School of Informatics

Informatics Research Review
Supervised and unsupervised image-to-image translation with GANs

[redacted]

January 2021

Abstract

We review the use of GANs in supervised and unsupervised image-to-image translation, focusing on methods that pushed the boundaries of this vast field and became the foundation upon which most recent models are developed. Emphasis is given to models that achieved general purpose and multimodal translation. Our findings suggest that recent studies focus on producing more realistic and diverse generated images, mainly in an unsupervised setting, trying to bridge the gap between supervised and unsupervised translation. Finally, we argue that better evaluation methods for translation models must be developed, as current techniques are not adequately reliable.

Date: Wednesday 20th January, 2021

Supervisor: [redacted]

1 Introduction

There exist numerous and important applications that require the mapping of a given image from its original domain X to a target domain Y, exactly like we would translate a phrase from English to French. For example, adding colour to a greyscale image could be perceived as the translation of an image x from the domain of greyscale images X to a corresponding output image y of the target domain of RGB images Y ($x : X \rightarrow y : Y$). Other examples are mapping edge-maps to images, day scenes to night scenes, semantic labels to images and many more. This broad sub-field of computer vision and graphics is known as image-to-image translation and can be divided into sub-categories with various criteria.

A straightforward way of categorizing such techniques and methods concerns the existence of paired data. More specifically, many algorithms in this field require that a dataset exists with paired images from the input X and the target Y domains (e.g. the Night2Day [1] dataset consists of pairs of the same scene at day and night). If that is the case, image-to-image translation is considered supervised. However, this is not always possible. Not only can the pairing of images be time-consuming and expensive, but it can also be impossible in specific applications. For example, in tasks like object transfiguration (e.g. replacing a horse with a zebra in an image) pairs from the two domains do not exist.

In the past few years, *Generative Adversarial Networks* [2] (GANs) have revolutionized both supervised and unsupervised image-to-image translation. GANs were first presented in 2014 [2] and since then a plethora of studies has been conducted to improve this revolutionary architecture [3, 4]. Even though many variations of GANs exist the principle is always the same. A model, termed *Generator* G, produces samples and another model, termed the *Discriminator* D, classifies the samples as fake or real (i.e. as products of the Generator or as original samples, respectively). In the original GAN [2] the generator samples the input z from a latent space. The generative model's objective is to fool the discriminator by generating images that are indistinguishable from the real ones. The two models are simultaneously trained with adversarial goals, which corresponds to a minimax two-player game. This is implemented with the value function $V(G, D)$ (Equation 1), where real data x are sampled from the $p_{data}(x)$ distribution and the inputs z are sampled from the $p_z(z)$ distribution. The discriminator's goal is to maximize Equation 1, since D's output is a probability (scalar) that must be high for real data and low for generated data. On the other hand, the generative model's objective is to minimize Equation 1, by making the discriminator's output $D(G(z))$ as high as possible. In the original GAN the G and D were multilayer perceptrons and during training they were updated cyclically, one gradient step at a time.

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

Even though GANs have produced remarkable results in the past years, some issues still remain. A problem that is still open and greatly affects image-to-image translation is mode collapse [5]. The data distribution has multiple modes (i.e. areas with high concentration of samples) and the generator must model this distribution without missing any mode. However, it is likely that the discriminator will get trapped in a local minimum during optimization and the generator will learn to deceive the discriminator with samples from only specific modes. In such an example, the other modes will collapse and the generator will not generate samples from them.

The use of GANs in image-to-image translation has led to remarkable results and in the past few years alone dozens of new methods and variations of the original GAN architecture have been

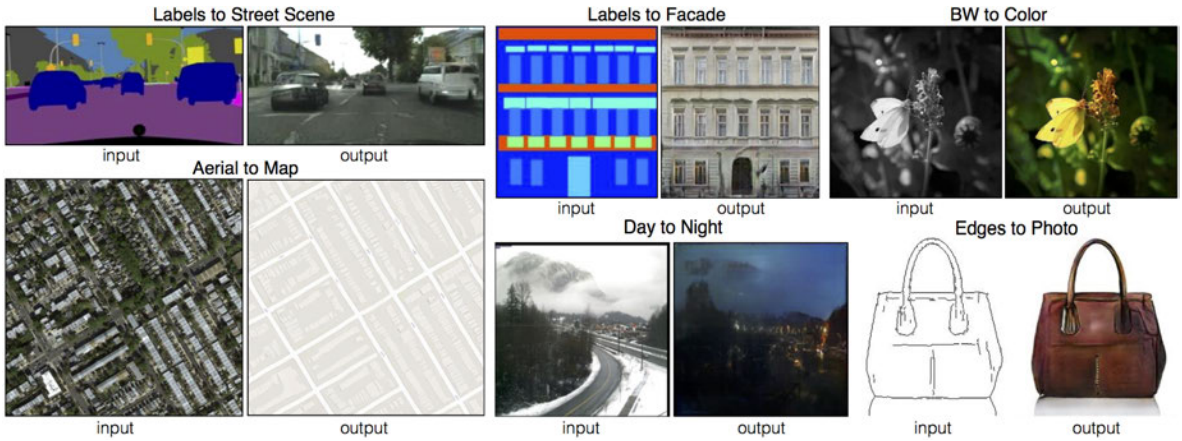


Figure 1: Supervised image-to-image translation results, generated by the Pix2Pix [6] framework.

developed. This might overwhelm newcomers and thus it is beneficial that a concise summary of the most significant techniques be presented. Such an effort had not been made until the last half of 2020 [7, 8]. However, these approaches either exclude qualitative and quantitative results, provide minimum information about important techniques or categorize the models based on their application, which makes little sense, since one of the main advantages of GANs is that they can lead to general-purpose methods.

In this paper, we review the use of GANs in supervised and unsupervised image-to-image translation, by focusing on models that pushed the boundaries of the field and became the foundation upon which most newer methods are based. Our main focus is how GANs led to methods that achieve general purpose and multimodal translation. Our findings suggest that generating realistic and diverse images remains an open problem and that the gap between the performance of supervised and unsupervised techniques is yet to be bridged, with the latter showing weaker results. Finally, we argue that the current evaluation methods are not adequately reliable. For example, small changes in the evaluation procedure between researches leads to completely different results for the same models, which prevents from an objective comparison between all the existing translation models. After all, not all translation methods are evaluated on the same datasets and that makes it even more difficult to compare their performance.

2 Image-to-Image Translation with GANs

In this section, models that have revolutionized image-to-image translation will be analysed and techniques that were based on them to produce state-of-the-art results will be briefly presented. The models are divided into two categories based on their type of translation, supervised or unsupervised, and in each category we focus in specific obstacles and how can they be overcome. Firstly, a major issue is the development of techniques that lead to general-purpose models (i.e. translation models that are not tailored for specific applications). Additionally, another significant challenge is generating diverse outputs and modeling a distribution of possible generated images (i.e. multimodal translation). Finally, methods will be briefly presented that unlock image-to-image translation for multiple domains, using but one model. Before diving into the techniques, some useful evaluation methods will be summarized that play a significant role in

generative processes.

2.1 Qualitative and Quantitative Evaluation Methods

It is impossible to talk about image-to-image translation and not use some qualitative and/or quantitative method for evaluating the results. In this section, a brief presentation of the most common metrics will be made. In table 1, the most widely used datasets in image-to-image translation are presented along with the corresponding translation task.

- **Amazon Mechanical Turk** (AMT): AMT is a service that has been extensively used by researchers to outsource the evaluation of generated data by humans. Generated and real images are uploaded and evaluators, known as Turkers, either have to distinguish the fake images or simply evaluate the visual quality of the data. A disadvantage of this method is that giving feedback to the Turkers when distinguishing between real and fake images, has a drastic effect on their evaluation [3]. Additionally, each research uses a different number of Turkers which can have an impact on the evaluation score.
- **Fully Convolutional Networks** (FCN): A common quantitative approach is the FCN-score [6, 9, 10, 11, 12]. In this process, a FCN model is trained with real images (e.g. for semantic segmentation) and then evaluates the generated ones. If the generated images are realistic and similar to the real data, then the output of the network for the generated images and its corresponding ground truth must be close. Usually it includes three different scores: (1) per-pixel accuracy, (2) per-class accuracy and (3) intersection-over-union (IOU). This evaluation method has a significant disadvantage in practice, which is that in every research a different network architecture is used for calculating the score. This makes it impossible to compare FCN-scores between different researches.
- **Learned Perceptual Image Patch Similarity**(LPIPS) [13]: This metric is responsible for reporting the diversity in generated results. To achieve that the average weighted L2 distance is measured between features of randomly-sampled pairs of generated images. The main advantage of this metric is that it correlates well with human perceptual judgement.

2.2 Supervised Image-to-Image Translation with GANs

2.2.1 General purpose image-to-image translation

Researching conditional GANs (cGANs) paved the road for the first general-purpose framework in supervised image-to-image translation called Pix2Pix [6], which produced state of the art results for various applications. The difference between a regular GAN and a cGAN is that the latter learns the mapping not just from random noise z to the output y, but also from an image x that belongs in the X domain. For the generator, a "U-Net" architecture [22] with skip connections was used, which assists features that are significant in both the X and Y domain to flow through the network. This way, domain-invariant information (e.g. the geometric structure of objects) remains intact. In order to encourage the generation of crisp images, an architecture termed PatchGAN [6] was used as discriminator, which classifies $N \times N$ patches of images, thus motivating high-frequency structures. The Pix2Pix model revolutionized image-to-image translation by producing great results, not only in numerous benchmark datasets but also in custom tasks by the twitter community.

Task	Dataset	No. images	Resolution	Paired
Map ↔ Aerial photo	Maps [6]	1K	512×512	✓
Day ↔ Night scene	Night2Day [1]	8.5K	various	✓
Semantic labels ↔ photo	Cityscapes [14]	5K	1024×2048	✓
Architectural labels ↔ photo	CMP Facades [15]	600	various	✓
Edges ↔ photo	Edges2Shoes[16, 17]	50K	256×256	✓
	Edges2Handbags [18]	137K	256×256	✓
Thermal → color photo	MPD [19]	95.328	640×480	✓
BW → color photo	ImageNet [6, 20]	1.2M	256×256	✓
Face attribute translation	Celeb-A [21]	200K	various	-
Season transfer	summer2winter [9]	2K	various	-
Object transfiguration	horse2zebra apple2oranges [9]	2K	various	-
Painter Style transfer	cezanne2photo ukiyoe2photo vangogh2photo [9]	2K	various	-

Table 1: Datasets for different supervised and unsupervised image-to-image translation tasks.

The objective of Pix2Pix consists of two parts. The first one (Equation 2) is the adversarial loss of the cGAN, which is the same as Equation 1, but now the discriminator is conditioned on the input. Additionally, in order to encourage the generated image to be similar to the ground truth, the $L1$ norm is added to the objective (Equation 3), which unlike the $L2$ norm, leads to less blurry results.

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_y[\log D(x, y)] + \mathbb{E}_{y,z}[\log (1 - D(x, G(z)))] \tag{2}$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \tag{3}$$

Combining the losses, the final objective is:

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{4}$$

Using AMT on a "Real or Fake" test, Pix2Pix managed to deceive only $6.1\% \pm 1.3\%$ of the Turkers that generated images were real for the *photo* → *map* task and only $18.9\% \pm 2.5\%$ for the *map* → *photo*. In Table 2, the Pix2Pix FCN-scores are presented for the *label* ↔ *photo* translation task on the Cityscapes [14] dataset, where the results for *label* → *photo* are relatively close to the one's of ground truth images.

Task w/ Pix2Pix	Per-pixel acc.	Per-class acc.	Class IoU
photo → *label*	0.83	0.36	0.29
label → *photo*	0.66	0.23	0.17
label → *photo* (GT)	0.80	0.26	0.21

Table 2: FCN-score on the Cityscapes [14] dataset for the Pix2Pix framework, using FCN-8s [23] for semantic segmentation.

Framework	Per-pixel acc.	Class IoU
Pix2Pix	0.78	0.39
Pix2PixHD	**0.83**	**0.63**
(GT)	0.84	0.68

Table 3: Comparison of FCN-scores on the Cityscapes [14] dataset for the Pix2Pix and Pix2PixHD frameworks, using PSPNet [24] for semantic segmentation.

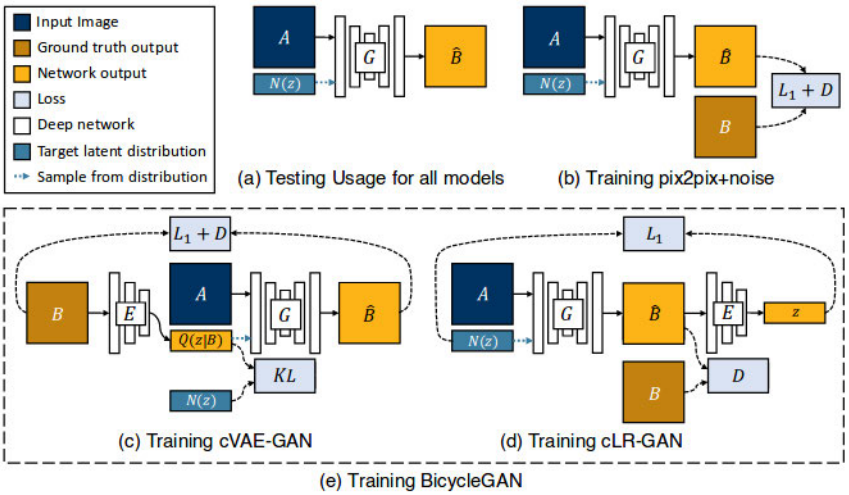


Figure 2: Training BicycleGAN [25]: (a) trained model, (b) baseline model for comparison, (c) training cVAE-GAN ($B \rightarrow z \rightarrow \hat{B}$), (d) training cLR-GAN ($z \rightarrow \hat{B} \rightarrow \hat{z}$), (e) training hybrid model BicycleGAN

Altering Pix2Pix to incorporate a coarse-to-fine generator ($G = \{G_1, G_2, .., G_N\}$) and a multi-scale discriminator ($D = \{D_1, D_2, .., D_N\}$), photo-realistic images of higher resolution can be generated. Pix2PixHD [10] enforced this technique, with 2 sub-networks for the generator and 3 discriminators, and modified the loss function to include feature matching between intermediate outputs of the discriminator for real and generated images. The result was images of 2048×1024 resolution.

Pix2PixHD seems to outperform Pix2Pix , both in FCN-scores (Table 3) and AMT qualitative results, since 93.8% of Turkers preferred Pix2PixHD results over Pix2Pix results. However, from Tables 2, 3, we can see that using a different FCN model changes the evaluation scores of Pix2Pix drastically.

2.2.2 Multimodal image-to-image translation

Both Pix2Pix and Pix2PixHD generate relatively realistic images but lack diversity in their results. Even though Pix2Pix applied noise to induce stochasticity in the generated images, it did not seem to work, since the cGAN learned to ignore the noise. This issue in supervised image-to-image translation was firstly overcome by the BicycleGAN [25] implementation, a multimodal solution that models a distribution of possible results (e.g. given a night image of a scene as input, many possible day images of the scene will be produced).

To achieve this, BicycleGAN learned a mapping from the latent space to the output and back, to form a bijection. This way, the model is discouraged from generating the same output from

Method	AMT Fooling Rate (%)	LPIPS Distance
Random Real Images	50.0%	0.265 ± .007
Pix2Pix+noise	27.93% ± 2.40 %	0.013 ± .000
BicycleGAN	**34.33% ± 2.69 %**	**0.111 ± .002**

Table 4: **AMT** (Fake vs Real) and **LPIPS** distance results for BicycleGAN, Pix2Pix and random real images, on the *maps* $\rightarrow$ *photos* translation task.

different samples in latent space. The implementation consists of two processes (Figure 2): (i) encoding with a conditional variational autoencoder GAN (cVAE-GAN) the ground-truth into a latent space and feeding it to the generator to reconstruct this image and (ii) sampling from the latent space to generate an image and then trying to reconstruct the latent vector, using a conditional latent regressor GAN (cLR-GAN). The BicycleGAN's architecture is strongly influenced by Pix2Pix, using as generator a "U-Net" with skip connections and two PatchGAN discriminators.

The objective of BicycleGAN is implemented by combining the objectives of the two processes. More specifically, the cVAE-GAN (i.e. encoding the ground-truth image to latent code and afterwards reconstructing the image) corresponds to the loss functions $\mathcal{L}^{VAE}_{cGAN}(G, D, E)$ and $\mathcal{L}^{VAE}_{L1}(G, E)$, which are identical to Equations 2 and 3, with the difference that the encoder is part of the loss function and z is the latent code and not a noise vector. Additionally, in order to force the latent distribution to be close to a Gaussian distribution the $\mathcal{L}_{KL}$ was added. Regarding the cLR-GAN (i.e. generating an image from sampled latent code and recovering it back), the $L1$ loss $\mathcal{L}^{latent}_{L1}(G, E)$ between the original and the recovered latent code was used along with the original adversarial loss from Equation 2, to encourage realistic results. Thus, the objective of BicycleGAN is:

$$G^*, E^* = \arg\min_{G,E}\max_{D} \mathcal{L}^{VAE}_{cGAN}(G, D, E) + \lambda\, \mathcal{L}^{VAE}_{L1}(G, E) + \mathcal{L}_{cGAN}(G, D) + \lambda_{latent}\, \mathcal{L}^{latent}_{L1}(G, E) + \lambda_{KL}\, \mathcal{L}_{KL}(E) \quad (5)$$

Pix2Pix was used as baseline in the experiments of BicycleGAN. In Table 4, the percentage of fake images that deceived humans is presented, along with the LPIPS distance score, for generated images from the Pix2Pix and the BicycleGAN. Even though neither is close to the fooling rate of "random real images", it seems that generated images from BicycleGAN are 5% more likely to deceive Turkers than those produced by Pix2Pix. Concerning the diversity of generated images, it is clear from the LPIPS distance that Pix2Pix does not excel in generating diverse results. This confirms that cGANs learn to ignore the added noise and thus output diversity is insignificant. On the other hand, the BicycleGAN model has no trouble producing images that are more diverse and realistic compared to Pix2Pix. Again, the weakness of AMT can be observed by comparing Tables 2 and 4, where for the same task the AMT scores are quite different.

A more advanced multimodal approach is CEGAN [26]. The key difference between BicycleGAN and CEGAN is that the latter model uses the discriminator to classify samples in the latent space, rather than the image space. This produced both more realistic and diverse results.

2.3 Unsupervised Image-to-Image Translation with GANs

2.3.1 General purpose image-to-image translation

As was the case with supervised image-to-image translation, no general purpose technique existed for unpaired data, until GANs were sufficiently studied. Early works (e.g. CoGAN [11], DTN [27] and SimGAN [12]) suffered from limitations and were not effective general purpose models. The development of CycleGAN [9] not only led to a reliable framework that was not tailored to a specific task, but also introduced cycle consistency to image-to-image translation with GANs, which became the foundation for many other models in this field.

By training a mapping $G : X \rightarrow Y$, it is not guaranteed that the input x will always be translated meaningfully to the output y, which in many cases can be caused by the mode collapse phenomenon. To deal with this issue, CycleGAN trains simultaneously the mapping $F : Y \rightarrow X$ along with G, by enforcing a cycle consistency loss [28], so that $F(G(x)) \approx x$ and $G(F(y)) \approx y$. This technique produces great results in many research fields (e.g. an English to French translator should ideally be able to return the original sentence in the inverse process). To implement this[1], two adversarial discriminators D_Y and D_X try to distinguish the translated $G(x)$ and $F(y)$ images. This process is presented in Figure 3, along with how cycle consistency loss is implemented in both $G : X \rightarrow Y$ and $F : Y \rightarrow X$ mappings. Concerning the generator architecture, it was adopted from [29], while a PatchGAN [6] was used as discriminator.

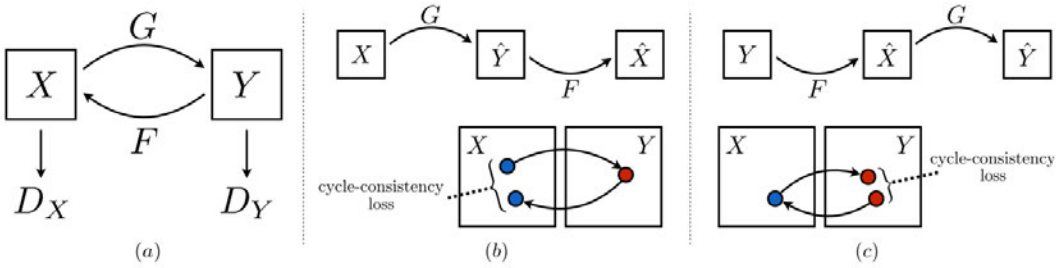


Figure 3: (a) CycleGAN $G : X \rightarrow Y$ and $F : Y \rightarrow X$ mappings, (b) forward cycle consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ and (c) backward cycle consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$.

The objective of CycleGAN consists of two adversarial losses, for the G mapping: $\mathcal{L}_{GAN}(G, D_Y, X, Y)$ and F mapping: $\mathcal{L}_{GAN}(G, D_X, Y, X)$, based on Equation 1. Additionally, the forward and backward cycle consistency loss $\mathcal{L}_{cyc}(G, F)$ is added, to reduce the difference between $F(G(x))$ and x and between $G(F(y))$ and y, using the $L1$ norm. So, the final objective is:

$$G^* = \arg \min_G \max_D \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(G, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \tag{6}$$

In Table 5, FCN-scores are presented for CycleGAN and other relevant models. CycleGAN managed to outperform all other unsupervised methods, while at the same time being a general-purpose model. However, it did not succeed in outperforming the supervised Pix2Pix framework. Once again, the FCN scores in Tables 2, 3, 5 do not match. Other similar, concurrent models

[1]BicycleGAN, which was developed after CycleGAN, implements a similar method in a supervised setting.

Task	Model	Per-pixel acc.	Per-class acc.	Class IoU
label → *photo*	CoGAN	0.40	0.10	0.06
	SimGAN	0.20	0.10	0.04
	CycleGAN	**0.52**	**0.17**	**0.11**
	Pix2Pix	0.71	0.25	0.18
photo → *label*	CoGAN	0.45	0.11	0.08
	SimGAN	0.47	0.11	0.07
	CycleGAN	**0.58**	**0.22**	**0.16**
	Pix2Pix	0.85	0.40	0.32

Table 5: Comparison of FCN-scores on the Cityscapes[14] dataset between the CycleGAN and the CoGAN, SimGAN and Pix2Pix models, using the FCN-8s network for semantic segmentation.

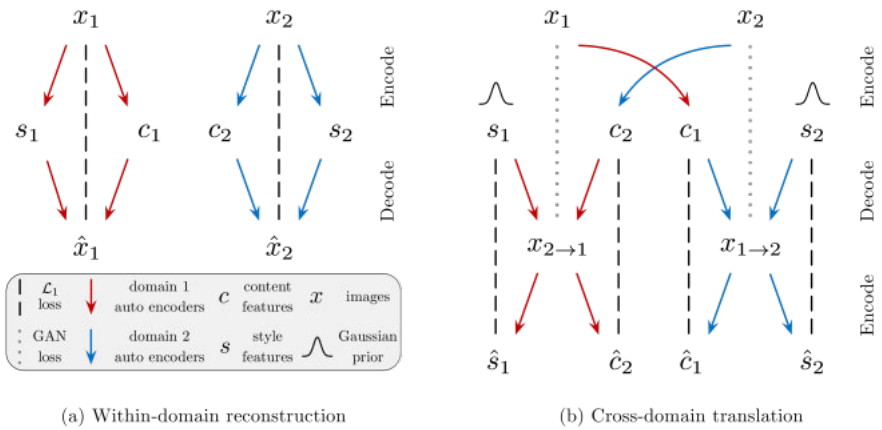
are DualGAN [30] and DiscoGAN [31] but none produced better results than CycleGAN. UNIT [32] was another concurrent research, with results close to the ones of CycleGAN. UNIT is based on the assumption that a pair of images in different domains can be mapped to the same latent code, in a shared latent space. All of these models utilized GANs to achieve unsupervised translation that was not task-specific. However, this applies only to cases were images are to be translated in a specific domain. That is why StarGAN [33] was proposed to perform multi-domain translation with only one generator (e.g. changing the hair colour and the facial expression of a face photo).

2.3.2 Multimodal image-to-image translation

Even though the previously presented unsupervised translation models produce relatively realistic results they still lack the ability to generate diverse outputs from a single input. To achieve multimodal unsupervised translation the MUNIT [34] and the DRIT [35] frameworks were developed. Both of these approaches implement a disentangled representations approach, by assuming that the images can be decomposed into a domain-invariant content space, which is shared between the two domains and corresponds to the spatial structure, and into a domain-specific attribute space, which corresponds to the style of the image. In both cases a form of cycle-consistency was applied, inspired by CycleGAN.

Since the approaches of the two models are similar, we will focus on the implementation of MUNIT, a model based on UNIT, which was briefly mentioned in the previous section. MUNIT contains two autoencoders, which consist of an encoder E_i and a decoder G_i for each $\mathcal{X}_i (i = 1, 2)$ domain. The images are decomposed into a content code c_i and a style code s_i in latent space. To translate an image x_1 from the $\mathcal{X}_1$ to the $\mathcal{X}_2$ domain the extracted content code c_1 is passed through the decoder of the other domain's autoencoder, along with a $s2$ style latent code, which is sampled from the target domain. This process is presented in Figure 4. A form of cycle consistency was applied, termed style-augmented, with the goal of learning to reconstruct the original image from the generated one, given the style s of the original image. In contrast with the cycle consistency applied in CycleGAN, this approach does not lead to deterministic translation.

The objective of MUNIT consists of a bidirectional reconstruction loss and an adversarial loss. The first loss ensures that the encoders and decoders are inverses of each other, by learning to reconstruct both the image (image → latent code → image, Equation 7) and the content/style latent code that has been sampled from the latent distribution (latent code → image → latent



(a) Within-domain reconstruction (b) Cross-domain translation

Figure 4: MUNIT overview.

code, Equations 8, 9), as shown in Figure 4. The goal of the adversarial loss is that translated images will be indistinguishable from the real ones, as shown in Equation 10 for the $\mathcal{X}_2$ domain. In these equations, $q(s_2)$ is the prior $\mathcal{N}(0, I)$.

$$\mathcal{L}^{x_1}_{recon} = \mathbb{E}_{x_1\sim p(x_1)}[\|G_1(E^c_1(x_1), E^s_1(x_1)) - x_1\|_1] \tag{7}$$
$$\mathcal{L}^{c_1}_{recon} = \mathbb{E}_{c_1\sim p(c_1), s_2\sim q(s_2)}[\|E^c_2(G_2(c_1, s_2)) - c_1\|_1] \tag{8}$$
$$\mathcal{L}^{s_2}_{recon} = \mathbb{E}_{c_1\sim p(c_1), s_2\sim q(s_2)}[\|E^s_2(G_2(c_1, s_2)) - s_2\|_1] \tag{9}$$
$$\mathcal{L}^{x_2}_{GAN} = \mathbb{E}_{x_2\sim p(x_2)}[\log D_2(x_2)] + \mathbb{E}_{c_1\sim p(c_1), s_2\sim q(s_2)}[\log(1 - D_2(G_2(c_1, s_2)))] \tag{10}$$

Combining Equations 7 - 10 for both autoencoders, the final objective is:

$$G^*(E_1, E_2, G_1, G_2, D_1, D_2) = \arg \min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} \mathcal{L}^{x_1}_{GAN} + \mathcal{L}^{x_2}_{GAN} +$$
$$\lambda_x(\mathcal{L}^{x_1}_{recon} + \mathcal{L}^{x_2}_{recon}) + \lambda_c(\mathcal{L}^{c_1}_{recon} + \mathcal{L}^{c_2}_{recon}) + \lambda_s(\mathcal{L}^{s_1}_{recon} + \mathcal{L}^{s_2}_{recon}) \tag{11}$$

In Table 6, we can see that MUNIT generates not only more diverse but also more realistic images than CycleGAN and its predecessor UNIT. Additionally, even though it does not seem to outperform its corresponding supervised multimodal method, BicycleGAN, their results are comparable. More recent techniques have managed to take multimodal translation even further by simultaneously achieving multi-domain translation. Such models are StarGANv2 [36], which is the multimodal version of StarGAN, DRIT++ [37] and GMM-UNIT [38], which are the multi-domain versions of DRIT and MUNIT, respectively.

3 Summary & Conclusion

Image-to-image translation is the task of mapping an image x from a domain X to an output image y of the domain Y. The methods used in this vast field can be divided into two categories based on the existence of paired data. Should paired data exist the translation is considered supervised and if not the translation is considered unsupervised. In general, having paired data is rare and expensive, which is why research mostly focuses on developing advanced

	edges → shoes		edges → handbags	
Model	Quality	Diversity	Quality	Diversity
UNIT	37.4 %	0.011	37.3 %	0.023
CycleGAN	36.0 %	0.010	40.8 %	0.012
MUNIT	**50.0 %**	**0.109**	**50.0 %**	**0.175**
BicycleGAN	56.7 %	0.104	51.2 %	0.140

Table 6: Comparison of diversity (LPIPS) and quality (human preference) scores between UNIT, CycleGAN, MUNIT and the supervised translation model BicycleGAN, on the edges → shoes [16] and edges → handbags [18] datasets.

methods that generate realistic and diverse results without such datasets. Both supervised and unsupervised image-to-image translation were revolutionized by the invention of GANs. Using this architecture led to models that not only generate more realistic images but also achieve **general-purpose** (i.e. not tailored to a specific application), **multimodal** (i.e. able to generate diverse results from a single input) and **multi-domain** (i.e. able to translate images to various domains) image-to-image translation.

In a supervised setting, Pix2Pix was one of the first models that achieved general purpose translation and it is still used as baseline in many researches. Lacking diversity in generated images, BicycleGAN was proposed to achieve multimodal translation, outperforming Pix2Pix both in realism and diversity. In an unsupervised setting, CycleGAN and UNIT were successful general purpose models that also introduced cycle consistency in image-to-image translation, a method that inspired many future models. DRIT and a new version of UNIT, named MUNIT, not only produced more realistic results than the two previous models, but also achieved unsupervised multimodal translation. More advanced models, like StarGANv2, DRIT++ and GMM-UNIT achieve general purpose multimodal translation along with multi-domain translation.

The purpose of this review was to present and analyze the supervised and unsupervised models that pushed the boundaries of image-to-image translation and became the foundation upon most of the recent state-of-the art methods are based. This paper may serve as a guide to beginners who might find this vast field overwhelming, due to the great number of different methods. Our findings suggest that:

1. For the past few years research is mostly focused on unsupervised methods because paired datasets are rare and expensive. However, more research is required to bridge the gap between the performance of corresponding supervised and unsupervised techniques.

2. Furthermore, even though tremendous steps have been made towards producing diverse outputs, all GAN-based approaches are prone to suffer from mode-collapse. Tackling this issue and generating more realistic results is still and open problem.

3. Current evaluation metrics are not adequately reliable. Concerning quantitative evaluation, many papers present different results for the same models, by altering the evaluation process. For example, Pix2Pix scored 0.66% per-pixel accuracy on label → photo in the original paper and 0.78% on the Pix2PixHD paper, because a different segmentation network was used for the FCN-score. The same applies to qualitative results, where AMT scores might differ due to the subjective nature of the test. For example, Pix2Pix fooled roughly 19% of the Turkers on the photo → map task in the original paper and approximately 28% in the BicycleGAN paper. These issues and the fact that in each paper a different dataset is used for evaluation, make it impossible to objectively compare models.

References

[1] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 33(4), 2014.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[3] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2234–2242. Curran Associates, Inc., 2016.

[4] Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 1486–1494. Curran Associates, Inc., 2015.

[5] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160, 2017.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[7] Aziz Alotaibi. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 12(10):1705, Oct 2020.

[8] Yuan Chen, Yang Zhao, Wei Jia, Li Cao, and Xiaoping Liu. Adversarial-learning-based image-to-image transformation: A survey. *Neurocomputing*, 411:468 – 486, 2020.

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 469–477. Curran Associates, Inc., 2016.

[12] S. Yu, H. Dong, F. Liang, Y. Mo, C. Wu, and Y. Guo. Simgan: Photo-realistic semantic image manipulation using generative adversarial networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 734–738, 2019.

[13] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Radim Šára Radim Tyleček. Spatial pattern templates for recognition of objects with regular structure. In *Proc. GCPR*, Saarbrucken, Germany, 2013.

[16] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[17] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

[18] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[19] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[24] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[25] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017.

[26] F. Xiong, Q. Wang, and Q. Gao. Consistent embedded gan for image-to-image translation. *IEEE Access*, 7:126651–126661, 2019.

[27] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *CoRR*, abs/1611.02200, 2016.

[28] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. *CoRR*, abs/1604.05383, 2016.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[30] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876, 2017.

[31] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1857–1865. JMLR.org, 2017.

[32] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 700–708. Curran Associates, Inc., 2017.

[33] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[34] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[35] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018.

[36] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[37] Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10-11):2402–2417, November 2020.

[38] Yahui Liu, Marco De Nadai, Jian Yao, Nicu Sebe, Bruno Lepri, and Xavier Alameda-Pineda. Gmm-unit: Unsupervised multi-domain and multi-modal image-to-image translation via attribute gaussian mixture modeling, 2020.