

# Image-to-Image Translation: Methods and Applications

Yingxue Pang, Jianxin Lin, Tao Qin, *Senior Member, IEEE*, and Zhibo Chen\*, *Senior Member, IEEE*,

**Abstract**—Image-to-image translation (I2I) aims to transfer images from a source domain to a target domain while preserving the content representations. I2I has drawn increasing attention and made tremendous progress in recent years because of its wide range of applications in many computer vision and image processing problems, such as image synthesis, segmentation, style transfer, restoration, and pose estimation. In this paper, we provide an overview of the I2I works developed in recent years. We will analyze the key techniques of the existing I2I works and clarify the main progress the community has made. Additionally, we will elaborate on the effect of I2I on the research and industry community and point out remaining challenges in related fields.

**Index Terms**—image-to-image translation, two-domain I2I, multi-domain I2I, supervised methods, unsupervised methods, semi-supervised methods, few-shot methods

## I. INTRODUCTION

**I**MAGINE if you take a selfie and want to make it more artistic as a drawing from a cartoonist, how can you automatically achieve that with a computer? This type of research work can be broadly deemed the image-to-image translation (I2I) ([1], [2]) problem. In general, the goal of I2I is to convert an input image  $x_A$  from a source domain  $A$  to a target domain  $B$  with the intrinsic source content preserved and the extrinsic target style transferred. For example, one can take selfie images as the source domain and “translate” them to desired artistic style images given some cartoons as target domain references, as shown in Fig. 1. To this end, we need to train a mapping  $G_{A \rightarrow B}$  that generates image  $x_{AB} \in B$  indistinguishable from target domain image  $x_B \in B$  given the input source image  $x_A \in A$ . Mathematically, we can model this translation process as

$$x_{AB} \in B : x_{AB} = G_{A \rightarrow B}(x_A). \quad (1)$$

From the above basic definition of I2I, we see that converting an image from one source domain to another target domain can cover many problems in image processing, computer graphics, computer vision and so on. Specifically, I2I has been broadly applied in semantic image synthesis [3], [4], [5], [6], [7], image segmentation [8], [9], [10], style transfer [2], [11], [12], [13], [14], image inpainting [15], [16], [17], [18], [19], 3D pose estimation [20], [21], image/video colorization [22],

Yingxue Pang, Jianxin Lin, and Zhibo Chen are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui, 230026, China. (e-mail: pangyx@mail.ustc.edu.cn; linjx@mail.ustc.edu.cn; chenzhibo@ustc.edu.cn.)

Tao Qin are with Microsoft Research Asia. (e-mail:taoqin@microsoft.com)

\* Corresponding author.

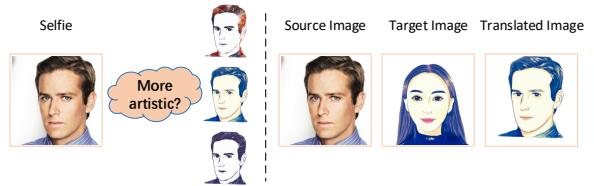


Fig. 1: An example of image-to-image translation (I2I) for illustration. (Left): How to make your selfie more artistic as drawings from cartoonists? This type of research work can be broadly deemed the I2I problem. (Right): You can take a selfie as a source image and a cartoon as a target reference to “translate” it into desired artistic style image.

[23], [24], [25], [26], [27], image super-resolution [28], [29], domain adaptation [30], [31], [32], cartoon generation [33], [34], [35], [36], [37], [38] and image registration [39]. We will analyze and discuss these related applications in detail in Section VI.

In this paper, we aim to provide a comprehensive review of the recent progress in image-to-image translation research works. To the best of our knowledge, this is the first overview paper to cover the analysis, methodology, and related applications of I2I. In detail, we present our survey with the following organization:

- First, we briefly introduce the two most representative and commonly adopted generative models, as well as some well-known evaluation metrics, applied for image-to-image translation, and then we analyze how these generative models learn to represent and acquire the desired translation results.
- Second, we categorize the I2I problem into two main sets of tasks, i.e., two-domain I2I tasks and multi-domain I2I tasks, in which numerous I2I works have appeared for each set of I2I tasks and brought far-reaching influence on other research fields, as shown in Fig. 2.
- Last but not least, we provide a thorough taxonomy of the I2I applications following the same categorizations of I2I methods, as illustrated in Table. V.

In general, our paper is organized as follows. Section I provides the problem setting of the image-to-image translation task. Section II introduces the generative models used for I2I methods. Section III discusses the works on the two-domain I2I task. Section IV focuses on works related to the multi-domain I2I task. Then, Section VI reviews the various and fruitful applications of I2I tasks. Summary and outlook are

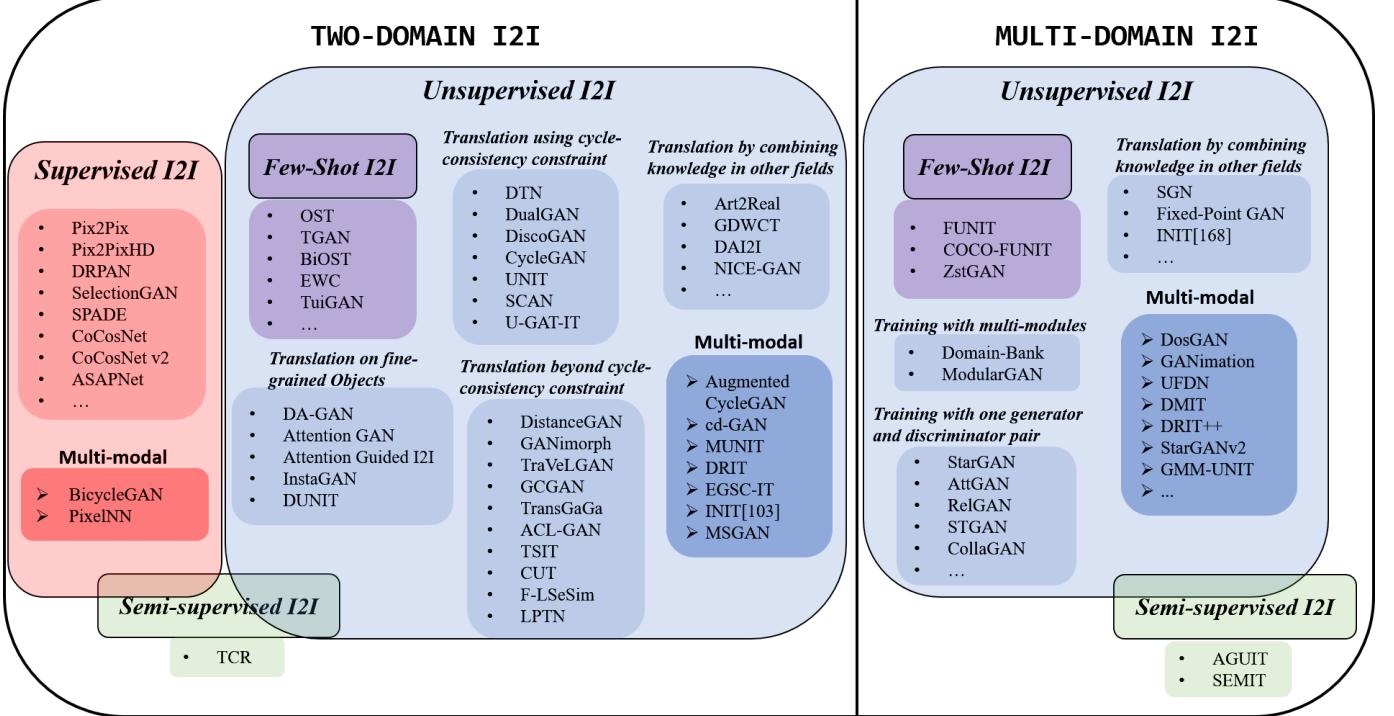


Fig. 2: An overview of image-to-image translation methods. This figure shows the relationship between different methods and where they intersect with each other.

given in Section VII.

## II. THE BACKBONE OF I2I

Because an I2I task aims to learn the mappings between different image domains, how to represent these mappings to generate the desirable results is explicitly related to the generative models. The generative model [40], [41], [42] assumes that data is created by a particular distribution that is defined by two parameters (i.e., a Gaussian distribution) or non-parametric variants (each instance has its own contribution to the distribution), and it approximates that underlying distribution with particular algorithms. This approach enables the generative model to generate data rather than only discriminate between data (classification). For instance, the deep generative models have shown substantial performance improvements in making predictions [43], estimating missing data [44], compressing datasets [45] and generating invisible data. In an I2I task, a generative model can model the distribution of the target domain by producing convincing “fake” data, namely, the translated images, that appear to be drawn from the distribution of the target domain.

However, considering the length of this article and the difference in research foci, we inevitably omit those generative models that are vaguely connected with the theme of I2I, such as deep Boltzmann machines (DBMs) [46], [47], [48], deep autoregressive models (DARs) [49], [50], [51] and normalizing flow models (NFM) [52], [?], [53]. Therefore, we will briefly introduce two of the most commonly used and efficient deep generative models in I2I tasks, variational autoencoders (VAEs) [54], [55], [56], [57], [58], [52], [59], [60], [61], [62]

and generative adversarial networks (GANs) [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], as well as the intuition behind them. Both models basically aim to construct a replica  $x = g(z)$  for generating the desired samples  $x$  from the latent variable  $z$ , but their specific approaches are different. A VAE models data distribution by maximizing the lower bound of the data log-likelihood, whereas a GAN tries to find the Nash equilibrium between a generator and discriminator.

On the other hand, after obtaining the translated results from the generative model, we need subjective and objective metrics for evaluating the quality of translated images. Therefore, we will also briefly present common evaluation metrics in the I2I problem.

### A. Variational AutoEncoder

Inspired by the Helmholtz machine [54], the variational autoencoder (VAE) [55], [56] was initially proposed for a variational inference problem in deep latent Gaussian models.

As shown in Fig 3, a VAE [55], [56] adopts a recognition model (encoder)  $q_\phi(z|x)$  to approximate the posterior distribution  $p(z|x)$  and a generative model (decoder)  $p_\theta(x|z)$  to map the latent variable  $z$  to the data  $x$ . Specifically, a VAE trains its generative model to learn a distribution  $p(x)$  to be near the given data  $x$  by maximizing the log-likelihood function  $\log p_\theta(x)$ :

$$\begin{aligned} \log p_\theta(x) &= \sum_{i=1}^N \log p_\theta(x_i), \\ \log p_\theta(x_i) &= \log \int p_\theta(x_i|z)p(z)dz. \end{aligned} \quad (2)$$

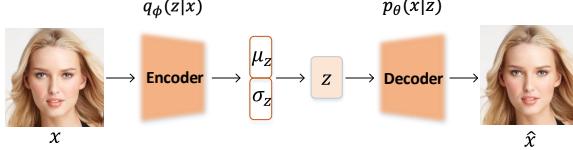


Fig. 3: The structure of a VAE

Stochastic gradient ascent (SGA) combined with the naive Monte Carlo gradient estimator (MCGE) can be used to find the optimal solution in Eqn.(2). However, it often fails because of the highly skewed samples  $p_\theta(x|z)$  that exhibit a very high variance. A VAE therefore introduces the recognition model  $q_\phi(z|x)$  as a multivariate Gaussian distribution with a diagonal covariance structure:

$$q_\phi(z|x) = \mathcal{N}(z|\mu_z(x, \phi), \sigma_z^2(x, \phi)I). \quad (3)$$

Eqn.(2) can be rewritten as:

$$\log p_\theta(x_i) = \mathcal{L}(x_i, \theta, \phi) + D_{KL}[q_\phi(z|x_i)||p_\theta(z|x_i)]. \quad (4)$$

where  $D_{KL}$  denotes the KL divergence that is non-negative, and  $\theta$  and  $\phi$  are neural network parameters. Naturally, we can obtain a variational lower bound on the log-likelihood:

$$\log p_\theta(x_i) \geq \mathcal{L}(x_i, \theta, \phi). \quad (5)$$

Hence, a VAE differentiates and optimizes the lower bound  $\mathcal{L}(x_i, \theta, \phi)$  instead of  $\log p_\theta(x_i)$ . Here is the final objective function of a VAE:

$$\mathcal{L}(x_i, \theta, \phi) = E_{z \sim q_\phi(z|x_i)}[\log p_\theta(x_i|z)] - D_{KL}[q_\phi(z|x_i)||p_\theta(z|x_i)]. \quad (6)$$

As stated in [73], VAEs provide more stable training than generative adversarial networks (GANs) [63] and more efficient sampling mechanisms than autoregressive models [57], [58]. However, several practical and theoretical challenges of VAEs remain unsolved. The main drawback of variational methods is their tendency to strike an unsatisfactory trade-off between the sample quality and the reconstruction quality because of the weak approximate posterior distribution or overly simplistic posterior distribution. The studies in [52], [59], [60] enrich the variational posterior to alleviate the blurriness of generated samples. Tomczak et al. [61] proposed a new prior, VampPrior, to learn more powerful hidden representations. In addition, [62] claimed that the inherent over-regularization induced by the KL divergence term in the VAE objective often leads to a gap between  $\mathcal{L}(x_i, \theta, \phi)$  and the true likelihood.

Generally, with the development of VAEs, this type of generative model constitutes one well-established approach for I2I tasks [16], [74], [75], [76], [77]. Next, we will introduce another important generative model, generative adversarial networks, which have been widely used in multiple I2I models [1], [78], [2], [79], [80].

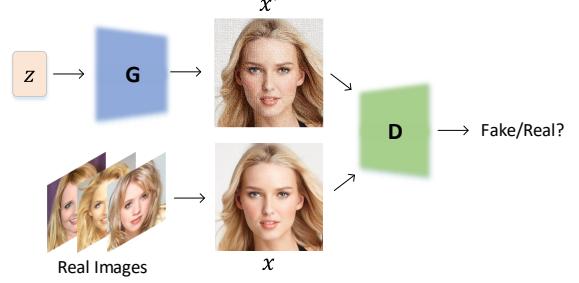


Fig. 4: The structure of unconditional GANs, where  $z, G$  and  $D$  denote the random noise, generator, and discriminator, respectively.

### B. Generative Adversarial Networks

The main idea of generative adversarial networks (GANs) [63], [64], [65] is to establish a zero-sum game between two players, namely, a generator and discriminator, in which each player is represented by a differentiable function controlled by a set of parameters. Generator  $G$  tries to generate fake but plausible images, while discriminator  $D$  is trained to distinguish the difference between real and fake images. The solution of this game is to find a Nash equilibrium between the two players. In the following subsection, we will discuss the unconditional GANs, the conditional GANs and the way to train GANs.

1) *Unconditional GANs*: The original GAN proposed by [63] can be considered an unconditional GAN. It adopts the multilayer perceptron (MLP) [81] to construct a structured probabilistic model taking latent noise variables  $z$  and observed real data  $x$  as inputs. Because the convolutional neural network (CNN) [82] has been demonstrated to be more effective than the MLP in representing image features, the studies in [66] proposed the deep convolutional generative adversarial networks (DGANs) to learn a better representation of images and improve the original GAN performance.

As illustrated in Fig 4, the generator  $G$  inputs a random noise  $z$  sampled from the model's prior distribution  $p(z)$  to generate a fake image  $G(z)$  to fit the distribution of real data as much as possible. Then, the discriminator  $D$  randomly takes the real sample  $x$  from the dataset and the fake sample  $G(z)$  as input to output a probability between 0 and 1, indicating whether the input is a real or fake image. In other words,  $D$  wants to discriminate the generated fake sample  $G(z)$  while  $G$  intends to create samples to confuse  $D$ . Consequently, the objective optimization problem is as shown below:

$$\min_G \max_D \mathcal{L}(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (7)$$

where  $x$  denotes the real data,  $z$  denotes the random noise vector, and  $G(z)$  are the fake samples generated by the generator  $G$ .  $D(x)$  indicates the probability that  $D$ 's input is real, and  $D(G(z))$  denotes the probability that  $D$  discriminates between the input generated by  $G$ .

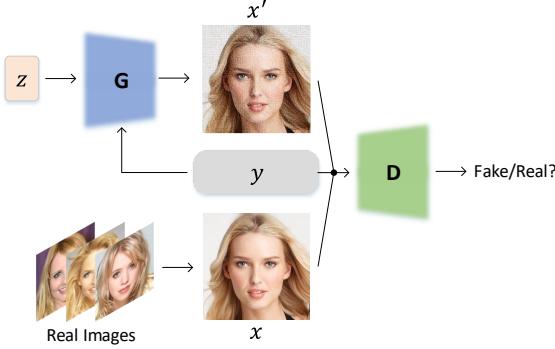


Fig. 5: The structure of conditional GANs, where  $z$ ,  $G$  and  $D$  denote the random noise, generator, and discriminator, respectively. Conditional GANs usually add additional information  $y$  (such as data labels, text or attributes of images) to the generator and discriminator to generate desirable results.

2) *Conditional GANs*: In the unconditional GAN, there is no control of what we want to generate because the only input is the random noise vector  $z$ . Therefore, [67] proposed adding additional information  $y$  concatenated with  $z$  to generate image  $G(z|y)$  shown in Fig 5. The conditional input  $y$  can be any information, such as data labels, text and attributes of images. In this way, we can use the additional information to adjust the generated results in a desirable direction. The objective function is described as:

$$\min_G \max_D \mathcal{L}(D, G) = E_{x \sim p_{data}(x)}[\log D(x|y)] + E_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \quad (8)$$

Note that the real data is also under the control of the same conditional variable  $y$ , i.e.,  $D(x|y)$ .

3) *The Way to Train GANs*: In the training process, GAN updates the parameters of  $G$  along with  $D$  using gradient-based optimization methods, such as stochastic gradient descent (SGD), Adam [83] and RMSProp [84]. The entire optimization goal is achieved when  $D$  cannot distinguish between the generated sample  $x' = G(z)$  and real sample  $x$ , i.e., when the Nash equilibrium is found in this status. In practice, the training of GANs is often trapped in mode collapse, and it is difficult to achieve convergence.

To address the stability problem, many recent studies focus on finding new cost functions with smoother non-vanishing or non-exploding gradients everywhere. WGAN[64] proposes a new cost function using the Wasserstein distance to address the mode collapse problem appearing in naive GAN [63], and WGAN-GP[68] uses a gradient penalty instead of the weight clipping to enforce the Lipschitz constraint in WGAN. LSGAN[65] finds that optimizing the least squares cost function is identical to optimizing a Pearson  $\chi^2$  divergence. EBGAN[69] replaces the discriminator with an autoencoder and uses the reconstruction cost (MSE) to criticize the real and generated images. BEGAN[70] builds with the same EBGAN autoencoder concept for the discriminator but with different cost functions. RSGAN[71] measures the probability

that the real data is more realistic than the generated data, making the cost function relativistic. SNGAN[72] proposes a weight normalization technique called spectral normalization to stabilize the training of the discriminator.

### C. Evaluation Metrics

To reflect the visual quality of the translation performance more comprehensively, we also introduce some common evaluation metrics used in I2I, including subjective and objective metrics.

#### 1) Subjective image quality assessment:

- **AMT perceptual studies**: This test is a “real or fake” two-alternative forced choice experiment on the Amazon Mechanical Turk (AMT) used in many I2I algorithms [85], [1], [2], [86]. Turkers are presented a series of pairs of images, one real and one fake (generated by the I2I models). Participants are asked to choose the photo they think is real and then obtain the feedback to compute the scores.

#### 2) Objective image quality assessment:

- **Peak signal-to-noise ratio (PSNR)**: PSNR is one of the most widely used full-reference quality metrics. It reflects the intensity differences between the translated image and its ground truth. A higher PSNR score means that the intensity of two images is closer.
- **Structural similarity index (SSIM)** [87]: I2I uses SSIM to compute the perceptual distance between the translated image and its ground truth. The higher the SSIM is, the greater the similarity of the luminance, contrast and structure of two images will be.
- **Inception score (IS)** [88]: IS encodes the diversity across all translated outputs. It exploits a pretrained inception classification model to predict the domain label of the translated images. A higher score indicates a better translated performance.
- **Mask-SSIM and Mask-IS** [89]: These two metrics are the masked versions of SSIM and IS to reduce the background influence by masking it out. They are designed for evaluating the performance of person image generation task.
- **Conditional inception score (CIS)** [90]: CIS is modified from IS to better evaluate the multimodal I2I works. It encodes the diversity of the translated output conditioned on a single input image. A higher score indicates a better translated performance.
- **Perceptual distance (PD)** [91]: PD computes the perceptual distance between the translated image and corresponding source image. A lower PD score indicates that the contents of two images are more similar.
- **Fréchet inception distance (FID)** [92]: The FID measures the distance between the distributions of synthesized images and real images. A lower FID score means a better performance.
- **Kernel inception distance (KID)** [93]: The KID computes the squared maximum mean discrepancy between the feature representations of real and generated images in which feature representations are extracted from the

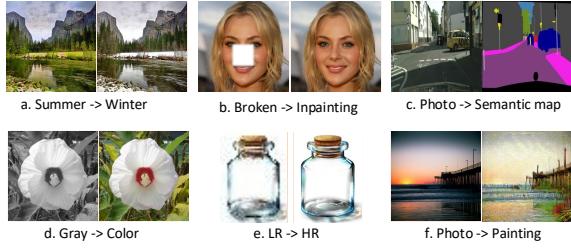


Fig. 6: Examples of two-domain I2I

Inception network [94]. A lower KID indicates more shared visual similarities between real and generated images.

- **Single image Fréchet inception distance (SIFID)** [95]: The SIFID captures the difference between the internal distributions of two images, which is implemented by computing the Fréchet inception distance (FID) between the deep features of two images. The SIFID is computed using the translated image and corresponding target image. A lower SIFID score indicates that the styles of two images are more similar.
- **LPIPS** [96]: LPIPS evaluates the diversity of the translated images and is demonstrated to correlate well with human perceptual similarity. It is computed as the average LPIPS distance between pairs of randomly sampled translation outputs from the same input. A higher LPIPS score means a more realistic, diverse translated result.
- **FCN scores** [1]: This metric is mostly used in the translation of *semantic maps*  $\leftrightarrow$  *real photos* (e.g., c. in Fig.6). It uses the FCN-8s architecture [97] for semantic segmentation to predict a label map for a translated photo and then compares this label map with ground truth labels with standard semantic segmentation metrics, such as per-pixel accuracy, per-class accuracy, and mean class intersection-over-union (class IOU). A higher score indicates a better translated result.
- **Classification accuracy** [74]: This metric adapts a classifier pretrained on target domain images to classify the translated images. The intuition behind this metric is that a well-trained I2I model would generate outputs that can be classified as an image from the target domain. A higher accuracy indicates that the model learns more deterministic patterns to be represented in the target domain.
- **Density and Coverage (DC)** [98] is the latest metric for simultaneously judging the diversity and fidelity of generative models. It measures the distance between real images and generated images by introducing a manifold estimation procedure. Higher scores indicate larger diversity and better coverage to the ground-truth domain, respectively.

### III. TWO-DOMAIN IMAGE-TO-IMAGE TRANSLATION

In this section, we focus on introducing the two-domain I2I methods. As shown in Fig. 6, two-domain I2I can solve many

problems in computer vision, computer graphics and image processing, such as image style transfer (f.) [2], [108], which can be used in photo editor apps to promote user experience and semantic segmentation (c.) [4], [5], which benefits the autonomous driving and image colorization (d.) [23], [27]. 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.) [28], [29]. 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 [109], [110], [111], 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 in Table I including method name, publication year, the type of training data, whether multi-modal or not and corresponding insights.

- **Supervised I2I** In the earlier I2I works [1], 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.
- **Unsupervised I2I** 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 (e.g., f. in Fig.6), it is almost impossible to collect massive amounts of labeled paintings that match the input landscapes. Hence, unsupervised methods [2], [11], [112] 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.
- **Semi-supervised I2I** 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 [113] or genomics [114]. Therefore, researchers consider introducing semi-supervised learning [115], [116], [117] 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.
- **Few-shot I2I** 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 [118], [119] and few-shot learning [120], [121], 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,

TABLE I: List of two-domain I2I methods including model name, publication year, the type of training data, whether multimodal or not and corresponding insights.

Method	Publication	Data	Multimodal	Insights
pix2pix	2017	paired	No	conditional GAN;
DRGAN	2018	paired	No	reviser module;
pix2pixHD [99]	2018	paired	No	high-resolution; multi-scale architecture;
SelectionGAN	2019	paired	No	controllable, user-specific generation;
SPADE	2019	paired	No	cross-view translation; attention selection;
SEAN	2020	paired	No	spatially-adaptive normalization layer;
CoCosNet	2020	paired	No	semantic region-adaptive normalization layer;
CoCosNetv2	2021	paired	No	dense semantic correspondence;
ASAPNet	2021	paired	No	PatchMatch;
BicycleGAN	2017	paired	Yes	pointwise; non-linear transformation; MLP;
PixelNN [100]	2018	paired	Yes	cVAE-GAN; CLR-GAN;
	2018	paired	Yes	nearest-neighbor approach;
				disentangled representation;
TCR	2020	paired+unpaired	No	transformation consistency regularization;
DTN	2016	unpaired	No	invariant representation; domain adaptation;
DualGAN/DiscoGAN/CycleGAN	2017	unpaired	No	cyclic loss;
UNIT	2017	unpaired	No	cyclic loss; shared latent space;
SCAN	2018	unpaired	No	cyclic loss; multi-stage generation;
U-GAT-IT	2019	unpaired	No	cyclic loss; cam attention; auxiliary classifier;
GANimorph	2018	unpaired	No	cyclic loss; large domain gaps; semantic segmentation;
TraVeLGAN	2019	unpaired	No	cyclic loss; large domain gaps; Siamese network;
TransGaGa	2019	unpaired	No	cyclic loss; large domain gaps; disentangled representation;
ACL-GAN [101]	2020	unpaired	No	large domain gaps; adversarial-consistency loss;
DistanceGAN	2017	unpaired	No	cyclic loss; large domain gaps; pretrained VGG; cascaded translation;
GCGAN	2019	unpaired	No	one-sided UI2I; pairwise distances matching;
CUT	2020	unpaired	No	one-sided UI2I; geometric transformation perservation;
[102]	2020	unpaired	No	one-sided UI2I; contrastive learning;
TSIT	2020	unpaired	No	one-sided UI2I; disentanglement; contrastive learning;
F-LSeSim	2021	unpaired	No	one-sided UI2I; self-similarity;
LPTN	2021	unpaired	No	one-sided UI2I; laplacian pyramid;
DAGAN	2018	unpaired	No	instance-level UI2I; attention;
attention-GAN/attention-guided I2I	2018	unpaired	No	instance-level UI2I; attention; object transfiguration;
InstaGAN	2018	unpaired	No	instance-level UI2I; attention; cyclic loss;
INIT[103]	2019	unpaired	Yes	instance-level UI2I; segmentation mask; cyclic loss;
DUNIT	2020	unpaired	Yes	instance-level UI2I; object+global; cyclic loss;
Art2real	2019	unpaired	No	instance-level UI2I; object detector; cyclic loss;
GDWCT	2019	unpaired	No	segmentation; memory bank;
RevGAN [104]	2019	unpaired	No	whitening-and-coloring transformation;
DAI2I	2020	unpaired	No	invertible neural networks;
NICE-GAN [78]	2020	unpaired	No	knowledge distillation;
[105]	2018	unpaired	Yes	domain adaptation;
cd-GAN/MUNIT/DRIT/EGSC-IT	2018	unpaired	Yes	introspective adversarial networks;
MSGAN [106]	2019	unpaired	Yes	domain-specific; domain-invariant;
DSMAP	2020	unpaired	Yes	augmented CycleGAN;
TGAN	2018	unpaired	Yes	disentangled representation;
MT-GAN	2019	unpaired	Yes	mode-seeking regularization;
EWC [107]	2020	unpaired	Yes	latent filter scaling;
OST	2021	unpaired	No	domain-specific mapping;
BiOST	2018	unpaired	No	few-shot UI2I; transfer learning;
TuiGAN	2019	unpaired	No	few-shot UI2I; meta-learning;
	2020	unpaired	No	few-shot UI2I; life-long learning;
			No	few-shot UI2I; distance consistency; anchor-based strategy;
			No	one-shot UI2I; sharing layers; selective backpropagation
			No	one-shot UI2I; sharing layers; selective backpropagation; bi-direction;
			No	one-shot UI2I; multi-scale; cyclic loss;

few- and one-shot I2I algorithms [18], [80], [122], [79] 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, as shown in Fig.6. However, in practice, the two-domain I2I is inherently ambiguous, as one input image may correspond to multiple possible outputs, namely, multimodal outputs, as shown in Fig.7. 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 in supervised (subsection III-A) and unsupervised settings (subsection III-B).

#### A. Supervised Image-to-Image Translation

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

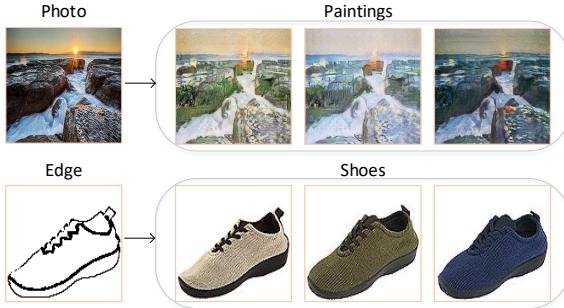


Fig. 7: Examples of multimodal outputs in two-domain I2I.

further divide the supervised I2I in two categories: methods with single-modal output and methods with multimodal outputs.

1) *Single-modal Output*: The idea of I2I can be traced back to Hertzmann et al.'s image analogies [123], 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. [1] 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 that the outputs cannot be distinguished from "real" images. The objective is:

$$\mathcal{L} = \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{\mathcal{L}_1}(G). \quad (9)$$

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. [86] 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. [124] argue that the adversarial training in pix2pix [1] 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. [99] take an important step toward addressing the controllable or user-specific generation based on pix2pix [1] via respecting the constraints provided by an external, user-provided guidance image.

Unfortunately, pix2pix [1] and its improved variants [86], [124], [99] 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. [125] 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 [4] 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 [5] therefore designs semantic region-adaptive normalization layer to alleviate the two shortcomings.

Having said that, Shaham et al. [126] claim that traditional I2I networks [1], [124], [4] 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. [127] proposed an exemplar-based I2I framework, CoCosNet, 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. [128] therefore proposed a GRU-assisted refinement module that applies PatchMatch in a hierarchy to first learn the full-resolution, 1024×1024, cross-domain semantic correspondence, namely CoCosNetv2.

2) *Multimodal Outputs*: As shown in Fig.7, 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.

Actually, this multimodal translation benefits from the solutions of *mode collapse problem* [129], [64], [68], in which the generator tends to learn to map different input samples to the same output. Thus, many multimodal I2I methods [16], [130] focus on solving the mode collapse problem to lead to diverse outputs naturally. BicycleGAN [16] became the first supervised multimodal I2I work by combining cVAE-GAN [131], [55], [132] and cLR-GAN [133], [134], [135] to systematically study a family of solutions to the mode collapse problem and generate diverse and realistic outputs.

Similarly, Bansal et al. [130] 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* [133], [136], [137], [138] 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. [100] 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.

### B. Unsupervised Image-to-Image Translation (UI2I)

UI2I uses two large but unpaired sets of training images to convert images from one representation to another. In this subsection, we follow the same categories in subsection III-A: single-modal output and multimodal outputs.

**1) Single-modal Output:** UI2I methods have been explored primarily by focusing on different issues. We will introduce those methods with single-modal output in the following four categories: translation using a cycle-consistency constraint, translation beyond a cycle-consistency constraint, translation of fine-grained objects and translation by combining knowledge in other fields.

- **Translation using a Cycle-consistency Constraint** In the beginning, researchers tried to find new frameworks or constraints to establish the I2I mapping without labels or pairings. Based on this motivation, the cycle-consistency constraint was proposed, as shown in Fig.8 and was proved to be an effective strategy for overcoming the lack of supervised pairing.

- **Translation beyond Cycle-consistency Constraint** However, while the cycle-consistency constraint can eliminate the dependence on supervised paired data, it tends to force the model to generate a translated image that contains all the information of the input image for reconstructing the input image. Approaches using cyclic loss are typically unsuccessful when the two domains require substantial clutter and heterogeneity instead of small, simple changes in low-level shape and context. Therefore, many UI2I methods focus on the translation beyond the cycle-consistency constraint, as shown in Fig.9, to solve the homogeneous limitation, as well as the large shape deformation problem between the source and target domains.

- **Translation of Fine-grained Objects** Most UI2I models using or beyond the cycle-consistency constraint tend to directly synthesis a new domain with the global target style translated and give little thought of the local objects or fine-grained instances during translation. However, in some application scenarios, such as virtual try-on, we may only need to change a local object, such as changing pants to a skirt with other parts unchanged, as shown in Fig.10. In this case, severe setbacks are incurred when the translation involves large shape changes of instances or multiple discrepant objects. Hence, research on applying instances or objects information in UI2I is a growing trend.

- **Translation by combining knowledge in other fields** In addition, some UI2Is try to improve the network efficiency or translation performance by combining knowledge from other research areas.

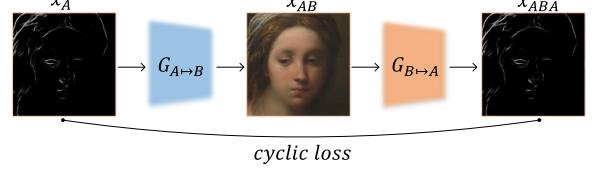


Fig. 8: Taking edge  $\leftrightarrow$  face translation as an example, we use a cycle-consistency constraint between a source image  $x_A$  and its cyclic reconstructed image  $x_{ABA}$ , termed cyclic loss through two translators  $G_{A \rightarrow B}$  and  $G_{B \rightarrow A}$ .

#### a) Translation using a Cycle-consistency Constraint:

The popularly known strategy for tackling an unsupervised setting is to use the cycle-consistency constraint (cyclic loss) shown in Fig. 8. Cyclic loss uses two translators,  $G_{A \rightarrow B}$  and  $G_{B \rightarrow A}$ , to define a cycle-consistency loss between the source image  $x_A$  and its reconstruction  $x_{ABA}$  when the pairs are not available, and the objective can be written as:

$$\mathcal{L}_{cyc} = \mathcal{L}(x_A, G_{B \rightarrow A}(G_{A \rightarrow B}(x_A))). \quad (10)$$

Taigman et al. [139] present a domain transfer network (DTN) for unpaired cross-domain image generation by assuming constant latent space between two domains, which could generate images of the target domains' style and preserve their identity. Similar to the idea of dual learning in neural machine translation, DualGAN [112], DiscoGAN [11] and CycleGAN [2] are proposed to train two cross-domain transfer GANs with two cyclic losses at the same time. Liu et al. [74] propose UNIT to make a shared latent space assumption that a pair of corresponding images in different domains can be mapped to the same latent code in a shared latent space. They show that the shared-latent space constraint implies the cycle-consistency constraint. Li et al. [140] claim that these single-stage unsupervised approaches are difficult to use for translating two-domain images with high-resolution or a substantial visual gap. They hence propose a stacked cycle-consistent adversarial network (SCAN) to decompose the single complex image translation process into multistage transformations. More recently, Kim et al. [37] proposed U-GAT-IT to incorporate a novel attention module to force the generator and discriminator to focus on more important regions via the auxiliary classifier.

#### b) Translation beyond Cycle-consistency Constraint:

To address the challenging shape deformation problem (i.e., large domain gaps) in I2I shown in Fig. 9, Gokaslan et al. [141] propose GANimorph to reframe the discrimination problem from determining real or fake images into a semantic segmentation task of finding real or fake regions of the image with dilated convolutions. Amodio et al. [142] introduce TraVeLGAN to address the challenge. In addition to the generator and discriminator, they add a Siamese network to define a transformation vector between two images of each domain and minimize the distance between the two vectors. The Siamese network guides the generator such that each original image shares semantics with its generated version. Wu et al. [77] present TransGaGa to solve the large geometry variations in

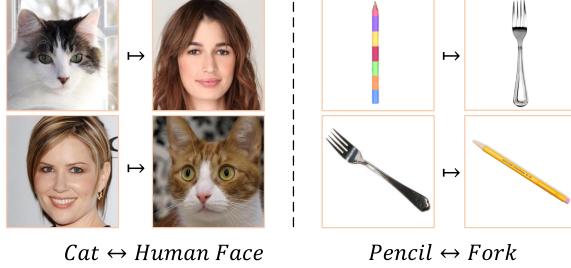


Fig. 9: Examples of I2I with large domain gaps, where the translated images are sufficiently realistic in the target domain and preserve semantic information learned from the source domain.

I2I. They disentangle each domain into a Cartesian product of the geometry space and appearance space by the VAE. In each space, they apply a bi-direction geometry transformation and appearance transformation with two transformers, respectively. Zhao et al. [143] argue that I2I can barely perform shape changes, remove objects or ignore irrelevant texture because of the strict pixel-level constraint of cycle-consistent loss. They propose ACL-GAN, which uses a novel adversarial-consistency loss to replace the cyclic loss to maintain the commonalities across two domains. Recently, Katzir et al. [101] mitigated shape translation in a cascaded, deep-to-shallow fashion, in which they exploited the deep features extracted from a pretrained VGG-19 and translated them at the feature level.

Moreover, some UI2I works try to design a one-side translation process to remove the cycle-consistency constraint. These methods usually take into account some kind of geometry distance as content loss between the original source image and translated results. Benaim et al. [144] propose DistanceGAN to achieve one-side translation by maintaining the distances between images within domains. Fu et al. [145] propose GC-GAN to preserve the given geometric transformation between the input images before and after translation. Zheng et al. [146] propose F-LSeSim to learn a domain-invariant representation to precisely express scene structure via self-similarity. Liang et al. [147] propose a Laplacian Pyramid Translation Network (LPTN) to achieve photorealistic I2I by decomposing the input into a Laplacian pyramid and translating on the low-frequency component. Park et al. [148] propose CUT to maximize the mutual information between the input-output pairs via contrastive learning [149] in a patch-based way rather than operating on entire images. Jiang et al. [150] propose a symmetrical two-stream framework (TSIT) to learn feature-level semantic structure information and style representation, and then they exploit the generator to fuse content and style feature maps from coarse to fine. Park et al. [102] also propose a swapping autoencoder for texture swapping by enforcing the output and reference patches to appear indistinguishable via the patch co-occurrence discriminator.

*c) Translation of Fine-grained Objects:* Some I2I works try to translate on a higher semantic level by replacing the local

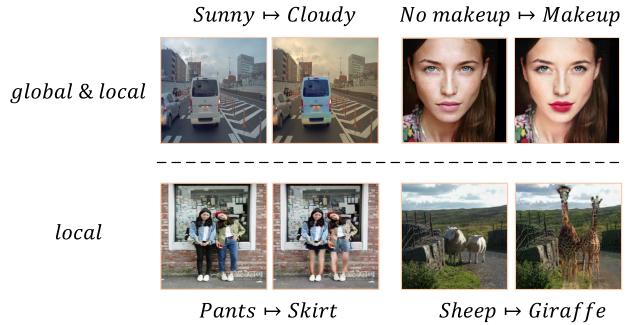


Fig. 10: Examples of I2I focusing on fine-grained objects. The top examples generate results with the global style (weather or foundation) translated and the local object instance (car or lipstick) changed. The bottom examples achieve remarkable translated results for fine-grained local objects.

texture information of object instances as well as the global style translated, as shown in Fig.10. Ma et al. [151] propose DAGAN to construct a deep attention encoder to enable the instance-level translation. Chen et al. [152] and Mejjati [13] almost simultaneously proposed attention GAN and attention-guided I2I, respectively, focusing on achieving an I2I translation of individual objects without altering the background, as shown in Fig.10 (local). Mo et al. [153] propose InstaGAN, which is the first work to solve multi-instance transfiguration tasks in UI2I. It uses the object segmentation masks to translate an image and the corresponding set of instance attributes while maintaining the permutation invariance property of instances. Shen et al. [103] propose the instance-aware I2I approach (INIT) to use the fine-grained local instances based on MUNIT [90] and DRIT [75]. DUNIT [154] incorporates an object detector within the I2I architecture used in DRIT[75] to leverage the object instances to reassemble the resulting representation.

*d) Translation by Combining Knowledge in Other Fields:* Tomei et al. [14] present Art2Real to translate artistic paintings to real photos using a weakly supervised segmentation model and memory banks. Inspired by the style transfer, Cho et al. [155] propose GDWCT, which extends whitening-and-coloring transformation (WCT) to I2I to achieve a highly competitive image quality. Chen et al. [104] use the knowledge distillation scheme to define a teacher generator and student discriminator. A distilling portable model is shown to achieve a comparable performance with substantially lower memory usage and computational cost. With the help of domain adaptation, Chen et al. [156] develop DAI2I to adapt a given I2I model trained on the source domain to a new domain, which improves the generalization capacity of existing models. RevGAN [157] interpolates the invertible neural networks (INNs) into I2I to reduce memory overhead, as well as increase the fidelity of the output. NICE-GAN [158] first reuses the discriminator for embedding from images to hidden vectors (as encoding) in which the discriminator is conducted using introspective adversarial networks (IANs). It derives a more compact and effective architecture for generating translated

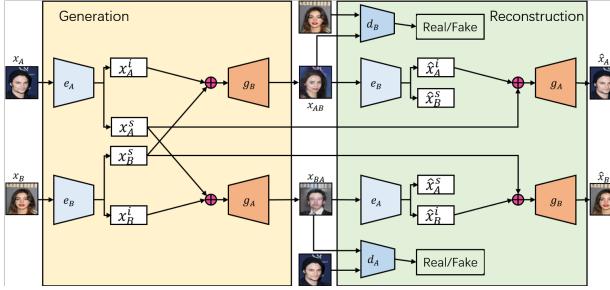


Fig. 11: The architecture of cd-GAN [159].

images.

2) *Multi-modal Outputs*: Kazemi et al. [78] show that shared-latent space assumptions only model the domain-invariant information across two domains and fail to capture the domain-specific information. They argue for learning a one-to-many UI2I mapping by extending CycleGAN to learn a domain-specific code for each domain jointly with a domain-invariant code. Similarly, Almahairi et al. [105] also claim that the mapping across two domains should be characterized as many-to-many instead of one-to-one. They therefore introduce the augmented CycleGAN model to capture the diversity of the outputs, i.e., multimodal, by extending CycleGAN’s training procedure to the augmented spaces.

Disentangled representations [133], [136], [137] also offer a solution to this problem in unsupervised settings. They inspire multimodal UI2I advances, such as cd-GAN [159], MUNIT [90], DRIT [75] and EGSC-IT [76], which were proposed almost simultaneously and present similar model designings, such as that of Fig.11. For example, DRIT [75] assumes embedding images onto two spaces, domain-invariant content space and domain-specific attribute space, via a content encoder and attribute encoder. Specifically, DRIT learns two objective losses, content adversarial loss and cross-cycle consistency loss. Through content adversarial loss, it applies weight sharing and a content discriminator to force content representation to be mapped onto the same shared space, as well as to guarantee the same content representations encode the same information for both domains. Then, with the constraint of cross-cycle consistency loss, it performs forward-backward translation by swapping domain-specific representations. At test time, DRIT can use different attribute vectors randomly sampled from domain-specific attribute space to generated diverse outputs.

However, these aforementioned methods still cannot solve the problem of target domain images that are content-rich with multiple discrepant objects. Shen et al. [103] therefore propose INIT to translate instance-level objects and background/global areas separately with different style codes. Chang et al. [160] declare that the shared domain-invariant content space in disentangled representations could limit the ability to represent content because these representations ignore the relationship between content and style. They present DSMAP to leverage two extra domain-specific mappings to remap the content features from shared domain-invariant content space to two independent domain-specific content spaces for two domains.

Other attempts to address multimodal UI2I are proposed by Mao et al. [161] and Alharbi et al. [106]. The study in [161] uses MSGAN to employ a mode-seeking regularization method to solve the mode collapse problem in cGANs, and the proposed regularization method can be readily integrated with an existing cGANs framework, such as DRIT [75], to generate more diverse translated images. In addition, [106] uses latent filter scaling (LFS) to perform multimodal UI2I, which is the first multimodal UI2I framework that does not require autoencoding or reconstruction losses for the latent codes or images.

### C. Semi-Supervised Image-to-Image Translation

Semi-supervised I2I draws much attention in some special applications, such as old movie restoration or artistic reconstructions. In these scenarios, one needs few human-labeled data for guidance and abundant, unlabeled other data for automatic translation. Unpaired data, when used in conjunction with a small amount of paired data, can produce considerable improvement in translation performance.

Mustafa et al. [113] first study the applicability of semi-supervised learning in a two-domain I2I setting. They introduce a regularization term, transformation consistency regularization (TCR), to force a model’s prediction to remain unchanged for the perturbed (geometric transform) input sample and its reconstruction version. In detail, they train an I2I mapping model  $f_\theta$  by minimizing the supervised loss  $\mathcal{L}_s$  with paired source-target images  $(x_i, y_i)$ :

$$\mathcal{L}_s = mse(y_i, f_\theta(x_i)). \quad (11)$$

Then, they leverage unsupervised data to regularize the model’s predictions over varied forms of geometric transformations  $T_m$ . They make use of  $T_m$  to process the unlabeled input samples and feed these transformed samples into the I2I model  $f_\theta$  to obtain the disturbed outputs  $f_\theta(T_m(u_i))$ . On the other hand, they directly feed the unlabeled samples into  $f_\theta$  to acquire primary outputs and apply geometric transformations  $T_m$  onto them to obtain another type of perturbed outputs  $T_m(f_\theta(u_i))$ . The TCR of unlabeled data guarantees the consistency between the two outputs to learn more about the inherent structure of the source and target domain distributions. The detailed unsupervised TCR regularization loss  $\mathcal{L}_{us}$  is as follows:

$$\mathcal{L}_{us} = mse(T_m(f_\theta(u_i)), f_\theta(T_m(u_i))). \quad (12)$$

Their method can use unlabeled data and less than 1% of labeled data to complete several I2I tasks, such as image colorization, image denoising and image super-resolution.

### D. Few-Shot Image-to-Image Translation

Existing I2I models cannot translate images from very few (even one) training examples of the source and target domains. In contrast, humans can learn from very limited exemplars to obtain extraordinary learning results. For example, a child can recognize what a "zebra" and "rhino" are with only a few pictures. Inspired by the rapid learning ability of humans, researchers expect that after the machine learning model has

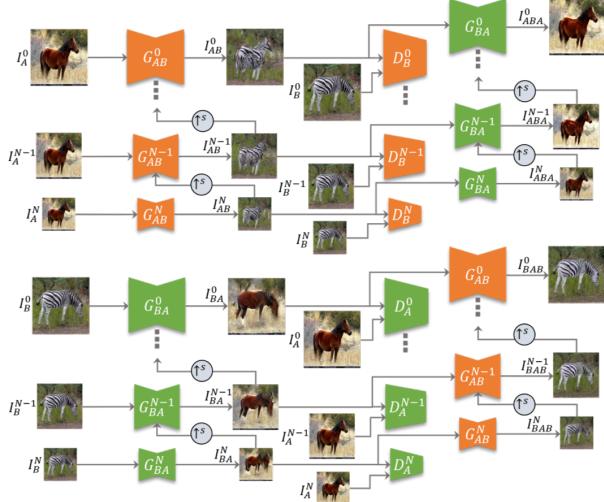


Fig. 12: The architecture of TuiGAN [80].

learned a large amount of data in a certain category, for new categories, it only needs a few samples to learn quickly. In other words, few-shot I2I wants to solve the transfer capability or generalization ability of the I2I model given very few samples.

Drawing inspiration from domain adaptation or transfer learning, some methods solve the few-shot translation by adapting a pretrained-network trained on large-scale source domain to the target domain with a few images only. Wang et al. [162] propose Transferring GAN (TGAN) to successfully combine transfer learning with GAN. It transfers images from source domain to target domain with a pretrained network when data is limited. Lin et al. propose MT-GAN [122] to incorporate prior from previous domain translation tasks when assuming a new domain translation task from a perspective of meta-learning. Likewise, Li et al. [163] propose to utilize EWC to adapt the weights of pretrained network on source domain to a new target domain. Ojha et al. [107] achieve few-shot adaptation by a novel cross-domain distance consistency loss and an anchor-based strategy.

An extreme scenario of few-shot I2I is one-shot I2I. Benaim et al.[164] propose OST to solve the one-shot cross-domain translation problem, which aims to learn a unidirectional mapping function given a single source image and a set of images from the target domain. By sharing layers between the two autoencoders and selective backpropagation, OST enforces the same structure on the encoding of both domains, which benefits the translation. As an extension of OST, Cohen et al. [165] propose BiOST to translate in bi-direction without a weight sharing strategy via a feature-cycle consistency term.

In contrast to the one-shot setting in [164], [165] that uses a single image from the source domain and a set of images from the target domain, Lin et al. propose TuiGAN [80] to achieve one-shot UI2I with only two unpaired images from two domains. They train TuiGAN in a coarse-to-fine manner using the cycle-consistency constraint shown in Fig.12. In detail, they design two pyramids of generators and discriminators to transfer the domain distribution of the input image to the target



Fig. 13: Illustration of the dataset used in multi-domain I2I scenarios: Each domain usually means a set of images sharing the same attribute value. Images are from the CelebA dataset [169].

domain by progressively translating the image from coarse to fine. Using this progressive translation, their model can extract the underlying relationship between two images by continuously varying the receptive fields at different scales. All in all, TuiGAN represents a further step toward the possibility of unsupervised learning with extremely limited data.

#### IV. MULTI-DOMAIN IMAGE-TO-IMAGE TRANSLATION

In this section, we will discuss the I2I problem on multiple domains and list correlative algorithms in Table II covering model name, publication year, the type of training data, whether multi-modal or not and corresponding insights. We have discussed a series of attractive I2I works for translating two domains. However, these methods can only capture the relationship of two domains based on one model at a time. Given  $n$  domains, the network requires  $n \times (n-1)$  generators to be trained, which leads to an unavoidable burden. Moreover, it fails to fully use the entire training data from all the domains. Even if there exists global information learned from all the domains that can be applied to promote the translation performance, the network still only learns from two domains, and it is difficult to acquire that global multi-domain information. How to further reduce the network complexity and improve the efficiency to handle multiple domains remains unaddressed.

Therefore, researchers study the multi-domain I2I problem. It focuses on handling multiple domains using a single unified model in which multiple outputs contain different semantic contents or style textures. We divide multi-domain I2I research into three categories: unsupervised multi-domain I2I, semi-supervised multi-domain I2I and few-shot multi-domain I2I.

##### A. Unsupervised multi-domain Image-to-Image Translation

In this subsection, we introduce unsupervised multi-domain I2I (multi-domain UI2I) in two aspects: single-modal output and multimodal outputs. First, we explain how to achieve this translation with multiple domains. For example, the CelebA dataset [169] contains 40 facial attributes, and each domain usually means a set of images sharing the same attribute value in [170], [171], [172], [173]. We therefore can obtain numerous unpaired translation domains based on different

TABLE II: List of multi-domain I2I methods including model name, publication year, the type of training data, whether multimodal or not and corresponding insights.

Method	Publication	Data	Multimodal	Insights
Domain-Bank/ModularGAN	2018	unpaired	No	each domain each module;
StarGAN	2018	unpaired	No	unified single model; auxiliary classifier;
AttGAN	2019	unpaired	No	unified single model; auxiliary classifier; AE-GAN;
RelGAN/STGAN	2019	unpaired	No	auxiliary classifier; relative-attribute;
CollaGAN	2019	unpaired	No	auxiliary classifier; multiple inputs;
[166]	2019	unpaired	No	auxiliary domain; multi-path consistency;
SGN	2019	unpaired	No	sym-parameter;
Fixed-Point GAN	2019	unpaired	No	forced discriminator;
[167]	2019	unpaired	No	deliberation learning;
ADSPM	2019	unpaired	No	spontaneous motion;
INIT[168]	2020	unpaired	No	informative sample mining network; multihop training;
GANimation	2018	unpaired	Yes	action unit;
DosGAN	2019	unpaired	Yes	domain classifier;
UFDN	2018	unpaired	Yes	disentanglement;
DMIT	2019	unpaired	Yes	disentanglement; multi-mapping;
StarGANv2	2020	unpaired	Yes	disentanglement; multi-task discriminator; diversity regularization;
DRIT++	2020	unpaired	Yes	disentanglement; latent regression loss;
GMM-UNIT	2020	unpaired	Yes	disentanglement; Gaussian mixture model;
FUNIT	2019	unpaired	Yes	few-shot UI2I; multi-task discriminator;;
COCO-FUNIT	2020	unpaired	Yes	few-shot UI2I; multi-task discriminator; content leak;
AGUIT	2019	paired+unpaired	Yes	disentanglement; domain classifier; cyclic loss;
SEMIT	2020	paired+unpaired	No	few-shot I2I; pseudo-label;

attribute values, as shown in Fig.13. Notwithstanding the demonstrated success of unsupervised two-domain I2I (two-domain UI2I) in subsection III-B, when we have multiple unpaired domains, do these two-domain UI2I methods still work? The answer may be “no.” The typical problems are the efficiency and network burden in which these two-domain UI2I can only transfer one pair of different domains through one training. The multi-domain UI2I hence attracts much attention, and we will provide detailed illustrations from two aspects: multi-domain UI2I with single-modal output and with multimodal outputs.

1) *Single-modal Output*: In addition, a large variety of multi-domain UI2I methods with a single-modal output have been proposed to obtain image representations in an unsupervised way. We classify them into three categories: training with multimodules, training with one generator and discriminator pair and training by combining knowledge in other fields.

- **Training with multimodules** In earlier times, methods mainly designed complex multiple modules to address multi-domain UI2I by regarding it as a composition of several two-domain UI2Is, in which each module represents each domain information. Compared with directly applied two-domain UI2I methods, these works can train all the domains at one time, which saves much training time and many model parameters.

- **Training with one generator and discriminator pair** Unfortunately, methods training with multimodules can train translations between multiple domains at one time, but they still need to define multiple-domain modules to represent the corresponding domains. Is there a model that can train all domains at one time using the same module to process multiple-domain information? A more effective solution is to use an auxiliary label (binary or one-shot attribute vector) to represent domain information that leads to a more flexible translation. After randomly choosing the target domain label as conditional input,

we can translate the source domain input to this target domain without extra translators using one generator and discriminator pair.

- **Training by combining knowledge in other fields** Some algorithms try to introduce knowledge from other research areas to facilitate multi-domain UI2I. To some extent, these methods have indeed brought us new insight.

a) *Training with Multimodules*: Based on the shared-latent space assumption [74], Hui et al. [174] propose a unified framework named Domain-Bank. Given  $n$  domains, Domain-Bank obtains  $n$  pairs of translated results by training the network only once, while the two-domain I2I methods require the training of  $n$  models for translations between different pairs of domains. By leveraging several reusable and composable modules, Zhao et al. [175] propose ModularGAN to translate an image to multiple domains efficiently. They predefine an attribute set  $\mathbf{A} = \{A_1, \dots, A_n\}$  in which each attribute  $A_i$  represents meaningful inherent property of each domain with different attribute values.

b) *Training with One Generator and Discriminator Pair*: Choi et al. propose StarGAN [170], which fully proves the effectiveness of the auxiliary domain label by mappings between all available domains using only a single model. They design a special discriminator and introduce an auxiliary classifier on top of it, in which the discriminator not only justifies whether an image is a natural or fake one  $D_{src}(x_A)$  but also distinguishes which domain the input belongs to  $D_{cls}(x_A)$ :

$$D : x_A \mapsto \{D_{src}(x_A), D_{cls}(x_A)\}. \quad (13)$$

To translate input image  $x_A \in A$  to target domain B, StarGAN learns an adversarial loss and a cycle-consistency loss conditioned with input domain label  $c_A$  and target domain label  $c_B$ .

Through the three loss functions, StarGAN can achieve a scalable translation for multiple domains and obtain results with higher visual quality. Sharing an extremely similar idea,

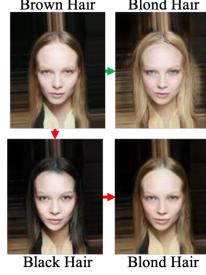


Fig. 14: The illustration of multipath consistency regularization [166] on the translation between different hair colors. Ideally, they consider that the direct translation (i.e., one-hop translation) from brown hair to blonde should be identical to the indirect translation (i.e., two-hop translation) from brown to black to blonde.

He et al. [171] propose AttGAN to address this problem. However, AttGAN uses an encoder-decoder architecture to model the relationship between the latent representation and the attributes.

The target-attribute constraint used in StarGAN and AttGAN fails to provide a more fine-grained control and often requires specifying the complete target attributes, even if most of the attributes are not changed. Wu et al. [172] and Liu et al. [173] therefore consider a novel attribute description termed relative-attribute that represents the desired change of attributes. They propose RelGAN and STGAN, respectively, to satisfy arbitrary image attribute editing with relative attributes. Given input domain attribute  $c_A$  and target domain attribute  $c_B$ , the relative attribute  $c$  is formulated as:

$$c = c_B - c_A. \quad (14)$$

StarGAN or AttGAN may perform worse when multiple inputs are required to obtain a desired output. Any missing input data will introduce a large bias and lead to terrible results. Therefore, CollaGAN [176] has been proposed to process multiple-inputs from multiple domains instead of only handling single-input and single-output.

Rather than introducing an auxiliary domain classifier, Lin et al. [166] propose introducing an additional auxiliary domain and constructing a multipath consistency loss for multi-domain I2I. Their work is motivated by an important property shown in Fig. 14, namely, the direct translation (i.e., one-hop translation) from brown hair to blonde should ideally be identical to the indirect translation (i.e., two-hop translation) from brown to black to blonde. Their multipath consistency loss evaluates the differences between direct two-domain translation  $A \mapsto C$  and indirect multiple-domain translations  $A \mapsto B \mapsto C$  with domain  $B$  as an auxiliary domain. The method regularizes the training of each task and obtains a better performance.

*c) Training by Combining Knowledge in Other Fields:* By expanding the concept of a multi-domain from data to the loss area, Chang et al. [177] introduce the sym-parameter to synchronize various mixed losses with input conditions. Siddiquee et al. [178] propose Fixed-Point GAN, which uses a trainable generator and a frozen discriminator to perform

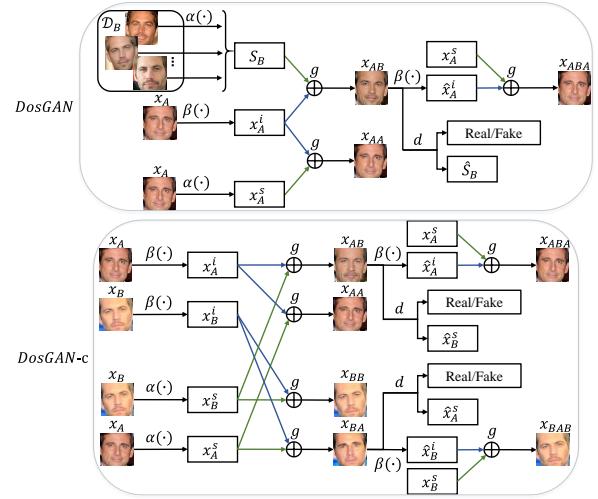


Fig. 15: The training architecture of DosGAN [180]: (1) top: DosGAN for unpaired I2I; (2) bottom: DosGAN-c for unpaired conditional I2I.

fixed-point translation learning. He et al. [167] propose deliberation learning for I2I by adding a polishing step on the output image. Cao et al. [168] propose the informative sample mining network (INIT) to analyze the importance of sample selection and select the informative samples for multihop training. Wu et al. [179] propose ADSPM to learn attribute-driven deformation by a spontaneous motion (SPM) estimation module and a refinement part (R) with much consideration for geometric transform.

*2) Multimodal Outputs:* However, all of the multi-domain approaches mentioned above still learn a deterministic mapping between two arbitrary domains. Researchers therefore consider addressing the multi-domain I2I, as well as the outputs of multimodal results.

Lin et al. observe that if the pretrained CNN network can accurately classify the domain of an image, then the output of the second-to-last layer of the classifier should well capture the domain information of this image. Combining this network with a domain classifier, they propose domain-supervised GAN (DosGAN) [180] to use the domain label as an explicit supervision and pretrain a deep CNN to predict which domain an image is from. The detailed training architecture is shown in Fig.15.

In addition, GANimation [181] is proposed to generate anatomically aware facial animation. It continuously synthesizes anatomical facial movements by controlling the magnitude of activation of each action unit (AU).

By exploiting the disentanglement assumption, UFDN [32], DMIT[182], StarGAN v2 [183], DRIT++ [75] and GMM-UNIT [184] are proposed to perform multimodal outputs in a multi-domain UI2I setting. For example, DRIT++ consists of two content encoders  $\{E_A^c, E_B^c\}$ , two attribute encoders  $\{E_A^a, E_B^a\}$ , two generators  $\{G_A, G_B\}$ , two discriminators  $D_A, D_B$  and a content discriminator  $D_{adv}^c$ . Through weight sharing and a content discriminator  $D_{adv}^c$ , the network can achieve representation disentanglement with content adver-

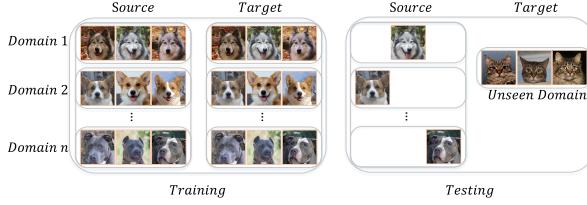


Fig. 16: Illustration of the dataset used in few-shot multi-domain I2I scenarios: The training set consists of multiple domains in which the source and target images are randomly sampled from arbitrary  $n$  domains; given very few images of the target unseen domain (unseen in the training process), few-shot multi-domain I2I aims to translate a source content image (randomly sampled from  $n$  domains) into an image analogous to this unseen domain.

sarial loss. Then, it leverages cross-cycle consistency loss for forward and backward translations. In addition to these two losses, these methods also use domain adversarial loss, self-reconstruction loss, latent regression loss and an extra mode-seeking regularization to effectively improve the sample diversity and visual quality.

### B. Semi-Supervised multi-domain Image-to-Image Translation

Li et al. [185] propose an attribute guided I2I (AGUIT) model that is the first work to handle multimodal and multi-domain I2I with semi-supervised learning. AGUIT is trained following three steps. The first step is representation decomposition, which extracts content and style features with two encoders, a content discriminator and a label predictor, and the style code includes a noise part and an attribute part. The second step is reconstruction and translation using AdaIN [186] and a discriminator and a domain classifier. The third step is consistency reconstruction with cycle consistency loss and feature consistency loss. AGUIT is trained in a training set containing labeled images mixed with unlabeled images so that it can translate attributes well. By going one step further to reduce the amount of labeled data required in the training process and source domain, Wang et al. [187] propose SEMIT to address the challenging problem combined with few-shot I2I. SEMIT initially applies semi-supervised learning via a noise-tolerant pseudo-labeling procedure to assign pseudo-labels to the unlabeled training data. Then, it performs UI2I using adversarial loss, classification loss and reconstruction loss with only a few labeled examples during training.

### C. Few-Shot multi-domain Image-to-Image Translation

Although prolific, the aforementioned successful multi-domain I2I techniques can hardly rapidly generalize from a few examples. In contrast, humans can learn new tasks rapidly using what they learned in the past. Given a static picture of a butterfly, you can easily imagine it flying similar to a bird or a bee after watching a video of a flock of birds or a swarm of bees in flight. Hence, few-shot multi-domain I2I attracts much attention.

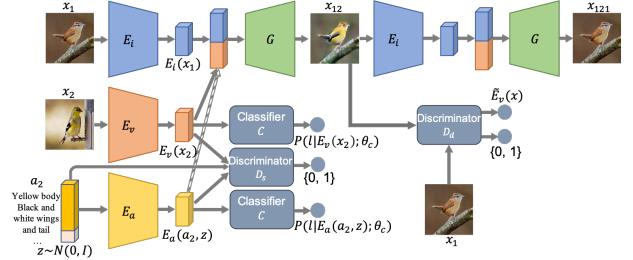


Fig. 17: The architecture of ZstGAN [79].



Fig. 18: Qualitative comparison on single modal two-domain I2I methods. Here we show the examples of edge→shoes.

Liu et al. [18] seek a few-shot UI2I algorithm, FUNIT, to successfully translate source images to analogous images of the target class with many source class images but few target class images available. FUNIT first trains a multiclass UI2I with multiple classes of images, such as those of various animal species, based on a few-shot image translator and a multitask adversarial discriminator. In the test time, it can translate any source class image to analogous images of the target class with a few images from a novel object class (namely, the unseen target class).

However, FUNIT fails to preserve domain invariant appearance information in the content image because of the severe influence of the style code extracted from the target image, namely, the *content loss* problem. Saito et al. [188] therefore proposed COCO-FUNIT to redesign a content-conditioned style encoder that interpolates content information into a style code.

Moreover, Lin et al. found that the current I2I works translate from random noise, which, unlike humans, cannot easily adapt acquired prior knowledge to solve new problems. They hence proposed the unsupervised zero-shot I2I (ZstGAN) [79] shown in Fig.17. ZstGAN uses meta-learning to transfer translation knowledge from seen domains to unseen classes using a translator trained on seen domains to translate images of unseen domains with annotated attributes.

## V. EXPERIMENTAL EVALUATION

In this section, we evaluate twelve I2I models on two tasks, including seven two-domain algorithms on edge-to-shoes translation task and five multi-domain algorithms on attribute manipulation task. We train all the models following their default settings as original papers except the same dataset



Fig. 19: Qualitative comparison on multi-modal two-domain I2I methods. Here we show the examples of edge→shoes, where \* indicates additionally injecting noise vectors to the translation network.

and implementation environments. The selection criteria of methods mainly takes into account algorithm categories and publication years. All experimental codes come from the public official version.

#### A. Datasets

**UT-Zap50K** We utilize the UT-Zap50K dataset [189] to evaluate the performance of two-domain I2I methods. The number of training pairs is 49826 where each pair consists of a shoes image and its corresponding edge map. And the number of testing images is 200. We resize all images to  $256 \times 256$  for training and testing. In unsupervised setting, images from source domain and target domain are not paired.

**CelebA** We employ the CelebFaces Attributes (CelebA) dataset [169] to compare the performance of multi-domain I2I methods. It contains 202,599 face images of celebrities with 40 *with/without* attribute labels for each image. We randomly divide all images into training set, validation set and test set with ratio 8 : 1 : 1. Next, we center-crop the initial  $178 \times 218$  size images to  $178 \times 178$ . Finally, after resizing all images to  $128 \times 128$  by bicubic interpolation, we construct the multiple domains dataset using the following attributes: Black hair, Blond hair, Brown hair, gender (male/female), and age (young/old).

#### B. Metrics

We evaluate both the visual quality and the diversity of generated images using Fréchet inception distance (FID), Inception score (IS) and Learned Perceptual Image Patch Similarity (LPIPS).

**Fréchet inception distance (FID)** [92] is computed by measuring the mean and variance distance of the generated and real images in a deep feature space. A lower score means a better performance. (1) For single-modal two-domain setting, we directly compared the mean and variance of generated and real sets. (2) For multi-modal two-domain setting, we sample the same testing set 19 times. Then compute the FID for each testing set and average the scores to get the final FID score. (3) For single-modal multi-domain setting, we compute the

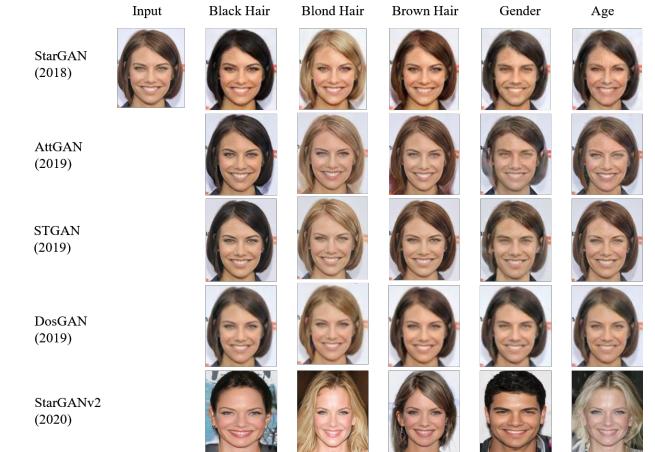


Fig. 20: Qualitative comparison on single modal multi-domain I2I methods. Here we show the examples of 5 attributes.

FID score in each domain and then average the scores. (4) For multi-modal multi-domain setting, we first sample each image in each domain 19 times. Then we compute the average FID scores within each domain. Finally, we average the scores again to get the result.

**Inception score (IS)** [88] encodes the diversity across all translated outputs. It exploits a pretrained inception classification model to predict the domain label of the translated images. A higher score indicates a better translated performance. The evaluation process is just as similar as FID.

**Learned Perceptual Image Patch Similarity (LPIPS)** [96] evaluates the diversity of the translated images and is demonstrated to correlate well with human perceptual similarity. It is computed as the average LPIPS distance between pairs of randomly sampled translation outputs from the same input. Specifically, we sample 100 images with 19 pairs of outputs (randomly sample two style vectors or inputs added random Gaussian noise). We then compute the distance between two generated results and get an average. A higher LPIPS score means a more realistic, diverse translated result.

#### C. Results

A fair comparison is only possible by keeping all the parameters consistent. That said, it is difficult to declare that one algorithm has an absolute superiority over the others. Besides model design itself, there are still many factors influencing the performance, such as training time, batch size and iteration times, FLOPs and number of parameters, etc. Therefore, our conclusion only build on current experimental settings, models and tasks.

**Two-domain I2I** We qualitatively and quantitatively compare pix2pix [1], BicycleGAN [16], CycleGAN [2], U-GAT-IT [37], GDWCT [155], CUT [148] and MUNIT [90] in single-modal and multi-modal setting respectively.

The single-modal qualitative comparisons are shown in Fig. 18 where two supervised methods pix2pix and BicycleGAN achieve better FID, IS and LPIPS scores than unsupervised methods CycleGAN, U-GAT-IT and GDWCT. Without any supervision, the newest method CUT gets the best FID

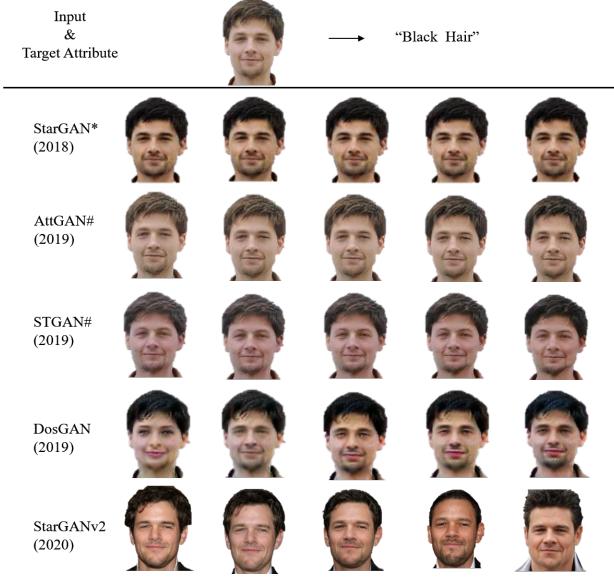


Fig. 21: Qualitative comparison on multi-modal two-domain I2I methods. Here we show the examples of 5 domain translation, where \* indicates additionally injecting noise vectors to the translation network, # denotes the linear interpolation between two attributes: Brown-Hair→Black-Hair.

and IS scores in Table. III than the rest methods including supervised and unsupervised. There could be a couple reasons for that. First, the backbone of CUT, namely StyleGAN, is a strong powerful GAN model for image synthesis compared to others. Besides, the contrastive learning they used is an effective content constraint for translation.

As for the multi-modal setting shown in Fig. 19, we inject the Gaussian noise into the input of pix2pix, CycleGAN, U-GAT-IT, GDWCT and CUT to get multi-modal results. However, they can hardly generate the diverse outputs. On the contrary, the multi-modal algorithms BicycleGAN and MUNIT can acquire multi-modal and realistic results. Also supervised method BicycleGAN achieves 0.047 more LPIPS scores than unsupervised method MUNIT as shown in Table. III.

**Multi-domain I2I** We qualitatively and quantitatively compare StarGAN [170], AttGAN [171], STGAN [173] DosGAN [180] and StarGANv2 [183] in single-modal and multi-modal setting respectively.

In Fig. 20, all methods can successfully achieve multiple domains translation. However, StarGAN and AttGAN generate obvious visible artifacts while DosGAN leads to blurry results. The results of STGAN are pretty excellent whereas StarGANv2 can generate realistic and vivid translated results by changing the image style latent code. Table IV shows that StarGANv2 acquires the best FID and IS scores. Similarly, there are also many factors contributed to such result including stronger GAN backbone, more effective training strategies, higher quality dataset, etc.

We also conduct the multimodal multi-domain I2I experiments for comparison. In detail, we additionally inject noise vectors to StarGAN for multi-modal translation. As for

AttGAN and STGAN, we apply the linear interpolation between two attributes: Brown-Hair→Black-Hair as multi-modal results. As shown in Fig 21 and Table IV, StarGAN fails to generate diverse outputs and get the worst LPIPS score despite the injected randomness. AttGAN and STGAN can generate multi-modal results, but the gap is very small meanwhile DosGAN performs worse translation quality. In comparison, StarGANv2 can generate totally different modality leading to the best LPIPS score.

**Conclusion** Generally, supervised methods usually produce better translated results than unsupervised methods on similar network structure. However, in some special cases, supervised methods do not always perform better than unsupervised methods such as CUT which benefits from the development of network architecture (StyleGAN) and more effective training strategies (contrastive learning). As reported in Table. III and Table. IV, choosing an updated model to train I2I task may be a good idea because such model is usually trained with some of the latest training strategies and well-designed network architecture. Moreover, the high-quality dataset plays a crucial role in I2I task.

## VI. APPLICATION

In this section, we review the various and fruitful applications of I2I shown in Table V. We classify the main applications following the taxonomy of I2I methods.

For realistic-looking image synthesis, the related I2I works tend to generate photos of real-world scenes given different forms of input data. A typical task involves translating a semantic segmentation mask [4], [5], [6], [7], [241], [242] into real-world images, that is, semantic synthesis. Person image synthesis, including virtual try-on [191], [192], [193], [194], [195], [196], [243], [244], [245], [246], [247] and skeleton/keypoint-to-person translation [127], [128], [248] learns to translate an image of a person to another image of the same person with a new outfit as well as diverse poses by manipulating the target clothes or poses. In addition, sketch-to-image translation [1], [86], [124], [130], [156], [78], [105], [90], [198], [75], [249] text-to-image translation [151], [161], [182], [79], [250], [21], audio-to-image translation [239] and painting-to-image translation [102], [14], [75], [164] aim to translate human-drawn sketches, text, audio and artwork paintings to realistic images of the real world. Before I2I, most methods relied on the retrieval of existing photographs and copying the image patches to the corresponding location in an inefficient and time-consuming manner.

Using I2I for image manipulation focuses on altering or modifying an image while keeping the unedited factors unchanged. Semantic manipulation tries to edit the high-level semantics of an image, such as the presence and appearance of objects (image composition) [200] with or without makeup [201]. Attribute manipulation [171], [179], [178] varies the binary representations or utilizes the landmark to edit image attributes, such as the gender of the subject [251], the color of hair [155], [183], the presence of glasses [252] and the facial expression [253], [254], [255], and performs image relabeling [200] as well as gaze correction and animation in the

TABLE III: The average FID, IS, LPIPS scores of different two-domain I2I methods trained on UT-Zap50K dataset [189] in task edge→shoes. The best scores are in bold.

Category	Supervised I2I		Unsupervised I2I					
	single-modal	multi-modal	single-modal			CUT	MUNIT	
Method	pix2pix	BicycleGAN	CycleGAN	U-GAI-IT	GDWCT	CUT	MUNIT	
Publication	2017	2017	2017	2019	2019	2020	2018	
FID ( $\downarrow$ )	65.09	64.23	73.76	91.33	79.56	<b>50.03</b>	75.31	
IS ( $\uparrow$ )	$3.08 \pm 0.39$	$2.96 \pm 0.36$	$2.66 \pm 0.25$	$2.86 \pm 0.31$	$2.69 \pm 0.39$	<b><math>3.21 \pm 0.77</math></b>	$2.33 \pm 0.25$	
LPIPS ( $\uparrow$ )	0.064	<b>0.237</b>	0.038	0.028	0.017	0.019	0.190	

TABLE IV: The average FID, IS, LPIPS scores of different multi-domain I2I methods trained on CelebA dataset [169] in 5 domains including Black hair, Blond hair, Brown hair, gender (male or female) and age (young or old). In addition to the final average metric scores, we also report two domains results (Black hair and gender) for reference. The best scores are in bold.

Category	Method	Publication	FID (Black Hair)	FID (Gender)	FID ( $\downarrow$ )	IS (Black Hair)	IS (Gender)	IS ( $\uparrow$ )	LPIPS (Black Hair)	LPIPS (Gender)	LPIPS ( $\uparrow$ )	
Unsupervised I2I	single- modal	StarGAN	2018	101.66	96.96	94.39	$1.494 \pm 0.167$	$1.506 \pm 0.295$	$1.497 \pm 0.158$	0.009	0.010	0.011
		AttGAN	2019	87.80	83.29	74.48	$1.136 \pm 0.056$	$1.228 \pm 0.086$	$1.231 \pm 0.121$	0.023	0.027	0.021
	multi- modal	STGAN	2019	86.80	95.37	82.41	$1.417 \pm 0.205$	$1.636 \pm 0.461$	$1.568 \pm 0.304$	0.065	0.043	0.036
		DosGAN	2019	68.38	71.35	73.36	$1.617 \pm 0.285$	$1.556 \pm 0.284$	$1.568 \pm 0.206$	0.064	0.055	0.061
		StarGANv2	2020	42.89	43.86	<b>40.52</b>	$1.537 \pm 0.233$	$1.674 \pm 0.440$	<b>1.586 <math>\pm 0.278</math></b>	0.414	0.377	<b>0.397</b>

wild [256]. Moreover, image/video retargeting [203] enables the transfer of sequential content from one domain to another while preserving the style of the target domain. Much of the I2I research focuses on filling in missing pixels, i.e., image inpainting [15], [16], [17], [18], [19] and image outpainting [204], but they treat different occluded images. Taking the image of a human face as an example, the image inpainting task produces visually realistic and semantically correct results from the input with a masked nose, mouth and eyes, while the image outpainting task translates a highly occluded face image that only has a nose, mouth and eyes.

I2I has made great contributions to artistic creation. In the past, redrawing an image in a particular form of art requires a well-trained artist and much time. In contrast, many I2I studies can automatically turn photo-realistic images into synthetic artworks without human intervention. Using the I2I methods for artistic creation can directly translate real-world photographic works into illustrations in children’s books [257], cartoon images [139], [112], [37], [141], [143], [151], [33], [38], [36], [35], [75], [160], [80], comics [34], [205] or a multichirography of Chinese characters [206]. Additionally, the style transfer task achieves remarkable success through I2I methods. It contains two main objectives: artistic style transfer [2], [11], [12], [74], [75], [90], [155], [108], [150], which involves translating the input image to the desired artistic style, such as that of Monet or van Gogh; and photo-realistic style transfer [112], [102], [2], [74], [150], [148], [103], [160], which must clearly maintain the original edge structure when transferring a style.

We can also exploit I2I for image restoration. The goal of image restoration is to restore a degraded image to its original form via the degradation model. Specifically, the image super-resolution task [28], [29] involves increasing the resolution of an image, which is usually trained with down-scaled versions of the target image as inputs. Image denoising [208], [207], [209] aims to remove artificially added noise from the images. Image deraining [210], [211], [212], image dehazing [214], [215], [216], [213], [217] and image deblurring [221], [218],

[219], [220] aim to remove optical distortions from photos that were taken out of focus or while the camera was moving, or from photos of faraway geographical or astronomical features.

Image enhancement is a subjective process that involves heuristic procedures designed to process an image to satisfy the human visual system. I2I proves its effectiveness in this field including image colorization and image quality improvement. Image colorization [22], [23], [24], [25], [26], [27] involves imagining the color of each pixel, given only its luminosity. It is trained on images with their color artificially removed. Image quality improvement [2], [222], [224], [223] focuses on producing noticeably fewer colored artifacts around hard edges and more accurate colors, as well as reduced noise in smooth shadows. Moreover, [258] Learns to fuse multi-focus image using I2I methods.

We also notice that two special types of data are used in I2I algorithms for particular tasks: remote sensing imaging for wildlife habitat analysis [225] and building extraction [259]; medical imaging for disease diagnosis [226], [227], [228], [207], dose calculation [229] and surgical training phantoms improvement [230].

I2I methods can also contribute to other visual tasks, such as transfer learning for reinforcement learning [231], image registration [39], domain adaptation [30], [31], [32], person re-identification [232], [233], [234], [235], [236], image segmentation [8], [9], [10], facial geometry reconstruction [237], 3D pose estimation [20], [21], neural talking head generation [238] and hand gesture-to-gesture translation [240].

## VII. SUMMARY AND OUTLOOK

In recent years, the image-to-image translation (I2I) task has achieved great success and benefited many computer visual tasks. I2I is attracting increasing attention because of its wide practical application value and scope. We therefore conduct this comprehensive review of the analysis, methodology, and related applications of I2I to clarify the main progress the community has made. In detail, we first briefly introduce the two most representative generative models that are widely used

TABLE V: Applications of I2I discussed in Section VI

TASK	TWO-DOMAIN I2I						MULTI-DOMAIN I2I							
	Supervised I2I		Unsupervised I2I		Semi-supervised I2I		Few-shot I2I		Unsupervised I2I		Semi-supervised I2I		Few-shot I2I	
	Single-modal Output	Multi-modal Outputs	Single-modal Output	Multi-modal Outputs	Single-modal Output	Single-modal Output	Single-modal Output	Multi-modal Outputs	Single-modal Output	Multi-modal Outputs	Single-modal Output	Multi-modal Outputs	Single-modal Output	
Semantic synthesis	[123], [1], [86], [124], [4], [5], [3], [127], [128]	[16]	[112], [2], [140], [148], [104], [190]	[78], [161], [106]	-	[122], [164]	[167]	-	-	-	-	-	-	
Person image synthesis (skeleton-to-person) (virtual try-on)	[127], [128]	-	[151], [191], [192], [193], [194], [195], [196]	-	-	-	-	-	-	-	-	-	-	
Sketch-to-image	[1], [86], [124], [127], [128]	[16], [130], [197]	[112], [198], [11], [2], [142], [156]	[78], [105], [90], [75], [106]	[199]	-	[168]	[184]	-	-	-	-	-	
Paint-to-image	-	-	[102], [14]	[75]	-	[164]	-	-	-	-	-	-	-	
Text-to-image	-	-	[151]	[161]	-	-	-	[182]	-	-	-	-	[79]	
Semantic manipulation	[4], [5], [6]	-	[200], [102], [201]	-	-	-	-	-	-	-	-	-	-	
Attribute manipulation	-	-	[74], [142], [143], [102], [155], [202]	[76]	-	[165]	[174], [175], [170], [171], [172], [173], [176], [166], [178], [168], [179]	[180], [181], [32], [182], [183], [184]	[185], [187]	[185]	[18]	[188]	-	
Retargeting	-	-	[203]	-	-	-	-	-	-	-	-	-	-	
Image inpainting	-	-	[15], [16], [17], [18], [19]	-	-	-	-	-	-	-	-	-	-	
Image outpainting	-	-	[204]	-	-	-	-	-	-	-	-	-	-	
Image-to-cartoon	-	-	[139], [112], [37], [141], [143], [151], [33], [38], [36], [35]	[75], [160]	-	[80]	-	[32], [108]	-	-	-	-	-	
Image-to-comics	-	-	[34], [205]	-	-	-	-	-	-	-	-	-	-	
Chinese character translation	-	-	[206]	-	-	-	-	-	-	-	-	-	-	
style transfer	artistic:[123], [1]	artistic:[16]	artistic:[2], [37], [150], [148], [155]; photo-realistic:[112], [102], [2], [74], [150], [148]	artistic:[160]; photo-realistic:[90], [75], [76], [103], [165], [106]	-	artistic:[164], [80], [177]; photo-realistic:[164], [165], [80]	artistic:[174], [171], [167]; photo-realistic:[171], [173], [167]	artistic:[108]; photo-realistic:[180], [182], [108]	-	-	-	-	-	
image super-resolution	[123]	[130]	-	-	[113]	-	-	-	-	-	-	-	-	
image denoising	[207]	-	[208], [209]	-	[113]	-	-	-	-	-	-	-	-	
image deraining	[210], [211]	-	[212]	-	-	-	-	-	-	-	-	-	-	
image dehazing	[213]	-	[214], [215], [216], [217]	-	-	-	-	-	-	-	-	-	-	
image deblurring	[218], [219], [220]	-	[221]	-	-	-	-	-	-	-	-	-	-	
image colorization	[1], [86], [22], [24], [25], [26]	-	[151], [23], [27]	-	[113]	-	-	-	-	-	-	-	-	
image quality improvement	[222], [223]	-	[2], [224]	-	-	-	-	-	-	-	-	-	-	
wildlife habitat analysis	}	-	[225]	-	-	-	-	-	-	-	-	-	-	
disease diagnosis	[207]	-	[226], [227], [228]	-	-	-	-	[176], [178]	-	-	-	-	-	
dose calculation	-	-	[229]	-	-	-	-	-	-	-	-	-	-	
surgical training	-	-	[230]	-	-	-	-	-	-	-	-	-	-	
phantoms improvement	-	-	[231]	-	-	-	-	-	-	-	-	-	-	
transfer learning	-	-	[39]	-	-	-	-	-	-	-	-	-	-	
image registration	-	-	[30], [31]	-	-	-	-	[32]	-	-	-	-	-	
domain adaptation	-	-	[232], [233], [234], [235], [236]	-	-	-	-	-	-	-	-	-	-	
person re-identification	-	-	[237]	-	-	-	-	-	-	-	-	-	-	
image segmentation	[1], [8], [9], [10]	-	[2], [80], [190]	-	-	-	-	-	-	-	-	-	-	
facial geometry reconstruction	-	-	[20], [21]	-	-	-	-	-	-	-	-	-	-	
3D pose estimation	-	-	[238]	-	-	-	-	-	-	-	-	-	-	
neural talking head generation	-	-	[239]	-	-	-	-	-	-	-	-	-	-	
Audio-to-image	[240]	-	-	-	-	-	-	-	-	-	-	-	-	
hand gesture-to-gesture translation	-	-	-	-	-	-	-	-	-	-	-	-	-	

as the backbone of I2I and some well-known evaluation metrics. Then, we elaborate on the methodology of I2I regarding two-domain and multi-domain tasks. In addition, we provide a thorough taxonomy of the I2I applications.

Looking forward, there are still many challenges in I2I, which need further explorations and investigations. The most iconic dilemma is the trade-off between network complexity and result quality with higher resolution. Similarly, the efficiency should also be considered when the I2I framework attempts to generate diverse and high-fidelity outputs. We believe that a more lightweight I2I network would attract more

attention for practical application. Moreover, it is an interesting research trend to generalize the aforementioned methods to domains beyond images, such as those of language, text and speech, termed cross-modality translation tasks. Overall, we hope that this article can serve as a basis for the development of better methods for I2I and inspire researchers in more domains in addition to images.

## REFERENCES

- [1] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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.
- [2] J.-Y. Zhu, T. Park, P. Isola, and A. 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, Y. Xu *et al.*, “MRI cross-modality neuroimage-to-neuroimage translation,” *arXiv preprint arXiv:1801.06940*, 2018.
- [9] X. Guo, Z. Wang, Q. Yang, W. Lv, X. Liu, Q. Wu, and J. Huang, “Gan-based virtual-to-real image translation for urban scene semantic segmentation,” *Neurocomputing*, vol. 394, pp. 127–135, 2020.
- [10] R. Li, W. Cao, Q. Jiao, S. Wu, and H.-S. Wong, “Simplified unsupervised image translation for semantic segmentation adaptation,” *Pattern Recognition*, vol. 105, p. 107343, 2020.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, “Toward multimodal image-to-image translation” in *Advances in Neural Information Processing Systems*, 2017, pp. 465–476.
- [17] 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.
- [18] 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.
- [19] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, and D. Lu, “Uctgan: Diverse image inpainting based on unsupervised cross-space translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [20] H.-Y. Fish Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, “Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [21] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, “Manigan: Text-guided image manipulation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [22] 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,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] J. Cao, O. Katzir, P. Jiang, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, “Dida: Disentangled synthesis for domain adaptation,” *CoRR*, vol. abs/1805.08019, 2018.
- [32] 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.
- [33] Y. Shi, D. Deb, and A. K. Jain, “Warpgan: Automatic caricature generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10762–10771.
- [34] M. Pesko, A. Svystun, P. Andruszkiewicz, P. Rokita, and T. Trzcinski, “Comixify: Transform video into comics,” *Fundamenta Informaticae*, vol. 168, no. 2–4, pp. 311–333, 2019.
- [35] Z. Zheng, C. Wang, Z. Yu, N. Wang, H. Zheng, and B. Zheng, “Unpaired photo-to-caricature translation on faces in the wild,” *Neurocomputing*, vol. 355, pp. 71–81, 2019.
- [36] Y. Chen, Y.-K. Lai, and Y.-J. Liu, “Cartoongan: Generative adversarial networks for photo cartoonization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9465–9474.
- [37] J. Kim, M. Kim, H. Kang, and K. H. Lee, “U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation,” in *International Conference on Learning Representations*, 2019.
- [38] X. Wang and J. Yu, “Learning to cartoonize using white-box cartoon representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8090–8099.
- [39] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, “Unsupervised multi-modal image registration via geometry preserving image-to-image translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [40] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, “A tutorial on energy-based learning,” *Predicting structured data*, vol. 1, no. 0, 2006.
- [41] J. Xu, H. Li, and S. Zhou, “An overview of deep generative models,” *IETE Technical Review*, vol. 32, no. 2, pp. 131–139, 2015.
- [42] A. Oussidi and A. Elhassouny, “Deep generative models: Survey,” in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2018, pp. 1–8.
- [43] H.-M. Chu, C.-K. Yeh, and Y.-C. Frank Wang, “Deep generative models for weakly-supervised multi-label classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 400–415.

- [44] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic image inpainting with deep generative models,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5485–5493.
- [45] M. Tschannen, E. Agustsson, and M. Lucic, “Deep generative models for distribution-preserving lossy compression,” in *Advances in Neural Information Processing Systems*, 2018, pp. 5929–5940.
- [46] R. Salakhutdinov and G. Hinton, “Deep boltzmann machines,” in *Artificial intelligence and statistics*, 2009, pp. 448–455.
- [47] R. Salakhutdinov, A. Mnih, and G. Hinton, “Restricted boltzmann machines for collaborative filtering,” in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 791–798.
- [48] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [49] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves et al., “Conditional image generation with pixelcnn decoders,” in *Advances in neural information processing systems*, 2016, pp. 4790–4798.
- [50] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” in *International Conference on Machine Learning*. PMLR, 2016, pp. 1747–1756.
- [51] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, “Pixelsnail: An improved autoregressive generative model,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 864–872.
- [52] L. Dinh, D. Krueger, and Y. Bengio, “Nice: Non-linear independent components estimation,” *arXiv preprint arXiv:1410.8516*, 2014.
- [53] A. Abdelhamed, M. A. Brubaker, and M. S. Brown, “Noise flow: Noise modeling with conditional normalizing flows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [54] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, “The helmholtz machine,” *Neural computation*, vol. 7, no. 5, pp. 889–904, 1995.
- [55] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *stat*, vol. 1050, p. 1, 2014.
- [56] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.
- [57] H. Larochelle and I. Murray, “The neural autoregressive distribution estimator,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 29–37.
- [58] M. Germain, K. Gregor, I. Murray, and H. Larochelle, “Made: Masked autoencoder for distribution estimation,” in *International Conference on Machine Learning*, 2015, pp. 881–889.
- [59] E. Nalisnick, L. Hertel, and P. Smyth, “Approximate inference for deep latent gaussian mixtures,” in *NIPS Workshop on Bayesian Deep Learning*, vol. 2, 2016, p. 131.
- [60] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 1530–1538.
- [61] J. Tomczak and M. Welling, “Vae with a vampprior,” in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1214–1223.
- [62] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” in *International Conference on Learning Representations*, 2018.
- [63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems (NIPS)*, 2014, pp. 2672–2680.
- [64] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*. PMLR, 2017, pp. 214–223.
- [65] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.
- [66] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- [67] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.
- [68] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.
- [69] J. J. Zhao, M. Mathieu, and Y. LeCun, “Energy-based generative adversarial networks,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- [70] D. Berthelot, T. Schumm, and L. Metz, “Began: boundary equilibrium generative adversarial networks,” *arXiv preprint arXiv:1703.10717*, 2017.
- [71] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard gan,” in *International Conference on Learning Representations*, 2018.
- [72] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *International Conference on Learning Representations*, 2018.
- [73] P. Ghosh, M. S. Sajjadi, A. Vergari, M. Black, and B. Scholkopf, “From variational to deterministic autoencoders,” in *International Conference on Learning Representations*, 2019.
- [74] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Advances in neural information processing systems*, 2017, pp. 700–708.
- [75] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, “Diverse image-to-image translation via disentangled representations,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.
- [76] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. Van Gool, “Exemplar guided unsupervised image-to-image translation with semantic consistency,” in *International Conference on Learning Representations*, 2018.
- [77] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, “Transgaga: Geometry-aware unsupervised image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8012–8021.
- [78] H. Kazemi, S. Soleymani, F. Taherkhani, S. Iranmanesh, and N. Nasrabadi, “Unsupervised image-to-image translation using domain-specific variational information bound,” in *Advances in neural information processing systems*, 2018, pp. 10348–10358.
- [79] 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.
- [80] 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.
- [81] S. Pal and S. Mitra, “Multilayer perceptron, fuzzy sets, and classification,” *IEEE transactions on neural networks*, vol. 3, no. 5, pp. 683–697, 1992.
- [82] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [83] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- [84] T. Tieleman and G. Hinton, “Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude,” COURSERA: Neural Networks for Machine Learning, 2012.
- [85] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666.
- [86] 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.
- [87] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [88] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Advances in neural information processing systems*, 2016, pp. 2234–2242.
- [89] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, “Pose guided person image generation,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [90] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.

- [91] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711.
- [92] 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,” in *Advances in neural information processing systems*, 2017, pp. 6626–6637.
- [93] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying mmd gans,” in *International Conference on Learning Representations*, 2018.
- [94] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [95] T. R. Shaham, T. Dekel, and T. Michaeli, “Singan: Learning a generative model from a single natural image,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4570–4580.
- [96] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [97] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [98] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable fidelity and diversity metrics for generative models,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 7176–7185.
- [99] 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.
- [100] 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.
- [101] O. Katzir, D. Lischinski, and D. Cohen-Or, “Cross-domain cascaded deep translation,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12347. Springer, 2020, pp. 673–689.
- [102] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. A. Efros, and R. Zhang, “Swapping autoencoder for deep image manipulation,” *arXiv preprint arXiv:2007.00653*, 2020.
- [103] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, “Towards instance-level image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3683–3692.
- [104] H. Chen, Y. Wang, H. Shu, C. Wen, C. Xu, B. Shi, C. Xu, and C. Xu, “Distilling portable generative adversarial networks for image translation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3585–3592.
- [105] A. Almahairi, S. Rajeshwar, A. Sordoni, P. Bachman, and A. Courville, “Augmented cyclegan: Learning many-to-many mappings from unpaired data,” ser. *Proceedings of Machine Learning Research*, J. Dy and A. Krause, Eds., vol. 80. Stockholmssässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 195–204.
- [106] Y. Alharbi, N. Smith, and P. Wonka, “Latent filter scaling for multi-modal unsupervised image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1458–1466.
- [107] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. J. Lee, E. Shechtman, and R. Zhang, “Few-shot image generation via cross-domain correspondence,” *arXiv preprint arXiv:2104.06820*, 2021.
- [108] 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.
- [109] R. Navigli, “Word sense disambiguation: A survey,” *ACM computing surveys (CSUR)*, vol. 41, no. 2, pp. 1–69, 2009.
- [110] 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.
- [111] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, “A survey on semi-, self- and unsupervised learning for image classification,” 2020.
- [112] 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.
- [113] 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.
- [114] 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.
- [115] 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.
- [116] 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.
- [117] 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.
- [118] 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.
- [119] 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.
- [120] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in neural information processing systems*, 2017, pp. 4077–4087.
- [121] 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.
- [122] J. Lin, Y. Wang, T. He, and Z. Chen, “Learning to transfer: Unsupervised meta domain translation,” *arXiv preprint arXiv:1906.00181*, 2019.
- [123] 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.
- [124] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.
- [125] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2417–2426.
- [126] 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.
- [127] 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.
- [128] 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.
- [129] I. Goodfellow, “Nips 2016 tutorial: Generative adversarial networks,” 2017.
- [130] A. Bansal, Y. Sheikh, and D. Ramanan, “Pixelnn: Example-based image synthesis,” in *International Conference on Learning Representations*, 2018.
- [131] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [132] 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.
- [133] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Advances in neural information processing systems*, 2016, pp. 2172–2180.
- [134] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” *arXiv preprint arXiv:1605.09782*, 2016.
- [135] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville, “Adversarially learned inference,” *stat*, vol. 1050, p. 2, 2016.

- [136] 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.
- [137] H. Kim and A. Mnih, “Disentangling by factorising,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.
- [138] 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.
- [139] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- [140] M. Li, H. Huang, L. Ma, W. Liu, T. Zhang, and Y. Jiang, “Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 184–199.
- [141] A. Gokaslan, V. Ramanujan, D. Ritchie, K. In Kim, and J. Tompkin, “Improving shape deformation in unsupervised image-to-image translation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 649–665.
- [142] M. Amadio and S. Krishnaswamy, “Travelgan: Image-to-image translation by transformation vector learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8983–8992.
- [143] Y. Zhao, R. Wu, and H. Dong, “Unpaired image-to-image translation using adversarial consistency loss,” in *European Conference on Computer Vision*. Springer, 2020, pp. 800–815.
- [144] S. Benaim and L. Wolf, “One-sided unsupervised domain mapping,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [145] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, “Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2427–2436.
- [146] C. Zheng, T.-J. Cham, and J. Cai, “The spatially-correlative loss for various image translation tasks,” *arXiv preprint arXiv:2104.00854*, 2021.
- [147] J. Liang, H. Zeng, and L. Zhang, “High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network,” *arXiv preprint arXiv:2105.09188*, 2021.
- [148] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 319–345.
- [149] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [150] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, “Tsist: A simple and versatile framework for image-to-image translation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 206–222.
- [151] S. Ma, J. Fu, C. W. Chen, and T. Mei, “Da-gan: Instance-level image translation by deep attention generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5657–5666.
- [152] X. Chen, C. Xu, X. Yang, and D. Tao, “Attention-gan for object transfiguration in wild images,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [153] S. Mo, M. Cho, and J. Shin, “Instagan: Instance-aware image-to-image translation,” in *International Conference on Learning Representations*, 2018.
- [154] D. Bhattacharjee, S. Kim, G. Vizier, and M. Salzmann, “Dunit: Detection-based unsupervised image-to-image translation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [155] W. Cho, S. Choi, D. K. Park, I. Shin, and J. Choo, “Image-to-image translation via group-wise deep whitening-and-coloring transformation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 639–10 647.
- [156] Y.-C. Chen, X. Xu, and J. Jia, “Domain adaptive image-to-image translation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [157] T. F. van der Ouderaa and D. E. Worrall, “Reversible gans for memory-efficient image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4720–4728.
- [158] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, “Reusing discriminators for encoding: Towards unsupervised image-to-image translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8168–8177.
- [159] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, “Conditional image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [160] H.-Y. Chang, Z. Wang, and Y.-Y. Chuang, “Domain-specific mappings for generative adversarial style transfer,” in *European Conference on Computer Vision*. Springer, 2020, pp. 573–589.
- [161] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, “Mode seeking generative adversarial networks for diverse image synthesis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1429–1437.
- [162] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, “Transferring gans: generating images from limited data,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 218–234.
- [163] Y. Li, R. Zhang, J. C. Lu, and E. Shechtman, “Few-shot image generation with elastic weight consolidation,” in *Advances in Neural Information Processing Systems*, 2020.
- [164] S. Benaim and L. Wolf, “One-shot unsupervised cross domain translation,” in *Advances in Neural Information Processing Systems*, 2018, pp. 2104–2114.
- [165] T. Cohen and L. Wolf, “Bidirectional one-shot unsupervised domain mapping,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1784–1792.
- [166] J. Lin, Y. Xia, Y. Wang, T. Qin, and Z. Chen, “Image-to-image translation with multi-path consistency regularization,” in *IJCAI*, 2019, pp. 2980–2986.
- [167] T. He, Y. Xia, J. Lin, X. Tan, D. He, T. Qin, and Z. Chen, “Deliberation learning for image-to-image translation.” in *IJCAI*, 2019, pp. 2484–2490.
- [168] J. Cao, H. Huang, Y. Li, R. He, and Z. Sun, “Informative sample mining network for multi-domain image-to-image translation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 404–419.
- [169] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [170] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stgan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [171] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “AttnGAN: Facial attribute editing by only changing what you want,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.
- [172] P.-W. Wu, Y.-J. Lin, C.-H. Chang, E. Y. Chang, and S.-W. Liao, “Relgan: Multi-domain image-to-image translation via relative attributes,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [173] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, “Stgan: A unified selective transfer network for arbitrary image attribute editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [174] L. Hui, X. Li, J. Chen, H. He, and J. Yang, “Unsupervised multi-domain image translation with domain-specific encoders/decoders,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2044–2049.
- [175] B. Zhao, B. Chang, Z. Jie, and L. Sigal, “Modular generative adversarial networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [176] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, “Collagan: Collaborative gan for missing image data imputation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [177] S. Chang, S. Park, J. Yang, and N. Kwak, “Sym-parameterized dynamic inference for mixed-domain image translation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [178] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, “Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

- [179] R. Wu, X. Tao, X. Gu, X. Shen, and J. Jia, "Attribute-driven spontaneous motion in unpaired image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [180] J. Lin, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo, "Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1254–1266, 2021.
- [181] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *The European Conference on Computer Vision (ECCV)*, September 2018.
- [182] X. Yu, Y. Chen, S. Liu, T. Li, and G. Li, "Multi-mapping image-to-image translation via learning disentanglement," in *Advances in Neural Information Processing Systems*, 2019, pp. 2994–3004.
- [183] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188–8197.
- [184] Y. Liu, M. De Nadai, J. Yao, N. Sebe, B. Lepri, and X. Alameda-Pineda, "Gmm-unit: Unsupervised multi-domain and multi-modal image-to-image translation via attribute gaussian mixture modeling," *arXiv preprint arXiv:2003.06788*, 2020.
- [185] X. Li, J. Hu, S. Zhang, X. Hong, Q. Ye, C. Wu, and R. Ji, "Attribute guided unpaired image-to-image translation with semi-supervised learning," *arXiv preprint arXiv:1904.12428*, 2019.
- [186] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
- [187] Y. Wang, S. Khan, A. Gonzalez-Garcia, J. van de Weijer, and F. S. Khan, "Semi-supervised learning for few-shot image-to-image translation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [188] K. Saito, K. Saenko, and M.-Y. Liu, "Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 382–398.
- [189] 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.
- [190] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, "Asymmetric gan for unpaired image-to-image translation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5881–5896, 2019.
- [191] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 199–207.
- [192] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," 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. 406–416.
- [193] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3408–3416.
- [194] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [195] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [196] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, "Towards multi-pose guided virtual try-on network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [197] J. Huang, J. Liao, and S. Kwong, "Semantic example guided image-to-image translation," *IEEE Transactions on Multimedia*, 2020.
- [198] W. Chen and J. Hays, "Sketchygan: Towards diverse and realistic sketch to image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9416–9425.
- [199] Z. Li, C. Deng, E. Yang, and D. Tao, "Staged sketch-to-image synthesis via semi-supervised generative adversarial networks," *IEEE Transactions on Multimedia*, 2020.
- [200] A. Shocher, Y. Gandelsman, I. Mosseri, M. Yarom, M. Irani, W. T. Freeman, and T. Dekel, "Semantic pyramid for image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [201] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 40–48.
- [202] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "Spa-gan: Spatial attention gan for image-to-image translation," *IEEE Transactions on Multimedia*, vol. 23, pp. 391–401, 2020.
- [203] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [204] L. Zhang, J. Wang, Y. Xu, J. Min, T. Wen, J. C. Gee, and J. Shi, "Nested scale-editing for conditional image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [205] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, and J. Wan, "Manganan: Unpaired photo-to-manga translation based on the methodology of manga drawing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2611–2619.
- [206] Y. Gao and J. Wu, "Gan-based unpaired chinese character image translation via skeleton transformation and stroke rendering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 646–653.
- [207] K. Armanios, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, vol. 79, p. 101684, 2020.
- [208] I. Manakov, M. Rohm, C. Kern, B. Schworm, K. Kortuem, and V. Tresp, "Noise as domain shift: Denoising medical images by unpaired image translation," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 3–10.
- [209] H. Touvron, M. Douze, M. Cord, and H. Jégou, "Powers of layers for image-to-image translation," *arXiv preprint arXiv:2008.05763*, 2020.
- [210] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3943–3956, 2020.
- [211] R. Li, L.-F. Cheong, and R. T. Tan, "Heavy rain image restoration: Integrating physics model and conditional adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1633–1642.
- [212] H. Zhu, X. Peng, J. T. Zhou, S. Yang, V. Chanderasekh, L. Li, and J.-H. Lim, "Single image rain removal with unpaired information: A differentiable programming perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9332–9339.
- [213] A. Dudhane, H. S. Aulakh, and S. Murala, "Ri-gan: An end-to-end network for single image haze removal," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2014–2023.
- [214] D. Engin, A. Genç, and H. Kemal Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 825–833.
- [215] Y. Cho, R. Malav, G. Pandey, and A. Kim, "Dehazegan: Underwater haze image restoration using unpaired image-to-image translation," *IFAC-PapersOnLine*, vol. 52, no. 21, pp. 82–85, 2019.
- [216] Y.-F. Chen, A. K. Patel, and C.-P. Chen, "Image haze removal by adaptive cyclegan," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1122–1127.
- [217] Y. Cho, H. Jang, R. Malav, G. Pandey, and A. Kim, "Underwater image dehazing via unpaired image-to-image translation," *International Journal of Control, Automation and Systems*, vol. 18, no. 3, pp. 605–614, 2020.
- [218] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8183–8192.
- [219] H. Liu, P. Navarrete Michelin, and D. Zhu, "Deep networks for image-to-image translation with mux and demux layers," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

- [220] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, “Deblurgan-v2: De-blurring (orders-of-magnitude) faster and better,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8878–8887.
- [221] T. Madam Nimisha, K. Sunil, and A. Rajagopalan, “Unsupervised class-specific deblurring,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 353–369.
- [222] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, “Dslr-quality photos on mobile devices with deep convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3277–3285.
- [223] E. de Stoutz, A. Ignatov, N. Kobyshev, R. Timofte, and L. Van Gool, “Fast perceptual image enhancement,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.
- [224] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6306–6314.
- [225] R. Zheng, Z. Luo, and B. Yan, “Exploiting time-series image-to-image translation to expand the range of wildlife habitat analysis,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 825–832.
- [226] R. Zhang, T. Pfister, and J. Li, “Harmonic unpaired image-to-image translation,” in *International Conference on Learning Representations*, 2018.
- [227] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, “Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 191–200.
- [228] K. Armanious, C. Jiang, S. Abdulatif, T. Küstner, S. Gatidis, and B. Yang, “Unsupervised medical image translation using cycle-medgan,” in *2019 27th European Signal Processing Conference (EU-SIPCO)*. IEEE, 2019, pp. 1–5.
- [229] S. Kaji and S. Kida, “Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging,” *Radiological physics and technology*, vol. 12, no. 3, pp. 235–248, 2019.
- [230] S. Engelhardt, R. De Simone, P. M. Full, M. Karck, and I. Wolf, “Improving surgical training phantoms by hyperrealism: deep unpaired image-to-image translation from real surgeries,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 747–755.
- [231] S. Gamiani and Y. Goldberg, “Transfer learning for related reinforcement learning tasks via image-to-image translation,” in *International Conference on Machine Learning*, 2019, pp. 2063–2072.
- [232] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [233] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camera style adaptation for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [234] Z. Zhong, L. Zheng, S. Li, and Y. Yang, “Generalizing a person retrieval model hetero- and homogeneously,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [235] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [236] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, “Adaptive transfer network for cross-domain person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [237] M. Sela, E. Richardson, and R. Kimmel, “Unrestricted facial geometry reconstruction using image-to-image translation,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [238] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [239] B. Duan, W. Wang, H. Tang, H. Latapie, and Y. Yan, “Cascade attention guided residue learning gan for cross-modal translation,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1336–1343.
- [240] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, “Gesturegan for hand gesture-to-gesture translation in the wild,” ser. MM ’18. New York, NY, USA: Association for Computing Machinery, 2018, p. 774–782.
- [241] H. Tang, S. Bai, and N. Sebe, “Dual attention gans for semantic image synthesis,” in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1994–2002.
- [242] H. Tang, X. Qi, D. Xu, P. H. Torr, and N. Sebe, “Edge guided gans with semantic preserving for semantic image synthesis,” *arXiv preprint arXiv:2003.13898*, 2020.
- [243] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [244] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, “Progressive pose attention transfer for person image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [245] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, “Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5904–5913.
- [246] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, “Xinggan for person image generation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 717–734.
- [247] H. Tang, S. Bai, P. H. Torr, and N. Sebe, “Bipartite graph reasoning gans for person image generation,” in *BMVC*, 2020.
- [248] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan, “Cycle in cycle generative adversarial networks for keypoint-guided image generation,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2052–2060.
- [249] H. Tang, X. Chen, W. Wang, D. Xu, J. J. Corso, N. Sebe, and Y. Yan, “Attribute-guided sketch generation,” in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.
- [250] M. Tao, H. Tang, S. Wu, N. Sebe, X.-Y. Jing, F. Wu, and B. Bao, “Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis,” *arXiv preprint arXiv:2008.05865*, 2020.
- [251] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. DENOYER, and M. A. Ranzato, “Fader networks: manipulating images by sliding attributes,” 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. 5967–5976.
- [252] O. Press, T. Galanti, S. Benaim, and L. Wolf, “Emerging disentanglement in auto-encoder based unsupervised image content transfer,” in *International Conference on Learning Representations*, 2018.
- [253] H. Tang, W. Wang, S. Wu, X. Chen, D. Xu, N. Sebe, and Y. Yan, “Expression conditional gan for facial expression-to-expression translation,” in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4449–4453.
- [254] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, “Every smile is unique: Landmark-guided diverse smile generation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7083–7092.
- [255] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, “Geometry guided adversarial facial expression synthesis,” in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM ’18. New York, NY, USA: Association for Computing Machinery, 2018, p. 627–635.
- [256] J. Zhang, J. Chen, H. Tang, W. Wang, Y. Yan, E. Sangineto, and N. Sebe, “Dual in-painting model for unsupervised gaze correction and animation in the wild,” in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1588–1596.
- [257] S. Hicsonmez, N. Samet, E. Akbas, and P. Duygulu, “Ganilla: Generative adversarial networks for image to illustration translation,” *Image and Vision Computing*, vol. 95, p. 103886, 2020.
- [258] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, “Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network,” *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1982–1996, 2019.
- [259] L. Ding, H. Tang, Y. Liu, Y. Shi, X. X. Zhu, and L. Bruzzone, “Adversarial shape learning for building extraction in vhr remote sensing images,” *arXiv preprint arXiv:2102.11262*, 2021.