HIERARCHICAL CONDITIONAL SEMI-PAIRED IMAGE-TO-IMAGE TRANSLATION FOR MULTI-TASK IMAGE DEFECT CORRECTION ON SHOPPING WEBSITES

Moyan Li⋆

University of Michigan
Ann Arbor, USA
moyanli@umich.edu

Jinmiao Fu⋆*, Shaoyuan Xu, Huidong Liu, Jia Liu, Bryan Wang*

Amazon
Seattle, USA
{jinmiaof,shaoyux,liuhuido,hliujia,brywan}@amazon.com

ABSTRACT

On shopping websites, product images of low quality negatively affect customer experience. Although there are plenty of work in detecting images with different defects, few efforts have been dedicated to correct those defects at scale. A major challenge is that there are thousands of product types and each has specific defects, therefore building defect specific models is unscalable. In this paper, we propose a unified Image-to-Image (I2I) translation model to correct multiple defects across different product types. Our model leverages an attention mechanism which hierarchically incorporates high-level defect groups and specific defect types to guide the network to focus on defect-related image regions. Evaluated on eight public datasets, our model reduces the Frechet Inception Distance (FID) by 24.6% in average compared with MoNCE, the state-of-the-art I2I method. Another practical challenge on shopping websites is the lack of high quality paired images. We extend our model to be semi-paired by leveraging both paired and unpaired data. Tested on a shopping website dataset to correct three image defects, our model reduces (FID) by 63.2% in average compared with WS-I2I, the state-of-the art semi-paired I2I method.

Index Terms— Image to Image translation, Computer Vision, Image defect auto-correction

1. INTRODUCTION

On shopping websites, product images provide customers with visual perception on the appearance of the products, thus play a critical role for customers' shopping decisions. However, images provided by sellers usually contain various kinds of defects, such as non-white background or watermark. Given the training data per defect (i.e., defective and non-defective images), it is straightforward to build an image classification model to detect defective images. However, reshooting defect-free images is expensive and time-consuming for most sellers, which motivates us to build an ML framework to correct image defects automatically.

Image-to-image (I2I) translation is a promising approach because it can transform images from a source domain (defective) to a target domain (non-defective). I2I has been widely used in image synthesis [1, 2], semantic segmentation [3, 4], image inpainting [5], etc. I2I algorithms can be broadly classified into three categories: paired I2I [6, 3], unpaired I2I [7, 8], and semi-paired I2I [9, 10]. In paired I2I, each image in the source domain is paired with an image in the target domain. In unpaired I2I, data from both domains are available but not paired. Semi-paired I2I leverages both paired and unpaired data.

There are two major limitations of applying I2I methods on the image defects correction of shopping websites: First, I2I methods mainly focus on transforming images from one source domain to one or multiple target domains. They cannot support a single model that transforms multiple source domains to their corresponding target domains. On shopping websites, there are thousands of product types (e.g., shirt) and each has specific defects (e.g., non-white background). Using the existing I2I methods, we need to train thousands of models to cover all the defects, which is unscalable. Second, defects on shopping websites are usually local, i.e., the defects only occupy a certain proportion of the whole image. Most local I2I methods, such as InstaGAN [11], require mask or bounding box labels, which are expensive to obtain. Although there are I2I methods [12, 13] that use attention modules to identify local regions without requiring labels, they do not perform well when the images are from multiple product types with different defects.

To enable the correction of multiple image defects with a single model while accurately locating the defect-related regions, we design a hierarchical attention module. The module leverages the high-level defect groups (e.g., background related) and specific defect types (e.g., non-white background and watermark are two specific defects within the background-related group) to guide the generator to focus on the defect-related regions when trained on images of different defects and product types. Furthermore, on shopping websites, due to the lack of human-audited paired data, the proposed model leverages synthesized paired data. To

⋆These authors contributed equally to this work.

improve the model's robustness against synthesized pairs of low quality, we design our model to be semi-paired by adopting cycle loss on unpaired data. In summary, our contributions include: (1) We propose a unified I2I pipeline to correct multiple image defects across different product types, (2) We hierarchically inject the high-level defect groups and specific defect types using attention modules, which guides the network to focus on defect-related local regions, (3) The proposed model can consume both paired data and unpaired data, while being robust against the synthesized pairs of low quality, (4) The proposed model achieves better FID scores on eight public datasets and a shopping website dataset compared with the state-of-the-art I2I methods.

2. METHOD

Our proposed framework is shown in Figure 1. It contains two generators G_{XY} and G_{YX} to transform images from domain X to Y and from Y to X respectively. For each pair of images (x, y) from domain X and Y, we train a discriminator D_Y to distinguish y and the transformed image $G_{XY}(x)$ using L1 loss. For each unpaired image x in domain X, we train a discriminator D_X to distinguish x and $G_{YX}(G_{XY}(x))$, the generated image by transforming x to domain Y and back to X. We insert the high level defect group g and specific defect type η to guide the two generators to focus on the defect related regions.

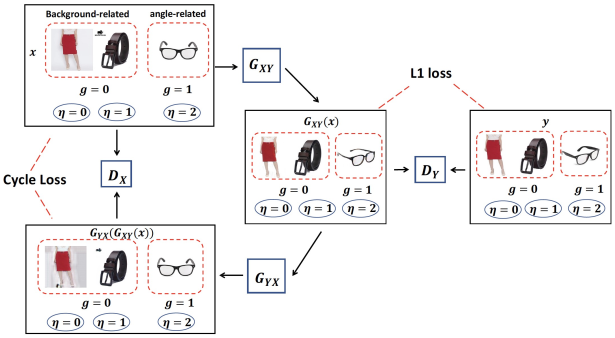

Fig. 1: The architecture of the proposed framework. The discriminators D_Y and D_X leverage the L1 loss and cycle loss as the training objective for paired and unpaired images respectively. The generators G_{XY} and G_{YX} leverage the high level defect group g and the specific defect type η to locate defect related regions.

2.1. Attention Guided Conditional Generator

To cover multiple defects with a single model, we propose an attention-guided conditional generator as shown in Fig.2. It hierarchically consumes the high-level defect group and specific defect type. For example, image defects such as non-white background and watermark are background-related and incorrect sunglasses angle is angle-related. The high-level defect group g (0 = background-related, 1 = angle-related) helps the generator decide whether to focus on distinguishing the background and foreground, or identifying the angle of the object. Furthermore, although non-white background and watermark belong to the same group, watermark usually only occupies a small region. Adding specific defect type η (0 = non-white background, 1 = watermark, 2 = incorrect sunglasses angle) helps the model tell the difference of the intra-group tasks. The red block in Fig. 2 illustrates the procedure of inserting g into the model. Let $I(x) \in \mathbb{R}^{c' \times h \times w}$ denote the feature map of an image x, we first project g into a c'-dimensional vector through a linear transformation with $tanh$ activation, and perform spatial duplication to broadcast it into the same dimension as $I(x)$. Subsequently, we use $M_c(x) \in \mathbb{R}^{c' \times h \times w}$ computed as $M_c(x) = \tanh(\text{Conv}_1(I(x) \odot H_g(x)))$ to measure the relevance of each element in $I(x)$ to g, where $\odot$ is the element-wise multiplication, and Conv_1 is a convolution layer with 1×1 kernel. Then $M'_g(x) \in \mathbb{R}^{h \times w}$, the summation of $M_c(x)$ across all the channels, represents the relevance of each spatial position in $I(x)$ to g. Finally, we denote $I_1(x) \in \mathbb{R}^{c' \times h \times w}$ as the element-wise multiplication between each channel of $I(x)$ and $M'_g(x)$ as the updated feature map, which scales each position of $I(x)$ by its relevance to g. The similar process of inserting specific defect type η is shown in the green box. The conditional labels g and η enable the model to cover multiple defects while focusing on the defect-related image regions.

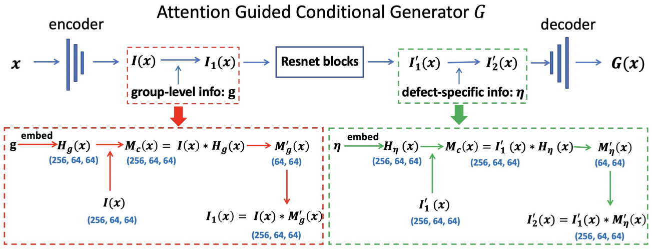

Fig. 2: Attention-guided conditional generator structure. The encoder first maps an image x into feature map $I(x)$. Then the high-level group label g is embedded into the model as shown in the red box. After several Resnet blocks, we embed the specific defect type into the model as shown in the green box. Finally, the decoder will output the generated image.

2.2. Semi-paired Structure

Another challenge on shopping websites is the lack of high-quality paired data. Although we can synthesize paired data, we cannot always guarantee the quality. For example, given an image of non-white background, we can synthesize its paired image of white background by detecting the foreground object and changing the background color to white. However, the synthesized image is not accurate if the foreground object is semi-transparent, of similar color of the background, or placed on another object (see the first line in Fig 6). The I2I model trained using such paired data will memorize the patterns of those low-quality pairs. To mitigate such effect, we leverage the cycle loss of unpaired data to ensure that the transformed images can be converted back.

2.3. Training Loss

The training loss consists of the following components. First, adversarial losses in Eq.1 and Eq.2 ensure the generated images look realistic. Specifically, we adopt the relativistic discriminator [14], i.e., $D_Y^{Rel}(y_1, y_2) = \text{sigmoid}(C(y_1) - C(y_2))$, to measure the probability that y_1 looks more realistic than y_2, where C refers to the non-transformed output of the discriminator. Eq.1 trains the discriminator to favor a real image y against a generative image $\hat{y}$, while Eq.2 trains the generator to generate a $\hat{y}$ that looks more realistic than y.

$$\mathcal{L}_{GAN}^{D}\left(G_{XY}, D_Y^{Rel}, X, Y\right) = \mathbb{E}_{y,\hat{y}}\left[\log D_Y^{Rel}(y, \hat{y})\right] \quad (1)$$

$$\mathcal{L}_{GAN}^{G}\left(G_{XY}, D_Y^{Rel}, X, Y\right) = \mathbb{E}_{y,\hat{y}}\left[\log D_Y^{Rel}(\hat{y}, y)\right] \quad (2)$$

We denote the sum of these two losses as $\mathcal{L}_{GAN}^{XY}$, i.e., the total adversarial loss from domain X to Y. Similarly, we denote $\mathcal{L}_{GAN}^{YX}$ as the total adversarial loss from Y to X. Second, for paired data, we use reconstruction loss (Eq.3) to minimize the distance between an image in the target domain and the image generated from the paired image in the source domain.

$$\mathcal{L}_{L1}\left(G_{XY}, G_{YX}\right) = \mathbb{E}_x \|(G_{XY}(x) - y\|_1 + \mathbb{E}_y \left\|G_{YX}(y) - x\right\|_1 \quad (3)$$

For unpaired data, we use cycle loss (Eq.4) to ensure the generated images can be transformed back to the source domain, which prevents the model from overfitting to the paired images of low quality.

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX}) = {} & \mathbb{E}_x\left[\|G_{YX}(G_{XY}(x)) - x\|_1\right] \\ & + \mathbb{E}_y\left[\|G_{XY}(G_{YX}(y)) - y\|_1\right]\end{aligned} \quad (4)$$

Besides, we add identity loss (Eq. 5), which encourages the generators to be close to an identity mapping when images from the target domain are fed to the generators [15, 7].

$$\mathcal{L}_{idt}\left(G_{XY}, G_{YX}\right) = \mathbb{E}_x \|(G_{YX}(x) - x\|_1 + \mathbb{E}_y \left\|G_{XY}(y) - y\right\|_1 \quad (5)$$

At last, the total loss is defined as

$$\mathcal{L}_{total} = \lambda_1\left(\mathcal{L}_{GAN}^{XY} + \mathcal{L}_{GAN}^{YX}\right) + \lambda_2\mathcal{L}_{L1} + \lambda_3\mathcal{L}_{cycle} + \lambda_4\mathcal{L}_{idt} \quad (6)$$

3. EXPERIMENTS

3.1. Datasets

3.1.1. Public Datasets

We evaluate our model on eight public datasets, including four paired datasets: (i) maps: 1,296 map-to-aerial photo paired images [3], (ii) facades (FA): 606 facade-to-segmentation paired images [16], (iii) edges2shoes (E2S): 50K paired images from UT Zappos50K dataset [17], (iv) edges2handbags (E2H): 20K Amazon Handbag images [18], and four unpaired datasets: (i) horse2zebra (H2Z): 1,267 horse images and 1,474 zebra images [19], (ii) apple2orange (A2O): 1,261 apple images and 1,529 orange images [19], monet2photo (M2P): 1,193 Monet's paintings and 7,038 photos [7], (iv) vangogh2photo (V2P): 800 Vangogh's paintings and 7,038 photos [7]. We split each dataset into training and test (80/20). We combine the four paired datasets where we assign maps and FA to the segmentation-related group (g = 0), and E2S and E2H to the colorization-related group (g = 1). Similarly, we combine the four unpaired datasets where we assign H2Z and A2O to the color-related group (g = 0), and M2P and V2P to the style-related group (g = 1).

3.1.2. Image Defects Dataset

We collected images with three different defects including non-white background (non-Wbg), watermark (WM) and incorrect sunglasses angle (in-SA) from a shopping website[1]. For non-Wbg and WM images, we use salient object detection [20] to detect the objects and clean up the background to construct paired images of white background or no watermarks. We construct the pairs of in-SA images using the main and secondary image of each sunglasses product. In total there are 2,703 non-Wbg image pairs, 4,465 WM image pairs and 8,070 in-SA image pairs. We assign these three specific defect types two high level defect groups where non-white background and watermark belongs to the background-related group ($g = 0$) and incorrect sunglasses angle belongs to the angle-related group ($g = 1$).

3.2. Training Details

For both datasets, we train the models for 300 epochs using Adam optimizer [12] with batch size 1. We set an initial learning rate of 0.0002, which is fixed for the first 150 epochs and decays linearly to zero afterwards. We set λ_1, λ_2, λ_3, λ_4 to be 1, 150, 10 and 10 following WS-I2I [9]. WS-I2I is a semi-paired I2I model which leverages the same loss functions as in this paper. The difference is WS-I2I cannot cover multiple tasks due to the lack of guidance from g and η. We use Frechet Inception Distance (FID) [21] as the evaluation metric, which measures the distance between the distributions of generated images and real images. Lower FID score means better performance.

3.3. Performance Comparison

On the combined four paired public data, we compare the proposed model with the baseline method Pix2Pix[3] and the state-of-the-art method MoNCE [22]. Since there are no unpaired data, we set λ_3 and λ_4 to 0. Table 1 (Paired) shows that our method is consistently better than Pix2Pix and MoNCE. As shown Figure 3, when trained on multiple tasks, MoNCE tends to transform an image from the source domain of a task to the target domain of another task (e.g., transforming a google map to a segmentation instead of an aerial-photo) due to the lack of guidance from g and η. On the combined four unpaired public data, we compare our method with the baseline method CycleGAN [7] and MoNCE [22]. Since there are no paired data, we set λ_2 to 0. We observe similar pattern as the paired data as shown in Table 1 (Unpaired) and Fig 4.

On the Image Defects Dataset to correct three image defects including non-Wbg, WM and in-SA, to avoid the model from memorizing the patterns of the low-quality synthesized

[1] We cannot reveal details on this dataset due to legal constraint.

Table 1: The comparison of FID scores on the paired (maps, FA, E2S and E2H) and unpaired (H2Z, A2O, M2P and V2P) datasets.

	Paired				Unpaired			
	Map	FA	E2S	E2H	H2Z	A2O	M2P	V2P
Baseline	301	266	166	111	185	288	197	141
MoNCE	200	198	131	99	135	203	178	135
Proposed	**107**	**155**	**86**	**89**	**90**	**170**	**145**	**113**

Table 2: FID scores of the proposed model and WS-I2I on correcting non-Wbg, WM and in-SA on the Image Defects Dataset

	non-Wbg	WM	in-SA
proposed	**22**	**20**	**66**
WS-I2I	55	89	138

pairs of non-Wbg and WM images, we configure our model to be semi-paired by combining the reconstruction loss (Eq. 3) of paired data with the cycle loss (Eq. 4) and the identity loss (Eq. 5) of non-Wbg and WM images. We compare the proposed model with WS-I2I [9], the state-of-the-art semi-paired I2I model to the best of our knowledge. As shown in Table 2, our proposed model performs better in all three tasks. Fig 5 shows some examples where our proposed model can transform the images correctly while WS-I2I cannot.

3.4. Ablation Study

We conduct the ablation study to demonstrate the effect of high-level group g, specific defect type η and unpaired data. Adding g and η significantly improves the model performance (by comparing the FID scores in the first line with the second and third line in Table 3). The results also demonstrate the benefit of leveraging unpaired data (by comparing the first line and the fourth line in Table 3). Fig 6 visualizes the benefit of using unpaired data when some of the paired data are of low quality. Specifically, when the object is of similar color of the background, placed on another object, or semi-transparent (first line in Fig 6), the synthesized paired images will be of low quality, which will negatively affect the model performance if trained using only paired data (second line in Fig 6). In comparison, our proposed model (third line in Fig 6) is robust against such situation.

4. CONCLUSION

In this paper, we propose a unified I2I translation framework to transform images from multiple source domains to their corresponding target domains. By hierarchically injecting high-level defect groups and specific defect types using attention modules, the framework can capture different levels of image defect patterns for better defect correction. The model is semi-paired so that it is robust against the low-quality paired data. Our framework is scalable to multiple image defects in various domains and can significantly improve FID compared to the state-of-the-art I2I methods.

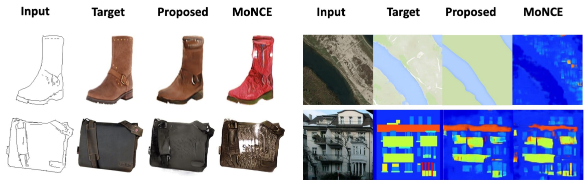

Fig. 3: The comparison between the proposed model and MoNCE on the combined four paired public data.

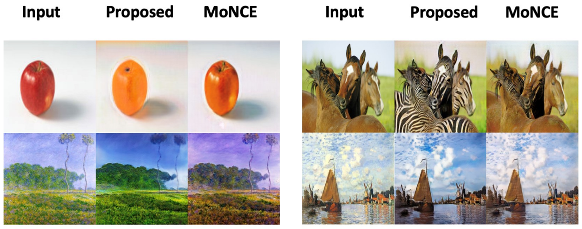

Fig. 4: The comparison between the proposed model and MoNCE on the combined four unpaired public data.

Table 3: FID scores of the proposed model, the proposed model w/o g, w/o g and η, and w/o unpaired data on correcting non-Wbg, WM and in-SA in the Image Defects Dataset

	non-Wbg	WM	in-SA
proposed	**22**	**20**	**66**
w/o g	27	25	70
w/o g, η	55	89	138
w/o unpaired data	43	53	68

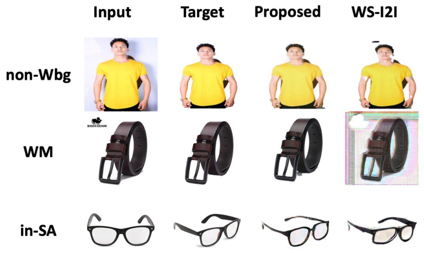

Fig. 5: The comparison between the proposed model with WS-I2I on the Image Defects Detection dataset.

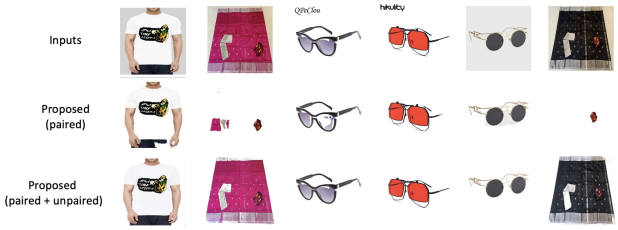

Fig. 6: Visualization of results on the Image Defects Dataset to compare the proposed model and the proposed model w/o unpaired data.

5. REFERENCES

[1] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 3501–3510.

[2] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.

[3] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[4] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8300–8311.

[5] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. Do, "Semantic image inpainting with perceptual and contextual losses," *arXiv preprint arXiv:1607.07539*, vol. 2, no. 3, 2016.

[6] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.

[7] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[8] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865.

[9] S. Shukla, L. Van Gool, and R. Timofte, "Extremely weak supervised image-to-image translation for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3368–3377.

[10] S. Tripathy, J. Kannala, and E. Rahtu, "Learning image-to-image translation using paired and unpaired training samples," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 51–66.

[11] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," *arXiv preprint arXiv:1812.10889*, 2018.

[12] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *arXiv preprint arXiv:1907.10830*, 2019.

[13] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[14] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.

[15] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," *arXiv preprint arXiv:1706.04987*, 2017.

[16] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *German conference on pattern recognition*. Springer, 2013, pp. 364–374.

[17] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 192–199.

[18] J. Zhu, P. Krähenbühl, E. Shechtman, and A. Efros, "Generative visual manipulation on the natural image manifold," in *European conference on computer vision*. Springer, 2016, pp. 597–613.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] M. S. Lee, W. Shin, and S. W. Han, "Tracer: Extreme attention guided salient object tracing network (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 12993–12994.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[22] F. Zhan, J. Zhang, Y. Yu, R. Wu, and S. Lu, "Modulated contrast for versatile image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18280–18290.