Damaged Photo Restoration with High Resolution through StyleGAN Prior

Tu Yang

***Abstract*—In the near 200 years since the invention of photography, we have been posed with the same problem: how do we prevent damage from accruing and ruining the quality of the image. Photos printed onto film and its precursor media can be damaged by exposure to the elements and age, not to mention the fragility of the materials themselves and their sensitivity to acute damage. While digital photographs have removed much of the potential problems including storage and protection, there remains a quality of blurriness inherent in digital photography that isn't in film. This is largely because a 35 mm piece of film is capable of capturing several times as much information as even 4k digital image capture devices. Thus, the individual failings of these two mediums for photography share a similar problem: how to restore or upscale the resolution and quality of those images. Introduced by the author researchers for their paper "Towards Real-World Blind Face Restoration with Generative Facial Prior" by Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan, GFP-GAN is a new GAN architecture designed to upscale the quality of human faces in damaged, aged, and otherwise low resolution photos. In practice, this has a restorative and upscaling affect on the quality of the images, and can be used in conjunction with other models to dramatically raise the quality of images [1].**

Fig. 1: .

I. Introduction

The make up of GFP-GAN is as follows:

First, a degradation removal module (in this case, a vanilla U-Net) takes the damaged photo and removes degradations while extracting latent features. This module notably extracts 2 types of features: the latent features to map the input image to the closest latent StyleGAN2 code, and multi-resolution spatial features for modulating the StyleGAN2 features[1]. Next, a pretrained StyleGAN2 model acts as the generative facial prior. Between the GAN and DRM, the latent features are transformed by several multi layer perceptrons into style vectors. These vectors are then used to produce intermediate convolutional features with the intent of using the spatial features to further modulate the final output. The Channel-Split Feature Transform allows the spatial features to be used to predict the transform parameters that can be used to scale and displace the features in the feature maps in the generator. This only occurs in some channels, so some features are allowed to pass through unchanged if the model doesn't see a need to change them. Finally, the generator perceptual reconstruction loss, adversarial loss, ID loss, and face component loss of the generated images are used to further refine the generated images until training is complete. In practice, this allows the GFP-GAN to radically restore and upscale the quality of the faces of damaged images. When combined with the author's previous work, REAL-ESRGAN, we can use these models to enhance photos far beyond the level of past attempts at the same challenge.

[1]https://blog.paperspace.com/restoring-old-photos-using-gfp-gan/

II. Related Work

Image to Image translation have been around for sometime before the invention of CycleGANs. One really interesting one is the work of Phillip Isola et al in the paper Image to Image Translation with Conditional Adversarial Networks where images from one domain are translated into images in another domain. The dataset for this work consists of aligned pair of images from each domain. Generative image models are well studied and fall into two categories: parametric and nonparametric. The non-parametric models often do matching from a database of existing images, often matching patches of images, and have been used in texture synthesis, super-resolution and in-painting. Parametric models for generating images has been explored extensively (for example on MNIST digits or for texture synthesis). However, generating natural images of the real world have had not much success until recently. A variational sampling approach to generating images has had some success, but the samples often suffer from being blurry. Another approach generates images using an iterative forward diffusion process. Generative Adversarial Networks generated images suffering from being noisy and incomprehensible. The maximum mean discrepancy (MMD) [1] is a measure of the difference between two distributions P and Q given by the supremum over a function space $\mathcal{F}$ of differences between the expectations with regard to two distributions. MMD has been used for deep generative models

[2], [3], [4] and model criticism [5]. WGAN [6] conducted a comprehensive theoretical analysis of how the Earth Mover (EM) distance behaves in comparison with popular probability distances and divergences such as the total variation (TV) distance, the Kullback-Leibler (KL) divergence, and the Jensen-Shannon (JS) divergence utilized in the context of learning distributions.Based on W-div, Wu et al. [7] introduce a Wasserstein divergence objective for GANs (WGAN-div), which can faithfully approximate W-div by optimization. CramerGAN [8] argues that the Wasserstein distance leads to biased gradients, suggesting the Cramér distance between two distributions. Other papers related to WGAN can be found in [9], [10], [11], [12], [13], [14].

Rather than utilizing a single unstructured noise vector z, InfoGAN [15] proposes to decompose the input noise vector into two parts: z, which is seen as incompressible noise; c, which is called the latent code and will target the significant structured semantic features of the real data distribution. Maximizing $I(c; G(z, c))$ means maximizing the mutual information between c and $G(z, c)$ to make c contain as much important and meaningful features of the real samples as possible. However, $I(c; G(z, c))$ is difficult to optimize directly in practice since it requires access to the posterior $P(c|x)$. Fortunately, we can have a lower bound of $I(c; G(z, c))$ by defining an auxiliary distribution $Q(c|x)$ to approximate $P(c|x)$. The discriminator of original GANs [16] is trained to maximize the log-likelihood that it assigns to the correct source [17]. Discriminators of most cGANs based methods [18], [19], [20], [21], [22] feed conditional information y into the discriminator by simply concatenating (embedded) y to the input or to the feature vector at some middle layer. cGANs with projection discriminator [23] adopts an inner product between the condition vector y and the feature vector.

Two-domain I2I can solve many problems in computer vision, computer graphics and image processing, such as image style transfer (f.) [24], [25], which can be used in photo editor apps to promote user experience and semantic segmentation (c.) [26], [27], which benefits the autonomous driving and image colorization (d.) [28], [29]. If low-resolution images are taken as the source domain and high-resolution images are taken as the target domain, we can naturally achieve image super-resolution through I2I (e.) [30], [31]. Indeed, two-domain I2I can be used for many different types of applications as long as the appropriate type and amount of data are provided as the source-target images. Therefore, we refer to the universal taxonomy in machine learning, such as the categorizations used in [32], [33], [34], and classify two-domain I2I methods into four categories based on the different ways of leveraging various sources of information: supervised I2I, unsupervised I2I, semi-supervised I2I and few-shot I2I, as described in following paragraph. We also provide the summary of these two-domain I2I methods including method name, publication year, the type of training data, whether multi-modal or not and corresponding insights.

In the earlier I2I works [35], researchers used many aligned image pairs as the source domain and target domain to obtain the translation model that translates the source images to the desired target images. Training supervised translation is not very practical because of the difficulty and high cost of acquiring these large, paired training data in many tasks. Taking photo-to-painting translation as an example, it is almost impossible to collect massive amounts of labeled paintings that match the input landscapes. Hence, unsupervised methods [24], [36], [37] have gradually attracted more attention. In an unsupervised learning setting, I2I methods use two large but unpaired sets of training images to convert images between representations. In some special scenarios, we still need a little expensive human labeling or expert guidance, as well as abundant unlabeled data, such as those of old movie restoration [38] or genomics [39]. Therefore, researchers consider introducing semi-supervised learning [40], [41], [42] into I2I to further promote the performance of image translation. Semi-supervised I2I approaches leverage only source images alongside a few source-target aligned image pairs for training but can achieve more promoted translated results than their unsupervised counterpart. Nonetheless, several problems remain regarding translation using a supervised, unsupervised or semi-supervised I2I method with extremely limited data. In contrast, humans can learn from only one or limited exemplars to achieve remarkable learning results. As noted by meta-learning [43], [44] and few-shot learning [45], [46], humans can effectively use prior experiences and knowledge when learning new tasks, while artificial learners usually severely overfit without the necessary prior knowledge. Inspired by the human learning strategy, few- and one-shot I2I algorithms [47], [48], [49], [50] have been proposed to translate from very few (or even one) in the limit unpaired training examples of the source and target domains. Although learning settings may differ, most of these I2I techniques tend to learn a deterministic one-to-one mapping and only generate single-modal output. However, in practice, the two-domain I2I is inherently ambiguous, as one input image may correspond to multiple possible outputs, namely, multimodal outputs. Multimodal I2I translates the input image from one domain to a distribution of potential outputs in the target domain while remaining faithful to the input. These diverse outputs represent different color or style texture themes (i.e., multimodal) but still preserve the similar semantic content as the input source image. Therefore, we actually view multimodal I2I as a special two-domain I2I and discuss it.

Supervised I2I aims to translate source images into the target domain with many aligned image pairs as the source domain and target domain for training. In this subsection, we further divide the supervised I2I in two categories: methods with single-modal output and methods with multimodal outputs. The idea of I2I can be traced back to Hertzmann et al.'s image analogies [51], which use a non-parametric texture model for a wide variety of "image filter" effects with an image pair input. More recent research on I2I mainly leverages the deep convolutional neural network to learn the mapping function. Isola et al. [35] first apply conditional GAN to an I2I problem by proposing pix2pix to solve a wide range of supervised I2I tasks. In addition to the pixelwise regression loss $\mathcal{L}_1$ between the translated image and the ground truth, pix2pix leverages adversarial training loss $\mathcal{L}_{cGAN}$ to ensure

Fig. 2: .

that the outputs cannot be distinguished from "real" images. The objective is: Pix2pix is also a strong baseline image translation framework that inspires many improved I2I works based on it, as described in following parts. Wang et al. [52] claim that the GAN loss and pixelwise loss used in pix2pix often lead to blurry results. They present discriminative region proposal adversarial networks (DRPANs) to address it by adding a reviser (R) to distinguish real from masked fake samples. Wang et al. [53] argue that the adversarial training in pix2pix [35] might be unstable and prone to failure for high-resolution image generation tasks. They propose an HD version of pix2pix that can increase the photo realism and resolution of the results to 2048×1024. Moreover, AlBahar et al. [54] take an important step toward addressing the controllable or user-specific generation based on pix2pix [35] via respecting the constraints provided by an external, user-provided guidance image. Unfortunately, pix2pix [35] and its improved variants [52], [53], [54] still fail to capture the complex scene structural relationships through a single translation network when the two domains have drastically different views and severe deformations. Tang et al. [55] therefore proposed SelectionGAN to solve the cross-view translation problem, i.e., translating source view images to target view scenes in which the fields of views have little or no overlap. It was the first attempt to combine the multichannel attention selection module with GAN to solve the I2I problem. What's more, SPADE [26] proposes the spatially-adaptive normalization layer to further improve the quality of the synthesized images. But SPADE uses only one style code to control the entire style of an image and inserts style information only in the beginning of a network. SEAN [27] therefore designs semantic region-adaptive normalization layer to alleviate the two shortcomings. With spatial or style guidance from keypoints map, [56], [57], [58] propose to generated images based on conditional GANs with supervised generation framework. Having said that, Shaham et al. [59] claim that traditional I2I networks [35], [53], [26] suffer from acute computational cost when operating on high-resolution images. They propose to design a more lightweight but efficient enough network ASAPNet for fast high-resolution I2I. Recently, Zhang et al. [60], [61] proposed an exemplar-based I2I framework to translate images by establishing the dense semantic correspondence between cross-domain images. However, the semantic matching process may lead to a prohibitive memory footprint when estimating a high-resolution correspondence. Zhou et al. [62] therefore proposed to reduce the memory cost with hierarchy PatchMatch or bilevel feature alignment while building correspondence.

Actually, this multimodal translation benefits from the solutions of *mode collapse problem* [63], [6], [64], in which the generator tends to learn to map different input samples to the same output. Thus, many multimodal I2I methods [65], [66] focus on solving the mode collapse problem to lead to diverse outputs naturally. BicycleGAN [65] became the first supervised multimodal I2I work by combining cVAE-GAN [67], [68], [69] and cLR-GAN [15], [70], [21] to systematically study a family of solutions to the mode collapse problem and generate diverse and realistic outputs. Similarly, Bansal et al. [66] proposed PixelNN to achieve multimodal and controllable translated results in I2I. They proposed a nearest-neighbor (NN) approach combining pixelwise matching to translate the incomplete, conditioned input to multiple outputs and allow a user to control the translation through on-the-fly editing of the exemplar set. Another solution for producing diverse outputs is to use *disentangled representation* [15], [71], [72], [73] which aims to break down, or disentangle, each feature into narrowly defined variables and encodes them as separate dimensions. When combining it with I2I, researchers disentangle the representation of the source and target domains into two parts: domain-invariant features *content*, which are preserved during the translation, and domain-specific features *style*, which are changed during the translation. In other words, I2I aims to transfer images from the source domain to the target domain by preserving *content* while replacing *style*. Therefore, one can achieve multimodal outputs by randomly choosing the *style* features that are often regularized to be drawn from a prior Gaussian distribution $N(0, 1)$. Gonzalez-Garcia et al. [74] disentangled the representation of two domains into three parts: the *shared* part containing common information of both domains, and two *exclusive* parts that only represent those factors of variation that are particular to each domain. In addition to the bi-directional multimodal translation and retrieval of similar images across domains, they can also transfer a domain-specific transfer and interpolation across two domains.

GANs can be extended to a conditional model [75] if both the discriminator and generator are conditioned on some extra information y. By comparing and, we can see that the generator of InfoGAN is similar to that of cGANs. However, the latent code c of InfoGAN is not known, and it is discovered by training. Furthermore, InfoGAN has an additional network Q to output the conditional variables $Q(c|x)$. Chrysos et al. [76] proposed robust cGANs. Thekumparampil et al. [77] discussed the robustness of conditional GANs to noisy labels. Conditional CycleGAN [78] uses cGANs with cyclic consistency. Mode seeking GANs (MSGANs) [79] proposes a simple yet effective regularization term to address the mode collapse issue for cGANs. GANs are also utilized to achieve image

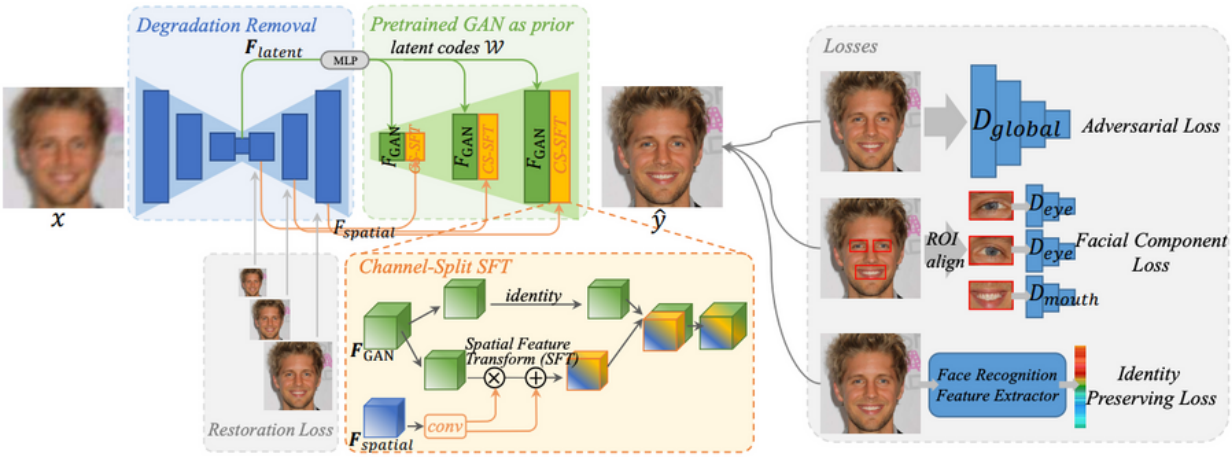

Fig. 3: .

composition [80], [81], [82], domain adaptation [83], [84], [85]. Based on cGANs, we can generate samples conditioning on class labels [17], [86], text [87], [88], [89], bounding box [56]. In [89], [90], text to photo-realistic image synthesis is conducted with stacked generative adversarial networks (SGAN) [91]. cGANs have been used for convolutional face generation [92], face aging [93], multi-modal image translation [55], [94], exemplar-based image synthesis [95], [60], synthesizing outdoor images having specific scenery attributes [96], natural image description [97], and scene manipulation [98], [99]. Image/video colorization [100], [28], [101], [102], [103], [29], image inpainting [104], [105], image super-resolution. Isola et al. [35] used cGANs and sparse regularization for image-to-image translation. The corresponding software is called pix2pix. In GANs, the generator learns a mapping from random noise z to $G(z)$. In contrast, there is no noise input in the generator of pix2pix. A novelty of pix2pix is that the generator of pix2pix learns a mapping from an observed image y to output image $G(y)$, for example, from a grayscale image to a color image. As a follow-up to pix2pix, pix2pixHD [53] used cGANs and feature matching loss for high-resolution image synthesis and semantic manipulation. With the discriminators, the learning problem is a multi-task learning problem. Image-to-image translation is a class of graphics and vision problems where the goal is to learn the mapping between an output image and an input image using a training set of aligned image pairs. When paired training data is available, reference [35] can be used for these image-to-image translation tasks. However, reference [35] can not be used for unpaired data (no input/output pairs), which was well solved by Cycle-consistent GANs (CycleGAN) [24]. CycleGAN is an important progress for unpaired data. It is proved that cycle-consistency is an upper bound of the conditional entropy [106]. CycleGAN can be derived as a special case within the proposed variational inference (VI) framework [107], naturally establishing its relationship with approximate Bayesian inference methods. Specifically, GAN models has been broadly applied in image synthesis [26], [27], [108], [109], The basic idea of DiscoGAN [36] and CycleGAN [24] is nearly the same. Both of them were proposed separately nearly at the same time. The only difference between CycleGAN [24] and DualGAN [110] is that DualGAN uses the loss format advocated by Wasserstein GAN (WGAN) rather than the sigmoid cross-entropy loss used in CycleGAN. Some other works utilize GAN to achieve image transfer [24], [36], [110], [111], [112].

III. Methods

The authors from ARC Tencent came up with GFP-GAN - a new method for real-world blind face restoration that leverages a pretrained GAN and spatial feature transform to restore facial details with a single forward pass. 1) Architecture overview: The proposed framework starts with a U-Net degradation removal module that aims to remove degradations (duh) and extract latent features for mapping the image to the closest StyleGAN-2 latent code as well as a set of multi-resolution spatial features for modulation of StyleGAN-2 intermediate feature maps. The predicted features are then used in a coarse-to-fine manner to synthesize the restored image. 2) Degradation removal module: A vanilla U-Net, except that the authors use L1 loss between the restored images at each resolution scale in the U-Net and the ground-truth image pyramid. The latent features are taken from the bottleneck layer, and the spatial features are obtained from the decoder part of the U-Net. 3) Generative Facial Prior: The latent features from the bottleneck layer of the U-Net are transformed by several MLPs (one per generator layer) to style vectors. The generator does not use the style vectors to output the RGB image right away. Instead, it produces intermediate convolutional features that can be further modulated by the spatial features. 4) Channel-Split Feature Transform: the spatial features are used to predict affine transform parameters (mean and std) that are used to modulate the feature maps in the generator by scaling and shifting them (just like AdaIN in the first StyleGAN). Interestingly the modulation is performed only on some of

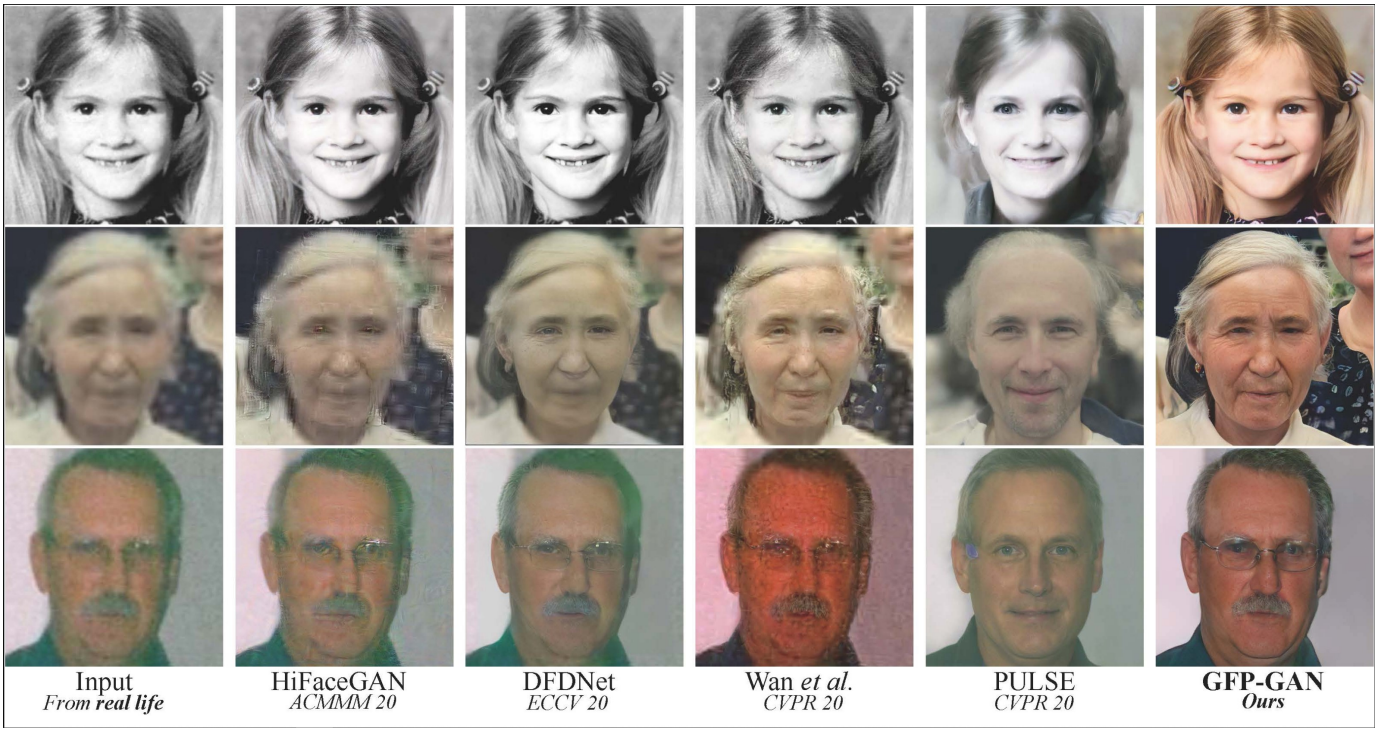

Fig. 4: .

the channels, letting the rest of the features pass through unchanged 5) Model Objectives: The loss is as follows: L1 & perceptual reconstruction loss, adversarial loss, ID loss (ArcFace), and face component loss, which is obtained from several small local discriminators trained on patches of eyes, mouth, etc.

Due to the expensive nature of image generation, it's recommended that you use this package with a GPU on your local or remote machine. We will now go through a quick tutorial for using the Gradient pre-made fork of the GFP-GAN repo to run the package on a remote instance. Log in to Gradient, and navigate to a project space in Gradient you would like to work in. Then, create a new notebook using the button on the top right. Because this package is written in PyTorch, select the PyTorch runtime and suitable GPU for your purposes. This should run fine on the Free GPUs we have available to all users, depending on supply. The final step is to toggle the advanced options at the bottom of the page. Be sure to paste the url for the pre-made fork of the GFP-GAN repo here. Now you can start the notebook. Once your Notebook is ready, open up the notebook "Run-GFPGAN.ipynb." You can use this notebook to run a simple demo using a pretrained GFP-GAN model instance provided by the creators of the repo. You can run all now to see the demo work on the provided sample images, but if you would like to use your own images: they need to be uploaded directly to Gradient.

When you hit run all, it will first install the needed library dependencies. Those in this first cell are all from the same team of researchers, and they facilitate one another. BasicSR is an open source tool kit for image and video restoration, facexlib packages a collection of ready made algorithms for working with facial features, and Real-ESRGAN works to enhance the backgrounds of damaged images much like GFP-GAN restores faces. The next code cell contains the remaining packages needed to ensure that our environment can run GFP-GAN. Finally, we can run the setup.py script to finish setting up our environment to run the generator. We also use a wget to get the pretrained GFP-GAN model provided by the authors for use. To actually run the generator, run the final cell in the notebook containing this command. It will output your newly restored images directly into the newly made results directory.

IV. Conclusion

This tutorial broke down the basic architecture of GFP-GAN, and demonstrated how to use GFP-GAN and its cousin package REAL-esrGAN to dramatically restore aged and damaged photos. While many people do photo restoration as a hobby, this may soon make such efforts much more sophisticated and less time consuming.

References

[1] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[2] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," *arXiv preprint arXiv:1505.03906*, 2015.

[3] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "Mmd gan: Towards deeper understanding of moment matching network," in *Neural Information Processing Systems*, 2017, pp. 2203–2213.

[4] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *International Conference on Machine Learning*, 2015, pp. 1718–1727.

[5] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton, "Generative models and model criticism via optimized maximum mean discrepancy," *arXiv preprint arXiv:1611.04488*, 2016.

[6] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[7] J. Wu, Z. Huang, J. Thoma, D. Acharya, and L. Van Gool, "Wasserstein divergence for gans," in *European Conference on Computer Vision*, 2018, pp. 653–668.

[8] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The cramer distance as a solution to biased wasserstein gradients," *arXiv preprint arXiv:1705.10743*, 2017.

[9] H. Petzka, A. Fischer, and D. Lukovnicov, "On the regularization of wasserstein gans," in *International Conference on Learning Representations*, 2018, pp. 1–24.

[10] F. Juefei-Xu, V. N. Boddeti, and M. Savvides, "Gang of gans: Generative adversarial networks with maximum margin ranking," *arXiv preprint arXiv:1704.04865*, 2017.

[11] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Interspeech*, 2017, pp. 3364–3368.

[12] J. Adler and S. Lunz, "Banach wasserstein gan," in *Neural Information Processing Systems*, 2018, pp. 6754–6763.

[13] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6306–6314.

[14] S. Athey, G. Imbens, J. Metzger, and E. Munro, "Using wasserstein generative adversial networks for the design of monte carlo simulations," *arXiv preprint arXiv:1909.02210*, 2019.

[15] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Neural Information Processing Systems*, 2016, pp. 2172–2180.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014, pp. 2672–2680.

[17] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International Conference on Machine Learning*, 2017, pp. 2642–2651.

[18] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using laplacian pyramid of adversarial networks," in *Neural Information Processing Systems*, 2015, pp. 1486–1494.

[19] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.

[20] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *IEEE International Conference on Computer Vision*, 2017, pp. 2830–2839.

[21] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[22] K. Sricharan, R. Bala, M. Shreve, H. Ding, K. Saketh, and J. Sun, "Semi-supervised conditional gans," *arXiv preprint arXiv:1708.05789*, 2017.

[23] T. Miyato and M. Koyama, "cgans with projection discriminator," *arXiv preprint arXiv:1802.05637*, 2018.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, 2017, pp. 2223–2232.

[25] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, pp. 1–16, 2020.

[26] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] 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 (CVPR)*, June 2020.

[28] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Infrared image colorization based on a triplet dcgan architecture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 18–23.

[29] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[30] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[31] Y. Zhang, S. Liu, C. Dong, X. Zhang, and Y. Yuan, "Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution," *IEEE transactions on Image Processing*, vol. 29, pp. 1101–1112, 2019.

[32] R. Navigli, "Word sense disambiguation: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 2, pp. 1–69, 2009.

[33] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[34] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self- and unsupervised learning for image classification," 2020.

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

[36] 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*, 2017, pp. 1857–1865.

[37] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[38] A. Mustafa and R. K. Mantiuk, "Transformation consistency regularization – a semi-supervised paradigm for image-to-image translation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 599–615.

[39] M. Shi and B. Zhang, "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," *Bioinformatics*, vol. 27, no. 21, pp. 3017–3023, 2011.

[40] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[41] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems*, 2015, pp. 3546–3554.

[42] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5049–5059.

[43] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 2365–2374.

[44] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 403–412.

[45] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[46] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[47] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[48] J. Lin, Y. Pang, Y. Xia, Z. Chen, and J. Luo, "Tuigan: Learning versatile image-to-image translation with two unpaired images," in *European Conference on Computer Vision*. Springer, 2020, pp. 18–35.

[49] J. Lin, Y. Wang, T. He, and Z. Chen, "Learning to transfer: Unsupervised meta domain translation," *arXiv preprint arXiv:1906.00181*, 2019.

[50] J. Lin, Y. Xia, S. Liu, T. Qin, and Z. Chen, "Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation," *arXiv preprint arXiv:1906.00184*, 2019.

[51] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 327–340.

[52] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, and B. Zheng, "Discriminative region proposal adversarial networks for high-quality image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[53] 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," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[54] B. AlBahar and J.-B. Huang, "Guided image-to-image translation with bi-directional feature transformation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[55] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2417–2426.

[56] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Neural Information Processing Systems*, 2016, pp. 217–225.

[57] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie, "Emlight: Lighting estimation via spherical distribution approximation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3287–3295.

[58] F. Zhan, Y. Yu, R. Wu, C. Zhang, S. Lu, L. Shao, F. Ma, and X. Xie, "Gmlight: Lighting estimation via geometric distribution approximation," *arXiv preprint arXiv:2102.10244*, 2021.

[59] T. R. Shaham, M. Gharbi, R. Zhang, E. Shechtman, and T. Michaeli, "Spatially-adaptive pixelwise networks for fast image translation," *arXiv preprint arXiv:2012.02992*, 2020.

[60] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5143–5153.

[61] F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, and C. Miao, "Unbalanced feature transport for exemplar-based image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[62] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, and F. Wen, "Full-resolution correspondence learning for image translation," *arXiv preprint arXiv:2012.02047*, 2020.

[63] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," 2017.

[64] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Neural Information Processing Systems*, 2017, pp. 5767–5777.

[65] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Neural Information Processing Systems*, 2017, pp. 465–476.

[66] A. Bansal, Y. Sheikh, and D. Ramanan, "Pixelnn: Example-based image synthesis," in *International Conference on Learning Representations*, 2018.

[67] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[68] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, p. 1, 2014.

[69] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[70] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.

[71] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[72] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.

[73] E. L. Denton and v. Birodkar, "Unsupervised learning of disentangled representations from video," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4414–4423.

[74] A. Gonzalez-Garcia, J. Van De Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Advances in neural information processing systems*, 2018, pp. 1287–1298.

[75] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[76] G. G. Chrysos, J. Kossaifi, and S. Zafeiriou, "Robust conditional generative adversarial networks," *arXiv preprint arXiv:1805.08657*, 2018.

[77] K. K. Thekumparampil, A. Khetan, Z. Lin, and S. Oh, "Robustness of conditional gans to noisy labels," in *Neural Information Processing Systems*, 2018, pp. 10 271–10 282.

[78] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Conditional cyclegan for attribute guided face image generation," *arXiv preprint arXiv:1705.09966*, 2017.

[79] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1429–1437.

[80] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "St-gan: Spatial transformer generative adversarial networks for image compositing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9455–9464.

[81] S. Azadi, D. Pathak, S. Ebrahimi, and T. Darrell, "Compositional gan: Learning image-conditional binary composition," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2570–2585, 2020.

[82] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion gan for image synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 3653–3662.

[83] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[84] J. Cao, O. Katzir, P. Jiang, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Dida: Disentangled synthesis for domain adaptation," 2018.

[85] A. H. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, "A unified feature disentangler for multi-domain image translation and manipulation," in *Advances in neural information processing systems*, 2018, pp. 2590–2599.

[86] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4467–4477.

[87] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*, 2016, pp. 1–10.

[88] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7986–7994.

[89] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 5907–5915.

[90] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[91] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5077–5086.

[92] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[93] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2089–2093.

[94] F. Zhan, C. Xue, and S. Lu, "Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9105–9115.

[95] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, and F. Wen, "Cocosnet v2: Full-resolution correspondence learning for image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 465–11 475.

[96] L. Karacan, Z. Akata, A. Erdem, and E. Erdem, "Learning to generate images of outdoor scenes from attributes and semantic layouts," *arXiv preprint arXiv:1612.00215*, 2016.

[97] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *IEEE International Conference on Computer Vision*, 2017, pp. 2970–2979.

[98] S. Yao, T. M. Hsu, J.-Y. Zhu, J. Wu, A. Torralba, B. Freeman, and J. Tenenbaum, "3d-aware scene manipulation via inverse graphics," in *Neural Information Processing Systems*, 2018, pp. 1887–1898.

[99] F. Zhan, S. Lu, and C. Xue, "Verisimilar image synthesis for accurate detection and recognition of texts in scenes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 249–266.

[100] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *arXiv preprint arXiv:1705.02999*, 2017.

[101] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–16, 2018.

[102] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061.

[103] Z. Xu, T. Wang, F. Fang, Y. Sheng, and G. Zhang, "Stylization-based architecture for fast deep exemplar colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9363–9372.

[104] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[105] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C.-C. J. Kuo, "Contextual-based image inpainting: Infer, match, and translate," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[106] C. Li, H. Liu, C. Chen, Y. Pu, L. Chen, R. Henao, and L. Carin, "Alice: Towards understanding adversarial learning for joint distribution matching," in *Neural Information Processing Systems*, 2017, pp. 5495–5503.

[107] L. C. Tiao, E. V. Bonilla, and F. Ramos, "Cycle-consistent adversarial learning as approximate bayesian inference," *arXiv preprint arXiv:1806.01771*, 2018.

[108] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[109] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[110] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *International Conference on Computer Vision*, 2017, pp. 2849–2857.

[111] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image-to-image translation," in *Advances in Neural Information Processing Systems*, 2018, pp. 3693–3703.

[112] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara, "Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5849–5859.