GUIDEGAN: ATTENTION BASED SPATIAL GUIDANCE FOR IMAGE-TO-IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Generative Adversarial Network (GAN) and numbers of its variants have been widely used to solve the image-to-image translation problem and achieved extraordinary results in both a supervised and unsupervised manner. However, most GAN-based methods suffer from the imbalance problem between the generator and discriminator in practice. Namely, the relative model capacities of the generator and discriminator do not match, leading to mode collapse and/or diminished gradients. To tackle this problem, we propose a GuideGAN based on attention mechanism. More specifically, we arm the discriminator with an attention mechanism so not only it estimates the probability that its input is real, but also does it create an attention map that highlights the critical features for such prediction. This attention map then assists the generator to produce more plausible and realistic images. We extensively evaluate the proposed GuideGAN framework on a number of image transfer tasks. Both qualitative results and quantitative comparison demonstrate the superiority of our proposed approach.

1 INTRODUCTION

Generative Adversarial Networks (GANs) have drawn much 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 super-resolution (Ledig et al., 2017), colorization (Cao et al., 2017), denoising (Yang et al., 2018) and style transfer (Zhang et al., 2017). 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 of trying to transfer a low-resolution image (source domain) to a corresponding high-resolution image (target domain). Existing literatures have shows that variants of GAN achieved impressive results in both a supervised and unsupervised setting. (Zhu et al., 2017; Liu et al., 2017; Wang et al., 2018; Isola et al., 2017; Choi et al., 2018; Huang et al., 2018)

Even with such great success, most GAN-based approaches are suffering from the imbalance between the generator and discriminator (Arjovsky & Bottou, 2017). In practice, the discriminator is usually too powerful for its task. Thus, the generator obtains very small gradients from discriminator and is hard to converge. Most state-of-the-art solutions are trying to find a new cost function or add some new regularization terms to the cost function, which mainly affect the generator (Arjovsky et al., 2017; Arjovsky & Bottou, 2017; Mao et al., 2017; Nowozin et al., 2016; Zhang et al., 2018; Hu et al., 2018). To address this problem from a different direction, we want to borrow some power from the discriminator by incorporating the attention mechanism to help the generator. In this paper, we propose that the critical locating areas are more significant in the translation. The generator should pay more attention to a particular area of the object rather than the whole image.

Imagine a student is learning how to draw a horse. 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 will provide some additional information. For instance, an error canvas circling each incorrect region. That is exactly our idea; we suggest that the student (generator) gains benefit from the second master (attention embedded discriminator). To achieve our goal in this paper, our main contribution is threefold:

- **A flexible attention-augmented discriminator**: such discriminator provides not only the probability of realness, but an attention map in the perspective of corresponding discriminator. Both trainable attention module and post hoc attention are implemented.
- **A unified GAN framework using attention map**: to improve the translation of the generator, we combine the attention map with corresponding raw input via two concatenate methods. 1) convert the input to a RGBA image by adding an alpha channel 2) residual element-wise production based on RAM. (Wang et al., 2017)
- **Extensive experiment validation on different benchmarks**: we provide extensive experimental validation of GuideGAN on different benchmarks; both the qualitative results and quantitative comparisons against state-of-the-art methods demonstrate the effectiveness of our approach.

To the best of our knowledge, we are the first to report image-to-image translation results using an attention embedded discriminator. Different with previous approaches, our framework strengthens the communication and guidance between the generator and discriminator. At a high level, the significance of our work is also on discovering that the attention information from auxiliary network affects the result of image-to-image translation, which we think would be influential to other related research in the future.

2 Relative Works

Generative Adversarial Network GANs have achieved impressive results in image translation tasks (Denton et al., 2015; Radford et al., 2015; Isola et al., 2017; Kim et al., 2017; Ledig et al., 2017). A classical GAN model 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 synthesised samples. Recently, various improvements to GANs have been proposed. For example, cost functions modification (Mao et al., 2017; Arjovsky et al., 2017), additional regulization terms (Zhu et al., 2017; Hoffman et al., 2017) and advanced training strategies (Gulrajani et al., 2017; Nowozin et al., 2016). Despite the success, little work has been done to collect more information from the discriminator and immigrate that information to the generator.

Image Translation Image to image translation can be considered as a generative process conditioned on an input image. *pix2pix* (Isola et al., 2017) was the first unified framework for supervised image-to-image translation based on conditional GAN (cGAN) (Mirza & Osindero, 2014). *TextureGAN* (Xian et al., 2018) solves the sketch-to-image problem using user defined texture patch and *ContextualGAN* (Lu et al., 2018) addresses the same problem by learning a joint distribution of the sketch and its image. More recently, Gonzalez-Garcia et al. (2018) adopted disentanglement representation to improve the rendering process and Tang et al. (2019) 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 be proposed to solve this unsupervised image-to-image translation problem. Cycle consistency loss was first proposed in *CycleGAN* (Zhu et al., 2017) and is widely used by other unsupervised image translation frameworks. UNIT (Liu et al., 2017) improves the translation with shared latent space assumption, which is the fundamental of MUNIT (Huang et al., 2018) that handles multi-modal translation. In contrast, our flexible framework can be applied on both supervised and unsupervised settings.

Attention Learning Generally, attention can be viewed as guidance to bias the allocation of available processing resources towards the most informative components of an input. Contemporary approaches are 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 (Simonyan et al., 2013; Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018). 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 (Wang et al., 2017; Hu et al., 2018; Woo et al., 2018).

Attention is also useful to solve the image-to-image translation problem. Ma et al. (2018) proposed the DA-GAN framework, which learns a deep attention encoder to discover the instance level correspondences. Mejjati et al. (2018) separates the instance and background using a trainable attention network. *InstaGAN* (Mo et al., 2018) incorporates the instance information, like segmentation

masks, to improve the multi-instance transfiguraiton. Even any attention mechanism producing an attention map can be used in our framework. In this paper, we implemented one representative attention model for post hoc and trainable attention. We directly compare against several state-of-the-art approaches in Section 4.

3 METHOD

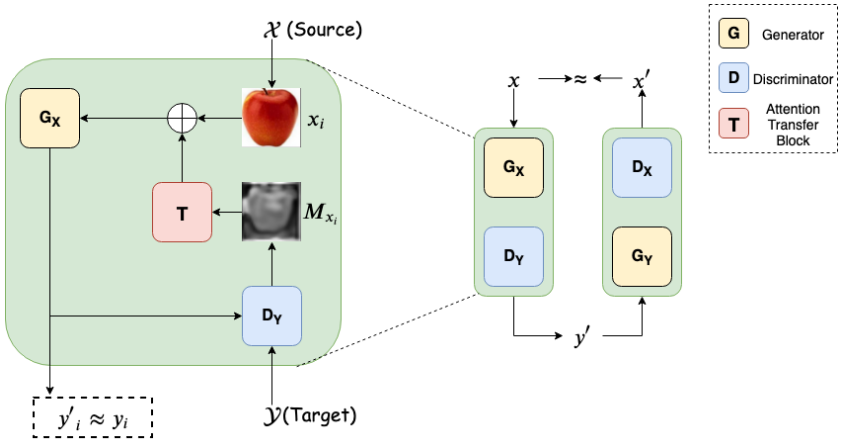

Figure 1: Overview of our framework. Left block is the standard GAN for supervised training. L1 loss between generated y'_i and corresponding ground truth y_i will be computed. Right side is the framework for unsupervised translation using cycle consistency. Ground truth y_i is not available and L1 loss between x and x' is calculated.

Consider images from two different domains, source domain $\mathcal{X}$ and target domain $\mathcal{Y}$. Data instances in source domain $x \in \mathcal{X}$ follows the distribution P_x, whereas instances in target domain $y \in \mathcal{Y}$ follow the distribution P_y. Notice that we do not have labels in both $\mathcal{X}$ and $\mathcal{Y}$. Our goal, in the problem setting of image-to-image translation, aims to learn mapping functions Gs across these two different image domains, $G_X : x \rightarrow y$ and/or $G_Y : y \rightarrow x$, such that differences between P_x and $G_Y \cdot P_y$, P_y and $G_X \cdot P_x$ are minimized. From the perspective of statistics, learning those two mapping functions can also be formulated as estimating the conditional distribution $P(x|y)$ and $P(y|x)$.

The main and unique idea of our approach is to incorporate the attention map generated by the discriminator, *i.e.*, augment a space of attention information A to the original input space X, to improve the image-to-image translation. The attention map can be further transformed to an extra alpha channel α (a mask channel with weight) or be interpreted as a pixel weight map. In this paper, different attention mechanisms and concatenation methods have been studied and achieve promising results based on a different task setting. Formally, our approach can be described as a joint-mapping learning from attention-augmented space $X \oplus A_X$ to Y, and $Y \oplus A_Y$ to X if cycle consistency applied, where $\oplus$ is the concatenate operation. Our method explicitly forces the generator to put more processing resources to attended areas so it can conduct a sharp and clear translation. Generally, this approach can be applied to any conditional GAN-based translation, hence, we call it GuideGAN. We will present the detail of our approach in the following subsections.

3.1 ARCHITECTURE

Our framework, as illustrated in Figure 1, is built upon GAN and attention mechanism. For the supervised learning setting, it consists of three components, a generator G_X, a discriminator D_Y and an attention transfer block T. It can be extended to unsupervised setting using CycleGAN framework easily, which now has five components: including two generators G_X and G_Y, two domain adversarial discriminators D_Y and D_X , and one shared attention transfer component T.

The training is based on each generator-discriminator pair. Considering a standard GAN, the generator G_X translates an image x_i in $\mathcal{X}$ to an image in domain $\mathcal{Y}$ and the discriminator D_Y tries to

distinguish whether its input is a real or fake image in domain $\mathcal{Y}$. Here, we denote $y'_i = G_X(x_i)$ as the output of generator, given x_i. Our attention embedded discriminator not only returns the probability of realness, $D_Y(y'_i) \in [0, 1]$, but also an attention map A_{x_i} that highlights the focusing area of D_Y. This attention map A_{x_i} will be passed to the attention transfer block T to create an alpha channel or a pixel weight map, depending on the concatenation method, which will be described in Section 3.3. For simplicity, the constructed term is denoted as M_{x_i} given A_{x_i}, despite its actual interpolation. Noteworthy is the input of our generator G_X is actually the concatenation of x_i and M_{x_i}, which is represented 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 noisy, have also be examined but have limited impact on the final result. The translation process of generator G_X can be formulated as:

$$y'^{(k+1)}_i = G_X(x_i \oplus T(D(y'^{(k)}_i)); \theta), k = 0, 1, 2, \dots \tag{1}$$

where k and $k + 1$ denote the index of iteration and θ is the parameter of G_X. We emphasize that the attention map from D_Y is crucial because it allows G_X to focus on informative areas. For example, if we only give the generator the raw input, G_X may waste its processing resources on some inessential locations and D_Y will 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 concatenating the raw input with M_x, the generator knows exactly where the discriminator is looking and allocates its processing resources properly on those areas. As illustrated in Figure 1, we can easily extend this framework to perform the unsupervised translation by adding another GAN component and enforcing cycle consistency.

3.2 Attention Map

Remember that our discriminator provides an extra attention map A_{x_i} for each image generated from x_i. Therefore, we consider both *post hoc attention mechanism* that does not change the capacity of the discriminator, and *trainable attention module*, which enhances the discriminator's distinguishing power.

Given input x, the post hoc attention map is constructed from the backward gradients, forward activation, or the mix of them. We use PatchGAN (Isola et al., 2017) as the bone of our discriminator. The network can be formulated as $D = \{l_0, l_1, \dots, l_m\}$ where l_i denotes i-th convolutional layer in the network, and $Act_D = \{a_1, a_2, \dots, a_m\}$ is the set of activation map of corresponding layer. This kind of attention map is sensitive to layer selection; different layer selection leads to different attention map (Mei et al., 2019). More specifically, if t is the layer we chose, the attention map can be described as:

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

where c is the number of channels in t-th layer and $g(\cdot)$ applies the min-max normalization. This attention map only requires minor computation and works surprisingly well in most cases, but it may not achieve promising results when handling complex images. On the contrary, a trainable attention module is suitable for such complex input since it simultaneously increases the capacity of generator and discriminator.

Our trainable attention module follows the same structure of the attention block in RAM (Wang et al., 2017). They built a very deep network with several such blocks, each containing two branches: mask branch and trunk branch. Mask branch cascades the input features through a bottom-up top-down architecture that mimics human attention. Trunk branch is applied as feature processing. Noteworthy is that each branch in their implementation contains several ***Resblock*** (He et al., 2016), which makes it infeasible in our framework. We built our discriminator using these two branch architectures, but the ***Resblock*** is replaced by one convolutional layer. First few layers of the discriminator extract the low-level information of the input, and passes it to following branches. Given the trunk branch output $T(x)$ with the input x, the mask branch learns an attention map $M(x)$ that softly weights the output of trunk branch. We put these two outputs together as:

$$E_{i,c} = (M_{i,c}(x) + 1) \times T_{i,c}(x) \tag{3}$$

where i ranges over all spatial positions and $c \in \{1, 2, ..., C\}$ is the index of channels. Finally, a few consecutive convolutional layers will do the final prediction based on E and attention map $\frac{1}{C}\sum_c M_c(x)$ constructed from mask branch output will be returned.

3.3 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 attention module in the RAM (Wang et al., 2017). We perform a residual element-wise multiplication between the attention map and original input. The reason of this operation is 1) Dot production with the attention range from zero to one will degrade the pixel value and cause fractional pixel problem (Mejjati et al., 2018) 2) Attention mask can potentially break good property of the raw input. This residual element-wise production can be formulated as:

$$x' = (M_x + 1) \times x \tag{4}$$

Another more intuitive concatenation is converting an RGB image to its RGBA version. RGBA, as a color space, stands for red-green-blue-alpha. Namely, it is the three-channel RGB color model supplemented with a 4-th alpha channel that indicates how opaque each pixel is. This concatenation somehow makes nonessential areas more transparent thus highlighting the crucial locations. Formally, this concatenation approach is described as:

$$x' = \{x_r, x_g, x_a, g(M_x; \phi)\} \tag{5}$$

where $g(; \phi)$ is a transfer function that maps attention map to alpha channel. Follow the standard image pre-processing step, this concatenation can also be applied on gray scale image. Gray scale image can be transformed into RGB image by repeating its intensity for each RGB channel.

3.4 TRAINING LOSS

Let's start with supervised translation. The adversarial loss of the generator G and its discriminator D can be expressed as:

$$L_{GAN}(G, D) = \mathbb{E}_{y \sim p_{data}(y)}[\log D(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D(G(x \oplus M_x)))] \tag{6}$$

which is the adversarial loss of vanilla GAN. G aims to minimize this objective while an adversary D tries to maximize it, *i.e.*, $\min_G \max_D L_{GAN}(G, D)$. However, this cost function is well known for its training difficulty. We adopt the modified least-squares loss, proposed in LSGAN (Mao et al., 2017), to further stabilize the training process and improve the quality of generated images. The adversarial loss now becomes:

$$L_{GAN}(G, D) = \mathbb{E}_{y \sim p_{data}(y)}[(D(y) - 1)^2] + \mathbb{E}_{x \sim p_{data}(x)}[(G(x \oplus M_x))^2] \tag{7}$$

Adversarial loss alone does not guarantee a sound translation. It is beneficial to mix traditional loss like L1 distance or L2 distance between synthesized image and ground truth. Based on the suggestion from pix2pix (Isola et al., 2017) that L1 loss encourages less blurry, we chose L1 loss as part of our training objective:

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

The final objective function in this setting is:

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

We can easily extend this framework to conduct unsupervised translation by adding another pair of generator and discriminator and enforcing cycle consistency. Assume the generator G_X simulates

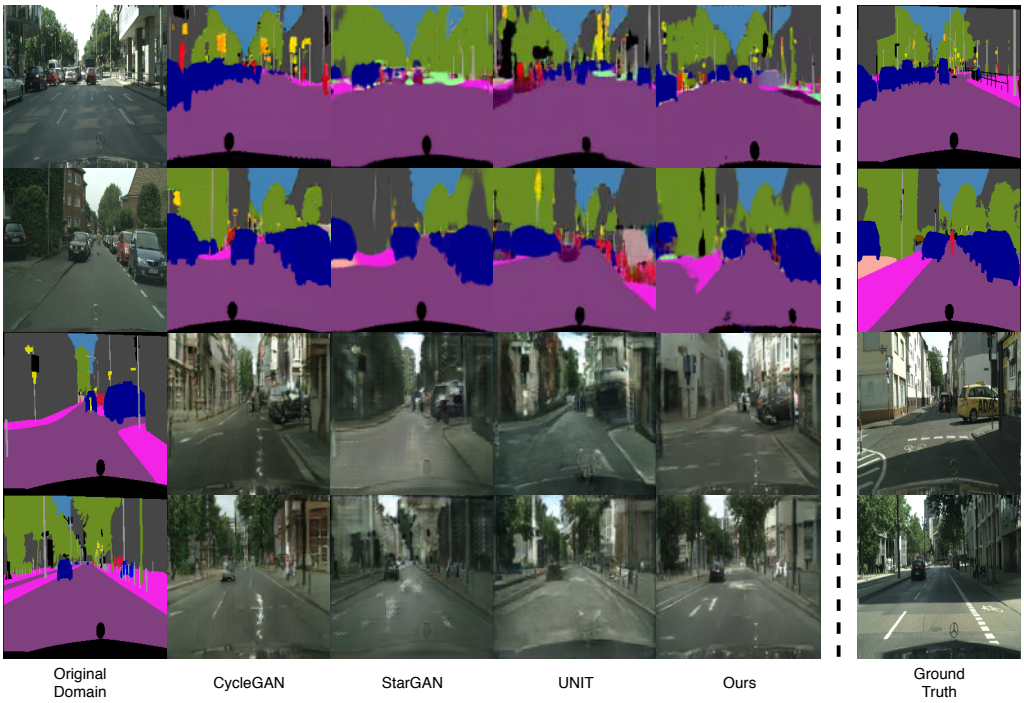

Figure 2: Different methods for mapping labels↔photos trained on Cityscapes images. Our results are generated from hoc attention plus residual multiplication.

the map function $G : X \rightarrow Y$ and discriminator D_Y are trying to distinguish between $G(x)$ and y, the objective of this GAN component is $L_{GAN}(G_X, D_Y)$. The generator G_Y and discriminator D_X is doing the same task in an opposite direction, their loss function is $L_{GAN}(G_Y, D_X)$. 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 $\hat{y}$ in domain $\mathcal{Y}$, we should get the same image x by applying $G_Y : Y \rightarrow X$. This behavior is formally presented as:

$$L_{cyc}(G_X, G_Y) = \mathbb{E}_{x \sim p_{data}(x)}[\|G_Y(G_X(x')) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G_X(G_Y(y')) - y\|_1] \quad (10)$$

The final objective function for the unsupervised translation is:

$$G_X^*, G_Y^* = \arg \min_{G_X, G_Y} \max_{D_X, D_Y} L_{GAN}(G_X, D_Y) + L_{GAN}(G_Y, G_X) + \lambda L_{cyc}(G_X, G_Y) \quad (11)$$

4 Experiments

4.1 Quantitative Comparison

We first quantitatively evaluate our method on *Cityscapes* dataset (Cordts et al., 2016). We compare our model with three unsupervised image translation approaches and two supervised translation models. UNIT (Liu et al., 2017), StarGAN (Choi et al., 2018) and CycleGAN (Zhu et al., 2017) are the baselines in an unsupervised setting and cGAN (Mirza & Osindero, 2014), pix2pix (Isola et al., 2017) are compared for supervised translation. We trained photo→label and label→photo tasks on the *Cityscapes* and compared the output label images with the ground truth using the standard metrics in the paper (Cordts et al., 2016).

We find that our method significantly outperforms the baselines in the experiment, especially when post-hoc attention and residual multiplication work together, as showed in Table 4.1. The image

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
StarGAN	0.47	0.16	0.11	**0.61**	0.21	0.17
Ours (post hoc)	**0.52**	**0.20**	**0.12**	0.60	**0.24**	**0.19**
Ours (RAM)	0.49	0.19	0.10	0.59	0.23	**0.19**

Table 1: FCN-scores for different methods, evaluated on Cityscapes 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
pix2pix	0.61	0.22	**0.16**	0.80	**0.43**	**0.32**
Ours(post hoc)	**0.63**	**0.23**	**0.16**	**0.81**	0.42	**0.32**
Ours(RAM)	**0.63**	0.22	**0.16**	0.75	0.40	0.30

Table 2: FCN-scores for different methods, evaluated on Cityscapes label↔photos in supervised setting

translation result is also presented in Figure 2. The significant improvement in the pixel-level accuracy came from the guidance of the attention map, which is expected. However, the improvement for class accuracy and Intersection over Union (IoU) are limited. Maybe the attention map only focuses on a few domain specific classes so the generator works too hard on those classes and ignores others. Perhaps small class number per image may be another potential reason of this phenomenon, since we cannot increase the accuracy for nonexistent class objects.

Figure 3: Translation result on *day2night* dataset using post hoc attention and residual multiplication. From left to right: real daytime images, fake night images, real night images, fake daytime images.

On the other hand, the improvement of the supervised translation is not as sharp as the unsupervised translation according to Table 4.1. We also present the result of original GAN that removes conditioning from the discriminator. Yet it still shows that we can further improve the translation result with little extra computation, especially when post hoc attention has been chosen. We think the main reason of this situation is the L1 loss between paired images. The generator receives two feedbacks when paired image is available. 1) The L1 loss between paired image and 2) The prediction from the discriminator. Remember that the idea behind our framework is letting the discriminator provide more useful information, but maybe the information from L1 loss is already sufficient.

4.2 Qualitative Result

We qualitatively evaluate our framework on several image translation benchmarks. *horse2zebra* and *apple2orange* are two datasets that have similar class objects from ImageNet (Deng et al., 2009). *day2night* is cropped from BDD110k (Yu et al., 2018) that contains 7870 images of daytime street

Figure 4: Translation result on *apple2orange* using post hoc attention and residual multiplication. From left to right: real apple images, fake orange images, real orange images, fake apple images.

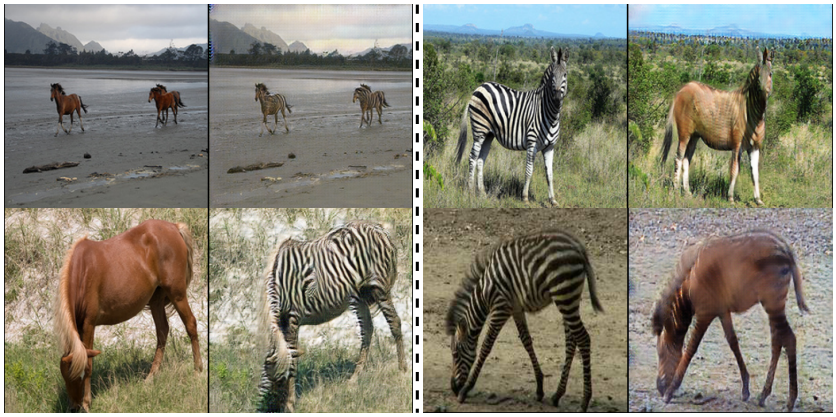

Figure 5: Our translation result on *horse2zebra* dataset using RAM attention and residual multiply concatenation. From left to right: real horse images, fake zebra images, real zebra images, fake horse images.

traffic sign and 8592 night street traffic sign images. Data were split into train and test randomly (80%/20% split). Different combinations of attention methods and concatenation strategy are examined and only the best result is presented here. More specifically, we achieved best result in *horse2zebra* using trainable attention module and alpha channel concatenation, the result is in Figure 5. While post hoc attention and residual multiply concatenation is the best for other two datasets, the result is in Figure 3 and 4.

5 Conclusion

we have proposed a novel method incorporating attention map from discriminator for image-to-image translation. The experiments on different datasets have shown successful translation in both supervised and unsupervised setting. We remark that our idea can apply on any GAN-based model with little modification, such as those baselines in this paper. Nonetheless, the results are sensitive to the selection of attention module and concatenation. Investigating the impact of different attention mechanism and new tasks could be an interesting research direction in the future.

REFERENCES

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. URL https://openreview.net/forum?id=Hk4_qw5xe.

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

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*, pp. 151–166. Springer, 2017.

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)*, pp. 839–847. IEEE, 2018.

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*, pp. 8789–8797, 2018.

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.

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*, pp. 248–255. Ieee, 2009.

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*, pp. 1486–1494, 2015.

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*, pp. 1287–1298, 2018.

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

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*, pp. 770–778, 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

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

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)*, pp. 172–189, 2018.

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*, pp. 1125–1134, 2017.

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*, pp. 1857–1865. JMLR. org, 2017.

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*, pp. 4681–4690, 2017.

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

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)*, pp. 205–220, 2018.

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*, pp. 5657–5666, 2018.

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*, pp. 2794–2802, 2017.

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.

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*, pp. 3693–3703, 2018.

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

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.

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

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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*, pp. 618–626, 2017.

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

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*, pp. 2417–2426, 2019.

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*, pp. 3156–3164, 2017.

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*, pp. 8798–8807, 2018.

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

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*, pp. 8456–8465, 2018.

Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.

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.

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

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)*, pp. 506–511. IEEE, 2017.

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*, pp. 2921–2929, 2016.

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*, pp. 2223–2232, 2017.

A IMPLEMENTATION DETAILS

For the unsupervised *Cityscapes* translation, we adopted the network architectures of CycleGAN (Zhu et al., 2017) as the basic of our proposed model. In specific, we adopted ResNet 6-blocks (He et al., 2016) generator and PatchGAN (Isola et al., 2017) discriminator. This ResNet generator contains 2 down-sampling blocks, 6 residual blocks and 2 up-sampling blocks. For the supervised translation, we adopted UNet-128 (Ronneberger et al., 2015) generator and a same PatchGAN discriminator. The PatchGAN discriminator is composed of 5 convolutional layers, including normalization and ReLU layers. As presented in Figure 6, to associate this PatchGAN discriminator with the RAM attention, we used the first one convolution layer for feature extractor, three consecutive convolution layers for trunk branch and the last one convolution layer for classifier. The mask branch is composed of two downsampling layers, two convolution layers and one upsampling layer. Moreover, we selected the 4-th convolution layer to create a post hoc attention map. All these attention maps have be resized to the shape of original input.

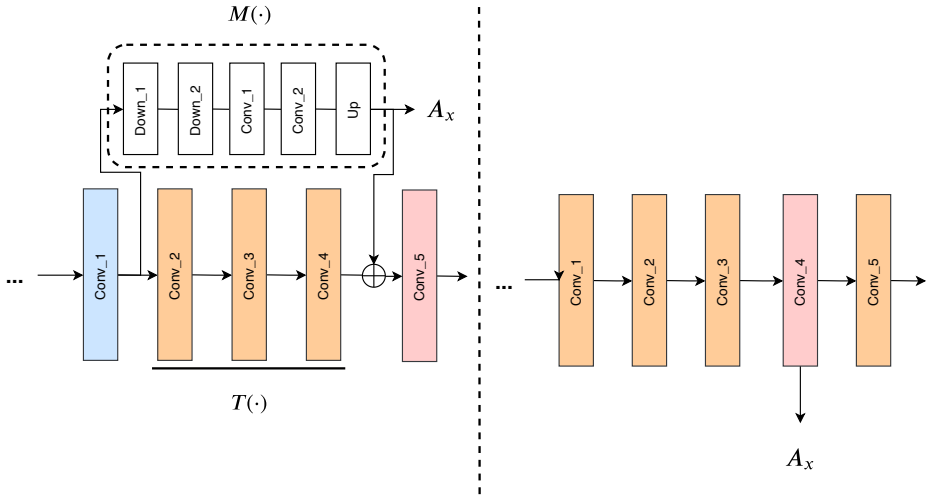

Figure 6: Left: RAM attention plus PatchGAN discriminator, the attention map is denoted as A_x; Right: Post hoc attention plus Patch discriminator, the attention map A_x is generated from 4-th conv layer.

horse2zebra, *apple2orange* and *day2night* tasks are performed under the unsupervised setting. For this three tasks, we adopted ResNet 9-blocks generator and aforementioned PatchGAN discriminator. Similar to prior works, we applied Instance Normalization (IN) for both generators and discriminators. In the preprocessing step, we resized the input image to 143×143 then randomly cropping back to 128×128 for all cityscapes related tasks. We resized the input image to 286×286 then randomly cropping back to 256×256 for the rest tasks.

For all the experiments, we simply set the weight factor of the GAN loss to 10 and the weight factor of L1 loss to 10 for our objective. For example, our implementation uses following objective for supervised training.

$$G^* = \arg\min_G \max_D 10L_{GAN}(G, D) + 10L_{L1}(G)$$

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. Similar to CycleGAN, we kept learning rate for first 100 epochs and linearly decayed to 0 for next 50 epochs for *apple2orange* and Cityscapes related tasks, and kept learning rate for first 50 epochs and linearly decayed to 0 for next 100 epochs for *horse2zebra* and *day2night* datasets.

B ATTENTION MAP DURING TRAINING

We present some intermediate training results with its attention maps in Figure 7, Figure 8 and Figure 9. The white areas in the attention map indicates that region is important. Please note that the attention map indicates the behavior of the discriminator thus some of them may not make sense from human's perspective.

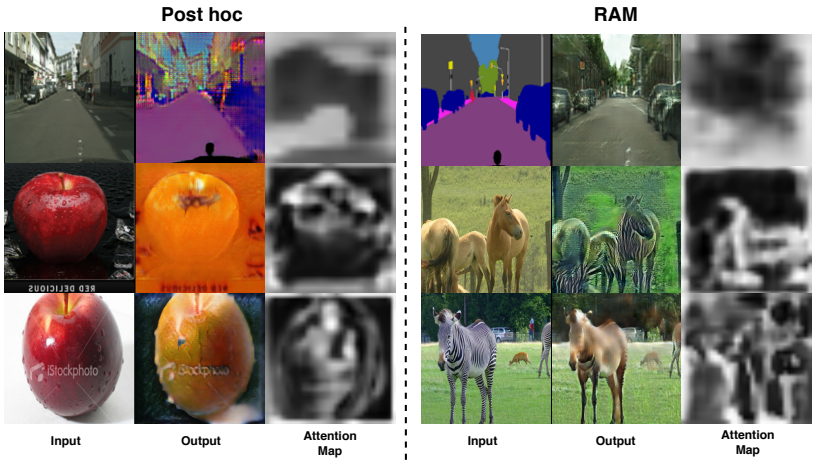

Figure 7: Inputs, outputs and corresponding attention maps at training epoch 10. Left: attention map generated by the post hoc attention; Right: attention map generated by RAM attention mechanism.

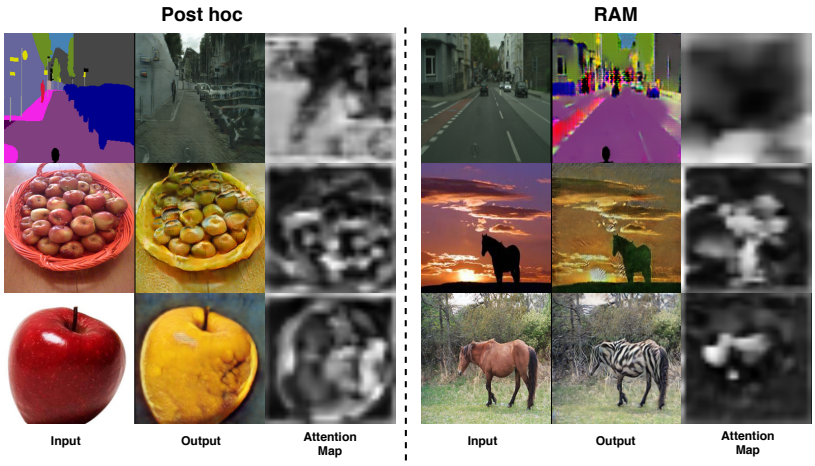

Figure 8: Inputs, outputs and corresponding attention maps at training epoch 50. Left: attention map generated by the post hoc attention; Right: attention map generated by RAM attention mechanism.

C MORE TRANSLATION RESULTS

We also provide more translation results in this section.

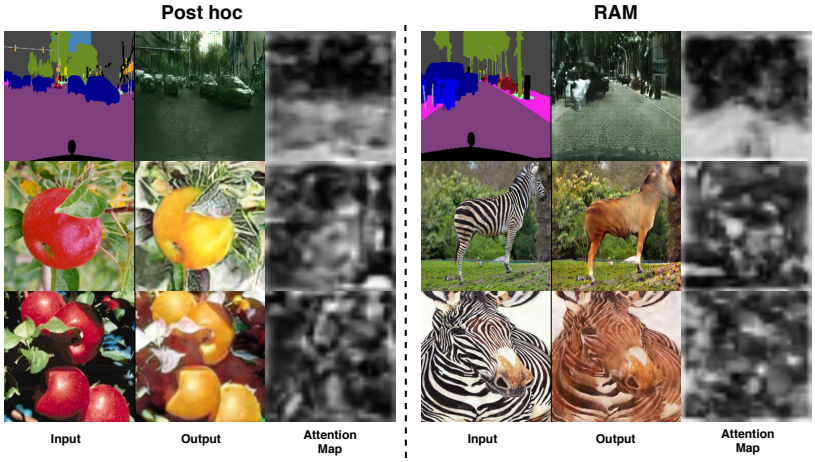

Figure 9: Inputs, outputs and corresponding attention maps at training epoch 100. Left: attention map generated by the post hoc attention; Right: attention map generated by RAM attention mechanism.

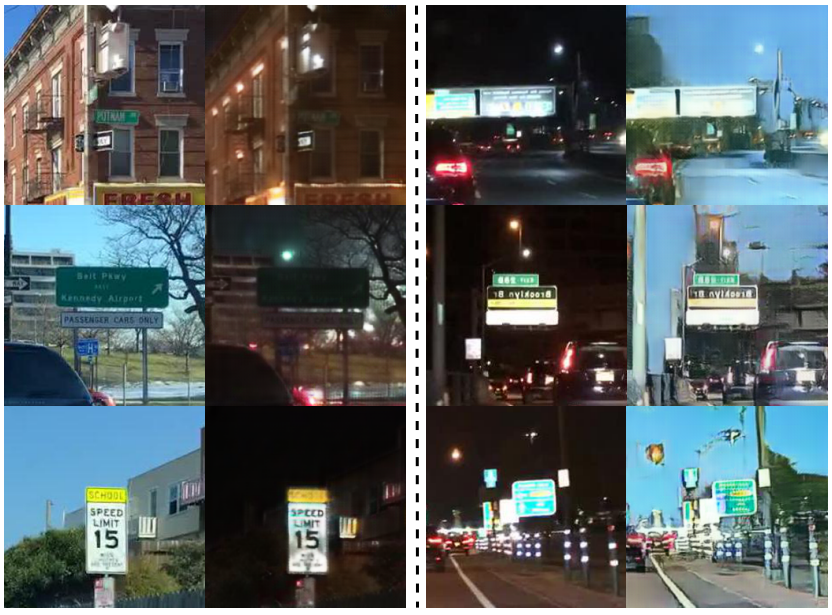

Figure 10: Additional translation results on *day2night* dataset using post hoc attention and residual multiplication. From left to right: real daytime images, fake night images, real night images, fake daytime images.

Figure 11: Additional translation results on *apple2orange* using post hoc attention and residual multiplication. From left to right: real apple images, fake orange images, real orange images, fake apple images.

Figure 12: Additional translation results on *horse2zebra* dataset using RAM attention and residual multiply concatenation. From left to right: real horse images, fake zebra images, real zebra images, fake horse images.