symmetry

MDPI

Review

Deep Generative Adversarial Networks for Image-to-Image Translation: A Review

Aziz Alotaibi

College of Computers and Information Technology, Taif University, Taif 21974, Saudi Arabia; azotaibi@tu.edu.sa

Received: 3 August 2020; Accepted: 9 October 2020; Published: 16 October 2020

Abstract: Many image processing, computer graphics, and computer vision problems can be treated as image-to-image translation tasks. Such translation entails learning to map one visual representation of a given input to another representation. Image-to-image translation with generative adversarial networks (GANs) has been intensively studied and applied to various tasks, such as multimodal image-to-image translation, super-resolution translation, object transfiguration-related translation, etc. However, image-to-image translation techniques suffer from some problems, such as mode collapse, instability, and a lack of diversity. This article provides a comprehensive overview of image-to-image translation based on GAN algorithms and its variants. It also discusses and analyzes current state-of-the-art image-to-image translation techniques that are based on multimodal and multidomain representations. Finally, open issues and future research directions utilizing reinforcement learning and three-dimensional (3D) modal translation are summarized and discussed.

Keywords: image-to-image translation; generative adversarial networks; adversarial learning; deep generative model; deep learning

1. Introduction

With the rapid advancement in deep learning algorithms, the tasks of analyzing and understanding digital images for many computer vision applications have drawn increasing attention in the recent years due to such algorithms' extraordinary performance and availability of large amounts of data. Such algorithms directly process raw data (e.g., an RGB image) and obviate the need for domain experts or handcrafted features [1–3]. The powerful ability of deep feature learning to automatically utilize complex and high-level feature representations has significantly advanced the performance of state-of-the-art methods across computer applications, such as object detection [4], medical imaging [5,6], image segmentation [7], image classification [8], and face detection [9]. The underlying structure and distinctive (complex) features are both discovered via deep learning-based methods that can be classified further into discriminative feature-learning algorithms and generative feature-learning algorithms. Discriminative models focus on the classification-learning process by learning the conditional probability p (x|y) to map input x to class label y. One of the most popular methods used for image feature learning utilizes convolutional neural networks (CNN) for feature extraction and image classification. Examples include LeNet [8], AlexNet [10], VGGNet [11], and ResNet [12,13], all of which are supervised learning algorithms. On the other hand, generative models focus on the data distribution to discover the underlying features from large amounts of data in an unsupervised setting. Such models are able to generate new samples by learning the estimation of the joint probability distribution p (x,y) and predicting y [14] in contexts, such as image super-resolution [15,16], text-to-image generation [17,18], and image-to-image translation [19,20].

2. Deep Generative Models

Generative models can be categorized into traditional generative models based on machine learning algorithms and deep generative models that are based on deep learning algorithms [21]. Traditional generative models use various forms of probability density function to approximate the distribution [22], and cannot perform well on complex distributions. Such models include infinite Gaussian mixture models (GMM) [23], the hidden naive Bayes model (NBM) [24], and hidden Markov models (HMM) [25]. Deep generative models utilize techniques, such as stochastic backpropagation, deep neural networks, and approximate Bayesian inference, in order to generate new samples based on variational distributions from large-scale datasets [26–28]; examples of such models include the deep Boltzmann machine (DBM) [29], deep belief networks (DBN) [30], variational autoencoder (VAE) [31], and generative adversarial networks GAN [32]. The most dominant and efficient deep generative models of recent years have been VAE and GAN. A variational autoencoder learns the underlying probability distribution and generates a new sample that is based on Bayesian inference by maximizing the lower bound of the data's log-likelihood. In contrast, generative adversarial networks learn data distributions through the adversarial training process based on game theory instead of maximizing the likelihood. The GAN approach offers several advantages over VAE-based models: (1) the ability to learn and model complex data distributions and (2) the ability to efficiently generate sharp and realistic samples [21,33,34].

2.1. Generative Adversarial Networks

The generative adversarial network proposed by Goodfellow et al. in 2014 has been one of significant recent developments in the domain of unsupervised deep generative models [32]. Figure 1 illustrates the architecture of a typical GAN. A GAN is composed of two competing neural networks inspired by the two-player minmax game: a generative network, called a generator and denoted G, and a discriminative network, called a discriminator and denoted D. The generator network tries to generate realistic samples to fool the discriminator, while the discriminator tries to distinguish real samples from fake samples. Generator and discriminator networks can both be any algorithms as long as the generator has the ability to learn the data distribution of the training data and the discriminator has the ability to extract the feature to classify the output. For instance, the generator network can be a de-convolutional neural network, and the discriminator network can be a convolutional neural network or a recurrent neural network [21]. Thus, GANs can be used to generate multi-dimensional data distributions, e.g., images. GANs have been used to make promising contributions in variety of difficult generative tasks [35], e.g., text-to-photo translation [18], image generation [36], image composition [37], and image-to-image translation [38]. Although GANs are one type of powerful deep generative models, the training of GANs suffers from several issues, such as mode collapse and training instability [39], as discussed in Section 7.1.

2.2. Image-To-Image Translation

The idea of image-to-image translation goes back to Hertzmann et al.'s image analogies [40], a proposed non-parametric model using a pair of images to achieve image transformation [41]. Many problems involving computer vision and computer graphics applications can be regarded as instances of the image-to-image translation problem. The task of Image-to-image translation is to learn the mapping from a given image (X) to a specific target image (Y), e.g., mapping grayscale images to RGB images. Learning the mapping from one visual representation to another requires an understanding of underlying features that are shared between these representations, such features are either domain-independent or domain-specific. Domain-independent features represent the underlying spatial structure and they should be preserved during translation (i.e., the content should be preserved when a natural image is translated to Van Gogh's style), while domain-specific features are related to the rendering of the structure and they could be changed during translation (i.e., if the style should be changed when translating the image to Van Gogh' styles) [42,43]. However, learning the mapping between two or multiple domains is a challenging task for two reasons. First, collecting a pair of images may be difficult or the relevant images might sometimes not exist. The second difficulty is that in performing a multi-model translation, whereby one input image maps to multiple outputs. In recent years, GANs and their variants have been used to provide state-of-the-art solutions to image-to-image translation problems. This article classifies the proposed solutions according to two image-to-image translation settings, supervised and unsupervised Image-to-Image Translation setting, as explained in Section 5.

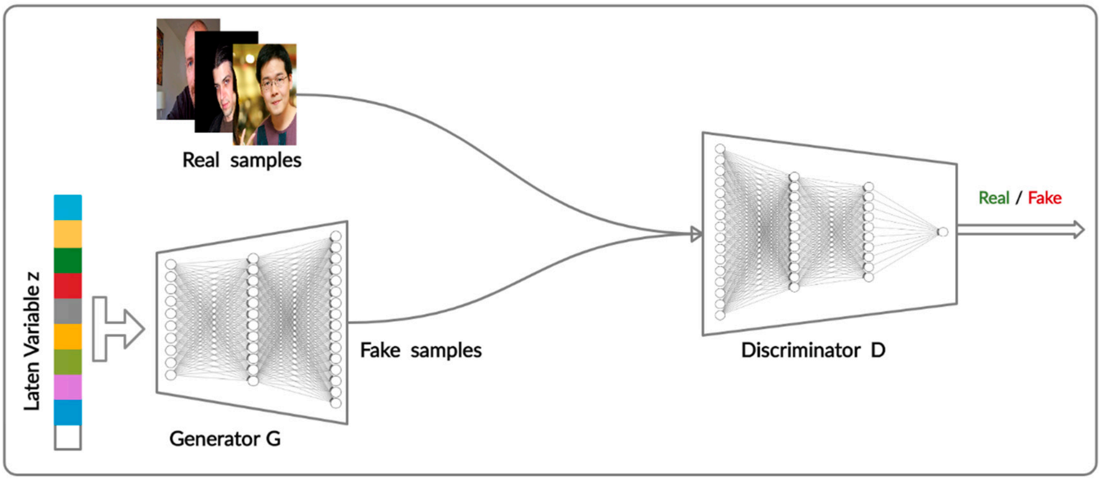

Figure 1. Generative Adversarial Network Architecture.

2.3. Definitions

In this section, notation, abbreviations, and concepts used throughout this survey are explained in order to facilitate the understanding of the topic. Lists of commonly used abbreviations and notations are shown in Tables 1 and A1, respectively. Concepts that are related to both generative adversarial networks and image-to-image translation are explained in what follows [42,44,45].

- **Attribute**: a meaningful feature, such as hair color, gender, size or age.
- **Domain**: a set of images sharing similar attributes.
- **Unimodal image-to-image translation**: a task in which the goal is to learn a one-to-one mapping. Given an input image in the source domain, the model learns to produce a deterministic output.
- **Multimodal image-to-image translation**: aims to learn a one-to-many mapping between the source domain and the target domain with the goal of enabling the model to generate many diverse outputs.
- **Domain-independent features**: those pertaining to the underlying spatial structure, known as the content code.
- **Domain-specific features**: those pertaining to the rendering of the structure, known as the style code.
- **Image generation**: a process of directly generating an image from a random noise vector.
- **Image translation**: a process of generating an image from an existing image and modifying it to have specific attributes.
- **Paired image-to-image translation**: source images X and the corresponding images Y are provided as a training set of aligned image pairs, as shown in Figure 2a,c.
- **Unpaired image-to-image translation**: a source image X and a corresponding image Y are from two different domains, as shown in Figure 2b,d.

Table 1. Commonly Used Notation.

Notation	Explanation
P_{data}	Real Sample
P_g	Fake sample
(z)	Random noise vector
p(x|y)	Conditional probability
P(x,y)	Joint probability
$G\left(z, \theta^{(G)}\right)$	Generator Network
$D\left(x, \theta^{(D)}\right)$	Discriminator Network
D(y)	Discriminator output
$\mathcal{L}_{\mathcal{D}}$	Discriminator loss
$\mathcal{L}_G$	Generator loss

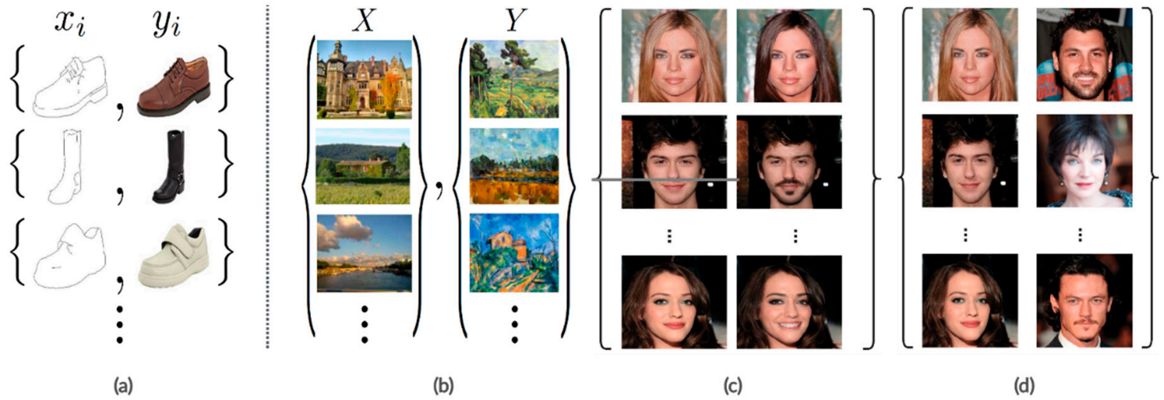

Figure 2. (**a**,**c**) are paired Images and (**b**,**d**) are unpaired Images [41,46].

2.4. Motivation and Contribution

With development in deep learning, numerous approaches have been proposed in order to improve the quality of image synthesis. Most recently, image synthesis with deep generative adversarial networks has attracted many researchers' attention due to their ability to capture the probability distribution as compared to traditional generative models. Several research papers have proposed performing image synthesis while using adversarial networks. Several articles [21,22,34,47–49] have recently surveyed generative adversarial networks and GAN variants, including image synthesis as an application. Other surveys [50,51] covered image synthesis with GANs and partially discussed image-to-image translation. The effort closest to this survey is that of [50], where authors discussed image synthesis, including few image-to-image translation methods.

To the best of our knowledge, image-to-image translation with GANs has never been reviewed previously. Thus, this article provides a comprehensive overview of image-to-image translation using GAN algorithms and variants. A general introduction to generative adversarial networks is given, and GAN variants, structures, and objective functions are demonstrated. The image-to-image translation approaches are discussed in detail, including the state-of-the-art algorithms, theory, applications, and open challenges. The image-to-image translation approaches are classified into supervised and supervised types. The contributions of this article can be summarized, as follows:

1. This review article provides a comprehensive review including general generative adversarial network algorithms, objective function, and structure.
2. Image-to-image translation approaches are classified into supervised and unsupervised types with in-depth explanations.
3. This review article also summarizes the benchmark datasets, evaluation metric, and image-to-image translation applications.

4. Limitations, open challenges, and directions for future research are among the topics discussed, illustrated, and investigated in depth.

This paper's structure can be summarized, as follows. In Section 2, GANs and variants of GAN architectures are demonstrated. GAN objective functions and GAN structure are discussed in Sections 3 and 4, respectively. Section 5 introduces and discusses both supervised and unsupervised image-to-image translation techniques. In Section 6, image-to-image translation applications, including the topics of datasets, practical applications, and evaluation metric, are illustrated and summarized. Discussion and Directions of future research utilizing reinforcement learning and three-dimensional (3D) models are discussed and summarized in Section 7. The last section concludes this review paper.

3. Generative Adversarial Networks' Algorithms

In this section, the basic principle of a typical GAN and general GAN variants that are related to general models and image synthesis are presented.

3.1. Fully Connected GAN

A generative adversarial network (GAN) [32] is a deep generative model that learns to capture the real data distribution using an adversarial process. A GAN typically consists of two components, generator G and discriminator D, as shown in Figure 1. The generator network and the discriminator network are both trained simultaneously in a completive way using the backpropagation algorithm inspired by the min–max game designed to reach a Nash equilibrium. These two networks are implemented as separate deep neural networks, where the goals of the generator are (1) to generate new samples and (2) to deceive the discriminator, and the goal of the discriminator is to estimate the probability to distinguish the real data distribution from a fake data distribution. The generator takes a random noise vector z (following a Gaussian distribution) as input and outputs a generated sample G(z) without any access to real samples. The discriminator takes both a real sample P_{data} and a generated sample P_g as input and predicts the probability of D(x) or D(G(x)) [39,52], as shown in Figure 1.

3.2. Conditional GAN

The GAN models have no control over the generated data, especially in cases of data with more than one labeled class. Therefore, extra information is needed to guide the direction of the distribution to a specific labeled class in order to direct the generated results to more than one labeled class. To this end, a conditional generative adversarial network (CGAN) has been introduced [53] to control the data generation process in a supervised manner. A CGAN combines random noise z and y into a joint hidden representation of real data x, along with conditional variable y; e.g., G (z,y) is used to direct the generated process, where variable y is an additional parameter.

$$min\ maxV(D,G) \ = \ \mathbb{E}_{x\sim Pdata(x)}[log\ D(x\big|y)] + \mathbb{E}_{x\sim Pz(x)}[log\ (1\ -\ D(G(z\big|y)))] \tag{1}$$

Conditional variable y could be text or a number that turns the GAN model into a supervised GAN model. CGAN can be used with images, sequence models, and other models. CGAN is used to model complex and large-scale datasets that have different labels by adding conditional information y to both the generator and discriminator.

3.3. Information GAN

The authors of InfoGAN [54] proposed learning the disentangled representations by maximizing mutual information in an unsupervised manner. In InfoGAN, the input to the generator is decomposed into two parts: the incompressible noise vector G(z) and latent variable c. Similar to CGAN, the generator uses latent code c; however, the latent code c of InfoGAN is unknown and it is to be discovered through training. InfoGAN is implemented by adding a regularization term to the original GAN's objective function.

$$\min \max V_I(D, G) = V(D, G) - \lambda I(c; G(z, c)) \tag{2}$$

where $V_I(D, G)$ is the loss function of GAN, λI (c;G(z,c)) is the mutual information, and Lambda is a constant. InfoGAN maximizes the mutual information between the generator's output G(z,c) and latent code c to discover the meaningful features of the real data distribution. However, mutual information $\lambda I(c; G(z, c))$ requires access to the posterior probability p(c|x), which makes it difficult to directly optimize [22,47,49,55]. Later, other InfoGAN variants were proposed, such as the semi-supervised InfoGAN (ss-InfoGAN) [56] and the causal InfoGAN [57]

3.4. BigGAN

Brock et al. [58] propose BigGAN, a class-conditional GAN that is based on the self-attention GAN (SAGAN) model [43]. BigGAN is trained on ImageNet at the 128 × 128 resolution to generate natural images with high fidelity and variety. The BigGAN model is based on scaling up the GAN models to improve the quality of generated samples by (1) adding orthogonal regularization to the generator, (2) increasing the number of batch size and many parameters, (3) normalizing the generator's weight using spectral normalization, and (4) introducing a truncation trick in order to control the variety of the generated samples.

4. GAN Objective Functions

The goal of the objective function used in GANs and its variants is to measure and minimize the distance between the real sample distribution and generated sample distribution. Although GANs have been successfully used in many tasks, there have been many problems that are caused by objective functions, such as gradient vanishing, and model collapse. These problems cannot be solved by modifying the GANs' structure. Reformulating the objective function has been proven to alleviate these problems. Therefore, many objective functions have been proposed and categorized in order to improve the quality and diversity of the generated sample and avoid the limitations of the original GAN and its variants, as shown in Table 2. Various objective functions are explained and discussed in what follows.

Table 2. Overview of popular generative adversarial network (GAN) objective functions.

Subject	Details	Reference
Objective function	f-divergence	GAN [32], LSGAN [59], f-GAN [60]
	Integral Probability Metric (IMP)	Fisher GAN [61], WGAN [62], McGAN [63], GMMN [64],MMGAN [65]
Autoencoder	Energy function	EBGAN [66], BEGAN [67], MAGAN [68]

Adversarial Loss. The original GAN consists of generator $G\left(z, \theta^{(G)}\right)$ and discriminator $D\left(x, \theta^{(D)}\right)$ competing against each other. GANs utilize the sigmoid cross entropy as a loss function for discriminator D, and use the minmax loss and a non-saturated loss with generator G. D is a differential function whose inputs and parameters are x and $\theta^{(D)}$, respectively, which outputs a single scalar D(y). D(y) represents the probability that input (x) belongs to the real data P_{data} rather than the generated data P_g. Generator G is a differential function whose input and parameters are z and $\theta^{(G)}$, respectively. The discriminator and the generator both have separate loss functions, as shown in Table 3. Both update their parameters to achieve the Nash equilibrium, whereby discriminator $V\left(D, \theta^{(D)}\right)$ aims to maximize the probability of assigning the correct label to both training samples and the sample generated by G [32]. Generator $V\left(D, \theta^{(G)}\right)$ aims to minimize $\log(1 - D(G(z)))$ to deceive D. They are both trained simultaneously and inspired by the min–max game.

Table 3. Loss functions of both discriminators and generators.

Model	Generator Loss	Discriminator Loss
GAN [32]	$\mathcal{L}_{\mathrm{GAN}}\ (\mathrm{G}) = \mathbb{E}_{x \backsim P\ z\ (Z)}[log\ (1 - D(G(z)))]$	$\mathcal{L}_{GAN}\ (\mathrm{D}) = \mathbb{E}_{x \backsim \mathrm{P\ data\ (x)}}[\log\ \mathrm{D(x)}] + \mathbb{E}_{x \backsim \mathrm{P\ z\ (Z)}}[\log\ (1 -\ \mathrm{D(G(z))})]$
LSGAN [59]	$\mathcal{L}_{\mathrm{LSGAN}}\ (\mathrm{G}) = \mathbb{E}_{x \backsim \mathrm{P\ z\ (Z)}}\left[(\mathrm{D(G(z))} -\ \mathrm{c})^2\right]$	$\mathcal{L}_{\mathrm{LSGAN}}\ (\mathrm{D}) = \mathbb{E}_{x \backsim \mathrm{P\ data\ (x)}}\left[(\mathrm{D(x)} -\ \mathrm{b})^2\right] + \mathbb{E}_{x \backsim \mathrm{P\ z\ (Z)}}\left[(\mathrm{D(G(z))} - \mathrm{a})^2\right]$
WGAN [62]	$\mathcal{L}_{\mathrm{WGAN}}\ (\mathrm{G}) = \mathbb{E}_{x \backsim P\ z\ (Z)}[(1 - D(G(z)))]$	$\mathcal{L}_{WGAN}\ (\mathrm{D}) = \mathbb{E}_{x \backsim \mathrm{P\ data\ (x)}}[\mathrm{D(x)}] - \mathbb{E}_{x \backsim \mathrm{P\ z\ (Z)}}[(1 -\ \mathrm{D(G(z))})]$
EBGAN [66]	$\mathcal{L}_{\mathrm{EBGAN}}\ (\mathrm{G}) = \mathbb{E}_{x \backsim P\ z\ (Z)}\ [(1 - D(G(z)))]$	$\mathcal{L}_{\mathrm{EBGAN}}\ (\mathrm{D}) = D(x) + [\ m - D(G(z))]^{+}$
BEGAN [67]	$\mathcal{L}_{BEGAN}\ (\mathrm{G}) = \mathbb{E}_{x \backsim P\ z\ (Z)}[(1 - D(G(z)))]$	$\mathcal{L}_{\mathrm{BEGAN}}\ (\mathrm{D}) = D(x) - k_t D(G(z))$
MAGAN [68]	$\mathcal{L}_{\mathrm{MAGAN}}\ (\mathrm{G}) = \mathbb{E}_{x \backsim P\ z\ (Z)}[(1 - D(G(z)))]$	$\mathcal{L}_{\mathrm{MAGAN}}\ (\mathrm{D}) = D(x) + [\ m - D(G(z))]^{+}$

Wassertien GAN. Arjovsky et al. [62] propose WGAN, using what is sometimes called the Earth Mover's (ME) distance, to overcome the GAN model instability and mode collapse. WGAN uses the Wasserstein distance instead of that of Jensen–Shannan to measure the similarity between the real data distribution and generated data distribution. The Wasserstein distance can be used to measure the distance between probability distributions P_{data} (x) and P_g (x) even if there is no overlap, where (P_{data}, P_g) denotes the set of all joint distributions between the real distribution and the generated distribution [34,47,52]. WGAN applies a weight clipping to enforce the Lipschitz constraint on the discriminator. However, WGAN may suffer from gradient vanishing or exploding due to the use of weight clipping. The discriminator in WGAN is utilized as a regression task to approximate the Wasserstein distance instead of being a binary classifier [49]. It should be noted that WGAN does not change the GAN structure, but instead enhances parameter learning and model optimization [22]. WGAN-GP [69] is a later proposal for stabilizing the training of a GAN by utilizing gradient penalty regularization [70].

Least Squares GAN. LSGAN [59] has been proposed to overcome the vanishing gradient problem that is caused by the minimax loss and the non-saturated loss in the original GAN model. LSGAN adopts the least squares or L2 loss function instead of the sigmoid cross-entropy loss function used in the original GAN. As shown in Table 3, a and b codes are the labels for generated and real samples, respectively, while c represents the value that G wishes D to believe for generated samples. There are two benefits of implementing LSGAN over the original GAN: first, LSGAN can generate high-quality samples; second, LSGAN allows for the learning process to be more stable.

Energy-based GAN. EBGAN [66] has been proposed to model the discriminator as an energy function. This model also uses the autoencoder architecture to first estimate the reconstruction error and, second, to assign a lower energy to the real samples and a higher energy to the generated samples. The output of the discriminator in EBGAN goes through a loss function to shape the energy function. Table 3 shows the loss function, where m is the positive margin, and $[.]^{+} = \max(0, .)$. The EBGAN framework exhibits better convergence and scalability, which result in generating high-resolution images [66].

Margin adaption GAN. MAGAN [68] is an extension of EBGAN that uses the hinge loss function, in which margin m is adapted automatically while using the expected energy of the real data distribution. Hence, margin m is monotonically reduced over time. Unlike EBGAN, MAGAN converges to its global optima, where both real and generated samples' distributions match exactly.

Boundary Equilibrium GAN. BEGAN [67] is an extension of EBGAN that uses an autoencoder as the discriminator. BEGAN's objective function computes the loss based on the Wasserstein distance. Using the proportional control theory, the authors of BEGAN propose an equilibrium method for balancing the generator and the discriminator during training without directly matching the data distributions [47].

5. GAN Structure

The typical GAN is based on a multilayer perceptron (MLP), as mentioned above. Subsequently, structures of various types have been proposed to either solve GANs' issues or address a specific application, as explained in what follows.

Deep convolutional GAN. DCGAN [71] is one of recent major improvements in the field of computer vision and generative modeling. It combines a GAN with a convolutional neural network (CNN). DCGAN has been proposed to stabilize GANs in order to train deep architectures to generate high-quality images. DCGANs have set some architectural constraints to train the generator and discriminator networks for unsupervised representation learning in order to resolve the issues of training instability and the collapse of the GAN architecture. First, DCGANs replace a spatial pooling layer with strided convolution and fractionally-strided convolution to allow for both the generator and the discriminator to learn its own spatial downsampling and spatial upsampling. Second, batch normalization (BN) is used to stabilize learning by normalizing the input and solve the vanishing gradient problem. BN is mainly applied to prevent the deep generator from collapsing all samples to the same points. Third, eliminating the fully connected hidden layers that would otherwise be on the top CNN increases model stability. Finally, DCGANs use both ReLU and LeakyReLU to allow the model to learn quickly and perform well. ReLU activation function is used in all generator layers, except the last layer, which uses the tanh activation function; additionally, LeakyReLU activation functions are used in all discriminator layers.

Self-Attention GAN. SAGAN [72] has been proposed to incorporate a self-attention mechanism into a convolutional GAN framework to improve the quality of generated images. A traditional convolution-based GAN has difficulty in modeling some image classes when trained on large multi-class image datasets due to the local receptive field. SAGAN adapts a self-attention mechanism to different stages of both the generator and the discriminator to model long-range and multi-level dependencies across image regions. SAGAN uses three techniques to stabilize the training of a GAN. First, it applies spectrally to both the generator and discriminator to improve performance and reduce the amount of computations that are performed during training. Second, it uses the Two Time-scale Update Rule (TTUR) for both the generator and discriminator to speed up the training of the regularized discriminator. Third, it integrates the conditional batch normalization layers into the generator and a projection into the discriminator. SAGAN utilizes the hinge loss as the adversarial loss and uses the Adam optimizer to train the model that achieves the state-of-the-art performance on class-condition image synthesis.

Progressive Growing GAN. Karras et al. [73] proposed PGGAN to generate large high-quality images and to stabilize the training process. The PGGAN architecture is based on growing both the generator and the discriminator progressively, starting from small-size images, and then adding new blocks of layers to both the generator and the discriminator. These new blocks of layers are incrementally enlarged in order to discover the large-scale structure and achieve high resolution. Progressive training has three benefits: stabilizing the learning process, increasing the resolution, and reducing the training time.

Laplacian Pyramid GAN. Denton et al. [74] proposed LAPGAN to generate high-quality images by combining a conditional GAN model with a Laplacian pyramid representation. A Laplacian pyramid is a linear invertible image representation that consists of a low-frequency residual based on a Gaussian pyramid. LAPGAN is a cascade of convolutional GANs with k levels of the Laplacian pyramid. The approach of LAPGAN uses multiple generators and discriminator networks and proceeds, as follows: in the beginning, the image is downsampled by a factor of two at each k level of the pyramid. Subsequently, the image is upsampled in a backward pass by a factor of two to reconstruct the image and then return it to its original size while the image acquires noise generated by a conditional GAN at each layer. LAPGAN is trained through unsupervised learning, and each level of a Laplacian pyramid is trained independently and evaluated while using both log-likelihood and human evaluation.

VAE-GAN. Larsen et al. [75] combine a variational autoencoder with a generative adversarial network (VAE-GAN) into an unsupervised generative model and train them jointly to produce high-quality generated images. VAE and GAN are implemented by assigning the VAE decoder to the

GAN generator and combining the VAE's objective function with an adversarial loss [47], where the element-wise reconstruction metric is replaced with a feature-wise reconstruction metric in order to produce sharp images.

6. Image-to-Image Translation Techniques

In this section, image-to-image translation methods are classified into supervised and unsupervised types, as shown in Figure 3. Supervised and unsupervised methods are both discussed in what follows.

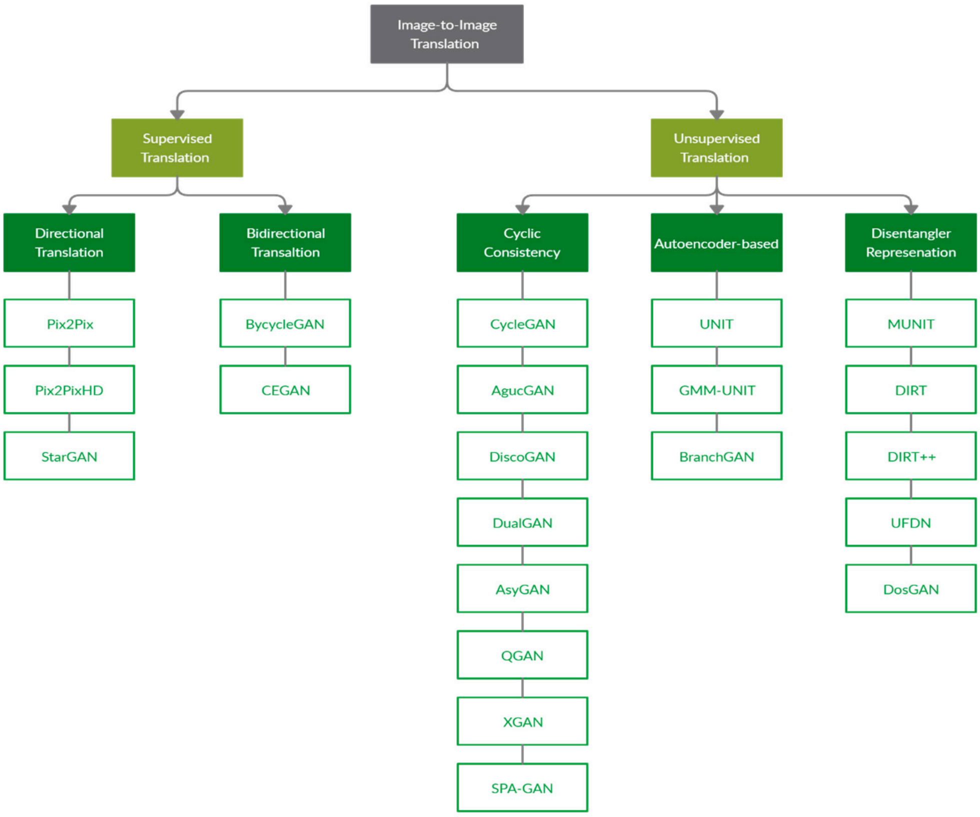

Figure 3. Image-to-image translation taxonomy based on technique.

6.1. Supervised Translation

A supervised method requires a set of pairs of corresponding images (x,y) in different domains (X,Y); for each image x ∈ X and a corresponding image from another domain y € Y, the method learns a probability distribution while using the translator G: X→Y. In some cases, supervised translation utilizes domain images that are conditioned on class labels or source images to generate a high-quality image, as shown in Figure 4a. Supervised translation is further divided into directional and bidirectional translation, as explained in the following sections.

6.1.1. Directional Supervision

Pix2Pix [76] is a supervised image-to-image translation approach that is based on a conditional generative adversarial network. Pix2Pix requires paired images to learn one-to-one mapping and uses two datasets; one dataset is used as input, and the other is used as condition input. An example is translating a semantic label x to a realistic-looking image, and Figure 5 shows an edge-to-photo translation. The generator uses a "U-Net"-based architecture that relies on skip connections to each layer. In contrast, the discriminator uses a convolution-based "PatchGAN" as a classifier. The objective function of Pix2Pix uses cGAN with the L1 norm instead of L2, which leads to less blurring. Although image-to-image translation methods that are based on cGAN such as Pix2Pix enable a variety of translation applications, the generated images are still limited to being of low resolution and blurry. Wang et al. [77] proposed Pix2pixHD to increase the resolution of the output images to 2048*1024 by utilizing a coarse-to-fine generator, an architecture based on three multiscale discriminators, and a robust adversarial learning objective function. It is worth noting that previous studies have been limited to two domains. Thus, StarGAN [44] has been proposed as a unified GAN for multi-domain image-to-image translation using only a single generator and a discriminator. The generator is trained to translate an input image (x) to an output image (y), conditioning on domain label information c, G(x,c)→y. To learn the mapping among k domain k(k-1) between multiple databases, mask vector m is utilized to control domain labels, ignore unknown labels, and focus on a particular label that belongs to a specific dataset. StarGAN requires a single generator and a discriminator to train multiple databases simultaneously by adding an auxiliary classifier on the top of the discriminator to control multiple domains and by applying cycle consistency to the generator [51].

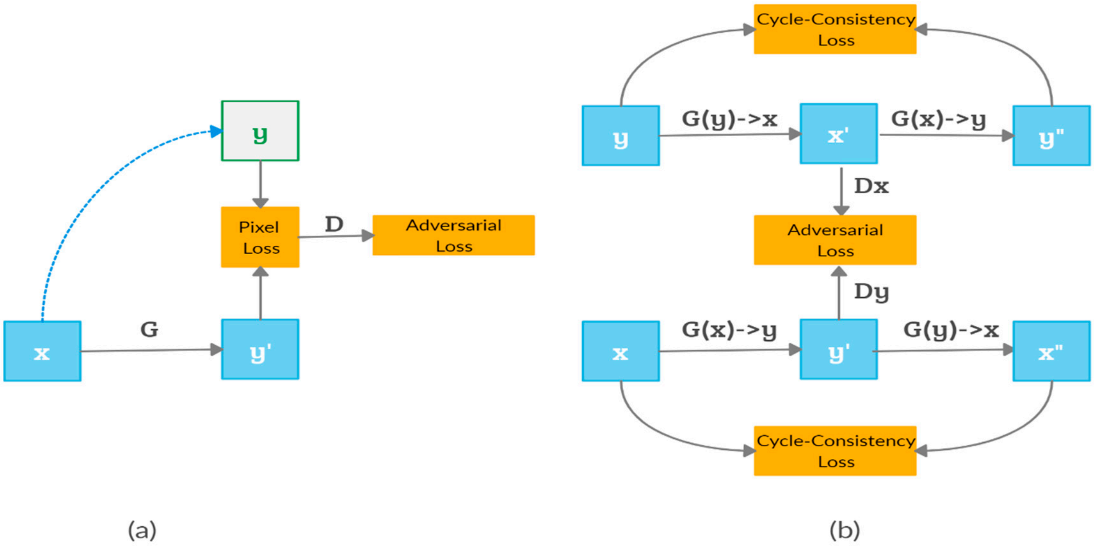

Figure 4. Comparison between supervised and unsupervised image-to-image translation methods. (**a**) Supervised methods, such as Pix2Pix and BicycleGAN. (**b**) Unsupervised methods, such as CycleGAN, DualGAN, and DiscoGAN.

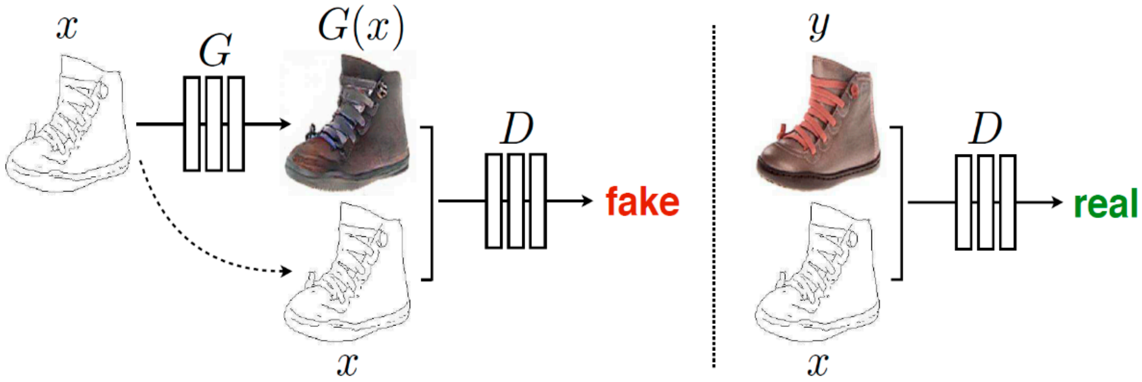

Figure 5. Example of supervised image-to-image translation, edge →photos [76].

6.1.2. Bidirectional Supervision

BicycleGAN [78] is a multi-model cross-domain translation method that requires paired training images. It combines a conditional variational autoencoder (cVAE) with a conditional latent regressor (cLR) to generate realistic and diverse outputs. BicycleGAN enforces a bijective connection between the latent encoding space and output to tackle the mode collapse problem [79]. This model is built based on

LSGAN, which uses a U-Net for the generator and two PatchGANs for the discriminator. BicycleGAN uses the least-squares loss function instead of the cross-entropy loss to stabilize the training process. In addition, CEGAN [80] has been proposed as a novel image-to-image translation approach to learning multi-model mapping that is based on conditional generation models for generating realistic and diverse images. CEGAN captures the distribution of possible multiple modes of results by enforcing tight connections between the latent space and the real image space. The model consists of generator G, discriminator D, and encoder E. In this model, unlike other proposed GAN methods, the discriminator is used to distinguish between real and fake samples in the latent space instead of the real image space, in order to reduce the impact of redundancy and noise and produce realistic-looking images.

6.2. Unsupervised Translation

Unsupervised image-to-image translation aims at learning the mapping between two or more domains without paired images and it has recently been explored intensively due its ability to learn the cross-mapping in image-to-image translation. Many methods have been proposed to perform unsupervised translation and alleviate the issue of limited training samples. These methods are classified into three groups: cyclic consistency-based, autoencoder-based, and that of methods using a disentangled representation, as explained in the following sections.

6.2.1. Unsupervised Translation with Cycle Consistency

In order to overcome the unpaired image-to-image translation problems, cyclic losses are used to preserve key attributes between the input and the translated image using three kinds of losses: adversarial loss, cyclic consistency loss and reconstruction loss. Reconstruction loss is used to regularize the translation to be close to the identity mapping that consists of two mapping functions, where the first learns the forward cycle mapping from the input domain to the target domain, and the second function, namely, a backward function mapping, learns the inverse of the forward mapping [81]. Both of the functions are trained simultaneously while using two loss functions, namely, adversarial losses and cyclic consistency losses, as shown in Figure 5. Zhu et al. [41] propose an unsupervised image-to-image translation method that learns the mapping between a collection of unpaired images from two different domains while using two generators and two discriminators. CycleGAN is a symmetric structure that has two mapping functions G: X→Y and F: Y→X and two adversarial discriminators D_y and D_x, as shown in Figure 4b. Generator G maps the input from domain X to Y, whereas generator F maps input from Y to X. Discriminator D(y) distinguishes y from G(x), whereas discriminator D(x) distinguishes x from F(y). The CycleGAN's objective function contains two terms: adversarial losses and cyclic consistency losses. The adversarial losses are used in order to match the distribution of generated images to that of target images:

$$L_{GAN}(G,D_Y,\ X,Y) = \mathbb{E}_{y \backsim P\ data\ (y)}[log\, D_Y(y)] + \mathbb{E}_{x \backsim P\ data\ (x)}[log\ (1 - D_Y(G(x)))] \tag{3}$$

$$L_{GAN}(F,D_X,\ Y,X) \ = \mathbb{E}_{x \backsim P\ data\ (x)}[log\, D_x(x)] + \mathbb{E}_{y \backsim P\ data\ (y)}[log\ (1 - D_x(G(y)))] \tag{4}$$

The cyclic consistency losses consist of forward and backward cyclic consistency terms, aiming to prevent the learned mappings G and F from contradicting each other:

$$L_{cyc}(G,F) = \mathbb{E}_{x \backsim P\ data\ (x)}\Big[\big\|F(G(x))\ -\ x\big\|1\Big] + \mathbb{E}_{y \backsim P\ data\ (y)}\Big[\big\|G(F(y))\ -\ y\big\|1\Big] \tag{5}$$

The full objective function is

$$L(G,F,Dx,\ Dy)\ =\ L_{GAN}(G,D_Y,\ X,Y)\ + L_{GAN}(F,D_X,\ Y,X)\ +\ \lambda L_{cyc}(G,f) \tag{6}$$

CycleGAN has been successfully applied to several image-to-image translation tasks, including object transfiguration, collection style transfer, season transfer, photo enhancement, etc. However, CycleGAN is only able to learn a one-to-one mapping. A recently proposed augmented cycleGAN [82] introduces an idea that is similar to cycleGAN, but it learns a many-to-many mapping between two domains in a unsupervised setting by adding auxiliary latent codes to each domain.

Similarly, DiscoGAN [83] and DualGAN [84] have been proposed at the same time to tackle the unpaired image-to-image translation problem based on cyclic consistency. However, DiscoGAN learns to discover relations among different domains using the reconstruction losses, while DualGAN is based on dual learning using the Wasserstein GAN. More recently, AsymGAN [85] has been proposed; it uses an asymmetric framework to model unpaired image-to-image translation between asymmetric domains by adding an auxiliary variable (aux). The aux is used to learn the extra information between the poor and rich domains. The difficulties involving complexity between different domains and imbalanced information are resolved and better balanced using this variable. AsymGAN is able to generate diverse and higher-quality outputs by utilizing different aux variables, and it has been proven to generate high-quality and diverse images, unlike CycleGAN and the related approaches that focus on unsupervised pixel-level image-to-image translation. Chen et al. [86] propose a unified quality-aware GAN (QGAN) framework that is based on a quality loss for unpaired image-to-image translation. The QGAN design involves two detailed implementations of the quality-aware loss—a classical quality assessment loss and an adaptive perceptual quality-aware loss—in addition to the adversarial loss. XGAN [43] is a proposed dual adversarial autoencoder for unsupervised image-to-image translation based on a semantic consistency loss. XGAN captures the shared feature representation of both domains to learn the joint feature-level information rather than pixel-level information. A semantic consistency loss is used in both domains' translation to preserve the semantic content of the image across domains. SPA-GAN [87] introduces an effective spatial attention mechanism for unsupervised image-to-image translation tasks. SPA-GAN computes the spatial attention maps in the discriminator and directly feeds it back to the generator. This forces the generator to focus on the most discriminative areas in image-to-image translation. SPA-GAN also introduces a feature map loss term, defined as the difference of feature maps of real and fake images, to encourage the generators to preserve domain-specific features during image translation, which leads to more realistic output images.

6.2.2. Unsupervised Translation with Autoencoder-Based Models

An autoencoder consists of both an encoder and a decoder that converts the input images into a latent representation. This compressed vector is fed to the generator in order to generate high-quality images.

UNIT [88] is a combination framework of CoGAN and VAEs for unsupervised image-to-image translation. It is based on the shared latent space assumption. The weight-share constraint is used to enforce the shared latent space to generate corresponding images in two domains. The UNIT framework consists of two encoders, two generators, and two discriminators. However, UNIT requires the two domains to have similar patterns in order to perform well. [88]. The authors of the coupled generative adversarial network (CoGAN) [89] method proposed using a pair of GANs to learn the joint distribution of multi-domain images in an unsupervised manner. CoGAN uses GAN pairs with identical structure to match the needed number of domains. Each pair of GANs is forced to share the weights of the first few layers and the last few layers, which enables CoGAN to learn joint distributions of multiple domains from samples drawn separately from marginal domain distributions.

More recently, Liu et al. [90] proposed an unsupervised image-to-image translation method that disentangled the representation content from the domain attributes. The attribute latent space was modeled by a Gaussian mixture model (GMM); thus, the model was named GMM-UNIT. In this model, each Gaussian component in a mixture was associated with a domain. There are two main advantages of GMM-UNIT: first, allowing for multi-model and multi-domain translation and, second, allowing for interpolation between domains and extrapolation to unseen domains and translation. Unlike DIRT++, GMM-UNIT entailed a proposal of a continuous domain encoding that allowed generating images with zero- or few-shot generation. Zhou et al. [91] proposed an unsupervised mutual image-to-image translation model, called BranchGAN, based on a single-encoder-dual-decoder architecture for two domains. BranchGAN transfers one image from one domain to another domain by exploiting the shared distribution of the two domains with the same encoder. It uses a reconstruction loss, an encoding loss, and an adversarial loss to train the model to learn the joint distribution of two image domains.

6.2.3. Unsupervised Translation with the Disentangled Representation

Several recent unsupervised image-to-image translation approaches learn the disentangled representation that models the factors of data variations by a content encoder and a style encoder [13]. Huang et al. [42] proposed a multimodal unsupervised image-to-image translation (MUNIT) framework that assumes that the latent space of images can be decomposed into two: a content space and a style space. MUNIT consists of two autoencoders, and the latent code of each autoencoder is factorized by content code and style code. The content code encodes the underlying spatial structure that should be preserved during translation, while style code represents the rendering of the structure that is not contained in the input image. In order to translate an image to a target domain and produce diverse and multimodal outputs, its content code is recombined with a different style code sampled from the style space in the target style space. Lee et al. [79] proposed the diverse image-to-image translation (DIRT) method that is based on a disentangled representation framework for producing diverse outputs without paired training images. DIRT decomposes the latent space into a shared content space and domain-specific attribute spaces. Weight sharing and a content discriminator strategy are both used to disentangle the content and attributes' representations by applying a novel cross-cycle consistency loss and a content adversarial loss. Later, DIRT was extended to DRIT++ by implementing two steps: incorporating a mode-seeking regularization term to alleviate the mode collapse problem, which helps to improve sample diversity and, second, generalizing the two-domain model to handle multi-domain image-to-image translation problems [90].

Liu et al. [92] proposed a compact model, called the Unified Feature Disentanglement Network (UFDN), which learns a domain-invariant representation from data across multiple domains and it is able to perform continuous cross-domain image translation and manipulation.

Lin et al. [93] proposed an unpaired image-to-image translation framework, called domain-supervised GAN (DosGAN). It consists of a domain-specific feature extractor, a domain-independent feature extractor, and an image generator for extracting better domain-specific features and translating images. It treats domain information as explicit supervision to capture each domain's specific characteristic for unconditional or conditional image-to-image translation.

7. Image-to-Image Translation Applications

Image-to-image translation techniques have been successfully applied to a wide variety of real-world applications, as described below. This section summarizes three significant aspects in the respective subsections: benchmark datasets, evaluation metrics, and practical applications.

7.1. Datasets

There are several benchmark datasets that are available for image synthesis tasks that can be utilized to perform image-to-image translation tasks. Such datasets differ in image counts, quality, resolution, complexity, and diversity, and they allow researchers to investigate a variety of practical applications such as facial attributes, cartoon faces, semantic applications, and urban scene analysis. Table 4 summarizes the selected benchmark datasets.

Table 4. List of public image datasets for image-to-image translation benchmarks.

Dataset	Source	Year	Total	Classes	Application	Citations
CelebA(CelebFaces)	[94]	2015	202,599	10177	Facial attributes	[44,79,82,83,85, 88–91,93,95,96]
RaFD	[97]	2010	8040	67	Facial expressions	[35,44,70]
CMP Facades	[98]	2013	606	12	Façade images	[35,76,80,85,91, 99]
Facescrub	[100]	2014	106,863	153	Faces	[87,93]
Cityscapes	[101]	2016	70,000	30	Semantic	[41,76,77,80,85, 87,88,91,95,99]
Helen Face	[102]	2012	2330	-	Face Parsing	[77,85]
CartoonSet	[43]	2018	10,000	-	Cartoon Faces	[43]
ImageNet	[103]	2009	3.2 m	12 subtrees	Diverse	[76,87]

7.2. Evaluation Metrics

To quantitatively and qualitatively evaluate the performance of image-to-image translation, several evaluation metrics that are related to such translation are reviewed and discussed in what follows.

- The inception score (IS) [104] is an automated metric for evaluating the visual quality of generated images by computing the KL divergence between the conditional class distribution and the marginal class distribution via inception networks. IS aims to measure the image quality and diversity. However, the IS metric has two limitations: (1) a high sensitivity to small changes and (2) a large variance of scores [105].
- The Amazon Mechanical Turk (AMT) is used to measure the realism and faithfulness of the translated images that are based on human perception. Workers ("turkers") are given an input image and translated images and are instructed to choose or score the best image based on quality and perceptual realism. The number of validated turkers varies by experiment.
- The Frechet inception distance (FID) is used to construct the FID score [106] that is used to evaluate the quality of the generated images and measure the similarity between two different datasets [80]. It is used to measure the distance between the generated images' distribution and the real image distribution by computing the Frechet inception distance using the inception network. FID very accurately captures the distribution and it is considered to be more consistent than IS with noise level. Lower FID values indicate better quality of the generated images' sample [107].
- The kernel inception distance (KID) [108] is an improved measure of GAN convergence that has a simple unbiased estimator with no unnecessary assumptions regarding the form of the activations' distribution. KID involves a computation of the squared maximum mean discrepancy between representations of reference and generated distributions [87]. A lower KID score signifies better visual quality of generated images
- The learned perceptual image patch similarity (LPIPS) distance [109] measures the image translation diversity by computing the average feature distance between the generated images. LPIPS is defined as a weighed L2 distance between deep features of two images. A higher LPIPS value indicates greater diversity among the generated images.
- Fully Convolutional Networks (FCN) [110] can be used to compute the FCN-score that uses the FCN model as a performance metric in order to evaluate the image quality by segmenting the generated image and comparing it with the ground truth label using a well-trained segmentation FCN model. A smaller value of the FCN-score between the generated image and ground truth means better performance. The FCN-score is calculated based on three parts: per-pixel accuracy, per-class accuracy, and class intersection-over-union (IOU).

7.3. Practical Applications

Many computer vision and graphics problems can be regarded as image-to-image translation problems. Transferring an image from one scene to another can be viewed as cross-domain image-to-image translation, whereas a one-to-many translation is called multimodal image-to-image translation. There are many image-to-image translation applications, such as transferring a summer scene to a winter scene. In this section, only four widely known applications are covered: super-resolution, style transfer, object transfiguration, and medical imaging.

7.3.1. Super-Resolution

Super-resolution (SR) refers to the process of translating a low-resolution source image to a high-resolution target image. GANs have recently been used to solve super-resolution problems in an end-to-end manner [111–114] by treating the generator as an SR model to output a high-resolution image and using the discriminator as a binary classifier. SRGAN [115] adds a new distributional loss term in order to generate an upsampled image with the resolution increased fourfold and it is based on the DCGAN architecture with a residual block. ERSGAN [15] has been proposed to improve the overall perceptual quality of SR results by introducing the residual-in-residual dense block (RRDB) without batch normalization in order to further enhance the quality of generated images.

7.3.2. Style Transfer

Style transfer is the process of rendering the content of an image with a specific style while preserving the content, as shown in Figure 6. The earlier style transfer models could only generate one image and transfer it according to one style. However, recent studies attempt to transfer image content according to multiple styles that are based on a perceptual loss. In addition, with advancement in deep generative models, adversarial losses can also be used to train the style model to make the output image indistinguishable from images in the targeted style domain. Style transfer is a practical application of image-to-image translation. Chen et al. [116] propose an adversarial gated networks, called Gated-GAN, to transfer multiple styles while using a single model based on three modalities: an encoder, a gated transformer, and a decoder. GANILLA [117] is a proposed novel framework with the ability to better balance between content and style.

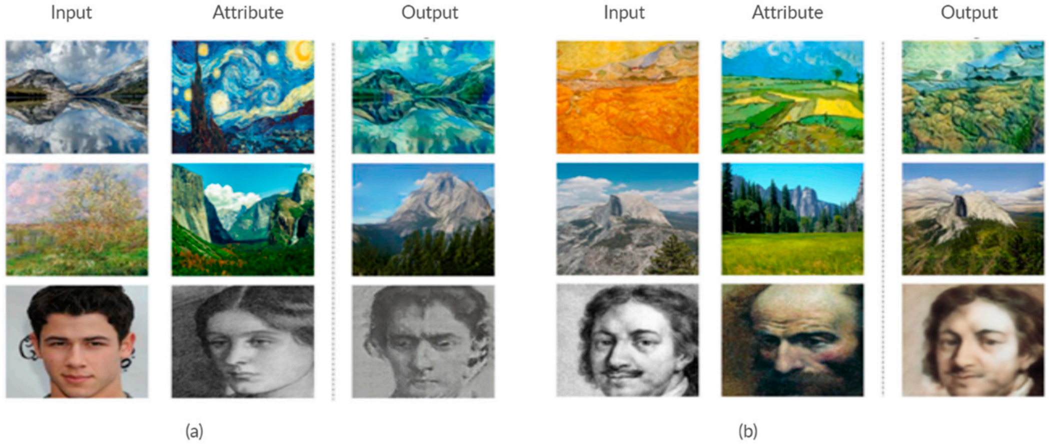

Figure 6. Style transfer applications with (**a**) inter-domain attribute transfer and (**b**) intra-domain attribute transfer [95].

7.3.3. Object Transfiguration

Object transfiguration aims to detect the object of interest in an image and then transform it into another object in the target domain while preserving the background regions, e.g., transforming an apple into an orange, or a horse into a zebra. Using GANs has been explored in object transfiguration to perform two tasks: (1) to detect the object of interest in an image and (2) to transform the object into a target domain. Attention GANs [118,119] are mostly used for object transfiguration; such a model consists of an attention network, a transformation network, and a discriminative network.

7.3.4. Medical Imaging

GANs have recently been used in medical imaging for two purposes: first, for discovering the underlying structure of the training data to generate new samples in order to overcome the privacy constraints and the lack or limited quantity of available positive cases, and, second, for detecting abnormal images through the discriminator [120]. MedGAN [121] is a medical image-to-image translation framework that is used to enhance further technical post-processing tasks that require globally consistent image properties. This framework utilizes style transfer losses to match the translated output with the desired target image according to style, texture, and content, as shown in Figure 7.

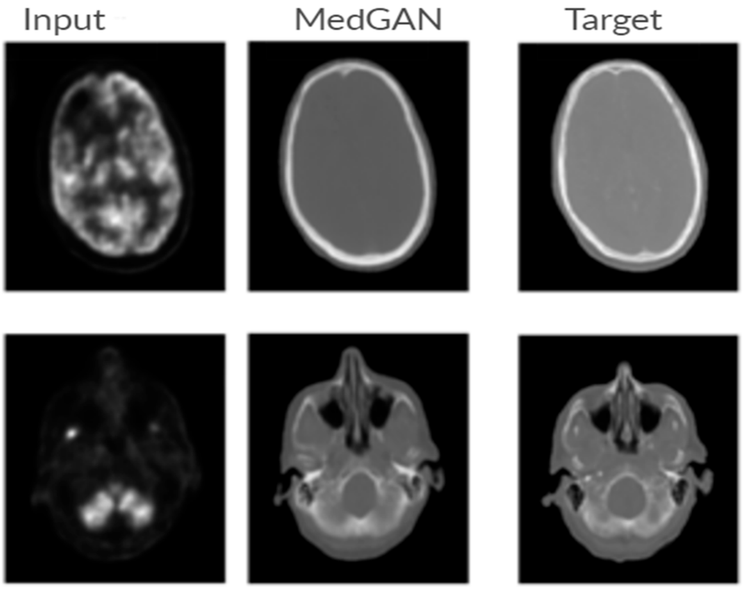

Figure 7. Medical image-to-image translation application using MedGAN [121].

8. Discussion and Directions for Future Research

Table 5 shows an overview of recent studies' structure and mechanisms to further investigate and compare the state-of-the-art image-to-image translation methods. There are two ways of performing image-to-image translation, namely, by supervised or unsupervised methods using either paired or unpaired datasets, as mentioned in the above classification. Although paired datasets have been used to obtain higher-quality outputs by conditional generative network models, it is expensive and difficult to collect such data for certain domains, or sometimes such data may not exist. Therefore, many recent studies [41,83,84,88,89] propose methods for unsupervised image-to-image translation, using unpaired images. However, all of these approaches are only capable of learning deterministic translation of one-to-one mapping: e.g., each translation model associates a single input image with a single output image. Therefore, modeling more relationships across domains is more complex, and it is difficult to learn the underlying distribution between two different domains, known as the many-to-many mapping. The approaches of MUNIT and DIRT enable multi-modal translation by decomposing the latent code into a domain-invariant content space and a domain-specific style space; this leads to mapping ambiguity, since there is more than one proper output and many degrees of freedom in changing the outputs. Many proposed methods have attempted to overcome the limitations of image-to-image translation; however, there are three main open challenges that have not been fully addressed due to several reasons, as investigated and explained in the following section.

Table 5. Comparison of the state-of-the-art methods of image-to-image translation.

Method	Unpaired Images	Multi-Domain Translation	Multi-Modal Translation	Unified Structure	Bidirectional Translation	Shared Representation	Feature Disentanglement
Pix2Pix [76]	-	-	-	-	-	-	-
BicycleGAN [78]	-	-	✓	-	✓	✓	-
StarGAN [44]	✓	✓	-	✓	✓	-	-
CycleGAN [41]	✓	-	-	-	✓	-	-
UNIT [88]	✓	-	-	-	✓	✓	-
GMM-UNIT [90]	✓	✓	✓	✓	✓	-	✓
MUNIT [42]	✓	-	✓	-	✓	✓	✓
DIRT [79]	✓	-	✓	-	✓	✓	✓
DIRT++ [95]	✓	✓	✓		✓	✓	✓
UFDN [92]	✓	✓	-	✓	✓	✓	✓

8.1. Open Challenges

- Mode Collapse

Image-to-image translation with GANs and GAN variants usually suffers from the mode collapse issue that occurs when the generator only generates the same output, regardless of whether it uses a single input or operates on multiple modes, e.g., when two images I_1 = G(c, z_1) and I_2 = G(c, z_2) are likely to be mapped to the same model [107]. There are two types of mode collapse: inter-mode and intra-mode collapse. Inter-mode collapse occurs when the expected output is known, e.g., if digits (0–9) are used and the generator keeps generating the same number to fool the discriminator. In contrast, intra-mode collapse usually happens if the generator only learns one style of the expected output to fool the discriminator. Many proposals have recently been made to alleviate and avoid mode collapse; the sample approaches include LSGAN [59], using a mode-seeking regularization term [107], and cycle consistency [84,95]. However, the mode collapse problem still has not been completely solved and it is considered to be one of the open issues of image-to-image translation tasks.

- Lack of evaluation metrics

As mentioned above, several evaluation methods have been proposed [104,106,108,109] to measure and assess the quality of the translated images and investigate the strengths and limitations of the used models. These evaluation measures can be categorized into quantitative and qualitative. They have been further explored in [122], where the difference between both of the techniques has been investigated in depth. Metrics of success of image-to-image translation usually evaluate the quality of generated images while using a limited number of test images or user studies. The evaluation of a limited number of test images must consider both style and content simultaneously, which is difficult to do [117]. In addition, user studies are based on human judgment, which is a subjective metric [1]. However, there is no well-defined evaluation metric, and it is still difficult to accurately assess the quality of generated images, since there is no strict one-to-one correspondence between the translated image and the input image [1].

- Lack of diversity

Image-to-image translation diversity is related to the quality of diverse generated outputs utilizing multi-modal and multi-domain mapping, as mentioned above. Several approaches [42,78,79] injected a random noise vector into the generator in order to model a diverse distribution in the target domain. One of the existing limitations of image-to-image translation is the lack of diversity of generated images due to the lack of regularization between the random noise and the target domain [79]. The DIRT [79] and DIRT++ [95] methods have been proposed to improve the diversity of generated images; however, generating diverse outputs for producing high quality and diverse images has not yet been fully explored.

8.2. Directions of Future Research

Deep reinforcement learning (LR) has recently drawn significant attention in the field of deep learning. Deep LR has been applied to a variety of image processing and computer vision tasks: image super-resolution, image denoising, and image restoration [123–125]. A deep reinforcement learning framework can play a role in a GAN to generate high-quality output, whereby the generator can be utilized as an agent and the discriminator's results as the reward signal. The generator should be trained as an agent and rewarded every time that it fools the discriminator, whereas the discriminator's training process should be the same as proposed in general GANs to distinguish the generated distribution from the real distribution. Moreover, reconstructing a 3D shape from either a two-dimensional (2D) object or sketches has been a challenging task due to the lack of a proposed and explored 3D GAN in image-to-image translation. In addition, there is a lack of available 3D datasets. Exploring 3D GANs and volumetric convolutional networks can play a very important role in the future in generating 3D Images. Furthermore, cybersecurity applications should utilize image-to-image translation with GAN to design reliable and efficient systems. The image steganography that is based on GAN should be further investigated and developed in order to overcome critical cybersecurity issues and challenges.

9. Conclusions

Image-to-image translation with GANs has made huge success in computer vision applications. This article presents a comprehensive overview of GAN variants that are related to image-to-image translation based on algorithms, the objective function and structure. Recent state-of-the-art image-to-image translation techniques, both supervised and unsupervised, are surveyed and classified. In addition, benchmark datasets, evaluation metrics and practical applications are summarized. This review paper covers open issues that are related to mode collapse, evaluation metrics, and lack of diversity. Finally, reinforcement learning and 3D models have not been fully explored and are suggested as future directions towards better performance on image-to-image translation tasks. In the future, quantum generative adversarial network for image-to-image translation will be further explored and implemented in order to overcome complex problems related to image generation.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Abbreviated names of GAN models.

GAN Model	Full Name	Publication Year	Authors
GAN	Generative Adversarial Network	2014	Goodfellow et al. [32]
CGAN	Conditional GAN	2014	Mirza, M. & Osindero, S. [53]
LAPGAN	Laplacian Pyramid GAN	2015	Denton et al. [74]
DCGAN	Deep convolutional GAN	2016	Radford et al. [71]
InfoGAN	Information-Maximizing GAN	2016	Chen et al. [54]
CoGAN	Coupled GAN	2016	Liu et al. [89]
VAE-GAN	Variational encoder-decoder GAN	2016	Larsen et al. [75]
WGAN	Wasserstein GAN	2017	Arjovsky et al. [62]
WGAN-PG	Wasserstein GAN with a Gradient Penalty	2017	Gulrajani et al. [69]
BEGAN	Boundary Equilibrium GAN	2017	Berthelot et al. [67]
EBGAN	Energy-Based GAN	2017	Zhao et al. [66]
MAGAN	Margin Adaption GAN	2017	Wang et al. [68]
CycleGAN	Cycle-Consistent GAN	2017	Zhu et al. [41]
MMDGAN	Maximum Mean Discrepancy GAN	2017	Li et al. [65]
DiscoGAN	Discover Cross-Domain GAN	2017	Kim et al. [83]
LSGAN	Least-Squares GAN	2017	Mao et al. [59]
ACGAN	Auxiliary Classifier GAN	2017	Odena et al. [126]
Pix2Pix	Pixel-to-Pixel	2017	Isola et al. [76]
DualGAN	Dual Learning GAN	2017	Yi et al. [84]
UNIT	Unsupervised Image-to-Image Translation	2017	Lui et al. [88]
SRGAN	Super-Resolution GAN	2017	Leding et al. [115]

Table A1. *Cont.*

GAN Model	Full Name	Publication Year	Authors
PROGAN	Progressive Growing GAN	2018	Karras et al. [73]
Pix2PixHD	Pixel-to-Pixel High-Resolution	2018	Wang et al. [77]
MUNIT	Multimodal Unsupervised Image-to-Image Translation	2018	Huang et al. [42]
DRIT	Diverse Image-to-Image Translation	2018	Lee at al. [79]
UFDN	Unified Feature Disentangler	2018	Liu et al. [92]
AguGAN	Augmented Cycle GAN	2018	Almahairi et al. [82]
BigGAN	Large-Scale (Big) GAN	2019	Brock et al. [58]
SAGAN	Self-Attention GAN	2019	Zhang et al. [72]
CEGAN	Consistent Embedded GAN	2019	Xiong et al. [80]
MSGAN	Mode Seek GAN	2019	Mao et al. [107]
QGAN	Quality-Aware GAN	2019	Chen et al. [86]
DRIT++	Diverse Image-to-Image Translation	2019	Lee et al. [95]
AsyGAN	Asymmetric GAN	2019	Li et al. [85]
RelGAN	Relative Attributes GAN	2019	Wu et al. [96]
Gated-GAN	Adversarial Gated Network	2019	Chen et al. [116]
DOSGAN	Domain-Supervised GAN	2019	Lin et al. [93]
SPA-GAN	Spatial Attention GAN	2020	Emami et al. [87]
GMM-UNIT	Gaussian Mixture Modeling UNIT	2020	Liu et al. [90]
GANILLA	GAN for Image-to-Illustration Translation	2020	Hicsonmez et al. [117]
XGAN	Cross GAN	2020	Royer et al. [43]

References

1. Huang, H.; Yu, P.S.; Wang, C. An introduction to image synthesis with generative adversarial nets. *arXiv* **2018**, arXiv:1803.04469.
2. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
5. Suzuki, K. Overview of deep learning in medical imaging. *Radiol. Phys. Technol.* **2017**, *10*, 257–273. [CrossRef]
6. Zhao, D.; Zhu, D.; Lu, J.; Luo, Y.; Zhang, G. Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16. *Symmetry* **2018**, *10*, 519.
7. Ma, B.; Ban, X.; Huang, H.; Chen, Y.; Liu, W.; Zhi, Y. Deep learning-based image segmentation for Al-La alloy microscopic images. *Symmetry* **2018**, *10*, 107. [CrossRef]
8. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
9. Alotaibi, A.; Mahmood, A. Deep face liveness detection based on nonlinear diffusion using convolution neural network. *SignalImage Video Process* **2017**, *11*, 713–720. [CrossRef]
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
11. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
13. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
14. Ng, A.Y.; Jordan, M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 841–848.
15. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 63–79.
16. Li, C.; Wang, L.; Cheng, S.; Ao, N. Generative Adversarial Network-Based Super-Resolution Considering Quantitative and Perceptual Quality. *Symmetry* **2020**, *12*, 449. [CrossRef]
17. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. *arXiv* **2016**, arXiv:1605.05396.
18. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
19. Bhattacharjee, D.; Kim, S.; Vizier, G.; Salzmann, M. DUNIT: Detection-Based Unsupervised Image-to-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4787–4796.
20. Venkateswara, H.; Chakraborty, S.; Panchanathan, S. Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations. *IEEE Signal Process. Mag.* **2017**, *34*, 117–129. [CrossRef]
21. Cao, Y.-J.; Jia, L.-L.; Chen, Y.-X.; Lin, N.; Yang, C.; Zhang, B.; Liu, Z.; Li, X.-X.; Dai, H.-H. Recent advances of generative adversarial networks in computer vision. *IEEE Access* **2018**, *7*, 14985–15006. [CrossRef]
22. Wang, K.; Gou, C.; Duan, Y.; Lin, Y.; Zheng, X.; Wang, F.-Y. Generative adversarial networks: Introduction and outlook. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 588–598. [CrossRef]
23. Rasmussen, C.E. The infinite Gaussian mixture model. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 6 December 2020; pp. 554–560.
24. Jiang, L.; Zhang, H.; Cai, Z. A novel Bayes model: Hidden naive Bayes. *IEEE Trans. Knowl. Data Eng.* **2008**, *21*, 1361–1371. [CrossRef]
25. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
26. Maaløe, L.; Sønderby, C.K.; Sønderby, S.K.; Winther, O. Auxiliary deep generative models. *arXiv* **2016**, arXiv:1602.05473.
27. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.-L.; Chen, S.-C.; Iyengar, S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–36. [CrossRef]
28. Oussidi, A.; Elhassouny, A. Deep generative models: Survey. In Proceedings of the 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2–4 April 2018; pp. 1–8.
29. Salakhutdinov, R.; Hinton, G. Deep boltzmann machines. In Proceedings of the Artificial Intelligence and Statistics, Clearwater, FL, USA, 16–19 April 2009; pp. 448–455.
30. Hinton, G.E. Deep belief networks. *Scholarpedia* **2009**, *4*, 5947. [CrossRef]
31. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
33. Abbasnejad, M.E.; Shi, Q.; van den Hengel, A.; Liu, L. A generative adversarial density estimator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 10782–10791.
34. Wang, Z.; She, Q.; Ward, T.E. Generative adversarial networks in computer vision: A survey and taxonomy. *arXiv* **2019**, arXiv:1906.01529.

35. Tang, H.; Xu, D.; Liu, H.; Sebe, N. Asymmetric Generative Adversarial Networks for Image-to-Image Translation. *arXiv* **2019**, arXiv:1912.06931.
36. Regmi, K.; Borji, A. Cross-view image synthesis using conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 3501–3510.
37. Lin, C.-H.; Yumer, E.; Wang, O.; Shechtman, E.; Lucey, S. St-gan: Spatial transformer generative adversarial networks for image compositing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 9455–9464.
38. Mo, S.; Cho, M.; Shin, J. Instagan: Instance-aware image-to-image translation. *arXiv* **2018**, arXiv:1812.10889.
39. Kahng, M.; Thorat, N.; Chau, D.H.P.; Viégas, F.B.; Wattenberg, M. Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 1–11. [CrossRef]
40. Hertzmann, A.; Jacobs, C.E.; Oliver, N.; Curless, B.; Salesin, D.H. Image analogies. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 12–17 August 2001; pp. 327–340.
41. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
42. Huang, X.; Liu, M.-Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
43. Royer, A.; Bousmalis, K.; Gouws, S.; Bertsch, F.; Mosseri, I.; Cole, F.; Murphy, K. Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 33–49.
44. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 8789–8797.
45. Zhao, B.; Chang, B.; Jie, Z.; Sigal, L. Modular generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 150–165.
46. Tao, R.; Li, Z.; Tao, R.; Li, B. ResAttr-GAN: Unpaired Deep Residual Attributes Learning for Multi-Domain Face Image Translation. *IEEE Access* **2019**, *7*, 132594–132608. [CrossRef]
47. Hong, Y.; Hwang, U.; Yoo, J.; Yoon, S. How generative adversarial networks and their variants work: An overview. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–43. [CrossRef]
48. Pan, Z.; Yu, W.; Yi, X.; Khan, A.; Yuan, F.; Zheng, Y. Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access* **2019**, *7*, 36322–36333. [CrossRef]
49. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv* **2020**, arXiv:2001.06937.
50. Wang, L.; Chen, W.; Yang, W.; Bi, F.; Yu, F.R. A State-of-the-Art Review on Image Synthesis with Generative Adversarial Networks. *IEEE Access* **2020**, *8*, 63514–63537. [CrossRef]
51. Wu, X.; Xu, K.; Hall, P. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Sci. Technol.* **2017**, *22*, 660–674. [CrossRef]
52. Gonog, L.; Zhou, Y. A review: Generative adversarial networks. In Proceedings of the 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 19–21 June 2019; pp. 505–510.
53. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
54. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2172–2180.
55. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
56. Spurr, A.; Aksan, E.; Hilliges, O. Guiding infogan with semi-supervision. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Skopje, Macedonia, 18–22 September 2017; pp. 119–134.

57. Kurutach, T.; Tamar, A.; Yang, G.; Russell, S.J.; Abbeel, P. Learning plannable representations with causal infogan. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 8733–8744.
58. Brock, A.; Donahue, J.; Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv* **2018**, arXiv:1809.11096.
59. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
60. Nowozin, S.; Cseke, B.; Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 271–279.
61. Mroueh, Y.; Sercu, T. Fisher gan. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2513–2523.
62. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv* **2017**, arXiv:1701.07875.
63. Mroueh, Y.; Sercu, T.; Goel, V. Mcgan: Mean and covariance feature matching gan. *arXiv* **2017**, arXiv:1702.08398.
64. Li, Y.; Swersky, K.; Zemel, R. Generative moment matching networks. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1718–1727.
65. Li, C.-L.; Chang, W.-C.; Cheng, Y.; Yang, Y.; Póczos, B. Mmd gan: Towards deeper understanding of moment matching network. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2203–2213.
66. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. *arXiv* **2016**, arXiv:1609.03126.
67. Berthelot, D.; Schumm, T.; Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv* **2017**, arXiv:1703.10717.
68. Wang, R.; Cully, A.; Chang, H.J.; Demiris, Y. Magan: Margin adaptation for generative adversarial networks. *arXiv* **2017**, arXiv:1704.03817.
69. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5767–5777.
70. Pan, Z.; Yu, W.; Wang, B.; Xie, H.; Sheng, V.S.; Lei, J.; Kwong, S. Loss Functions of Generative Adversarial Networks (GANs): Opportunities and Challenges. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *4*, 500–522. [CrossRef]
71. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
72. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 7354–7363.
73. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
74. Denton, E.L.; Chintala, S.; Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 1486–1494.
75. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1558–1566.
76. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
77. Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 8798–8807.
78. Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 465–476.

79. Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; Yang, M.-H. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–51.
80. Xiong, F.; Wang, Q.; Gao, Q. Consistent Embedded GAN for Image-to-Image Translation. *IEEE Access* **2019**, *7*, 126651–126661. [CrossRef]
81. Tripathy, S.; Kannala, J.; Rahtu, E. Learning image-to-image translation using paired and unpaired training samples. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 51–66.
82. Almahairi, A.; Rajeswar, S.; Sordoni, A.; Bachman, P.; Courville, A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv* **2018**, arXiv:1802.10151.
83. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. *arXiv* **2017**, arXiv:1703.05192.
84. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2849–2857.
85. Li, Y.; Tang, S.; Zhang, R.; Zhang, Y.; Li, J.; Yan, S. Asymmetric GAN for unpaired image-to-image translation. *IEEE Trans. Image Process.* **2019**, *28*, 5881–5896. [CrossRef]
86. Chen, L.; Wu, L.; Hu, Z.; Wang, M. Quality-aware unpaired image-to-image translation. *IEEE Trans. Multimed.* **2019**, *21*, 2664–2674. [CrossRef]
87. Emami, H.; Aliabadi, M.M.; Dong, M.; Chinnam, R. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Trans. Multimed.* **2020**. [CrossRef]
88. Liu, M.-Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 700–708.
89. Liu, M.-Y.; Tuzel, O. Coupled generative adversarial networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 469–477.
90. Liu, Y.; De Nadai, M.; Yao, J.; Sebe, N.; Lepri, B.; Alameda-Pineda, X. GMM-UNIT: Unsupervised Multi-Domain and Multi-Modal Image-to-Image Translation via Attribute Gaussian Mixture Modeling. *arXiv* **2020**, arXiv:2003.06788.
91. Zhou, Y.-F.; Jiang, R.-H.; Wu, X.; He, J.-Y.; Weng, S.; Peng, Q. Branchgan: Unsupervised mutual image-to-image transfer with a single encoder and dual decoders. *IEEE Trans. Multimed.* **2019**, *21*, 3136–3149. [CrossRef]
92. Liu, A.H.; Liu, Y.-C.; Yeh, Y.-Y.; Wang, Y.-C.F. A unified feature disentangler for multi-domain image translation and manipulation. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 2590–2599.
93. Lin, J.; Chen, Z.; Xia, Y.; Liu, S.; Qin, T.; Luo, J. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef]
94. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
95. Lee, H.-Y.; Tseng, H.-Y.; Mao, Q.; Huang, J.-B.; Lu, Y.-D.; Singh, M.; Yang, M.-H. Drit++: Diverse image-to-image translation via disentangled representations. *Int. J. Comput. Vis.* **2020**, 1–16. [CrossRef]
96. Wu, P.-W.; Lin, Y.-J.; Chang, C.-H.; Chang, E.Y.; Liao, S.-W. Relgan: Multi-domain image-to-image translation via relative attributes. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5914–5922.
97. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
98. Tyleček, R.; Šára, R. Spatial pattern templates for recognition of objects with regular structure. In Proceedings of the German Conference on Pattern Recognition, Saarbrücken, Germany, 3–6 September 2013; pp. 364–374.
99. Shen, Z.; Huang, M.; Shi, J.; Xue, X.; Huang, T.S. Towards instance-level image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 3683–3692.
100. Ng, H.-W.; Winkler, S. A data-driven approach to cleaning large face datasets. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 343–347.

101. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
102. Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In Proceedings of the European Conference on Computer Vision, Providence, RI, USA, 16–21 June 2012; pp. 679–692.
103. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
104. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.
105. Shmelkov, K.; Schmid, C.; Alahari, K. How good is my GAN? In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 213–229.
106. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
107. Mao, Q.; Lee, H.-Y.; Tseng, H.-Y.; Ma, S.; Yang, M.-H. Mode seeking generative adversarial networks for diverse image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 1429–1437.
108. Bińkowski, M.; Sutherland, D.J.; Arbel, M.; Gretton, A. Demystifying mmd gans. *arXiv* **2018**, arXiv:1801.01401.
109. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 586–595.
110. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
111. Huang, H.; He, R.; Sun, Z.; Tan, T. Wavelet domain generative adversarial network for multi-scale face hallucination. *Int. J. Comput. Vis.* **2019**, *127*, 763–784. [CrossRef]
112. Zhang, W.; Liu, Y.; Dong, C.; Qiao, Y. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3096–3105.
113. Wang, Y.; Perazzi, F.; McWilliams, B.; Sorkine-Hornung, A.; Sorkine-Hornung, O.; Schroers, C. A fully progressive approach to single-image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 864–873.
114. Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 701–710.
115. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
116. Chen, X.; Xu, C.; Yang, X.; Song, L.; Tao, D. Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Trans. Image Process.* **2018**, *28*, 546–560. [CrossRef] [PubMed]
117. Hicsonmez, S.; Samet, N.; Akbas, E.; Duygulu, P. GANILLA: Generative adversarial networks for image to illustration translation. *Image Vis. Comput.* **2020**, *95*, 103886. [CrossRef]
118. Ma, S.; Fu, J.; Wen Chen, C.; Mei, T. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 5657–5666.
119. Chen, X.; Xu, C.; Yang, X.; Tao, D. Attention-gan for object transfiguration in wild images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 164–180.
120. Yi, X.; Walia, E.; Babyn, P. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* **2019**, *58*, 101552. [CrossRef]

121. Armanious, K.; Jiang, C.; Fischer, M.; Küstner, T.; Hepp, T.; Nikolaou, K.; Gatidis, S.; Yang, B. MedGAN: Medical image translation using GANs. *Comput. Med. Imaging Graph.* **2020**, *79*, 101684. [CrossRef]
122. Borji, A. Pros and cons of gan evaluation measures. *Comput. Vis. Image Underst.* **2019**, *179*, 41–65. [CrossRef]
123. Furuta, R.; Inoue, N.; Yamasaki, T. Fully convolutional network with multi-step reinforcement learning for image processing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 3598–3605.
124. Kosugi, S.; Yamasaki, T. Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. *arXiv* **2019**, arXiv:1912.07833. [CrossRef]
125. Ganin, Y.; Kulkarni, T.; Babuschkin, I.; Eslami, S.; Vinyals, O. Synthesizing programs for images using reinforced adversarial learning. *arXiv* **2018**, arXiv:1804.01118.
126. Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.