We believe that it's majorly due to the strong regularization enforced by the L1 norm. Note that *pix2pix* and our framework share $\lambda = 100$ in Eq. 9, but FAL has $\lambda = 10$ in their implementation. This may explain why FAL, as a recurrent modification of *pix2pix*, got worse performance. It also further justified that L1 loss may sufficient for the supervised case already.

5. Conclusion

This work argues for spatial attention, which unveils the regions of an image for the discriminator to determine whether that image is real or fake, can significantly improve the performance of GANs on image-to-image translation tasks. It is noteworthy that no additional supervision is needed to generate this attention map. Our method not only shows compelling improvement on both unsupervised and supervised learning tasks compared to state-of-the-art algorithms, but also demonstrates an insightful investigation to the behaviors of GANs. We further remark that our idea can apply on any GAN-based model with little modification. According to our experiment, we observe that the performance of our proposed framework is sensitive to the selection of attention module and concatenation method. Investigating the impact of different attention mechanisms and new tasks could be an interesting research direction in the future.

6. Acknowledgment

The research reported herein was supported in part by NSF awards DMS-1737978, DGE-2039542, OAC-1828467, OAC-1931541, DGE-1906630; and an IBM faculty award (Research).

References

[1] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[4] Zdzislaw Bubnicki. *Modern control theory*, volume 2005925392. Springer, 2005.

[5] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 151–166. Springer, 2017.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[7] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[8] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[11] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia*, 2020.

[12] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, pages 1287–1298, 2018.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[15] Ahsanul Haque, Zhuoyi Wang, Swarup Chandra, Yupeng Gao, Latifur Khan, and Charu Aggarwal. Sampling-based distributed kernel mean matching using spark. In *2016 IEEE Big Data*, pages 462–471, 2016.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[19] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[20] Minyoung Huh, Shao-Hua Sun, and Ning Zhang. Feedback adversarial learning: Spatial feedback for improving generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1476–1485, 2019.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[22] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017.

[23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[24] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[25] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 205–220, 2018.

[26] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2018.

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[28] Xiaoguang Mei, Erting Pan, Yong Ma, Xiaobing Dai, Jun Huang, Fan Fan, Qinglei Du, Hong Zheng, and Jiayi Ma. Spectral-spatial attention networks for hyperspectral image classification. *Remote Sensing*, 11(8):963, 2019.

[29] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3693–3703, 2018.

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

[31] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. In *ICLR 2019*. ICLR committee, 2019.

[32] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[36] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019.

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[39] Zhuoyi Wang, Yigong Wang, Yu Lin, Evan Delord, and Khan Latifur. Few-sample and adversarial representation learning for continual stream mining. In *Proceedings of The Web Conference 2020*, pages 718–728, 2020.

[40] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[41] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018.

[42] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

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

[44] Lvmin Zhang, Yi Ji, Xin Lin, and Chunping Liu. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 506–511. IEEE, 2017.

[45] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.