Unpaired Salient Object Translation via Spatial Attention Prior

Xianfang Zeng[a], Yusu Pan[a], Hao Zhang[a], Mengmeng Wang[a], Guanzhong Tian[a] and Yong Liu[a,b]

[a]Institute of Cyber-System and Control, Zhejiang University, Hangzhou 310027, China
[b]State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Generative model
Adversarial learning
Image translation
Spatial attention

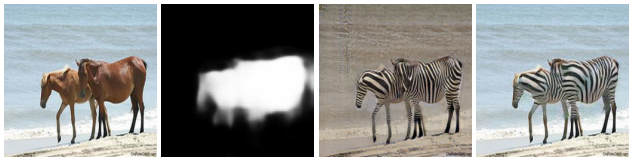
ABSTRACT

With only set-level constraints, unpaired image translation is challenging in discovering the correct semantic-level correspondences between two domains. This limitation often results in false positives such as significantly changing color and appearance of the background during image translation. To address this limitation, we propose the Spatial Attention-Aware Generative Adversarial Network (SAAGAN), a novel approach to jointly learn salient object discovery and translation. Specifically, our generator consists of (1) spatial attention prediction branch and (2) image translation branch. For attention branch, we extract spatial attention prior from a pre-trained classification network to provide weak supervision for object discovery. The proposed attention loss can largely stabilize the training process of attention-guided generator. For translation branch, we revise classical adversarial loss for salient object translation. Such a discriminator only distinguish the distribution of the object between two domains. What is more, we propose a fake sample augmentation strategy to provide extra spatial information for discriminator. Our approach allows simultaneously locating the attention areas in each image and translating the related areas between two domains. Extensive experiments and evaluations show that our model can achieve more realistic mappings compared to state-of-the-art unpaired image translation methods.

1. Introduction

In this work, we consider the challenging task of unpaired salient object translation [24, 6]. Namely, the goal is to learn to transfer prominent objects and ignore the background during the image translating process. More specifically, we consider this problem under the unpaired setting where aligned samples or location annotations are not available. For instance, when translating horses into zebras, the algorithm only draws the particular black-white stripes on the horses while keeping everything else unchanged (see Figure 1). Such ability holds promise to an abundance of applications, e.g., image colorization, video editing, data augmentation and augmented reality. This model can also be used as pre-training for supervised object translation with few annotations.

Existing unpaired image translation approaches [21, 48, 15, 7] typically build upon Generative Adversarial Networks (GANs) [9], which encourages the distribution of generated images close to the distribution of target domain. The limitation in those algorithms is that the input image is viewed as an entire and its spatial structure is ignored during the translation process. In other words, each pixel in the input image is equal when being fed to the discriminator and generator, which disobeys the human intuition that an image usually consists of meaningful objects and meaningless background. As shown in Figure 1(c), the image translation methods are agnostic to the objects in the input image and bundle the objects together with the background as data distribution, significantly changing the color and appearance of the background during translation.



(a) (b) (c) (d)

Figure 1: Salient object translation in horse→zebra pair. Given an image (a), our model simultaneously predicts a spatial attention map (b) and translates image to the target domain (c). In the final result (d), we fuse input image and translated image through the predicted attention map. Best viewed in color.

Compared to the image translation task, salient object translation faces an extra challenge, i.e., how to find the semantic correspondence between two domains. In this work, we decompose the salient object translation into two fundamental issues: (1) how to locate object areas in each image with only domain/class annotations. (2) how to translate images between domains with attention map. Overall, we propose a novel attention-guided generator which consists of a spatial attention branch and an image translation branch to tackle the above issues simultaneously. The attention branch predicts the object attention map, whose value means the probability that a pixel belongs to the foreground. The translation branch converts the input image to target domain. To obtain the final result, the translated image and the input image are fused by the estimated spatial attention map.

Very recently, some attention-based methods have been proposed to tackle the unpaired object translation. AttentionGAN [6] and AGGAN [24] add an auxiliary attention network trained with the generator to locate the object areas. Those methods have evaluated the effects of the spatial at-

zzlongjuanfeng@zju.edu.cn (X. Zeng); corenel@zju.edu.cn (Y. Pan); 21832043@zju.edu.cn (H. Zhang); mengmengwang@zju.edu.cn (M. Wang); gztian@zju.edu.cn (G. Tian); yongliu@iipc.zju.edu.cn (Y. Liu)
ORCID(s):

tention map on object translation task. However, a major problem in those methods is the instability of attention network. As discussed in [24], the mask from the attention network is always zero if training is not performed carefully. The training of attention-based framework suffers from two main challenges: (1) the training of auxiliary attention network is not stable due to the trade-off between the adversarial loss and the cycle consistency loss. The generator can take shortcuts to 'cheat' and not perform the task as desired. (2) GAN training is notoriously unstable.

A distinguishing feature and a core novelty of SAAGAN compared to other attention-based methods is that we introduce spatial attention prior into image translation framework. To stabilize the training of attention branch, we borrow the class activation maps from some remarkable works [46, 47, 45, 30, 19, 27] in weakly supervised object localization community. They have shown that although the class activation maps may not cover the entire target object, it provides strong visual cues related to the input image, i.e., object prior. In this paper, we extract the class activation maps from classification network and then discretize it to a ternary mask to provide weak supervision for the attention branch. With the supervision of the adversarial loss [9], the cycle loss [48], and the attention loss, the attention branch is able to produce attention map which can cover the most discriminative areas and has a clear object boundary. Furthermore, we revise the original adversarial loss which is designed for image generation/translation. For the salient object translation task, we utilize the discretized class activation maps to weight the discriminator's output. Instead of minimizing whole-image distribution, this discriminator only minimizes the distribution of relevant parts between two domains. For the second challenge, we propose fake sample augmentation, which samples images from the neighborhood of the fake. It can stabilize the initial training of GAN since the augmented images provide extra information in object discovery.

In summary, our contributions can be presented as follows:

1) We propose a novel salient translation framework, which can jointly learn salient object discovery and translation. Perceptive and quantitative experiments have demonstrated that the proposed SAAGAN can generate photo realistic images.

2) We introduce an attention loss to stabilize the training of the attention branch. To the best of our knowledge, this is the first time that spatial attention priors are applied to image translation task. We also adapt the original adversarial loss to object translation task. The discriminator therefore ignore the background during translation.

3) We utilize the ternary mask to augment fake samples. This simple but effective augmentation is used to stabilize the initial training of GAN.

The rest of this paper is organized as follows. The related works of our method are discussed in Section 2. Section 3 introduces the details of the proposed framework. The experimental results and analysis are given in Section 4.

2. Related work

2.1. Image Translation

As a class of powerful generative model based on two-player game: a discriminator learns to distinguish the generated samples from real ones while a generator learns to generate fake samples that can fool the discriminator, GANs [9] have been successfully used for various computer vision tasks such as enhancing the security of deep networks [43], image generation [28, 26, 8], super-resolution imaging [17], as well as image translation [48, 15, 21, 24].

Build upon GANs, many image translation systems have achieved impressive results with paired training samples [12, 36]. For instance, imposing an L1 loss between the generated image and its ground-truth, pix2pix [12] achieves an incredible result in the sketch to image task. Given the semantic map of street, the cascaded refinement network proposed by Chen and Koltun [5] can synthesis a real street image. Later, many unsupervised image translation frameworks [21, 48, 15, 42] have been proposed to alleviate the problem of obtaining data pairs. For example, with the assumption of shared low-dimensional latent space in cross domains, Liu *et al.* propose UNIT [21] to learn the joint distribution between the source and target domains by combining variational autoencoders with CoGAN [22]. Further, without such assumption, [48, 15, 42] introduce the cycle consistency loss to preserve key attributes between the input and the translated image. However, all these frameworks are only capable of learning set-level relations between two domains.

Recently, some attention-guided methods have also been proposed for unpaired image translation. Ghislain *et al.* [32] reconstruct seen image with a generative adversarial network conditioned on the brain activity. Ma *et al.* propose DA-GAN [23] to translate image at instance-level in a highly-structured latent space, which relies on the extra location information to obtain meaningful correspondences between samples. AGGAN [24] and attention-GAN [6] learn semantic-level correspondences between two domains by leveraging an auxiliary attention network to predict spatial attention map. In contrast to those methods, we propose a more efficient model to address object localization and image translation simultaneously and introduces a novel regularization to stabilize its training process without additional annotations.

2.2. Weakly Supervised Object Localization

Weakly supervised object/saliency localization [3], as a related domain of spatial attention estimation, encodes the location of objects belonging to a given class. Jetley *et al.* [13] propose a trainable soft attention mechanism to get objects of interest during the training process, where the global feature is used as the query vector for estimating attention map. Simonyan *et al.* [31] introduce a technique to estimate the class saliency map, based on computing the gradient of the class score with respect to the input image. Shen *et al.* [46] propose the object-specific pixel gradient (OPG) to perform weakly supervised object localization, which performs in an iterative manner to localize

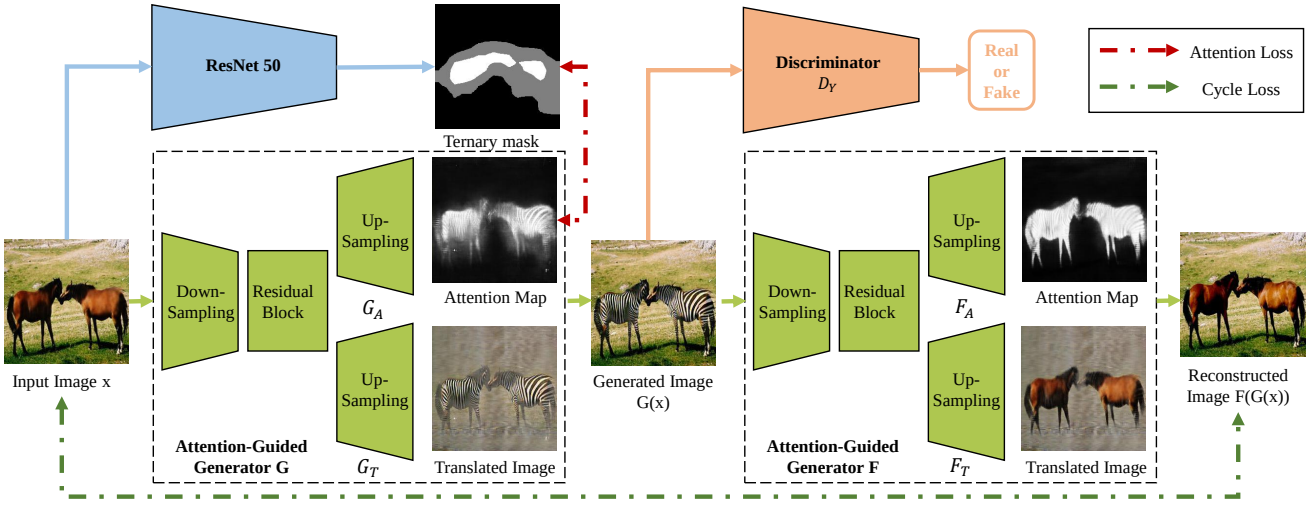


Figure 2: Overview of our method to jointly learn spatial attention estimation and image translation. The feed-ward process of attention-guided generators G and F. At first, x is fed into the source-to-target generator G, which consists of an attention branch G_A and a translating branch G_T. We get the translated image $G(x)$ by fusing x, $G_A(x)$ and $G_T(x)$, as formulated in Eq. (2). Then, $G(x)$ is fed into the target-to-source generator F to reconstruct the input image. Meanwhile, the attention-guided discriminator D_Y takes the $G(x)$ to discriminate its realistic. To stabilize the entire training process, the ternary mask $A(x)$ is introduced to constraint the attention map $G_A(x)$ focusing on object areas.

potential objects. Zhou *et al.* [46] propose Class Activation Mapping (CAM) for identifying discriminative regions based on a pre-trained image classification network. Similarly, instead of using the last layer weight, Selvaraju *et al.* [30] present a generalized CAM using the gradients of class scores to get attention map. Wei *et al.* [41] and Zhang *et al.* [44] introduce iterative adversarial learning into predicting attention map so that the generated map is able to cover the whole object with only image level annotations. Jiang *et al.* [14] propose a discriminative regional feature integration approach to detect salient object. Recently, a stage-wise approach, named SPG [45], incorporates high confident object regions to learn attention mask.

Another closely related area is weakly supervised salient object detection [38, 39]. Please refer to Wang *et al.* [37] for a survey. Wang *et al.* [40] propose a pyramid attention and salient edges module to discover saliency objects. Li *et al.* [18] extract salient objects utilizing contour knowledge. Inspired by those methods, we calculate the ternary mask from class activation maps to provide weak supervision on our attention estimation branch.

3. Method

In the task of unpaired salient object translation, we have two domains X and Y with unaligned training samples $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{M}$. Our goal is to learn mapping functions which simultaneously locate object areas in each input image and translate related areas between two domains. Figure 2 illustrates the mapping process of SAAGAN from X to Y to X. Since the inverse mapping from Y to X to Y is similar, *for simplicity, we will only describe the former mapping in following subsections.* The full framework involves one classification network to extract spatial attention prior, two attention-guided generators and two discriminators to synthesize realistic images. Below we present details of the five parts:

- The classification network C takes real images and extracts their class activation maps as spatial attention prior. We utilize it to provide weak supervision for generators' attention branch.
- The attention-guided generators $G : X \rightarrow Y$ and $F : Y \rightarrow X$ transfer the salient objects between two domains. Each generator consists of a spatial attention estimation branch and a image translation branch. In mapping G, we denote G_A with the attention branch, G_T with the translation branch. In mapping F, the same goes for F_A and F_T.
- We introduce two discriminators D_X and D_Y, where D_X aims to distinguish between images $\{x\}$ and generated images $\{F(y)\}$, while D_Y aims to distinguish between $\{y\}$ and $\{G(x)\}$.

3.1. Spatial Attention Prior

Class activation maps play the spatial attention prior role in our framework. They have been widely used in many tasks [44, 47, 1], offering a promising way to extract object localization information. We utilize it to provide weakly constraint for the attention branch. Since original CAM [46] can only visualize the class activation maps of the categories contained in the pre-trained model, we slightly modify pre-trained architecture by replacing the last fully connected layer with a 1×1 convolution layer. The output is then fed into a global average pooling followed by a softmax layer

for classification. Given an image x, the spatial attention prior is computed as

$$C(x) = f(x) \cdot W_{k,c}, \tag{1}$$

where $f(x)$ denotes the output feature maps of backbone network and $W_{k,c} \in \mathbb{R}^{K \times C}$ denotes the weight matrix of the 1×1 convolution layer. We adopt the ResNet50 [10] pre-trained on the ImageNet [29] as the backbone network and finetune the new layer with domain/class label in our experiments.

3.2. Attention-Guided Generator

For simultaneous object discovery and image translation, we propose the attention-guided generator. We first feed input image x into the source-to-target generator G, which consists of a spatial attention estimating branch G_A and a translation branch G_T. The attention branch G_A predicts a foreground attention map $G_A(x)$, which has the same shape as the input image and has a continuous value between [0,1] in each position. The higher the value, the more likely the pixel belongs to the foreground. The translation branch G_T converts the whole input image to a corresponding image in the target domain Y, denoted as $G_T(x)$. Our goal is to translate the foreground object and keep the background unchanged. Given x, $G_A(x)$ and $G_T(x)$, the final transformed image $G(x)$ is fused by

$$G(x) = \underbrace{G_A(x) \odot G_T(x)}_{foreground} + \underbrace{(1 - G_A(x)) \odot x}_{background}, \tag{2}$$

where $\odot$ means element-wise product. Subsequently, $G(x)$ is fed into target-to-source generator F to get the reconstructed image $F(G(x))$. In this mapping process, our objective contains four terms: adversarial loss [9], cycle consistency loss [48], attention loss, and smoothness loss.

3.2.1. Adversarial Loss

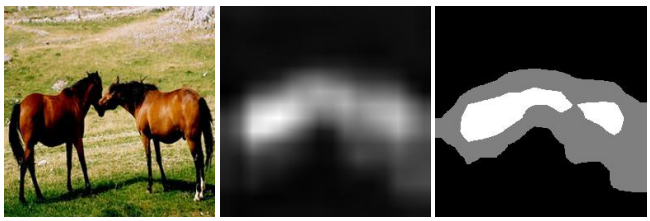
To make the generated images indistinguishable from real images, we adopt an adversarial loss [9]

$$\begin{aligned}\mathcal{L}_{adv}(G, D_Y) = &\mathbb{E}_{y \in Y}\big[\log D_Y(y)\big] + \\ &\mathbb{E}_{x \in X}\big[\log(1 - D_Y(G(x)))\big],\end{aligned} \tag{3}$$

where G tries to generate images $G(x)$ that look similar to images from domain Y, while D_Y aims to distinguish between translated samples $G(x)$ and real samples y.

3.2.2. Cycle Consistency Loss

By minimizing the adversarial loss, the generator G is enforced to generate realistic images. However, without ground-truth supervision, there is no constraint to guarantee that the translated images preserve the content of its input images. Leveraging the cycle consistency loss [48], we force the reconstructed images $F(G(x))$ to be identity of its input, i.e., $F(G(x)) \approx x$. Meanwhile, as introduced in [6], the regions of interest in the original image and the transformed image should be the same, i.e., $G_A(x) \approx F_A(G(x))$.



(a) (b) (c)

Figure 3: (a) Input image. **(b)** class activation maps. **(c)** Ternary mask. It is obtained by discretizing the class activation maps to three regions: the foreground, the background and the uncertain.

Thus, we define an enhanced cycle consistency loss as

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \in X}[&\|F(G(x)) - x\|_1 + \\ &\big\|G_A(x) - F_A(G(x))\big\|_1].\end{aligned} \tag{4}$$

3.2.3. Attention Loss

The estimated attention map plays a key role in Eq. (2). If $G_A(x)$ is all zeros, the translated images would be identical to the inputs. If the attention map $G_A(x)$ is filled with one, we will map the entire image to the target domain as what CycleGAN [48] does. In theory, by minimizing the adversarial loss and cycle consistency loss, we could get the attention map that focuses on the discriminative areas of the input. There are two reasons: (1) the adversarial loss encourages the attention map to cover the discriminative areas in the input otherwise the discriminator could easily found the drawback of the generated image; (2) the cycle consistency loss encourages the entire attention map to be zero as the loss of unattended areas would always be zero. In equilibrium, the attention map focuses on the discriminative areas and ignores the background. However, the attention map would easily collapse to zero [24] because the cycle consistency loss is always zero in this case.

To stabilize the entire training process, we extract the class activation maps from a pre-trained classification network as spatial attention prior. As shown in Figure 3, we discretize the class activation maps $C(x)$ into three regions: the foreground, the background and the uncertain. Particularly, in each class activation maps, the regions with very low response are considered as background, while the high activated regions are foreground and the rest regions are uncertain. This discretized mask is a lower bound of the object segmentation map. we assign different value for these three regions to provide weak supervision on the attention branch. We donate the ternary mask as $A(x)$, calculated as:

$$A(x) = \begin{cases} 0 & \text{if } C(x) < \delta_l \\ 1 & \text{if } C(x) > \delta_h \\ 0.5 & \text{if } \delta_l \le C(x) \le \delta_h \end{cases} \tag{5}$$

where δ_l and δ_h are thresholds to identify regions in class activation maps as background and foreground. The atten-

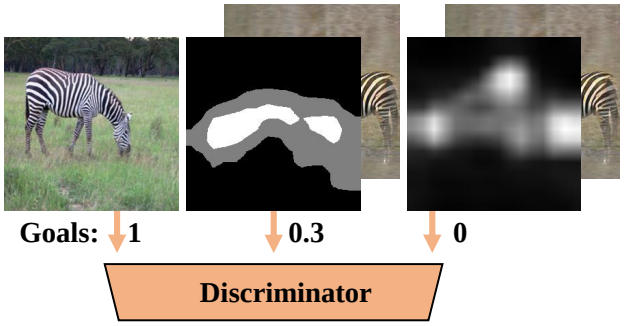


Figure 4: We leverage the ternary mask to augment fake images. Those images are sampled from the neighborhood of the fake, providing extra information for discriminator to learn object discovery. We assign them a smooth label.

tion loss is defined as

$$\mathcal{L}_{attn}(G_A) = -\mathbb{E}_{x\in X}[A(x)\log G_A(x) + (1-A(x))\log(1-G_A(x))], \quad (6)$$

where only the positions labeled as 0 or 1 in the ternary mask are served as pixel-level supervision. The pixels with values of 0.5 are temporarily ignored. The ignored pixels do not contribute to the loss and their gradients do not back-propagated.

3.2.4. Smoothness Loss

Since the attention map is used for combining the input and the translated image, we add a *Total Variation Regularization* on $G_A(x)$ to increase its smoothness. The smoothness term penalizes quickly-changing in attention map to avoid local salt and pepper noise. In particular, it decreases the total variation of attention map horizontally and vertically, denoted as

$$\mathcal{L}_s\left(G_A\right) = \mathbb{E}_{x\in X}\left[\left\|\nabla_u G_A(x)\right\|_1 + \left\|\nabla_v G_A(x)\right\|_1\right]. \quad (7)$$

3.3. Guiding the Discriminator for Further Stability

The original discriminator faces two difficulties in the salient object translation task. Firstly, the attention-guided generator only transform the attended regions. The discriminator constrained by Eq. (3) takes the whole image into consideration when distinguishing generated and real images. It creates an inconsistency between generator and discriminator. Secondly, with only real/fake labels, it is hard for discriminator learning locating discriminative areas.

3.3.1. Revised Adversarial Loss

Since the generator's gradient comes from the discriminator, the whole-image discriminator encourages the attention map to cover the entire input image. To overcome that limitation, we utilize the ternary mask to weight the adversarial loss and train the discriminator to ignore the background. Thus, we update the adversarial loss $\mathcal{L}_{adv}$ of Eq. (3) to

$$\mathcal{L}'_{adv}(G, D_Y) = \mathbb{E}_{y\in Y}\left[A(y)\log D_Y(y)\right] + \mathbb{E}_{x\in X}\left[A(x)\log(1-D_Y(G(x)))\right]. \quad (8)$$

Notice that we use PatchGAN [12] as the discriminator, which aims to classify whether the overlapping image patches are real or fake. We interpolate the weight mask to the same size as the discriminator's outputs and then leverage it as the weighting factor.

3.3.2. Fake Sample Augmentation

Since the real/fake labels contain too little information, it is hard for discriminator to learn locating salient object. Actually, the attention branch always predicts a zero attention map at early training phase. To stabilize initial training, we replace the estimated attention map $G_A(x)$ with the ternary mask $A(x)$ in Equation 2 to artificially augment fake samples. In other words, those fake samples are created by fusing $G_T(x)$ and x with $A(x)$. In this way, there are some inputs that help training the discriminator on object discovery. Figure 4 illustrates the training processing of discriminator. We assign those fake samples with a smooth label which is 0.3 in our experiments.

3.4. Optimization

We optimize two attention-guided generators (G, F) and two attention-guided discriminators (D_X, D_Y) during the training process. Especially, the weight of the classification network is fixed as we only utilize the pre-trained model to get the class activation maps. During the training process, we alternate optimizing between generator and discriminator. When two generators (G, F) are fixed, the discriminators D_X and D_Y are optimized to distinguish the fake photographs from the real ones. The corresponding objective functions of discriminator are written as follows:

$$\begin{aligned}\max_{D_Y}\mathcal{L}(G,F,D_X,D_Y) &= \max_{D_Y}\mathcal{L}'_{adv}(G,D_Y)\\ &= \mathbb{E}_{x\in X}\left[A(x)\log(1-D_Y(G(x)))\right]\\ &+ \mathbb{E}_{y\in Y}\left[A(y)\log D_Y(y)\right].\end{aligned} \quad (9)$$

$$\begin{aligned}\max_{D_X}\mathcal{L}(G,F,D_X,D_Y) &= \max_{D_X}\mathcal{L}'_{adv}(F,D_X)\\ &= \mathbb{E}_{y\in Y}\left[A(y)\log(1-D_X(F(y)))\right]\\ &+ \mathbb{E}_{x\in X}\left[A(x)\log D_X(x)\right].\end{aligned} \quad (10)$$

After one step of optimizing discriminator, we train the generator G and F to generate fake images that aims at fooling the discriminator. As the cycle consistency loss couples two generators together, we optimize them in one full ob-

jective:

$$\begin{aligned}\min_{G,F} \mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}'_{adv}(G, D_Y) + \mathcal{L}'_{adv}(F, D_X) \\ + \lambda_{cyc}(\mathcal{L}_{cyc}(G, F) + \mathcal{L}_{cyc}(F, G)) \\ + \lambda_{attn}(\mathcal{L}_{attn}(G_A) + \mathcal{L}_{attn}(F_A)) \\ + \lambda_s(\mathcal{L}_s(G_A) + \mathcal{L}_s(F_A)),\end{aligned} \tag{11}$$

where λ_{cyc}, λ_{attn} and λ_s are the hyper-parameters that control the relative importance of every loss term. Finally, we can define the following mini-max problem:

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y). \tag{12}$$

Additionally, we constrain our discriminator D, F to lie in D, which represents the set of 1-Lipschitz functions. The training procedure of the proposed SAAGAN method is shown in Algorithm 1.

Algorithm 1 Training procedure of the proposed SAAGAN.

Input: Initialized networks G, F, D_X, and D_Y; hyper-parameters λ_{cyc}, λ_{attn} and λ_s; a pre-trained classification network C; total training steps K.

Output: Optimized networks G, F, D_X, and D_Y.

1: **for** $i = 1$ to K **do**
2: Sample a data point x from X and y from Y.
3: Get class activation maps $C(x)$ and $C(y)$ by feeding real images to the pre-trained image classification network C.
4: Generate fake images $G(x)$ and $F(y)$ by feeding x to G and y to F.
5: Reconstruct input images, denoted as $F(G(x))$ and $G(F(y))$, by feeding $G(x)$ to F and $F(y)$ to G, .
6: Calculate weight masks $A(x)$, $A(y)$ from $C(x), C(y)$. Then optimize D_X and D_Y by solving Eq. 9 and Eq. 10.
7: Fix the parameters of D_X, D_Y and optimize G, F by solving Eq. 11
8: **end for**

3.5. Network Architecture

We next introduce architecture details of the proposed SAAGAN, which consists of two attention-guided generators, two attention-guided discriminators.

The generator builds upon CycleGAN [48], whose generator composes of three stride 2 convolutions, nine residual blocks, and two stride 2 transposed convolutions. We modify it by adding an attention branch after the forth residual block. The new attention branch composes of two residual blocks and two stride 2 transposed convolutions, aim at locating the objects of interest in the input image. Following [34], we also use instance norm and ReLU after each convolution layer.

Following [48], we adopt the PatchGAN architecture of [12] to classify whether local image patches are real or fake. Specifically, we stack the modules of convolution-BatchNorm-LeakyReLU 3 times, where the stride of convolution is 2 to increase filter's receptive field. Then, another convolution-BatchNorm-LeakyReLU block with one stride is added. After the last layer, a convolution operation is applied to map the features to a 2-D output, followed by a sigmoid function.

4. Experiments

This section provides a thorough experimental evaluation of our approach. We first compare our model against recent unpaired image translation methods both qualitatively and quantitatively. Then, we evaluate the effect of proposed attention loss, which is able to stabilize the training of the attention branch. Furthermore, We compare our full method against several ablations to study the effects of terms in our loss function. Finally, we conduct a comparison to supervised results for a more convincing experiment.

4.1. Setup and Evaluation Metrics

4.1.1. Datasets

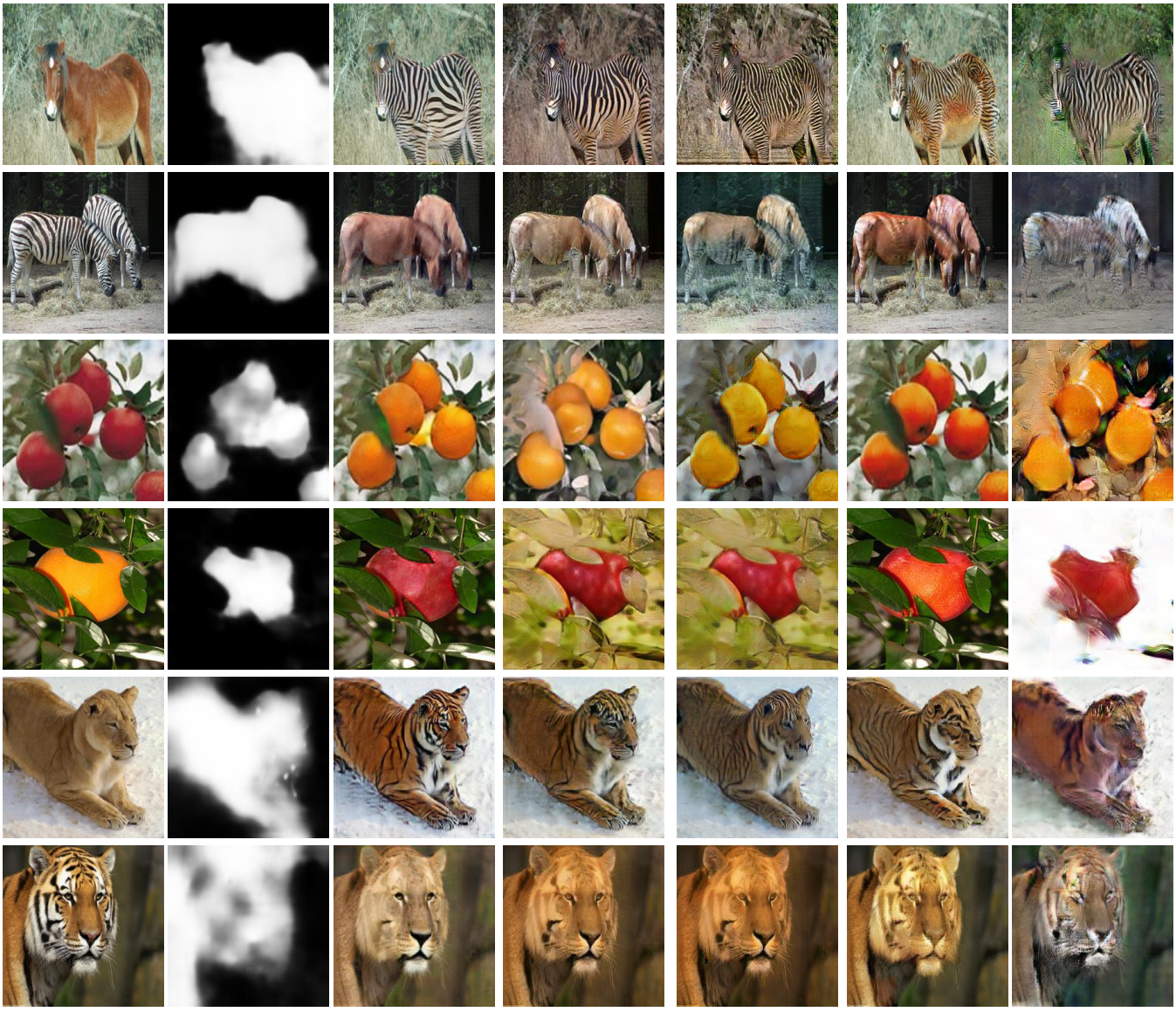
We evaluate the proposed SAAGAN on four different tasks: horse ↔ zebra, apple ↔ orange, tiger ↔ lion and bird transforms. The images for horse, zebra, apple and orange are provided by CycleGAN [48], and the images for tiger and lion are obtained from the corresponding classes in the ImageNet [29]. Besides, we use images of four classes from the CUB-200-2011 [35] dataset—Cardinal, Summer Tanager, Cape Glossy Starling and Indigo Bunting, to perform bird transform. These images contain objects at different scales across various backgrounds, which makes image translation task challenging. In supervised experiment, we perform the horse ↔ zebra task where the segmentation map is manually annotated. As a common convention, the samples are first scaled to 286×286, and then randomly flipped and cropped to 256×256 in the training process while the input images are directly scaled to 256×256 in the inference phase.

4.1.2. Implementation Details

For all the experiments, we use Adam solver with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and batch size 1. All networks are trained for 200 epochs except those on the bird transform task, where the networks are trained for 2000 epochs. We train all networks from scratch with an initial learning rate of 0.0002, keeping the same learning rate for the first half epochs and linearly decaying the rate to zero over the next half epochs. In the full loss function, the weight coefficients in Eq. (11) are set as $\lambda_{cyc} = 8$, $\lambda_{attn} = 1$, and $\lambda_s = 5$. We set the thresholds in Eq. (5) as $\delta_l = 0.2$ and $\delta_h = 0.7$. Following [16], we apply spectral normalization [25] for the discriminators to improve the training stability of generative adversarial networks.

4.1.3. Metrics

We use the Fréchet Inception Distance (FID) [11] and Kernel Inception Distance (KID) [2] to quantitatively evalu-



(a) Input (b) Attention (c) Ours (d) CycleGAN [48] (e) DiscoGAN [15] (f) AGGAN [24] (g) UNIT [21]

Figure 5: Salient object translation results and comparisons with baselines. We evaluate all the methods on three tasks: from top to bottom is horse ↔ zebra, apple ↔ orange and lion ↔ tiger (every two rows compose an experiment pair). Our method can learn a meaningful attention map from the class activation map. The predicted spatial attention map has a sharp object boundary and can cover the salient objects. Benefited from that, our method can generate more realistic results than all baselines.

ate our image translation framework and the Mean Absolute Error (MAE) [4] to assess the attention map predicted by our attention branch. FID has been shown to be consistent with human evaluation in assessing the realism and variation of the generated samples. It computes the Fréchet distance between feature representations of real and generated images. Such representations are extracted from the last hidden layer of the Inception architecture [33]. Lower FID value means a closer distance between synthetic and real data distributions. The Fréchet distance between the Gaussian with mean and covariance $(\hat{\boldsymbol{m}}, \hat{\boldsymbol{C}})$ obtained from p_{fake} and the Gaussian $(\hat{\boldsymbol{m}}_{\boldsymbol{w}}, \hat{\boldsymbol{C}}_{\boldsymbol{w}})$ obtained from p_{data} is given by:

$$FID = \left\|\hat{\boldsymbol{m}} - \hat{\boldsymbol{m}_{\boldsymbol{w}}}\right\|_2^2 + \mathrm{Tr}(\hat{\boldsymbol{C}} + \hat{\boldsymbol{C}_{\boldsymbol{w}}} - 2(\hat{\boldsymbol{C}}\hat{\boldsymbol{C}_{\boldsymbol{w}}})^{1/2}). \quad (13)$$

KID computes the squared maximum mean discrepancy (MMD) between feature representations of real and generated images, denote ad P_r and P_g. The kernel MMD between P_r and P_g for some fixed characteristic kernel function k is given by:

$$KID\left(P_r, P_g\right) = \mathbb{E}_{x,x'\sim P_r}\left[k\left(x, x'\right)\right] - 2\mathbb{E}_{x\sim P_r, y\sim P_g}[k(x, y)] + \mathbb{E}_{y,y'\sim P_g}\left[k\left(y, y'\right)\right]. \quad (14)$$

We use a polynomial kernel, $k(x, y) = \left(\frac{1}{d}x^\top y + 1\right)^3$ where d is the representation dimension.

As formated in Eq. (2), the generated attention map is

Table 1
The FID (lower is better) for our method and baselines. Abbreviations: (H)orse, (Z)ebra, (A)pple, (O)range, (T)iger, (L)ion, (C)ardinal, (I)ndigo Bunting, (S)ummer Tanager, (Ca)pe Glossy Starting.

Algorithm	$H \to Z$	$Z \to H$	$A \to O$	$O \to A$	$L \to T$	$T \to L$	$C \to I$	$I \to C$	$S \to CA$	$CA \to S$
UNIT	98.82	136.84	132.98	123.40	140.17	138.21	127.81	140.12	123.17	124.56
AGGAN	100.35	79.17	106.08	96.34	66.22	98.74	84.06	81.75	90.24	82.94
DiscoGAN	46.05	71.72	98.07	93.05	70.44	93.16	76.21	75.39	81.87	85.56
CycleGAN	44.61	55.79	98.77	98.26	68.05	77.87	83.19	66.49	80.07	68.83
Ours	**38.73**	**53.80**	**88.68**	**87.64**	**42.91**	**68.97**	**62.79**	**61.67**	**52.49**	**50.37**

Table 2
The KID × 100 for different image translation algorithms. Lower is better. Abbreviations: (H)orse, (Z)ebra, (A)pple, (O)range, (T)iger, (L)ion, (C)ardinal, (I)ndigo Bunting, (S)ummer Tanager, (Ca)pe Glossy Starting.

Algorithm	$H \to Z$	$Z \to H$	$A \to O$	$O \to A$	$L \to T$	$T \to L$	$C \to I$	$I \to C$	$S \to CA$	$CA \to S$
UNIT	18.97	22.91	18.48	20.79	21.59	21.35	19.24	16.63	16.32	15.38
AGGAN	12.62	8.27	14.88	9.96	10.47	9.25	9.80	10.98	12.82	13.28
DiscoGAN	9.09	9.83	9.24	11.98	10.88	13.73	14.62	15.85	14.79	8.25
CycleGAN	8.17	8.80	10.25	10.13	10.49	13.93	9.67	13.67	11.25	13.68
Ours	**6.78**	**6.69**	**6.52**	**9.38**	**9.39**	**7.37**	**7.75**	**5.99**	**5.42**	**5.75**

continuous to fuse generated images and inputs. Therefore, we calculate MAE between the continuous attention map $\hat{S}$ and the ground truth $\hat{T}$, to evaluate the accuracy of predicted attention map. MAE score is defined as:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |\hat{S}_{i,j} - \hat{T}_{i,j}|. \tag{15}$$

4.2. Image Translation Results

4.2.1. Results on ImageNet

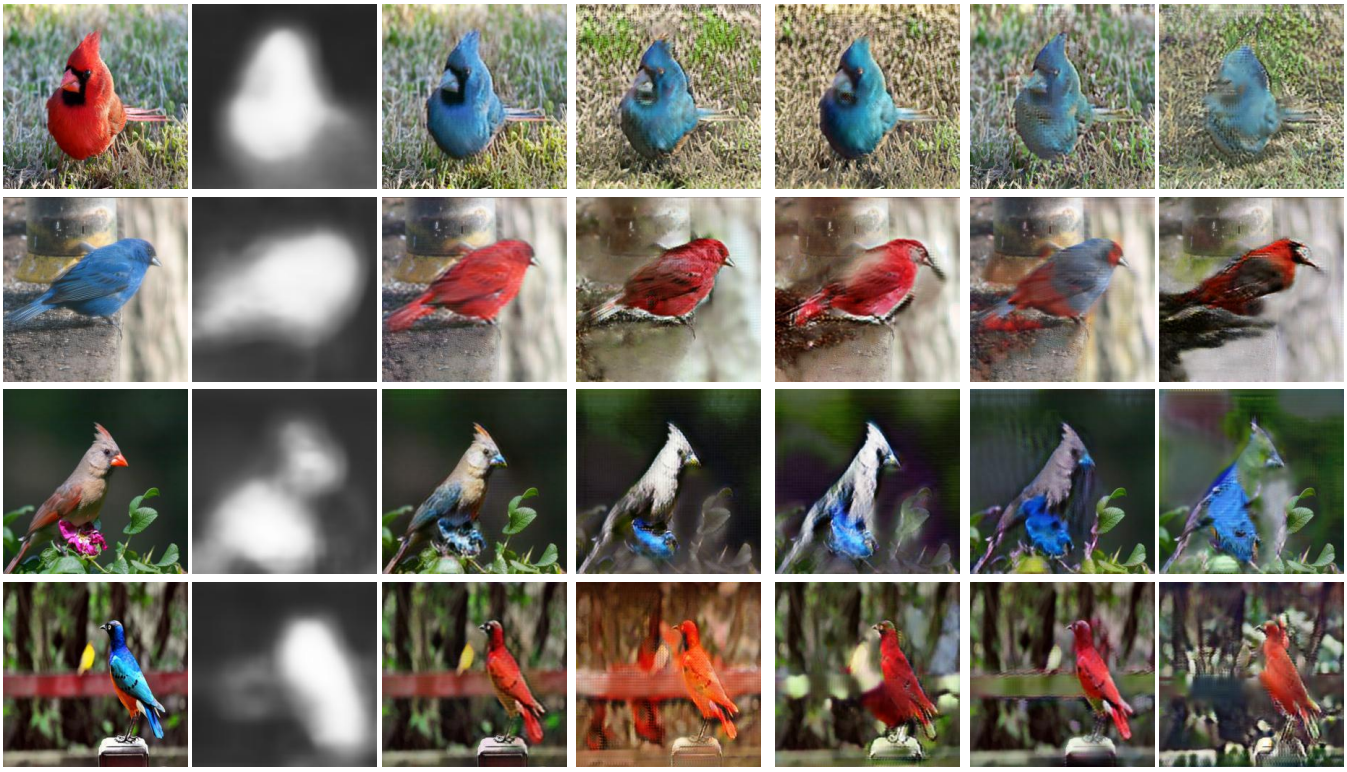
In this study, we compare some state-of-the-art image translation methods with our method. They are UNIT [21], CycleGAN [48], DiscoGAN [15] and AGGAN [24]. The UNIT model is an unsupervised image-to-image translation work consisting of two VAE-GANs with an assumption of fully shared latent space. CycleGAN consists of two residual translation networks, which is trained with an adversarial loss to learning the mapping between two different domain and regularizes the mapping via cycle consistency loss. DiscoGAN is a contemporaneous unsupervised image translation work with CycleGAN while DiscoGAN uses a standard GAN loss and CycleGAN uses a least-squared GAN loss. We also compare with Mejjati et al's unsupervised attention-guided image translation method, which also learns the attention map for image translation through combining adversarial loss and cycle loss. For all the baselines, we compare our results to the images generated by the official released models.

As illustrated in Figure 5, for each input image, we show the predicted attention map, the translated image and the baselines' output. It can be seen at Figure 5(b) that our approach is able to locate the area of objects and ignore the background. Among competing approaches, AGGAN is most similar to our approach since it also predicts an attention map to preserve the background. However, the results of AGGAN are bright-colored, which causes a distortion in image color and reduces the realness of generated images. For instance, the generated horse and orange are over flashy in the Figure 5(f). For the other baselines without attention mechanism, CycleGAN produces the best results in visual appearance. It is able to translate objects between two domains but fall in preserving the background. For instance, CycleGAN generates a realistic zebra image with appropriate black-white stripes at the first row of Figure 5, even though some elements of the background are changed. Comparing Figure 5(d) and Figure 5(e), we can see that the DiscoGAN performances similar or slightly worse results to CycleGAN. It is a reasonable result since these two methods share the same idea and only have a difference in implementation details. Meanwhile, failing to preserve the content of the inputs, UNIT gets the worst performance in our comparison experiments. We believe this result is related to the assumption of shared latent space, which enable UNIT to change the geometric shape of objects.

We further conduct a quantitative evaluation, which can be found in Table 1 and 2. We report the FID and KID value computed using generated samples and target domain. Our approach gets the lowest FID/KID value in all translations while CycleGAN is the second great performing method. DiscoGAN achieves a similar score to CycleGAN while UNIT obtains the worst result, which accords with the visual effects displayed in Figure 5. Moreover, AGGAN gets the second bad result for FID/KID, which may be caused by the color distortion in generated images.

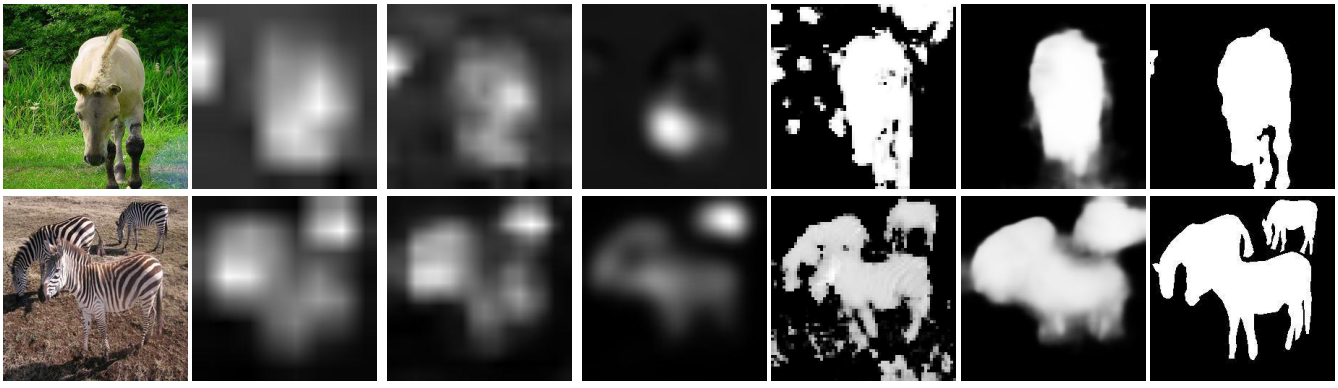
4.2.2. Results on CUB-200-2011

Since the classification network is pre-trained on ImageNet dataset, we further evaluate the scalability of SAA-GAN for translating the objects that are not contained in ImageNet. In this study, we use images of four classes from the CUB-200-2001 [35] dataset to perform bird transforms.



(a) Input (b) Attention (c) Ours (d) CycleGAN [48] (e) DiscoGAN [15] (f) AGGAN [24] (g) UNIT [21]

Figure 6: Salient object translation results on CUB-200-2011 dataset. Our method is able to focus on salient object and ignore the background during the image translation process, generating photo-realistic images.



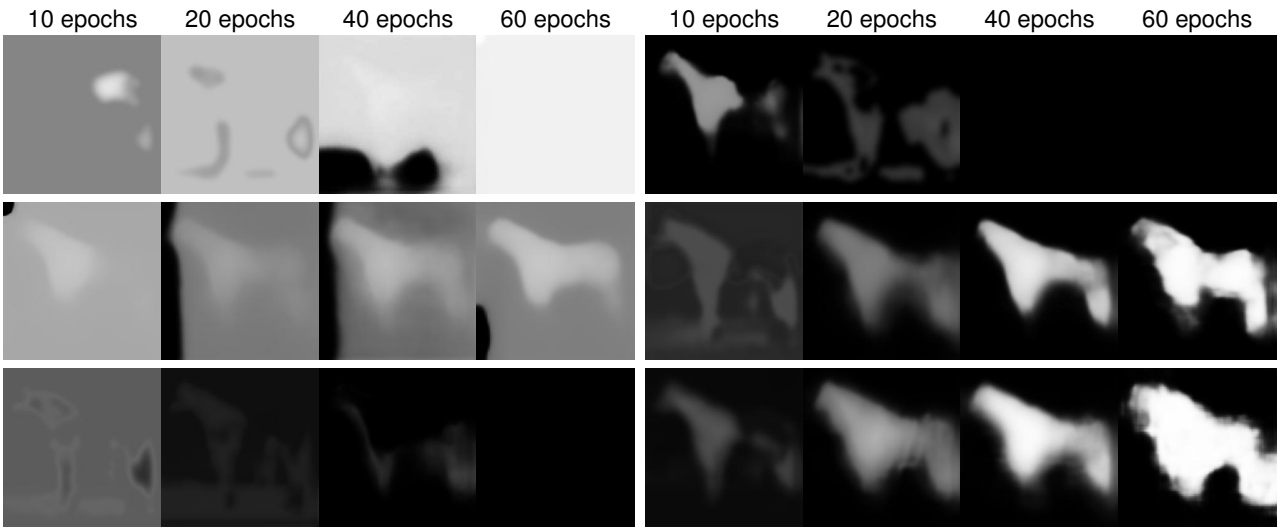
(a) Input (b) CAM [46] (c) Grad-CAM [30] (d) SPG [45] (e)AGGAN [24] (f) Ours (g) Ground Truth

Figure 7: Salient object estimation results and comparisons with baselines. All methods are trained with only class annotation.

Some example results are shown in Figure 6. We can see that SAAGAN generates extremely realistic images along with attention maps that focus on the objects of interesting. The idea of using the attention map in the image translation task has obvious advantages, as it leads to a clearly improved background region. As a comparison, CycleGAN translates the objects together with the background. The feature of the background sometimes is mapped to the object in the results of CycleGAN, e.g., the leaves are covered by the appearance of background in the third row of Figure 6(d). FID values are reported in Table 1. Our method outperforms baselines with a large margin.

4.3. Attention Results

We quantitatively evaluate our model's capability of predicting saliency maps. Following the same experimental procedure described in [6], we perform horse ↔ zebra task on MSCOCO dataset [20] where the images and corresponding annotations could be directly obtained. Furthermore, we observe that people often appear in the horse class in MSCOCO dataset. This data distribution is different from the ImageNet on which the comparisons are trained. For a fairly evaluation, we also evaluate horse ↔ zebra task on ImageNet [29] by manual annotating corresponding saliency maps. Results are shown in Table 3. We compare our ap-



(a) Tune the hyper-parameter of cycle consistency loss. (a) Tune the hyper-parameter of attention loss.

Figure 8: Visualization of the training processing of the attention branch. We investigate the effect of two key components by tune its hyper-parameter in Eq. 11. From top to bottom, (a) we fix $\lambda_{attn} = 0$ and tune λ_{cyc} to 0, 1 and 4; (b) we fix $\lambda_{cyc} = 8$ and vary λ_{attn} to 0, 1 and 4. As we can see, the training of attention branch is more stable after introducing the attention loss.

Table 3
MAE (lower is better) for different methods, evaluated on both MSCOCO and ImageNet.

Algorithms	MSCOCO [20]		ImageNet [29]	
	horse	zebra	horse	zebra
CAM	0.343	0.260	0.255	0.294
Grad-CAM	0.340	0.247	0.300	0.292
SPG	0.235	0.203	0.196	0.258
AGGAN	0.225	0.151	0.169	0.184
Ours	**0.201**	**0.140**	**0.146**	**0.158**

proach with CAM [46], Grad-CAM [30], SPG [45], AGGAN [24] and the ground truth. As can be seen, our method outperforms all the comparisons for MAE, which may benefit from adversarial learning. Figure 7 shows the visual results of estimated attention map. CAM and Grad-CAM are able to cover the object of interest but also parts of the background. With the supervision of self-produced attention seed, SPG can learn more confident patterns of foreground and background while failing in covering entire object. As a similar method to our method, the attention map generated by AGGAN is accurate but contains numerous noise points. As a comparison, our proposed approach can highlight nearly the entire object regions.

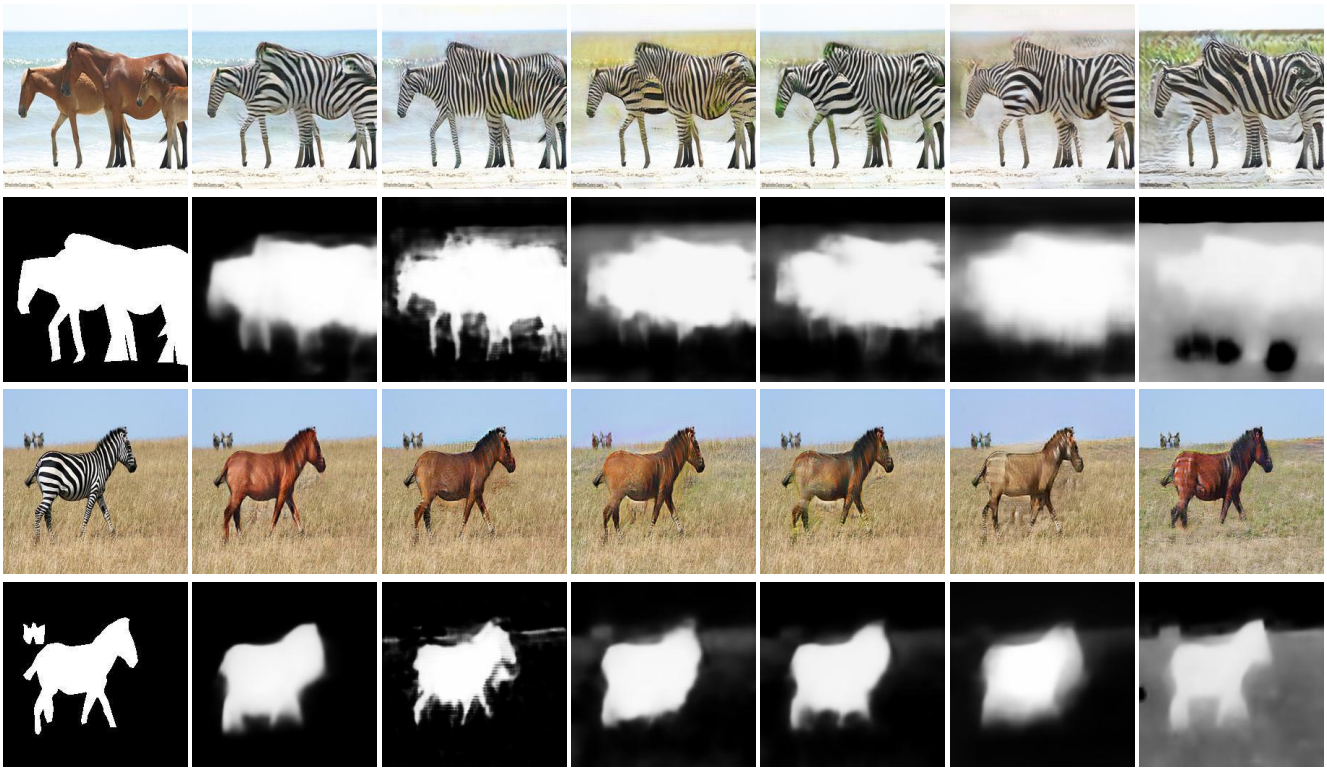
4.4. Training Stability

We introduce the attention loss as we noticed that training the attention branch is a trade-off between the cycle consistency loss and the adversarial loss. Figure 8(a) depicts such behavior: it can be seen that without the attention loss ($\lambda_{attn} = 0$ in Eq. (11)), generated attention maps are sensitive to the ratio of the cycle consistency loss to the adversarial loss. With only the adversarial loss supervision (first row in Figure 8(a)), the attention map would collapse to white. Such behavior can be explained by that the whole image is the discriminative area in the case of only two domains (wild horses live in green meadows while zebras live in dry landscapes). As the weight of the cycle consistency loss increases, the attention map includes less background area. In equilibrium, the attention map focuses on the objects of interesting and ignore the background. However, As discussed in [24], the attention map would collapse to black frequently.

As shown in Figure 8(b), the training stability of the attention branch is increased obviously after introducing the attention loss. Although generated attention maps change in shape with different λ_{attn}, these maps would not collapse to black with a high weight of the cycle consistency loss. The quantitative results are displayed on Table 4.

4.5. Ablation Study

To investigate the effect of each term in our full objective, we compare against ablations of our full loss on horse ↔ zebra task. The results are illustrated in Table 4 and shown in Figure 9. First, we test the effect of the smoothness loss, which reduces the absolute gradient of the attention map. As shown in Figure 9(c), removing the smoothness loss, the model ('w/o smooth') produces rougher attention maps and gets slightly higher MAE in both translating processes. Then, we evaluate the importance of our revised adversarial loss by replacing it with an original one. As illustrated in Figure 9(d), the original discriminator encourages the attention branch to attend the whole input image. Furthermore, we take out the fake sample augmentation ('w/o FSA'). Its MAE performance decrease with a large margin. The augmented images are sampled from the neighborhood of the fake, providing information for the discriminator to learn salient object discovery.



(a) Input & Truth (b) Full loss (c) W/o smooth (d) W/o RA (e) W/o FSA (e) W/o cycle (f) W/o attention

Figure 9: Different variants of our method for horse ↔ zebra mapping. We demonstrate our results in every two rows, one for image translation and the other for attention map estimation. From left to right: input image and its ground truth of attention map (input & Truth), full loss, without smoothness loss (w/o smooth), without revised adversarial loss (w/o RA), without fake sample augmentation (w/o FSA), without cycle consistency loss (w/o cycle), without attention loss (w/o attention).

Table 4
Ablation study: FID and MAE for different variants of our method, evaluated on horse ↔ zebra dataset.

Method	Horse → Zebra		Zebra → Horse	
	FID	MAE	FID	MAE
W/o smooth	45.33	0.162	56.12	0.169
W/o RA	47.50	0.193	58.72	0.174
W/o FSA	48.40	0.202	57.27	0.176
W/o cycle	60.22	0.180	60.59	0.183
W/o attention	50.84	0.403	58.23	0.351
Full loss	**38.73**	**0.146**	**53.81**	**0.158**

Table 5
The saliency object translation results with different accuracy in classification network. From top to bottom, we finetune the last layer of classification network for 1, 5, 10 epochs.

Top1 accuracy	Horse → Zebra		Zebra → Horse	
	FID	MAE	FID	MAE
0.9738	40.12	0.183	55.18	0.187
0.9879	39.63	0.168	55.81	0.173
0.9943	**38.73**	**0.146**	**53.81**	**0.158**

We remove the cycle consistency loss, whose motivation is ensuring 1-1 mapping, from the full loss. As a result, we get the highest FID in our ablation experiments, because the cycle loss enforces the generated image to preserve the structure of its input. Finally, we get the highest MAE in our ablation experiments by removing the attention loss, which provides weak supervision on the attention branch. What's more, the attention loss plays an important role in stabilizing the training process of the attention branch.

We also conduct an experiment to evaluate how the accuracy of classification network influences the final results. Results is shown on Table 5. The accuracy of classifier is an important factor for our model. The higher the classification accuracy, the better the final result.

4.6. Comparison to Supervised Results

We also perform a comparison to supervised results at the horse ↔ zebra task. We manual annotate segmentation maps of 1187 horse images and 1449 zebra images. In the supervised manner, the attention branch is constrained by the ground truth instead of the ternary mask $(A(x), A(y))$. As shown in Table 6, SAAGAN with supervision extremely outperforms unsupervised results at MAE score and, in the meantime, the supervised result gets slightly better FID value. With the ground truth of segmentation maps, the attention branch predicts attention maps more accurately, which further conduces to generate more realistic images. Though there is a gap at the accuracy of predicted attention map, unsupervised generated images are just slightly worse than

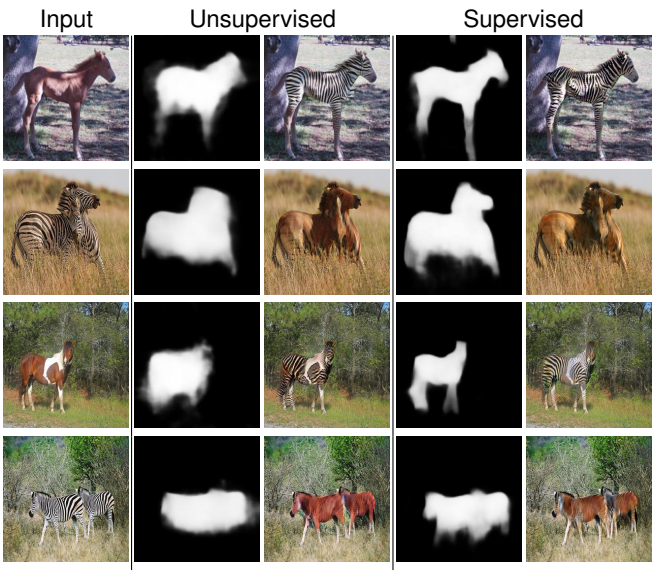


Figure 10: From left to right: input images, attention maps and translated images in an unsupervised manner, attention maps and translated images in a supervised manner.

Table 6
Comparison to supervised results, evaluated on horse ↔ zebra task.

	Task	Unsupervised	Supervised
FID	Horse → Zebra	38.73	34.81
	Zebra → Horse	53.81	49.14
MAE	Horse → Zebra	0.146	0.061
	Zebra → Horse	0.158	0.081

those in a supervised manner. It indicates that the ternary mask is an effective alternative in the case of no ground truth. Some samples are demonstrated in Figure 10 and it can be seen that SAAGAN is able to generate similar images to the supervised method and the later performs better at details. For instance, in the first row of Figure 10, the attention map of the unsupervised method does not cover the horse legs.

4.7. Failure Analysis

Our method relies on the spatial attention prior to provide weak supervision on the attention branch. As the prior is inferred from a classification network, it may fail in some cases. Figure 11 shows typical failure cases of our method. Although we add an attention mechanism in image translation, our model is still limited by some general computer vision problems, e.g., object occlusion or large-scale changes. In the first case, our model fails in dealing with object occlusion, and the attention map hence only includes one horse. The second case is related to errors in the attention mechanism when given a tiny image patch. The attention map does not cover the whole image as the input is completely different from a zebra class.

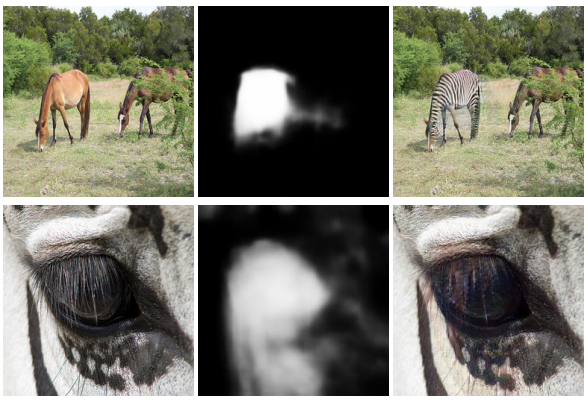


Figure 11: Typical failure cases of our method. Top: in the task of horse → zebra, our method fails in detecting the right horse in the input as it is obscured by trees. Bottom: our method also fails in this zebra → horse example as the tiny patch is completely different from a zebra class.

5. Conclusions

We introduce a novel framework for the salient object translation task. Our attention-guided generator allows simultaneously locating the attention areas in each image and translating the related areas between two domains. Two novel losses, the attention loss and the revised adversarial loss, are proposed to stabilize the training of the new added attention branch. Through a detailed visualization, we can see that the proposed attention loss is able to improve the training ability of the attention branch. We propose a fake sample augment strategy which utilizes the ternary mask to synthesize fake samples. It can stabilize the initial training of GAN since the augmented images provide extra information in object discovery. By leveraging spatial attention prior, the proposed method achieves superior performance in a variety of tasks demonstrated by both qualitative and quantitative experiments. The result indicates that the attention module is especially helpful to focus on the region of interest during the image translation, which further conduces to generate more realistic images.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant U1509210 and Key R&D Program Project of Zhejiang Province (2019C01004).

References

[1] Ahn, J., Cho, S., Kwak, S., 2019. Weakly supervised learning of instance segmentation with inter-pixel relations, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2209–2218.
[2] Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. arXiv:1801.01401 .
[3] Borji, A., Cheng, M.M., Hou, Q., Jiang, H., Li, J., 2014. Salient object detection: A survey. Computational Visual Media , 1–34.
[4] Borji, A., Cheng, M.M., Jiang, H., Li, J., 2015. Salient object de-

tection: A benchmark. IEEE Transactions on Image Processing 24, 5706–5722.

[5] Chen, Q., Koltun, V., 2017. Photographic image synthesis with cascaded refinement networks, in: IEEE International Conference on Computer Vision (ICCV), pp. 1511–1520.

[6] Chen, X., Xu, C., Yang, X., Tao, D., 2018. Attention-gan for object transfiguration in wild images, in: European Conference on Computer Vision (ECCV).

[7] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[8] Creswell, A., Bharath, A.A., 2018. Denoising adversarial autoencoders. IEEE Transactions on Neural Networks and Learning Systems , 1–17.

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 2672–2680.

[10] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

[11] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 6626–6637.

[12] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134.

[13] Jetley, S., Lord, N., Lee, N., Torr, P., 2018. Learn to pay attention, in: International Conference on Learning Representations (ICLR).

[14] Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S., 2013. Salient object detection: A discriminative regional feature integration approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2083–2090.

[15] Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks, in: International Conference on Machine Learning (ICML), pp. 1857–1865.

[16] Kurach, K., Lucic, M., Zhai, X., Michalski, M., Gelly, S., 2018. The gan landscape: Losses, architectures, regularization, and normalization. arXiv:1807.04720 .

[17] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690.

[18] Li, X., Yang, F., Cheng, H., Liu, W., Shen, D., 2018. Contour knowledge transfer for salient object detection, in: European Conference on Computer Vision (ECCV), pp. 355–370.

[19] Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J., 2016. Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing 25, 3919–3930.

[20] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European Conference on Computer Vision (ECCV), pp. 740–755.

[21] Liu, M.Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 700–708.

[22] Liu, M.Y., Tuzel, O., 2016. Coupled generative adversarial networks, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 469–477.

[23] Ma, S., Fu, J., Chen, C.W., Mei, T., 2018. Da-gan: Instance-level image translation by deep attention generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5657–5666.

[24] Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I., 2018. Unsupervised attention-guided image-to-image translation, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 3693–3703.

[25] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations (ICLR).

[26] Miyato, T., Koyama, M., 2018. cgans with projection discriminator. arXiv:1802.05637 .

[27] Qi, Z., Khorram, S., Li, F., 2019. Visualizing deep networks by optimizing with integrated gradients. arXiv:1905.00954 .

[28] Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 .

[29] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 211–252. doi:10.1007/s11263-015-0816-y.

[30] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision (ICCV), pp. 618–626.

[31] Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 .

[32] St-Yves, G., Naselaris, T., 2018. Generative adversarial networks conditioned on brain activity reconstruct seen images, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. pp. 1054–1061.

[33] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826.

[34] Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization.

[35] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001. California Institute of Technology.

[36] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[37] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., 2019a. Salient object detection in the deep learning era: An in-depth survey. arXiv:1904.09146 .

[38] Wang, W., Shen, J., Cheng, M.M., Shao, L., 2019b. An iterative and cooperative top-down and bottom-up inference network for salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5968–5977.

[39] Wang, W., Shen, J., Dong, X., Borji, A., Yang, R., 2019c. Inferring salient objects from human fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[40] Wang, W., Zhao, S., Shen, J., Hoi, S.C., Borji, A., 2019d. Salient object detection with pyramid attention and salient edges, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1448–1457.

[41] Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S., 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1568–1576.

[42] Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation, in: IEEE International Conference on Computer Vision (ICCV), pp. 2868–2876.

[43] Zhang, F., Chan, P.P., Biggio, B., Yeung, D.S., Roli, F., 2015. Adversarial feature selection against evasion attacks. IEEE Transactions on Cybernetics 46, 766–777.

[44] Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S., 2018a. Adversarial complementary learning for weakly supervised object localization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1325–1334.

[45] Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T., 2018b. Self-produced guidance for weakly-supervised object localization, in: European Conference on Computer Vision (ECCV).

[46] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929.

[47] Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J., 2018. Weakly supervised instance segmentation using class peak response, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3791–3800.

[48] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), pp. 2223–2232.