Realistic Image Generation via Cross-domain Information

Zhao Hongrui
ZheJiang University
Hangzhou,China
hrzhao@zju.edu.cn

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

This paper mainly solve the challenge task that synthesizing realistic images via cross-domain informations(e.g,attributes,texts,etc.).Recently,generative adversarial networks have made great breakthrough in the field of image generation.But if synthesizing image via text-to-image pipeline,part of the generated images will loss the constraint of shape.While synthesizing image via image-to-image pipeline,the shape will be constrained well,but loses semantic control to the image.We design a novel network structure which conbine text-to-image and image-to-image pipeline while only use semantic informations as priors. The image generated by our model not only can't be judged from ground truth,but also match the semantic information descriptions(e.g,shape,color,texture,etc) of object.The networks are based on generative adversarial networks,and we make a new segmentation dataset from CUB-200-2011 as the dataset for shape.

1. Introduction

Realistic image synthesis is a desirable and challenge problem in machine learning tasks.Synthesizing realisitc image has a lot of applications,such as style transfer,facilitating design,entertainment,etc.Deep convolutional neural networks[9] is a perfect tool for image synthesis,of which Generative Adeversarial Networks(GANs)[5]have greatly successed in synthesizing image in recent years.In particular,the proposal of DCGAN[13],SAGAN[6], WGAN[1]has highly improved the quality of image synthesis. The synthesized images not only can't be distinguished from real images,but also become more diverse.

Recently, text-to-image synthesis and image-to-image translation have made tremendous success.As for text-to-image synthesis,Reed et al.[14] solve this problem by a conditional-DCGAN based framework.But the resolution of the image generated by this method is only 64^2.The low resolution constrain the realistic of the object details,although the generated images are highly matched with the meanings of the text.Based on this method,AttnGAN[18] and HDGAN[20]stack another low-to-high resolution GAN to improve the quality of generated high-resolution images instead of using more upsampling layers to generate high-resolution images.These methods have made the generated images with more details and vivid object parts.But in some of the generated images,their objects are not salient.The shape of the object is distortion or absorbed in background.As for image-to-image translation,pix2pix[8] is a classic and successful image-to-image translation framework which uses GANs.It is a common framework for task:predict pixels from pixels.The tasks include labels to street scene,labels to facade,bw to color,aerial to map,day to night,edges to photo,etc.While image-to-image can reserve the structure of orginal image well,but often used in unsupervised learning.Dong et al.[4] do a well job in image-to-image translation controled by semantic information.

In analogy to how human painters draw,we decompose the problem of realistic image synthesis into two more tractable sub-problems with multi-stage GANs.Object shape is first generated by our Stage-I attribute2shape GAN.On the top of our Stage-I attribute2shape GAN,we stack Stage-II shape2image GAN to generate realistic images conditioned on Stage-I results and attributes except for shape.By conditioning on the Stage-I result and the attributes without shape again,Stage-II GAN constrained the shape of object well and draws more detail for the object.This is underlying reason why Stage-II GAN is able to generate images not only constrain the shape of object very well,and add detail information to the object.

2. Related Work

We review work in related areas such as image-to-image translation,text-to-image synthesis,and multi-modal conditional image generation.

Image-to-Image Translation:A lot of researchers leverage GANs for image-to-image translation,whose aim is to convert an input image from one domain to another do-

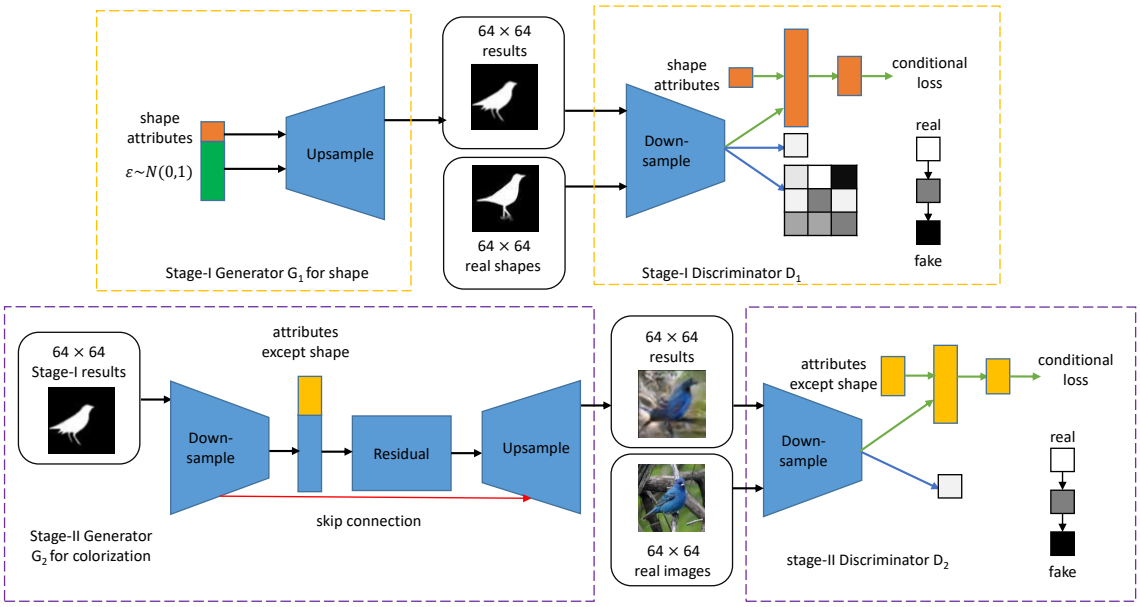


Figure 1. Network architecture of our proposed model.Stage-I GAN consists of a generator network and three discriminator networks.The generator has an deep convolutional generative adversarial network(DC-GAN)architecture and synthesizes shape segmentation conditioned on attributes about shape.One discriminator performs the discriminative task conditioned on shape attributes.Another two discriminators focus on global structures and local structures,respectively.Stage-II GAN is composed of a generator network and two discriminator networks.The generator is based on encoder-decoder architecture as "U-net" with skip connections between mirrored layers in the encoder and decoder stacks.Two discriminators perform the discriminative task conditioned on attributes except for shape and globally classify real or fake,respectively.

main using a training dataset of aligned image pairs.The pix2pix framework of Isola et al.[8]translates one image to another via conditional GANs can be a straightforward approach for different applications,such as generating photographs from sketches or from map of segmentation labels.SketchyGAN[3] synthesize realistic images from human drawn sketches.Chen et al.[2],Qi et al.[12],Wang et al.[17]generate high-resolution photo-realistic images from semantic label maps.Mo et al.[11]incorporates the instance information and improves multi-instance transfiguration.Zhu et al.[21]reveal execellent results on unsupervised image-to-image translation via introducing cycle-consistency losses.

Text-to-Image: The text-to-image problem uses text description as an input to generate an image.It has great advantages over other methods in that it can easily generate an image with the attributes that a user really wants,because text can express detailed high-level information on the appearance of an object with detailed attributions.Reed et al.[14] proposed a novel text-to-image generation model,and Zhang et al.[19],Zhang et al.[20] impoved the image quality later by stacking muliple GANs.Our problem setting is similar to text-to-image,but we use fine-grained attributes as input.

3. Method

To generate salient-object and photo-realistic images from attributes,we propose a simple but powerful Stacked Generative Adversarial Networks.It decomposes the attribute-to-image generative process into two stages.(see Figure 1).

- Stage-I attribute2shape:it drafts the orginal shape of the object conditioned on the given shape attributes with a random noise vector,producing a shape segmention.

- Stage-II shape2image:it combines shape segmention from stage-I with attributes except for shape,yielding a salient-object image with photo-realistic details.

3.1. Generative adversarial networks

Generative adversarial networks consis of two networks:a generator G and a discriminator D that compete with each other as a min-max game.The aim of the discriminator is to distinguish real images and fake images gener-

ated by generator,while the generator is trained to learn the real data distribution p_{data} by generating images that will fool the discriminator.Mathematically, the object G and D will be optimized by playing following min-max game on $V(D,G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_{data}(x)}}\left[\log D(x)\right] + \mathbb{E}_{\boldsymbol{z}\sim \boldsymbol{p_z(z)}}\left[\log(1-D(G(z)))\right],$$

[5] where noise prior z is sampled from a fixed latent distribution p_z(e.g.,uniform or Gausssian distribution) and x denote real image sampled from data distribution p_{data}.

GANs can be extended to Conditional GANs[10].Both the generator and discriminator receive extra conditional priors c.This enforce $G(z,c)$ to generate samples match the conditional priors c.

3.2. Stage-I attribute2shape

Instead of straightforward generating a image with object and details conditioned on the attributes,we decompose the process to first generate a object shape with our stage-I attribute2shape,which focuses on painting only a realisitc shape segmentation of object and correct details for the shape segmentation.

Markovian discriminator[8](PatchGAN):generating shape segmentation is a difficult task,a single global discriminator may over-emphasize certain biased local features and lead to artifacts which is pointed in [16](results are validated in experiments)

To tackle this problem and guarantee shape quality,we try to add local adversarial shape segmentation losses.We expect one discriminator to concentrate on global structure and another discriminators to concentrate on local shape details.The traditional GAN discriminator only force low-frequency correctness.In order to model high-frequencies,it is sensible to pay our attention to the structure in local image patches.Hence,we use a discriminator architecture named PatchGAN that only restrict attention to structure at the level of patches.This discriminator attempt to judge if each N×N patch in an image is real or fake.We perform this discriminator convolutionally through the image,averging all responses to provide the ultimately output of discriminator.In Section 4.1,we proof the importance of Markovian discriminator to get image of high fidelity.This because local details are important to the global expression of the shape.

$$L_{D_{GAN}} = \sum_{i=1}^{s} \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_{data}(x)}}\left[(D_i(x)-1)^2\right] + \mathbb{E}_{\boldsymbol{x_g}\sim \boldsymbol{p_G(\overline{a_s},z)}}\left[(D_i(x_g))^2\right].$$

Here,p_{data} and p_G denote the distributions of real and generated data respectively.$G(\overline{a_s},z)$ means that the output of the generator using the shape attribute a_s and noise z.The first term enforces the discriminator to output 1 for the true input,and the last term is to distinguish the fake image from the real one.Each discriminator D_i computes a $N_i \times N_i$ 2D probility map P_i for the local image loss,we control N_i accordingly to adjust the receptive field of each element in P_i, which distinguishes whether a corresponding local image patch is real or fake. The local GAN loss is come from pixel-to-pixel translation tasks[8].And we use mean-square loss(instead of the origin cross-entropy loss).P_i=1 refers to the global range(largest local).

Conditional discriminator[4]: We denote a_s as matching shape attributes,$\widehat{a_s}$ as mismatching shape attributes,and $\overline{a_s}$ as shape attributes used for synthesizing shape segmentation.s expresses the likelihood of attributes matching with an image x,and x_g denotes the synthesized image from generator $G(\overline{a_s},z)$.

In our method, we train the discriminator D_{cGAN} with three forms of input pairs,and the outputs of discriminator D_{cGAN} are the standalone likelihoods of the three forms:

- $s_r^+ \leftarrow D_{cGAN}(x,a_s)$ for real shape with matching shape attributes;
- $s_w^- \leftarrow D_{cGAN}(x,\widehat{a_s})$ for real shape with mismatching shape attibutes;
- $s_g^- \leftarrow D_{cGAN}(x_g,\overline{a_s})$ for synthesized shape with shape attibutes.

where + and - represent positive and negative examples respectively.

The term s_w^-,proposed by Reed et al.[14],makes the discriminator to generate more obvious image/attribute matching relationship,which enables the generator G to synthesize realistic images that better match the attributes.G synthesizes images via $s_g \leftarrow G(\overline{a_s},z)$ and is optimized adversarially with D_{cGAN}.

$$L_{D_{cGAN}} = \mathbb{E}_{x\sim p_{data}(x)}[(D_{cGAN}(x,a_s)-1)^2] + \mathbb{E}_{x\sim p_{data}(x)}[(D_{cGAN}(x,\widehat{a_s}))^2] + \mathbb{E}_{x_g\sim p_G(\overline{a_s},z)}[(D_{cGAN}(x_g,\overline{a_s}))^2].$$

Full Objective for Generator:

$$L_G = \mathbb{E}_{x_g\sim p_G(\overline{a_s},z)}[(D_{cGAN}(x_g,\overline{a_s})-1)^2 + \sum_{i=1}^{s}\lambda_i(D_i(x_g)-1)^2] + \lambda_0 \boldsymbol{D}_{KL}(N(\mu(\phi(\overline{a_s})),\Sigma(\phi(\overline{a_s}))) \parallel N(0,I)).$$

As is pointed out in stackgan[19].Therefore,we use $\mu(\phi(\overline{a_s}))$ and $\Sigma(\phi(\overline{a_s}))$

3.3. Stage-II shape2image

Skip Connections[8]:one significant feature of image-to-image translation task is that synthesis a high resolution output from high resolution input.What's more,this requires input and output are different in details,but are the same in entirely layout.Therefore,the output layout should be consistent with the layout of the input.The design of the generator in shape2image should take these problems into consideration.

Encoder-decoder network[7] is a common solution to problems in image-to-image translation fields.The input incrementally downsample by going through a set of layers and we get the output via reverse the process at a bottleneck layer.This is a classic framework of encoder-decoder network.In such a network,all information flow go through all the layers,without exception for bottleneck.The output is required to remain many low-frequencies information from the input in lots of image-to-image translation tasks.It is important for the network to progagate low-frequencies information directly through higher resolution layers.

We use skip connections as the globally pattern of a "U-Net"[15].This will make the low-frequencies information go through the generator can bypass the bottleneck.In particular,we use skip connections between each layer i and layer n-i,where n denote the amounts of all layers.Each skip connection directly concatenates entire channels at layer i with those at layer n-i.

Conditional discriminator: We denote a_o as matching attributes except for shape,$\widehat{a_o}$ as mismatching attributes except for shape,and $\overline{a_o}$ as attributes except for shape used for synthesizing realistic images.s expresses the likelihood of attributes except for shape matching with an image x,and x_g denotes the synthesized image from generator $G(\overline{a_o}, s)$,where s denote shape segmentation generated by Stage-I attribute2shape.

In our method, we train the discriminator D_{cGAN} with three forms of input pairs,and the outputs of discriminator D_{cGAN} are the standalone likelihoods of the three forms:

- $s_r^+ \leftarrow D_{cGAN}(x, a_o)$ for real image with matching attributes except for shape;
- $s_w^- \leftarrow D_{cGAN}(x, \widehat{a_o})$ for real image with mismatching attibutes except for shape;
- $s_g^- \leftarrow D_{cGAN}(x_g, \overline{a_o})$ for synthesized shape with attibutes except for shape.

where + and - represent positive and negative examples respectively.

G synthesizes images via $s_g \leftarrow G(\overline{a_o}, s)$ and is optimized adversarially with D_{cGAN}.

$$\begin{aligned} L_{D_{cGAN}} &= \mathbb{E}_{x\sim p_{data}(x)}[(D_{cGAN}(x, a_o) - 1)^2] \\ &+ \mathbb{E}_{x\sim p_{data}(x)}[(D_{cGAN}(x, \widehat{a_o}))^2] \\ &+ \mathbb{E}_{x_g\sim p_G(\overline{a_o}, s)}[(D_{cGAN}(x_g, \overline{a_o}))^2]. \end{aligned}$$

Full Objective for Generator:

$$\begin{aligned} L_G &= \mathbb{E}_{x_g\sim p_G(\overline{a_o}, s)}[(D_{cGAN}(x_g, \overline{a_o}) - 1)^2] \\ &+ \|G(\overline{a_o}, s) - y\|_1 \\ &+ \lambda_0 D_{KL}(N(\mu(\phi(\overline{a_o})), \Sigma(\phi(\overline{a_o})) \parallel N(0, I))) \end{aligned}$$

Since we have paired image data,we are able to provide direct supervision to the network L1-distance between generated images and ground truth images:

$$L_{sup}(G) = \|G(\overline{a_o}, s) - y\|_1$$

4. Experiments

We perform extensive experiments to evaluate the proposed methods.

Datasets. We evaluate our two stage conditional Gan for attribute-to-image synthesis on the CUB datasets.CUB contains 200 bird species with 11788 images.Since 80% of birds in this dataset have object-image size ratios of less than 0.5,as a preprocessing step,we crop all images to ensure that bounding boxes of birds have greater-than-0.75 object-image size ratios.For Stage-I attribute2shape,we create a better segmentation of 3715 images.For Stage-II shape2image,different from preform setting,we don't split CUB into class-disjoint trainging and test sets.

4.1. Analysis of the objective function in stage-I

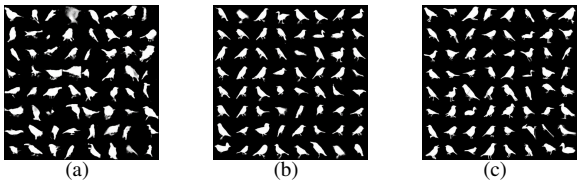


Figure 2. Demonstrate the importance of Markovian discriminator

We test the ability of Markovian discriminator only use 1×1 "PixelGAN" or use both 1×1 "PixelGAN" and 3×3 "PatchGAN" by Stage-I attribute2shape framework but on the set of non-condition generate.In Figure2,(a) is the result of 1×1 "PixelGAN",(b) is the result of 1×1 "PixelGAN" plus 3×3 "PatchGAN" and (c) is some examples of groundtruth.We can conclude that Markovian discriminator is critical to promote the details of shape.

As the results show in Figure3.The synthesis shape is consistent with the groundtruth in attribute if we only observe the image.But the attribute is not strongly correspond to the image.This is the problem of the dataset.

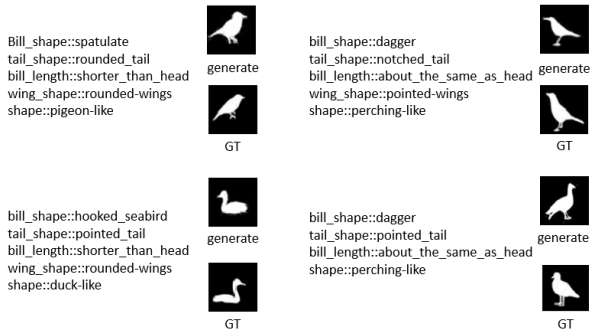
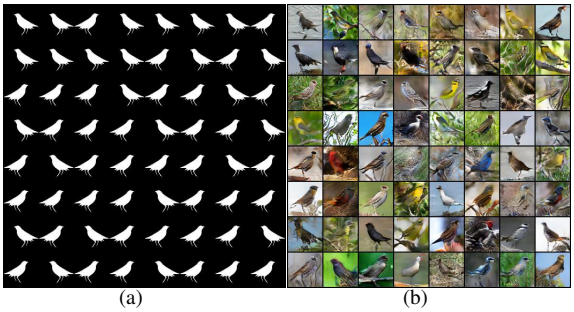


Figure 3. Example results by our Stage-I attribute2image

4.2. Analysis of the objective function in stage-II



(a) (b)

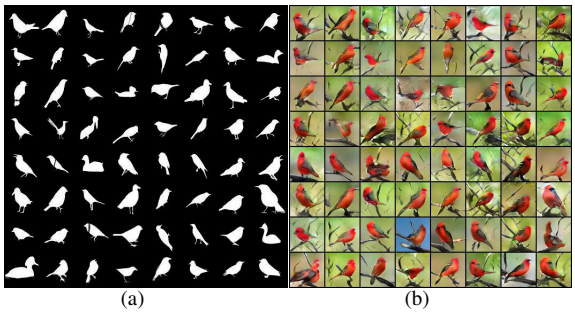
Figure 4.

As is shown in Figure4.If in stage-II shape2image,we use the same shape with different attributes except shape as input.The output image will have a salient object.Although the shape is the same,but the color of the object is different because we use different attributes except shape as input.

attribute name	attribute
wing_color	black
underparts_color	red
breast_color	red
throat_color	red
eye_color	black
forehead_color	red
under_tail_color	black
belly_color	red
primary_color	red
bill_color	black
crown_color	red

Table 1. attributes except shape used in Figure4.2

As is shown in Figure4.2.If in stage-II shape2image,we use the same attributes except shape with different shapes as input.The output image will have different salient object with the same color details.And some attributes except shape in Figure4.2 is in Table1.



(a) (b)

Figure 5.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. 2017.

[2] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[3] W. Chen and J. Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[6] Z. Han, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. 2018.

[7] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[9] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[10] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[11] S. Mo, M. Cho, and J. Shin. Instagan: Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019.

[12] X. Qi, Q. Chen, J. Jia, and V. Koltun. Semi-parametric image synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Computer Science*, 2015.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In

Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, pages 1060–1069. JMLR.org, 2016.

[15] O. Ronneberger. Invited talk: U-net convolutional networks for biomedical image segmentation. In K. H. Maier-Hein, geb. Fritzsche, T. M. Deserno, geb. Lehmann, H. Handels, and T. Tolxdorff, editors, *Bildverarbeitung für die Medizin 2017*, pages 3–3, Berlin, Heidelberg, 2017. Springer Berlin Heidelberg.

[16] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2242–2251, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[18] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[20] Z. Zhang, Y. Xie, and L. Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.