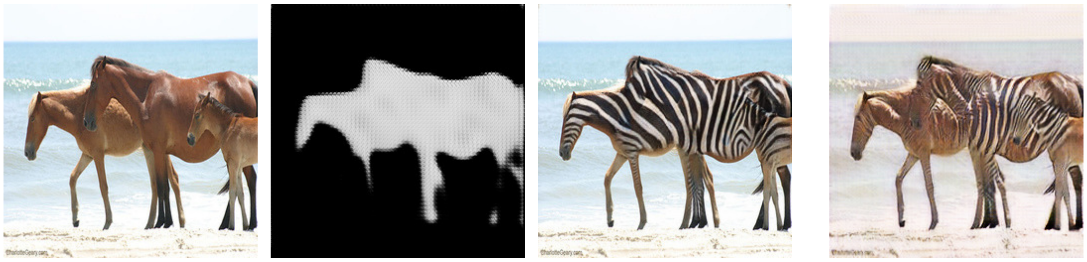
SHOW, ATTEND AND TRANSLATE: UNSUPERVISED IMAGE TRANSLATION WITH SELF-REGULARIZATION AND ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image translation between two domains is a class of problems aiming to learn mapping from an input image in the source domain to an output image in the target domain. It has been applied to numerous applications, such as data augmentation, domain adaptation, and unsupervised training. When paired training data is not accessible, image translation becomes an ill-posed problem. We constrain the problem with the assumption that the translated image needs to be perceptually similar to the original image and also appears to be drawn from the new domain, and propose a simple yet effective image translation model consisting of a single generator trained with a self-regularization term and an adversarial term. We further notice that existing image translation techniques Zhu et al. (2017a); Liu (2017) are agnostic to the subjects of interest and often introduce unwanted changes or artifacts to the input. Thus we propose to add an attention module to predict an attention map to guide the image translation process. The module learns to attend to key parts of the image while keeping everything else unaltered, essentially avoiding undesired artifacts or changes. The predicted attention map also opens door to applications such as unsupervised segmentation and saliency detection. Extensive experiments and evaluations show that our model while being simpler, achieves significantly better performance than existing image translation methods.

1 INTRODUCTION



(a) Input image (b) Predicted Attention Map (c) Final result (d) CycleGAN Zhu et al. (2017a)

Figure 1: Horse→zebra image translation. Our model learns to predict an attention map (b) and translates the horse to zebra while keeping the background untouched (c). By comparison, CycleGAN Zhu et al. (2017a) significantly alters the appearance of the background together with the horse (d).

Many computer vision problems can be cast as an image-to-image translation problem: the task is to map an image of one domain to a corresponding image of another domain. For example, image colorization can be considered as mapping gray-scale images to corresponding images in RGB space Zhang et al. (2016); style transfer can be viewed as translating images in one style to corresponding images with another style Gatys et al. (2016); Johnson et al. (2016); Gatys et al.

(2015). Other tasks falling into this category include semantic segmentation Long et al. (2015a), super-resolution Ledig et al. (2016), image manipulation Isola et al. (2016), etc. Another important application of image translation is related to domain adaptation and unsupervised learning: with the rise of deep learning, it is now considered crucial to have large labeled training datasets. However, labeling and annotating such large datasets are expensive and thus not scalable. An alternative is to use synthetic or simulated data for training, whose labels are trivial to acquire Zhu et al. (2017b); Tzeng et al. (2015a); Rusu et al. (2016); Richter et al. (2016); Qiu & Yuille (2016); Mahendran et al. (2016); Johnson-Roberson et al. (2017); Christiano et al. (2016). Unfortunately, learning from synthetic data can be problematic and most of the time does not generalize to real-world data, due to the data distribution gap between the two domains. Furthermore, due to the deep neural networks' capability of learning small details, it is anticipated that the trained model would easily over-fits to the synthetic domain. In order to close this gap, we can either find mappings or domain-invariant representations at feature level Bousmalis et al. (2016); Ganin et al. (2016); Long et al. (2015b); Sun et al. (2016); Tzeng et al. (2015b); Gretton et al. (2012); Caseiro et al. (2015); Ajakan et al. (2014); Kim et al. (2017) or learn to translate images from one domain to another domain to create "fake" labeled data for training Bousmalis et al. (2017); Zhu et al. (2017a); Liu et al. (2017); Ledig et al. (2016); Liu & Tuzel (2016); Yoo et al. (2016). In the latter case, we usually hope to learn a mapping that preserves the labels as well as the attributes we care about.

Typically there exist two settings for image translation given two domains X and Y. The first setting is supervised, where example image pairs x, y are available. This means for the training data, for each image $x_i \in X$ there is a corresponding $y_i \in Y$, and we wish to find a translator $G : X \rightarrow Y$ such that $G(x_i) \approx y_i$. Representative translation systems in the supervised setting include domain-specific works Eigen & Fergus (2015); Hertzmann et al. (2001); Laffont et al. (2014); Shih et al. (2013); Long et al. (2015a); Wang & Gupta (2016); Xie & Tu (2015); Zhang et al. (2016) and the more general Pix2Pix Isola et al. (2016); Wang et al. (2017). However, paired training data comes at a premium. For example, for image stylization, obtaining paired data requires lengthy artist authoring and is extremely expensive. For other tasks like object transfiguration, the desired output is not even well defined.

Therefore, we focus on the second setting, which is unsupervised image translation. In the unsupervised setting, X and Y are two independent sets of images, and we do not have access to paired examples showing how an image $x_i \in X$ could be translated to an image $y_i \in Y$. Our task is then to seek an algorithm that can learn to translate between X and Y without desired input-output examples. The unsupervised image translation setting has greater potentials because of its simplicity and flexibility but is also much more difficult. In fact, it is a highly under-constrained and ill-posed problem, since there could be unlimited many number of mappings between X and Y: from the probabilistic view, the challenge is to learn a joint distribution of images in different domains. As stated by the coupling theory Lindvall (2002), there exists an infinite set of joint distributions that can arrive the two marginal distributions in two different domains. Therefore, additional assumptions and constraints are needed for us to exploit the structure and supervision necessary to learn the mapping.

Existing works that address this problem assume that there are certain relationships between the two domains. For example, CycleGAN Zhu et al. (2017a) assumes cycle-consistency and the existence of an inverse mapping F that translates from Y to X. It then trains two generators which are bijections and inverse to each other and uses adversarial constraint Goodfellow et al. (2014) to ensure the translated image appears to be drawn from the target domain and the cycle-consistency constraint to ensure the translated image can be mapped back to the original image using the inverse mapping ($F(G(x)) \approx x$ and $G(F(y)) \approx y$). UNIT Liu et al. (2017), on the other hand, assumes shared-latent space, meaning a pair of images in different domains can be mapped to some shared latent representations. The model trains two generators G_X, G_Y with shared layers. Both G_X and G_Y maps an input to itself, while the domain translation is realized by letting x_i go through part of G_X and part of G_Y to get y_i. The model is trained with an adversarial constraint on the image, a variational constraint on the latent code Kingma & Welling (2013); Rezende et al. (2014), and another cycle-consistency constraint.

Assuming cycle consistency ensures 1-1 mapping and avoids mode collapses Salimans et al. (2016), both models generate reasonable image translation and domain adaptation results. However, there are several issues with existing approaches. First, such approaches are usually agnostic to the subjects of interest and there is little guarantee it reaches the desired output. In fact, approaches based on cycle-consistency Zhu et al. (2017a); Liu (2017) could theoretically find any arbitrary 1-1 mapping that satisfies the constraints, and this renders the training unstable and the results random. This is problematic in many image translation scenarios. For example, when translating from a horse image to a zebra image, most likely we only wish to draw the particular black-white stripes on top of the horses while keeping everything else unchanged. However, what we observe is that existing approaches Zhu et al. (2017a); Liu et al. (2017) do not differentiate between the horse/zebra from the scene background, and the colors and appearances of the background often significantly change during translation (Fig. 1). Second, most of the time we only care about one-way translation, while existing methods like CycleGAN Zhu et al. (2017a) and UNIT Liu (2017) always require training two generators of bijections. This is not only cumbersome but it is also hard to balance the effects of the two generators. Third, there is a sensitive trade-off between the faithfulness of the translated image to the input image and how similar it resembles the new domain, and it requires excessive manual tuning of the weight between the adversarial loss and the reconstruction loss to get satisfying results.

To address the aforementioned issues, we propose a simpler yet more effective image translation model that consists of a single generator with an attention module. We first re-consider what the desired outcome of an image translation task should be: most of the time the desired output should not only resemble the target domain but also preserve certain attributes and share similar visual appearance with input. For example, in the case of horse-zebra translation Zhu et al. (2017a), the output zebra should be similar to the input horse in terms of the scene background, the location and the shape of the zebra and horse, etc. In the domain adaptation task that translates MNIST LeCun et al. (2010) to USPS Denker et al. (1989), we expect the output is visually similar to the input in terms of the shape and structure of the digit such that it preserves the label. Based on such observation, our model proposes to use a single generator that maps X to Y and is trained with a self-regularization term that enforces perceptual similarity between the output and the input, together with an adversarial term that enforces the output to appear like drawn from Y. Furthermore, in order to focus the translation on key components of the image and avoid introducing unnecessary changes to irrelevant parts, we propose to add an attention module that predicts a probability map as to which part of the image it needs to attend to when translating. Such probability maps, which are learned in a completely unsupervised fashion, could further facilitate segmentation or saliency detection (Fig. 1). Third, we propose an automatic and principled way to find the optimal weight between the self-regularization term and the adversarial term such that we do not have to manually search for the best hyper-parameter.

Our model does not rely on cycle-consistency or shared representation assumption, and it only learns one-way mapping. Although the constraint is susceptible to oversimplify certain scenarios, we found that the model works surprisingly well. With the attention module, our model learns to detect the key objects from the background context and is able to correct artifacts and remove unwanted changes from the translated results. We apply our model on a variety of image translation and domain adaptation tasks and show that our model is not only simpler but also works better than existing methods, achieving superior qualitative and quantitative performance. To demonstrate its application in real-world tasks, we show our model can be used to improve the accuracy of face 3D morphable model Blanz & Vetter (1999) prediction by augmenting the training data of real images with adapted synthetic images.

2 Our Method

We begin by explaining our model for unsupervised image translation. Let X and Y be two image domains, our goal is to train a generator $G_\theta : X \rightarrow Y$, where θ are the function parameters. For simplicity, we omit θ and use G instead. We are given unpaired samples $x \in X$ and $y \in Y$, and the unsupervised setting assumes that x and y are independently drawn from the marginal distributions $P_{x \sim X}(x)$ and $P_{y \sim Y}(y)$. Let $y' = G(x)$ denote the translated image, the key requirement is that y' should appear like drawn from domain Y, while preserving the low-level visual characteristics of x. The translated images y' can be further used for other downstream tasks such as unsupervised learning. However, in our case, we decouple image translation from its applications.

Based on the requirements described, we propose to learn θ by minimizing the following loss:

$$\mathcal{L}_G = \ell_{adv}(G(x), Y) + \lambda \ell_{reg}(x, G(x)). \tag{1}$$

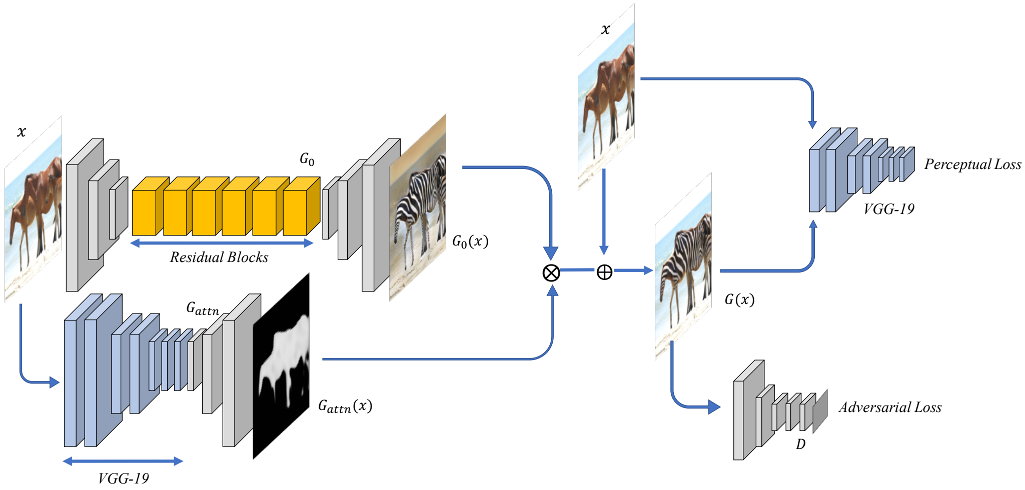


Figure 2: Model overview. Our generator G consists of a vanilla generator G_0 and an attention branch G_{attn}. We train the model using self-regularization perceptual loss and adversarial loss.

Here $G(x) = G_{attn}(x) \otimes G_0(x) + (1 - G_{attn}(x)) \otimes x$, where G_0 is the vanilla generator and G_{attn} is the attention branch. G_0 outputs a translated image while G_{attn} predicts a probability map that is used to composite $G_0(x)$ with x to get the final output. The first part of the loss, ℓ_{adv}, is the adversarial loss on the image domain that makes sure that $G(x)$ appears like domain Y. The second part of the losses ℓ_{reg} makes sure that $G(x)$ is visually similar to x. In our case, ℓ_{adv} is given by a discriminator D trained jointly with G, and ℓ_{reg} is measured with perceptual loss. We illustrate the model in Fig. 2.

The model architectures: Our model consists of a generator G and a discriminator D. The generator G has two branches: the vanilla generator G_0 and the attention branch G_{attn}. G_0 translates the input x as a whole to generate a similar image $G_0(x)$ in the new domain, and G_{attn} predicts a probability map $G_{attn}(x)$ as the attention mask. $G_{attn}(x)$ has the same size as x and each pixel is a probability value between 0-1. In the end, we composite the final image $G(x)$ by adding up x and $G_0(x)$ based on the attention mask.

G_0 is based on Fully Convolutional Network (FCN) and leverages properties of convolutional neural networks, such as translation invariance and parameter sharing. Similar to Isola et al. (2016); Zhu et al. (2017a), the generator G is built with three components: a down-sampling front-end to reduce the size, followed by multiple residual blocks He et al. (2016), and an up-sampling back-end to restore the original dimensions. The down-samping front-end consists of two convolutional blocks, each with a stride of 2. The intermediate part contains nine residual blocks that keep the height/width constant, and the up-sampling back-end consists of two deconvolutional blocks, also with a stride of 2. Each convolutional layer is followed by batch normalization and ReLU activation, except for the last layer whose output is in the image space. Using down-sampling at the beginning increases the receptive field of the residual blocks and makes it easier to learn the transformation at a smaller scale. Another modification is that we adopt the dilated convolution in all residual blocks, and set the dilation factor to 2. Dilated convolutions use spaced kernels, enabling it to compute each output value with a wider view of input without increasing the number of parameters and computational burden. G_{attn} consists of the initial layers of the VGG-19 network Simonyan & Zisserman (2014) (up to *conv3_3*), followed by two deconvolutional blocks. In the end it is a convolutional layer with sigmoid that outputs a single channel probability map. During training, the VGG-19 layers are warm-started with weights pretrained on ImageNet Russakovsky et al. (2015).

For the discriminator, we use a five-layer convolutional network. The first three layers have a stride of 2 followed by two convolution layers with stride 1, which effectively down-samples the networks three times. The output is a vector of real/fake predictions and each value corresponds to a patch of the image. Classifying each patch as real/fake introduces PatchGAN, and is shown to work better than the global GAN Zhu et al. (2017a); Isola et al. (2016).

Adversarial loss: Generative Adversarial Network Goodfellow et al. (2014) plays a two-player min-max game to update the network G and D. G learns to translate the image x to $G(x)$ which appears as if it is from Y, while D learns to distinguish $G(x)$ from y which is the real image drawn from Y. The parameters of D and G are updated alternatively. The discriminator D updates its parameters by maximizing the following objective:

$$\mathcal{L}_D = \log(D(y)) - \log(1 - D(G(x))). \tag{2}$$

The adversarial loss used to update the generator G is defined as:

$$\mathcal{L}_{adv}(G(x), Y) = -\log(-D(G(x))). \tag{3}$$

By minimizing the loss function, the generator G learns to create translated image that fools the network D into classifying the image as drawn from Y.

Self-regularization loss: Theoretically, adversarial training can learn a mapping G that produces outputs identically distributed as the target domain Y. However, if the capacity is large enough, a network can map the input images to any random permutations of images in the target domain. Thus, adversarial loses alone cannot guarantee that the learned function G maps the input to the desired output. To further constrain the learned mapping such that it is meaningful, we argue that G should preserve visual characteristics of the input image. In other words, the output and the input need to share perceptual similarities, especially regarding the low-level features. Such features may include color, edges, shape, objects, etc. We impose this constraint with the self-regularization term, which is modeled by minimizing the distance between the translated image y' and the input x: $\ell_{reg} = d(x, G(x))$. Here d is some distance function d, which can be ℓ_2, ℓ_1, SSIM, etc. However, recent research suggests that using perceptual distance based on a pre-trained network corresponds much better to human perception of similarity comparing with traditional distance measures Zhang et al. (2018). In particular, we defined the perceptual loss as:

$$\ell_{reg}(G(x), x) \qquad = \sum_{l=1,2,3} \frac{1}{H_l W_l} \sum_{h,w} (\| \, w_l \circ (\hat{F}(x)^l_{hw} - \hat{F}(G(x))^l_{hw}) \, \|^2_2). \tag{4}$$

Here $\hat{F}$ is VGG pretrained on ImageNet used to extract the neural features; we use l to represent each layer, and H_l, W_l are the height and width of feature $\hat{F}^l$. We extract neural features with $\hat{F}$ across multiple layers, compute the ℓ_2 difference at each location h, w of $\hat{F}^l$ and average over the feature height and width. We then scale it with layer-wise weight w_l. We did extensive experiments to try different combinations of feature layers and obtained the best results by only using the first three layers of VGG and setting w_1, w_2, w_3 to be $1.0/32, 1.0/16, 1.0/8$ respectively. This conforms to the intuition that we would like to preserve the low-level traits of the input during translation. Note that this may not always be true (such as in texture transfer), but it is a hyper-parameter that could be easily adjusted based on different problem settings. We also experimented with using different pre-trained networks such as AlexNet to extract neural features as suggested by Zhang et al. (2018) but do not observe much difference in results.

Training scheme: In our experiment, we found that training the attention branch and the vanilla generator branch is difficult as it is hard to balance the learned translation and mask. In our practice, we train the two branches separately. First, we train the vanilla generator G_0 without the attention branch. After it converges, we train the attention branch G_{attn} while keeping the trained generator G_0 fixed. In the end, we jointly fine-tune them with a smaller learning rate.

Adaptive weight induction: Like other image translation methods, the resemblance to the new domain and faithfulness to the original image is a trade-off. In our model, it is determined by the weight λ of the self-regularization term relative to the image adversarial term. If λ is too large, the translated image will be close to the input but does not look like the new domain. If λ is too small, the translated image would fail to pertain the visual traits of the input. Previous approaches usually decide the weight heuristically. Here we propose an adaptive scheme to search for the best λ: we start by setting $\lambda = 0$, which means we only use the adversarial constraint to train the generator. Then we gradually increase λ. This would lead to the increase of the adversarial loss as the output would shift away from Y to X, which makes it easier for D to classify. We stop increasing λ when the adversarial loss sinks below some threshold ℓ^t_{adv}. We then keep λ constant and continue to train the network until converging. Using the adaptive weight induction scheme avoids manual tuning of

λ for each specific task and gives results that are both similar to the input x and the new domain Y. Note that we repeat such process both when training G_0 and G_{attn}.

Analysis: Our model is related to CycleGAN in that if we assume 1-1 mapping, we can define an inverse mapping $F : Y \rightarrow X$ such that $F(G(x)) = x$. This satisfies the constraints of CycleGAN in that the cycle-consistency loss is zero. This shows that our learned mapping belongs to the set of possible mappings given by CycleGAN. On the other hand, although CycleGAN tends to learn the mapping such that the visual distance between y' and x is small possibly due to cycle-consistency constraint, it does not guarantee to minimize the perceptual distance between $G(x)$ and x. Comparing with UNIT, if we add another constraint that $G(y) = y$, then it is a special case of the UNIT model where all layers of the two generators are shared which leads to a single generator G. In this case, the cycle-consistency constraint is implicit as $G(G(x)) = G(x)$ and $\min d(x, G(x)) = \min d(x, G(G(x)))$. However, we observe that adding the additional self-mapping constraint for domain Y does not improve the results.

Even though our approach assumes the perceptual distance between x and its corresponding $y \in Y$ is small, our approach generalizes well to tasks where the input and output domains are significantly different, such as translation of photo to map, day to night, etc., as long as our assumption generally holds. For example, in the case of photo to map, the park (photo) is labeled as green (map) and the water (photo) is labeled as blue (map), which provides certain low-level similarities. Experiments show that even without the attention branch, our model produces results consistently similar or better than other methods. This indicates that the cycle-consistency assumption may not be necessary for image translation. Note that our approach is a meta-algorithm, and we could potentially improve the results by using new/more advanced components. For example, the generator and discriminator could be easily replaced with the latest GAN architectures such as LSGAN Mao et al. (2017), WGAN-GP Gulrajani et al. (2017), or adding spectral normalization Miyato et al. (2018). We may also improve the results by employing a more specific self-regularizaton term that is fine-tuned on the datasets we work on.

3 Results

We tested our model on a variety of datasets and tasks. In the following, we show the qualitative results of image translation, as well as quantitative results in several domain adaptation settings. In our experiments, all images are resized to 256x256. We use Adam solver Kingma & Ba (2014) to update the model weights during training. In order to reduce model oscillation, we update the discriminators using a history of generated images rather than the ones produced by the latest generative models Shrivastava et al. (2017): we keep an image buffer that stores the 50 previously generated images. All networks were trained from scratch with a learning rate of 0.0002. Starting from 5k iteration, we linearly decay the learning rate over the remaining 5k iterations. Most of our training takes about 1 day to converge on a single Titan X GPU.

3.1 Qualitative Results

Fig. 3 shows visual results of image translation of horse to zebra. For each image, we show the initial translation $G_0(x)$, the attention map $G_{attn}(x)$ and the final result $G(x)$ composited using $G_0(x)$ and x based on $G_{attn}(x)$. We also compare the results with CycleGAN Zhu et al. (2017a) and UNIT Liu (2017), and all models are trained using the same number of iterations. For the baseline implementation, we use the original authors' implementations. We can see from the examples that without the attention branch, our simple translation model G_0 already gives results similar or better than Zhu et al. (2017a); Liu (2017). However, all these results suffer from perturbations of background color/texture and artifacts near the region of interest. With the predicted attention map which learns to segment the horses, our final results have much higher visual quality, with the background keeping untouched and artifacts near the ROI removed (row 2, 4). Complete results of horse-zebra translations and comparisons are available online [1].

Fig. 4 shows more results on a variety of datasets. We can see that for all these tasks, our model can learn the region of interest and generate compositions that are not only more faithful to the input, but also have fewer artifacts. For example, in dog to cat translation, we notice most attention maps

[1] http://www.harryyang.org/img_trans

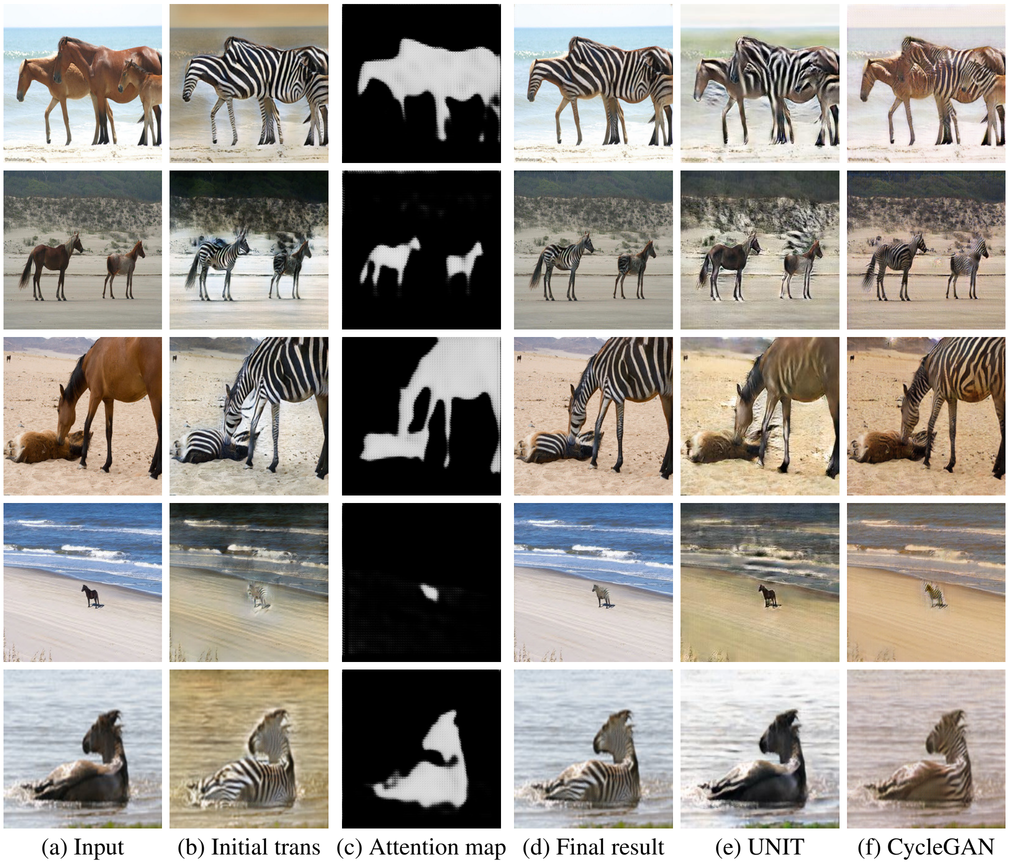


Figure 3: Image translation results of horse to zebra Isola et al. (2016) and comparison with UNIT and CycleGAN.

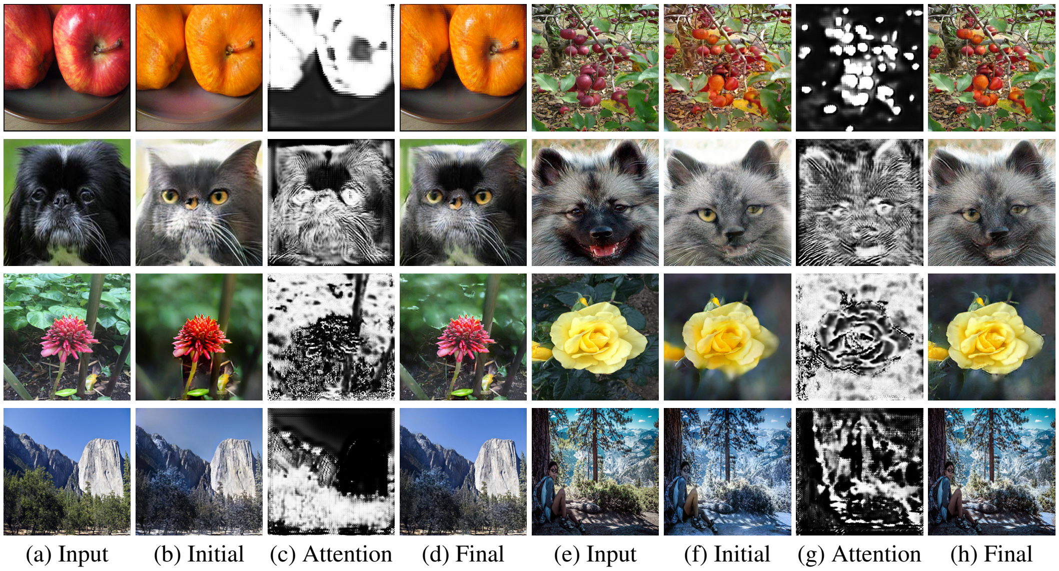


Figure 4: Image translation results on more datasets. From top to bottom: apple to orange Isola et al. (2016), dog to cat Parkhi et al. (2012), photo to DSLR Isola et al. (2016), yosemite summer to winter Isola et al. (2016).

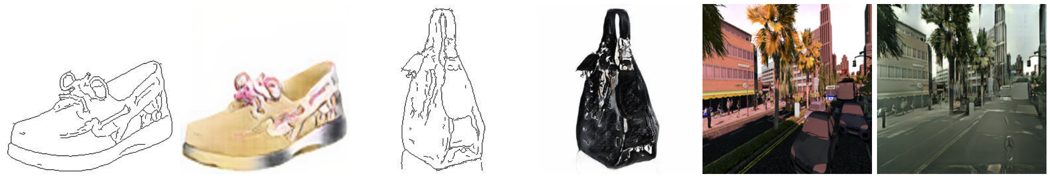


Figure 5: More image translation results. From left to right: edges to shoes Isola et al. (2016); edges to handbags Isola et al. (2016); SYNTHIA to cityscape Ros et al. (2016); Cordts et al. (2015). Given the source and target domains are globally different, the initial translation and final result are similar with the attention maps focusing on the entire images.

have large values around the eyes, indicating the eyes are key ROI to differentiate cats from dogs. In the examples of photo to DSLR, the ROI should be the background that we wish to defocus, while the initial translation changes the color of the foreground flower in the photo. The final result, on the other hand, learns to keep the color of the foreground flower. In the second example of summer to winter translation, we notice the initial result incorrectly changes color of the person. With the guidance of attention map, the final result removes such artifacts.

In a few scenarios, the attention map is less useful as the image does not explicitly contain region of interest and should be translated everywhere. In this case, the composited results largely rely on the initial prediction given by G_0. This is true for tasks like edges to shoes/handbags, SYNTHIA to cityscape (Fig. 5) and photo to map (Fig. 8). Although many of these tasks have very different source and target domains, our method is general and can be applied to get satisfying results.

To better demonstrate the effectiveness of our simple model, Fig. 6 shows several results before training with the attention branch and compares with baseline. We can see that even without the attention branch, our model generates better qualitative results comparing with CycleGAN and UNIT.

Method 1	Method 2	1 better	About same	2 better
Ours initial	CycleGAN	45.0%	39.3%	15.7%
	UNIT	82.7%	15.7	1.6%
Ours final	CycleGAN	70.7%	23.7%	5.6%
	UNIT	89.0%	10.7%	0.3%
	Ours initial	86.3%	11.0%	2.7%

Table 1: User study results.

User study: To more rigorously evaluate the performance, we perform a user study to compare the results. The procedure is as following: we asked for feedbacks from 10 users (all are graduate students). Each user is given 30 sets of images to compare. Each set has 5 images, which are the input, initial result (w/o attention), final result (with attention), CycleGAN results and UNIT results. In total there are 300 different image sets randomly selected from several image translation tasks. The images in each set are in random order. The user is then asked to rank the four results from highest visual quality to lowest. The user is fully informed about the task and is aware of the goal as to translate the input image into a new domain while avoiding unnecessary changes.

Table 1 shows the user-study results. We listed results of: CycleGAN vs ours initial/final; UNIT vs ours initial/final; and ours initial vs ours final. We can see that our results, even without applying the attention branch (*ours initial*), achieves higher ratings than CycleGAN or UNIT. The attention branch also significantly improves the results (*Ours final*). In terms of directly evaluating the effects of attention branch, ours final is overwhelmingly better than ours initial based on user rankings (Table 1 row 5).

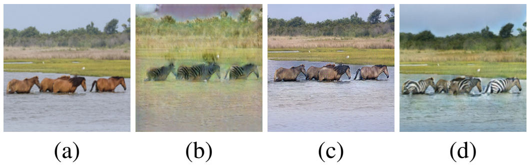
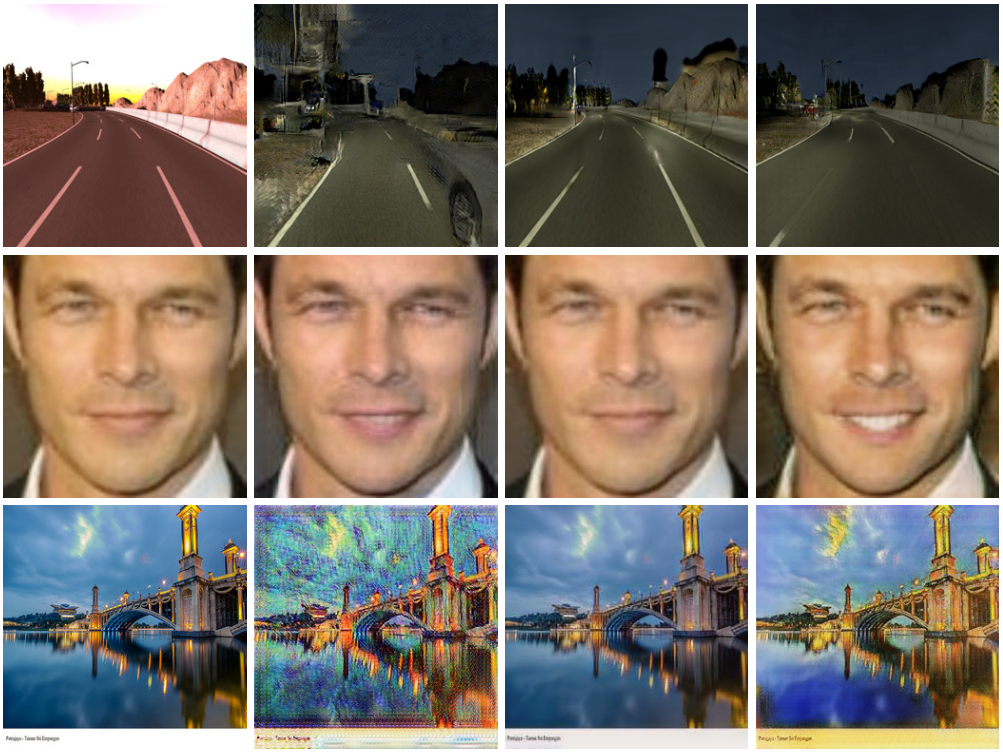


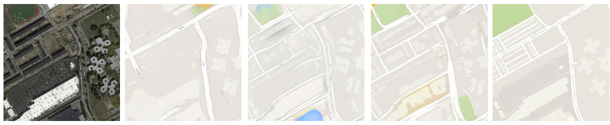
Figure 7: Effects of using different layers as feature extractors. From left to right: input (a), using the first two layers of VGG (b), using the last two layers of VGG (c) and using the first three layers of VGG (d).

Effects of using different layers as feature extractors: We experimented using different layers of VGG-19 as feature extractors to measure the perceptual loss. Fig. 7 shows visual example of the horse to zebra image translation results



(a) Input (b) CycleGAN (c) UNIT (d) Ours w/o attn

Figure 6: Comparing our results w/o attention with baselines. From top to bottom: dawn to night (SYNTHIA Ros et al. (2016)), non-smile to smile (CelebA Liu et al. (2015)) and photos to Vangoh Isola et al. (2016).

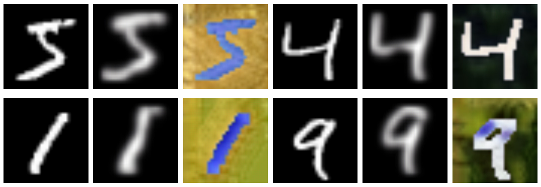


input Pix2Pix CycleGAN Ours GT

Figure 8: Unsupervised map prediction visualization.

Method	Accuracy
Pix2Pix Isola et al. (2016)	43.18%
CycleGAN Zhu et al. (2017a)	45.91%
Ours	**46.72%**

Table 2: Unsupervised map prediction accuracy.



(a) (b) (c) (d) (e) (f)

Figure 9: Visualization of image translation from MNIST (a),(d) to USPS (b),(e) and MNIST-M (c),(f).

Method	USPS	MNIST-M
CoGAN Liu & Tuzel (2016)	95.65%	-
PixelDA Bousmalis et al. (2017)	95.90%	98.20%
UNIT Liu et al. (2017)	95.97%	-
CycleGAN Zhu et al. (2017a)	94.28%	93.16%
Target-only	96.50%	96.40%
Ours	**96.80%**	**98.33%**

Table 3: Unsupervised classification results.

trained with different perceptual terms. We can see that only using high-level features as regularization leads to results that are almost identical to the input (Fig. 7 (c)) while only using low-level features as regularization leads to results that are blurry and noisy (Fig. 7 (b)). We find the balance by adopting the first three layers of VGG-19 as feature extractor which does a good job of image translation and also avoids introducing too many noise or artifacts (Fig. 7 (d)).

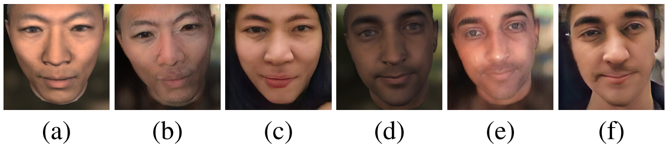


Figure 10: Visualization of rendered face to real face translation. (a)(d): input rendered faces; (b)(e): CycleGAN results; (c)(f): Our results.

Method	MSE
Baseline	2.26
CycleGAN Zhu et al. (2017a)	2.04
Ours	**1.97**

Table 4: Unsupervised 3DMM prediction results (MSE).

3.2 Quantitative Results

Map prediction: We translate images from satellite photos to maps with unpaired training data and compute the pixel accuracy of predicted maps. The original photo-map dataset consists of 1096 training pairs and 1098 testing pairs, where each pair contains a satellite photo and the corresponding map. To enable unsupervised learning, we take the 1096 photos from the training set and the 1098 maps from the test set, using them as the training data. Note that no attention is used here since the change is global and we observe training with attention yields similar results. At test time, we translate the test set photos to maps and again compute the accuracy. If the total RGB difference between the color of a pixel on the predicted map and that on the ground truth is larger than 12, we mark the pixel as wrong. Figure 8 and Table 2 show the visual results and the accuracy results, and we can see our approach achieves highest map prediction accuracy. Note that Pix2Pix is trained with paired data.

Unsupervised classification: We show unsupervised classification results on USPS Denker et al. (1989) and MNIST-M Ganin et al. (2016) in Figure 9 and Table 3. On both tasks, we assume we have access to labeled MNIST dataset. We first train a generator that maps MNIST to USPS or MNIST-M and then use the translated image and original label to train the classifier (we do not apply the attention branch here as we did not observe much difference after training with attention). We can see from the results that we achieve the highest accuracy on both tasks, advancing state-of-the-art. The qualitative results clearly show that our MNIST-translated images both preserve the original label and are also visually similar to USPS/MNIST-M.

3DMM face shape prediction: As a real-world application, we study the problem of estimating 3D face shape, which is modeled with the 3D morphable model (3DMM) Blanz et al. (2002). For a given face, the 3DMM encodes its shape with a 100 dimension vector. The goal of 3DMM regression is to predict the 100 dimension vector and we compare them with the ground truth using mean squared error (MSE). Tran et al. (2017) proposes to train a very deep neural network He et al. (2016) for 3DMM regression. However, in reality, the labeled training data for real faces are expensive to collect. We propose to use rendered faces instead, as their 3DMM parameters are readily available. We first rendered 200k faces as the source domain and use human selfie photo data of 645 face images we collected as the target domain. For test, we use our collected 112 3D-scanned faces as test data. For the purpose of domain adaptation, we first use our model to translate the rendered faces to real faces and use the results as the training data, assuming the 3DMM parameters stay unchanged. The 3DMM regression model structure is 102-layer Resnet He et al. (2016) as in Tran et al. (2017), and was trained with the translated faces. Figure 10 and Table 4 show the qualitative results and the final accuracy of 3DMM regression. From the visual results, we see that our translated face preserves the shape of the original rendered face and has higher quality than using CycleGAN. We also reduced the 3DMM regression error compared with baseline (where we trained on rendered faces and tested on real faces) and the CycleGAN results.

4 Conclusion

We propose to use a simple model with attention for image translation and domain adaption and achieve superior performance in a variety of tasks demonstrated by both qualitative and quantitative measures. We show that the attention module is particularly helpful to focus the translation on region of interest, remove unwanted changes or artifacts, and may also be used for unsupervised segmentation or saliency detection.

REFERENCES

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 202–207. IEEE, 2002.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2016.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 7, 2017.

Rui Caseiro, Joao F Henriques, Pedro Martins, and Jorge Batista. Beyond the shortest path: Unsupervised domain adaptation by sampling subspaces along the spline flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3846–3854, 2015.

Paul Christiano, Zain Shah, Igor Mordatch, Jonas Schneider, Trevor Blackwell, Joshua Tobin, Pieter Abbeel, and Wojciech Zaremba. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518*, 2016.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 1, pp. 3, 2015.

John S Denker, WR Gardner, Hans Peter Graf, Donnie Henderson, RE Howard, W Hubbard, Lawrence D Jackel, Henry S Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In *Advances in neural information processing systems*, pp. 323–331, 1989.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pp. 2414–2423. IEEE, 2016.

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

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

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.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 327–340. ACM, 2001.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 746–753. IEEE, 2017.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)*, 33(4):149, 2014.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

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. *arXiv preprint*, 2016.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Ming-Yu Liu. Unsupervised Image-to-Image Translation. `https://github.com/mingyuliutw/UNIT`, 2017. [Online; accessed 7-May-2018].

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pp. 469–477, 2016.

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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015a.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015b.

Aravindh Mahendran, Hakan Bilen, João F Henriques, and Andrea Vedaldi. Researchdoom and cocodoom: learning computer vision with games. *arXiv preprint arXiv:1610.02431*, 2016.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2813–2821. IEEE, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *European Conference on Computer Vision*, pp. 909–916. Springer, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, 2014.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2016.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3234–3243, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Andrei A Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. *arXiv preprint arXiv:1610.04286*, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics (TOG)*, 32 (6):200, 2013.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, pp. 6, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, volume 6, pp. 8, 2016.

Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1493–1502. IEEE, 2017.

Eric Tzeng, Coline Devin, Judy Hoffman, Chelsea Finn, Xingchao Peng, Sergey Levine, Kate Saenko, and Trevor Darrell. Towards adapting deep visuomotor representations from simulated to real environments. *CoRR, abs/1511.07111*, 2015a.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pp. 4068–4076. IEEE, 2015b.

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. *arXiv preprint arXiv:1711.11585*, 2017.

Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pp. 318–335. Springer, 2016.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision*, pp. 517–532. Springer, 2016.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pp. 649–666. Springer, 2016.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017a.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 3357–3364. IEEE, 2017b.