Neural Networks 148 (2022) 166–175

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Not every sample is efficient: Analogical generative adversarial network for unpaired image-to-image translation

Ziqiang Zheng [a], Jie Yang [b], Zhibin Yu [a,*], Yubo Wang [c,*], Zhijian Sun [d,e], Bing Zheng [a]

[a] Ocean University of China/ Sanya Oceanographic Institution, Ocean University of China, No. 238, Songling Road, Qingdao/Sanya, Shandong/Hainan, China
[b] University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[c] School of Life Science and Technology, Xidian University, Xi'an, Shanxi, China
[d] Crossocean of Suzhou technology, No. 218 East Qingdao Road, Suzhou, China
[e] Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 22 November 2021
Accepted 20 January 2022
Available online 29 January 2022

Keywords:
Unpaired image-to-image translation
Generative adversarial network
Metric learning
Analogical learning

ABSTRACT

Image translation is to learn an effective mapping function that aims to convert an image from a source domain to another target domain. With the proposal and further developments of generative adversarial networks (GANs), the generative models have achieved great breakthroughs. The image-to-image (I2I) translation methods can mainly fall into two categories: *Paired* and *Unpaired*. The former paired methods usually require a large amount of input–output sample pairs to perform one-side image translation, which heavily limits its practicability. To address the lack of the paired samples, CycleGAN and its extensions utilize the cycle-consistency loss to provide an elegant and generic solution to perform the unpaired I2I translation between two domains based on unpaired data. This thread of dual learning-based methods usually adopts the random sampling strategy for optimizing and does not consider the content similarity between samples. However, not every sample is efficient and effective for the desired optimization and leads to optimal convergence. Inspired by analogical learning, which is to utilize the relationships and similarities between sample observations, we propose a novel generic metric-based sampling strategy to effectively select samples from different domains for training. Besides, we introduce a novel analogical adversarial loss to force the model to learn from the effective samples and alleviate the influence of the negative samples. Experimental results on various vision tasks have demonstrated the superior performance of the proposed method. The proposed method is also a generic framework that can be easily extended to other I2I translation methods and result in a performance gain.

1. Introduction

Thanks to the proposal of generative adversarial networks (GANs) (Goodfellow et al., 2014), the image-to-image (I2I) translation technology has made great breakthroughs in various computer vision fields such as deraining (Zhang, Sindagi, & Patel, 2017), super-resolution (Johnson, Alahi, & Fei-Fei, 2016; Ledig et al., 2017), colorization (Su, Chu, & Huang, 2020) and so on (Bau et al., 2019; Chen et al., 2016; Luo, Liu, Guan, Yu, & Yang, 2020; Wang, Liu, Zhu, Yakovenko, et al., 2018; Zheng, Yu, Wu et al., 2021; Zheng, Yu, Zheng, Yang, & Shen, 2021). Many variants of GAN-based image translation algorithms were designed to achieve remarkable translation performance. The existing I2I translation methods can mainly fall into two categories: **Paired** (Isola, Zhu, Zhou, & Efros, 2017; Wang, Liu, Zhu, Tao, et al., 2018; Yi, Liu, Lai, & Rosin, 2019; Zheng, Wang, Yu, Zheng, & Zheng, 2018) and **Unpaired** (Huang, Liu, Belongie, & Kautz, 2018; Kim, Kim, Kang, & Lee, 2020; Lee, Tseng, Huang, Singh, & Yang, 2018; Yi, Zhang, Tan, & Gong, 2017; Zheng et al., 2019; Zheng, Wu, Han, & Shi, 2020; Zhu, Park, Isola, & Efros, 2017). The former paired I2I translation methods usually require a large amount of paired data to conduct precise one-side image translation. In consideration of the lack of the paired samples and labors to collect paired training data, Cycle-GAN methods (Kim, Cha, Kim, Lee, & Kim, 2017; Yi et al., 2017; Zhu et al., 2017) utilized the cycle-consistency loss to provide an elegant and generic solution to perform the unpaired I2I translation between two image domains based on unpaired data. Furthermore, to achieve the unpaired I2I translation among multiple domains, the representative StarGAN (Choi et al., 2018) extended CycleGAN (Zhu et al., 2017) and combined an additional classification loss to perform a conditional translation (Choi et al., 2018; Liu et al., 2019). However, all of these above-mentioned

* Corresponding authors.
E-mail addresses: yuzhibin@ouc.edu.cn (Z. Yu), ybwang@xidian.edu.cn (Y. Wang).

https://doi.org/10.1016/j.neunet.2022.01.013
0893-6080/

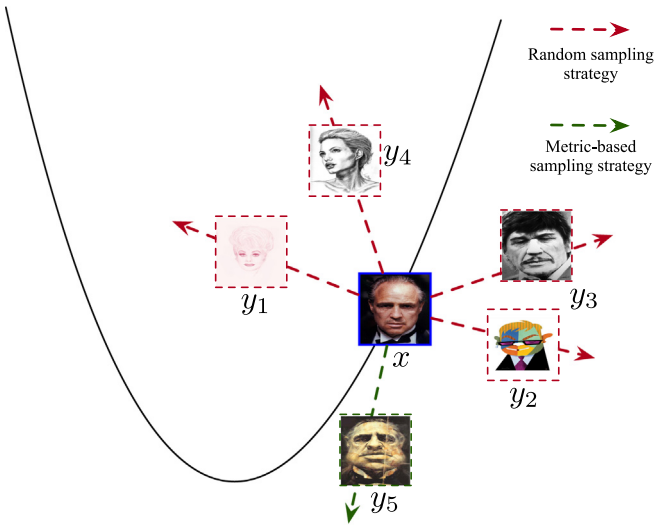

Fig. 1. The intuitive comparison between *random sampling strategy* of previous unpaired I2I methods and our *metric-based sampling strategy*. The random sampling strategy could result in wrong gradient descent directions (indicated by red arrows) due to the irrelevant content representations. In contrast, our method can learn reasonable feature representations and perform better because the metric-based sampling strategy can find the correct gradient descent direction (indicated by the green arrow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

methods did not consider an effective sampling strategy to choose effective and suitable samples for the adversarial learning to promote the unpaired I2I translation performance.

In general detail, given one sample x from the source domain $\mathbb{X}$, most of unpaired I2I translation methods (*e.g.,* CycleGAN) randomly sample an image y in the target domain $\mathbb{Y}$ and compute the loss constraint (*e.g.,* cycle-consistency loss) between the training data (x, y). This random sampling strategy assumes that every sample is efficient and equally important for optimizing the model. However, the assumption may not hold on most cases. In a lucky case, the training model may unearth the key clues what the models need to learn between x and y. Unfortunately, in most cases, x and y may have different feature representations. In other words, there is a content mismatch between x and y, which will result in the wrong gradient descent (indicated by the red arrows) of the model shown in Fig. 1. Thus, the model cannot obtain an optimal solution or catastrophically converge slowly due to the random gradient descent directions. The intrinsic property of the random sampling strategy cannot guarantee the training data with correct gradient descent direction is sampled. Besides, examples in the target domain are wrongly assumed to have equal importance during the training process. However, as illustrated in Fig. 1, we argue that not every sample is efficient for the convergence of the model and providing correct gradient descent direction. To address this issue, inspired by the **analogical learning** (Morrison et al., 2004; Vendetti, Matlen, Richland, & Bunge, 2015): *an artificial intelligence engine can perceive and utilize relationships and similarity between the observations and recall previous observations for scene understanding*, we propose a novel **Ana**logical **G**enerative **A**dversarial **N**etwork (**AnaGAN** for short) with an effective metric-based sampling strategy to select the efficient training sample for optimization shown in Fig. 1. The proposed AnaGAN could utilize the content similarity between samples and choose more effective samples from the target domain to teach the generator how to synthesize reasonable and plausible outputs.

An intuitive illustration of our method is shown in Fig. 2: an engine is reminded of some similar cases when analogical retrieval occurs (Gentner & Smith, 2013), which means to seek similar observations from memory. To measure the content similarity in a low dimensional semantic space, we combine the deep metric learning (Bellet, Habrard, & Sebban, 2013; Duan, Zheng, Lin, Lu, & Zhou, 2018; Kaya & Bilge, 2019) to rank the samples from $\mathbb{Y}$ when it sees an image x from $\mathbb{X}$. Then our AnaGAN utilizes the content similarity and relationship between x and the selected samples in $\mathbb{Y}$ for effective learning to obtain a better mapping function from $\mathbb{X}$ to $\mathbb{Y}$. Furthermore, we consider the analogical evaluation and combine a dual analogical adversarial loss to enhance the importance of the selected efficient samples and alleviate the influence of negative samples with the content mismatch. Through these operations, we can develop a universal I2I translation framework, which is superior to previous methods and achieve better translation performance without introducing extra network parameters.

In this paper, we propose a novel generic analogical generative adversarial network for unpaired I2I translation tasks. To alleviate the content mismatch in the random sampling strategy proposed in CycleGAN (Zhu et al., 2017), the proposed method has designed a simple and effective metric-based sampling strategy to promote the performance of the unpaired I2I translation task. Comprehensive experiments on various large-scale datasets are conducted and the translation performance has demonstrated the effectiveness of the proposed method. Besides, the proposed AnaGAN is also a generic framework that can be easily extended to other unpaired I2I solutions. To sum up, our main contributions are listed as follows.

- We propose a simple, novel and effective end-to-end unpaired I2I translation framework based on the analogical learning scheme and deep metric learning principle to boost the unpaired translation performance. The dual analogical adversarial loss is designed to force the model to learn from effective samples and alleviate the influence of the content mismatch.
- Our AnaGAN can heavily promote the unpaired translation performance through a more effective analogical learning. The comprehensive experiments have demonstrated the superior performance of the proposed method.
- The proposed method is a general training strategy, which could be easily extended to other unpaired I2I translation methods such as MUNIT and UNIT and obtain a remarkable performance gain without introducing any extra network parameters and computation cost.

2. Related work

2.1. CycleGAN solutions

Unpaired domain translation (Almahairi, Rajeswar, Sordoni, Bachman, & Courville, 2018; Chen, Pan, Yao, Tian, & Mei, 2019; Pizzati, Charette, Zaccaria, & Cerri, 2020; Tang, Liu, Xu, Torr, & Sebe, 2021; Wang, Du, & Guo, 2019) aims to transfer the shared knowledge from a source domain to another target domain using unpaired training data. For the pixel-level unpaired domain adaptation, Cycle-GAN based methods, which adopted the cycle-consistency constraint to perform unpaired I2I translation, (Huang et al., 2018; Lee et al., 2018; Liu, Breuel, & Kautz, 2017; Yi et al., 2017; Zhu et al., 2017) became popular. These Cycle-GAN based I2I translation methods have been utilized in various applications such as image style transfer (Gatys, Ecker, & Bethge, 2015; Johnson et al., 2016; Sanakoyeu, Kotovenko, Lang, & Ommer, 2018), night-to-day (Zheng et al., 2020), and image restoration (Li et al., 2018). These visual tasks aim to perform

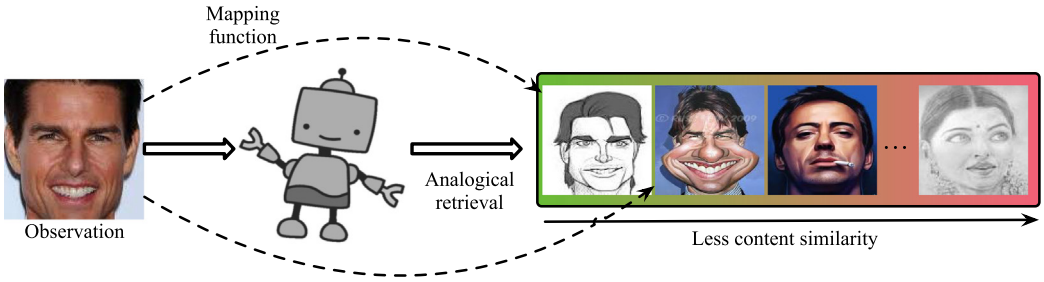

Fig. 2. The intuitive explanation of our AnaGAN. When observing a new photo, to achieve reasonable photo-to-caricature image translation, the agent would recall the seen caricature images and learn the mapping function between the photo observation and the top caricature images with most content similarities. Through mimicking, the model could generate more plausible caricature outputs.

transfer one image with a source style to the desired output with the target style. To promote the translation quality, some specially designed networks (Tang et al., 2021) were further designed. Training Cycle-GAN based I2I translation model requires sampling two images from both source and target domain to obtain gradient for model optimization. The random sampling strategy is widely applied during the training of the Cycle-GAN methods, in which each sample contributes to the optimization of the model equally. However, some negative samples may degrade the adversarial learning efficiency. To address this issue, Gomez et al. (2020) firstly proposed to combine a parallel image retrieval system to boost the disentanglement of the domain-specific and domain-invariant feature representations through multiple task learning. However, the iterative training process of Gomez et al. (2020) is cumbersome and requires a sophisticated design. Furthermore, Gong et al. proposed to formulate analogical image translation AIT (Gong, Dai, Chen, Li, & Van Gool, 2020) to utilize the content relationship between image pairs. In this work, we propose a novel, simple and effective metric-based sampling strategy based on analogical learning to utilize the content similarity between samples. Moreover, the proposed method is an unpaired I2I training framework, which alleviates the labor to collect paired samples in AIT (Gong et al., 2020).

2.2. Analogical learning

Analogy, which refers to the ability to perceive and utilize relational similarity between two objects, systems, or events, can establish an inference projection between two domains (Prade & Richard, 2017). Learning by analogy involves summarizing a relationship between two similar situations or events and generating further inferences driven by these commonalities (Cambria, Gastaldo, Bisio, & Zunino, 2015; Gentner, 2006; Vendetti et al., 2015). Analogical learning is regarded as a fundamental aspect of human cognition (Knowlton, Morrison, Hummel, & Holyoak, 2012). This crucial cognitive mechanism distinguishes human cognition from that of other intelligent species (Morrison et al., 2004). Analogical learning is also proved to be efficient in many computer vision tasks. Tao et al. developed a novel analogy-detail networks (termed ADNets) for object recognition (Tao, Hong, Shi, Chang, & Gong, 2020). Chen et al. proposed a hybrid analogical learning system to boost the visual relation detection process (Chen & Forbus, 2021). Gong et al. proposed AIT (Gong et al., 2020) and established "Analogical Image Translation" framework. AIT took advantage of analogical learning to build an I2I translation method and achieve a zero-shot image translation capability by coupling a supervised training scheme in a synthetic domain. In our paper, we seek to extend the idea of *Analogical Image Translation* to a more general unpaired I2I case. Unlike AIT, which still requires image pairs for training, a standard unpaired I2I translation framework is introduced in our paper. The forward $F(x)$ and backward $G(y)$ translation function target to utilize the content similarity between samples to achieve effective learning.

2.3. Deep metric learning

Deep metric learning (Bellet et al., 2013; Duan et al., 2018; Kaya & Bilge, 2019) aims to measure the similarity among samples and provide an optimal distance metric for different computer vision tasks. The deep metric learning had achieved remarkable performance on image retrieval (Li & Tang, 2015), text-to-image matching (Wei et al., 2020; Zhang & Lu, 2018) and recommendation system (Campo, Espinoza, Rieger, & Taliyan, 2018). To compute the similarity between samples, some attempts (Hu, Lu, & Tan, 2014; Kaya & Bilge, 2019) had been done to find a good distance metric to achieve superior results. To measure the distance between samples, euclidean distance (Danielsson, 1980), Mahalanobis distance (De Maesschalck, Jouan-Rimbaud, & Massart, 2000) and Kullback–Leibler (Elgammal, Duraiswami, & Davis, 2003) were designed to evaluate the distance in the projected space. Siamese (Bertinetto, Valmadre, Henriques, Vedaldi, & Torr, 2016) and Triplet (Hoffer & Ailon, 2015) networks are two commonly used network architectures with shared weights to measure the distance among different samples. The triplet loss is designed to enlarge the distance between the anchor and the negative sample while reducing the distance between the anchor and the positive sample. In this paper, considering that not every sample is equally important for the optimization, we take advantage of deep metric learning to choose similar samples between the source and target domain to enhance the training efficiency. Furthermore, inspired by the triplet loss (Hoffer & Ailon, 2015), we propose a dual analogical adversarial loss to force the model to learn from effective samples while alleviating the influence of negative samples.

3. Methods

3.1. Preliminary

As a generic unpaired I2I solution, the proposed AnaGAN is based on the previous CycleGAN solution (Huang et al., 2018; Zhu et al., 2017). We replace the random sampling strategy adopted in CycleGAN with a novel metric-based sampling strategy. Following the idea of analogical learning, we believe that the samples sharing high content similarity should help the model to learn an effective representation. In this work, the metric-based sampling strategy is proposed, in which the sample with high content similarity is utilized while the irrelevant samples are discarded. More specifically, we randomly sample a sample x (termed *anchor*) from the source domain $\mathbb{X}$, then we obtain the top k retrieval outputs: $\{y_i\}_1^k$ in the target domain $\mathbb{Y}$ to formulate the *positive pool*. The rest of the training samples from $\mathbb{Y}$ are regarded as the *negative pool*. Similarly, the same sampling strategy is also conducted in the source domain $\mathbb{X}$ to get the positive and negative pools based on y for the reverse direction. Furthermore, we design a dual analogical adversarial loss to force the model to learn from the effective samples and reduce the influence of the negative samples with the content mismatch.

3.2. Image translation based on dual learning

To achieve pixel-level unpaired domain translation between two visual domains: $\mathbb{X}$ and $\mathbb{Y}$ (*e.g.*, the photo image domain and another caricature image domain). Two reverse image translation functions are introduced: the forward translator F aims to generate $\tilde{y} = F(x)$ in the target domain $\mathbb{Y}$ while the reverse translator G targets to synthesize the counterpart reconstruction $\hat{x} = G(\tilde{y})$ in the original photo domain $\mathbb{X}$. To make $\tilde{y}$ look like the samples from $\mathbb{Y}$ as possible, the widely used adversarial loss is adopted:

$$\begin{aligned}\mathcal{L}_{adv}(F, D_{\mathbb{Y}}) =& \mathbb{E}_{\mathbb{Y}}[\log D_{\mathbb{Y}}(y)] + \mathbb{E}_{\tilde{y}}[\log(1 - D_{\mathbb{Y}}(\tilde{y}))], \\ & \text{with} \quad \tilde{y} = F(x),\end{aligned} \tag{1}$$

where $D_{\mathbb{Y}}$ is the domain-specific discriminator for the domain $\mathbb{Y}$. The reverse adversarial loss $\mathcal{L}_{adv}(G, D_{\mathbb{X}})$ is also computed to synthesize reasonable outputs for the reverse translation direction, where $D_{\mathbb{X}}$ is the domain-specific discriminator for domain $\mathbb{X}$. At the training stage, x and y are randomly selected in the CycleGAN (Zhu et al., 2017). There could be content mismatch between x and y of using the random sampling strategy. To preserve the content information as much as possible, the cycle-consistency loss $\mathcal{L}_{cyc}$ (Zhu et al., 2017) is adopted to link F and G:

$$\mathcal{L}_{cyc}(F, G) = \mathbb{E}_{\mathbb{X}}[\|G(F(x)) - x\|_1] \quad + \mathbb{E}_{\mathbb{Y}}[\|F(G(y)) - y\|_1], \tag{2}$$

Through the pixel-wise distance, we can preserve the content information after the unpaired I2I translation. Most CycleGAN solutions adopt the random sampling strategy during training, which is able to reduce variance and avoid overfitting. However, a large sample size of $\mathbb{X}$ and $\mathbb{Y}$ may not help unpaired I2I task, since the selected samples from $\mathbb{X}$ and $\mathbb{Y}$ can have totally different content representations (please refer to Section 4.5 for more detail). Thus, the randomly selected sample can harm the unpaired generative model as shown in Fig. 1. To alleviate the negative influence, we propose a novel metric-based sampling strategy.

3.3. Metric-based sampling strategy

Take a close look to our method, to obtain closely associated training samples from source and target domain, the metric-based sampling strategy is designed. We adopt a pre-trained ResNet50 (He, Zhang, Ren, & Sun, 2016) model Φ on ImageNet (Deng et al., 2009) to obtain the feature vectors after the global average pooling layer and compute the content similarity between samples from different domains. During the unpaired I2I translation process, the selection procedure actively selects k most similar samples $\{y_i\}_1^k$ from $\mathbb{Y}$ (with n training samples and $n \gg k$) based on content similarity with the anchor image x. To be more specific, given an anchor image x from $\mathbb{X}$, the network Φ select the top k samples $\{y_i\}_1^k$ from $\mathbb{Y}$ according to content similarity between $\Phi(x)$ and $\Phi(\{y_i\}_1^n)$ to formulate a subset of $\mathbb{Y}_p$ (*positive pool*). We formulate a *negative pool* $\mathbb{Y}_n$ using the left $n-k$ training samples from $\mathbb{Y}$. The examples y_p sampled from the subset of $\mathbb{Y}_p(y_p|x) : \{y_p^i|x\}_1^k \in \mathbb{Y}$ and y_n from $\mathbb{Y}_n(y_n|x) : \{y_n^i|x\}_1^{n-k} \in \mathbb{Y}$ are chosen for the unpaired training[1]. For arbitrary y_p and y_n, we have the following restriction:

$$\|\Phi(x) - \Phi(y_p)\|_1 < \|\Phi(x) - \Phi(y_n)\|_1, \tag{3}$$

which indicates that an arbitrary positive sample y_p is more similar to x than a negative y_n in the projected content space (see Fig. 3).

[1] For simplicity, we only illustrate the metric-based sampling strategy for the forward adversarial training and $\mathbb{X}_p(x_p|y)$ and $\mathbb{X}_n(x_n|y)$ are also sampled for the backward adversarial training following the same manner.

3.4. Dual analogical adversarial loss

As we discussed in Section 3.2, given a sample x, it is not efficient to train the network when the corresponding adversarial sample y is selected randomly. Inspired by analogical learning theory (Gentner & Smith, 2013), we argue that similar samples would have a better chance to guide the dual adversarial learning, which could result in an accurate gradient descent direction and achieve better translation performance. Thus, the similar sample (y_p) should have greater contributions than the dissimilar ones (y_n). Following this intuition, we redesign the adversarial loss function as:

$$\begin{aligned}\mathcal{L}_{adv}(F, D_{\mathbb{Y}})_{y_p \in \mathbb{Y}_n, y_n \in \mathbb{Y}_n} =& \alpha(\mathbb{E}_{\mathbb{Y}_p}[\log D_{\mathbb{Y}}(y_p|x)] + \mathbb{E}_{\tilde{y}}[\log(1 - D_{\mathbb{Y}}(\tilde{y}|x))]) \\ & + \mathbb{E}_{\mathbb{Y}_n}[\log D_y(y_n|x)] + \mathbb{E}_{\tilde{y}}[\log(1 - D_{\mathbb{Y}}(\tilde{y}|x))], \\ & \text{with} \quad \tilde{y} = F(x),\end{aligned} \tag{4}$$

where α is the parameter to balance the influence of the positive samples and negative examples. Higher α indicates that positive samples have a higher contribution for the unpaired adversarial training. In our experiments, we set $\alpha = 2.0$. More discussion and experiments about the choice of α could be found in Section 4.5. For a dual learning framework, the backward adversarial loss function can also be written as:

$$\begin{aligned}\mathcal{L}_{adv}(G, D_{\mathbb{X}})_{x_p \in \mathbb{X}_p, x_n \in \mathbb{X}_n} =& \alpha(\mathbb{E}_{\mathbb{X}_p}[\log D_{\mathbb{X}}(x_p|y)] + \mathbb{E}_{\tilde{x}}[\log(1 - D_{\mathbb{X}}(\tilde{x}|y))]) \\ & + \mathbb{E}_{\mathbb{X}_n}[\log D_{\mathbb{X}}(x_n|y)] + \mathbb{E}_{\tilde{x}}[\log(1 - D_{\mathbb{X}}(\tilde{x}|y))], \\ & \text{with} \quad \tilde{x} = G(y),\end{aligned} \tag{5}$$

and the final adversarial loss $\mathcal{L}_{adv}$ is described as:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}(F, D_{\mathbb{Y}}) + \mathcal{L}_{adv}(G, D_{\mathbb{X}}), \tag{6}$$

3.5. Final objective

The total loss of our method is a weighted sum of all the losses mentioned above:

$$\mathcal{L}(F, G, D_{\mathbb{X}}, D_{\mathbb{Y}}, \Phi) = \mathcal{L}_{adv} + \lambda \mathcal{L}_{cyc}, \tag{7}$$

where λ is the hyper-parameters to balance the different loss terms and we set $\lambda = 10$ following CycleGAN (Zhu et al., 2017). We optimize the final objective with Adam optimizer (Kingma & Ba, 2015). To be noted, Φ is frozen during the whole training process.

4. Experiments

4.1. Implementation details

We compare the proposed method with previous unpaired methods based on various vision tasks, including the defogging, photo-to-caricature and night-to-day tasks. We first choose two generic unpaired image-to-image translation methods: CycleGAN (Zhu et al., 2017) and MUNIT (Huang et al., 2018) for all vision tasks. For defogging task, we choose the paired Pix2pixHD (Wang, Liu, Zhu, Tao, et al., 2018) for comparison considering the paired samples exist for this task. Besides, we also perform ToDayGAN (Anoosheh, Sattler, Timofte, Pollefeys, & Van Gool, 2019) for comparison. As for photo-to-caricature task, besides MUNIT (Huang et al., 2018) and CycleGAN (Zhu et al., 2017), two specially designed photo-to-caricature methods: CariGAN (Cao, Liao, & Yuan, 2018), DualPathGAN (Zheng et al., 2019) are chosen for comparison. Finally, we choose ToDayGAN (Anoosheh et al., 2019), TSIT (Jiang et al., 2020) for the night-to-day translation

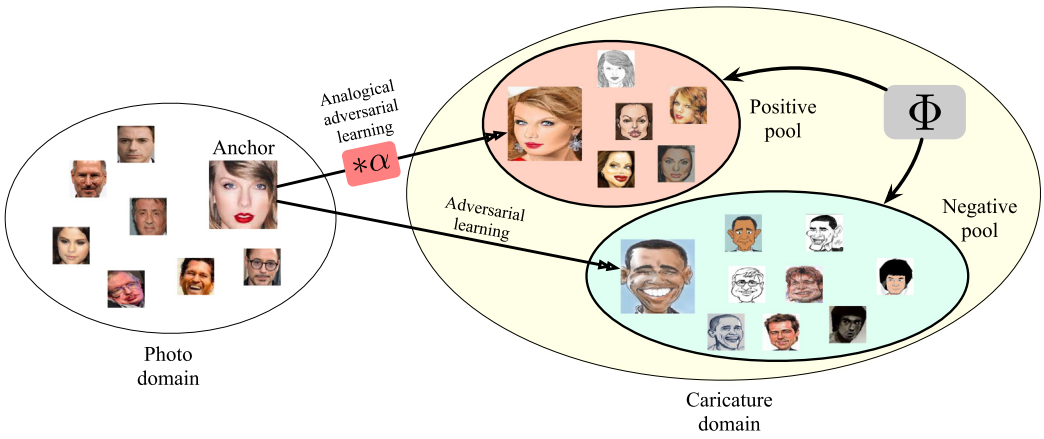

Fig. 3. The illustration of our triplet-like adversarial loss. We encourage the model to learn from the effective samples with similar content representations and suppress the influence of the negative samples with content mismatch.

task. We follow the official instructions of those methods and make a fair setting for comparison. Please note that the proposed AnaGAN has the same network architecture and parameters as CycleGAN. We choose $\alpha = 2.0$ and $k = 50$ in our experiments.

4.2. Evaluation metrics

Fréchet Inception Distance (FID) (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017) is defined as the distance between the generated sample distribution and real distribution. FID is a consistent and robust measurement for evaluating the generated images (Borji, 2019; Lucic, Kurach, Michalski, Gelly, & Bousquet, 2018), which can be calculated by:

$$\mathrm{FID} = \|\mu_x - \mu_g\|_2^2 + \mathrm{Tr}\left(\sum_x + \sum_g - 2(\sum_x \sum_g)^{\frac{1}{2}}\right), \tag{8}$$

where $(\mu_x, \sum_x)$ and $(\mu_g, \sum_g)$ are mean and covariance of the sample embeddings from the data distribution and model distribution. A lower FID score indicates higher generated image quality.

LPIPS (Learned Perceptual Image Patch Similarity (Zhang, Isola, Efros, Shechtman, & Wang, 2018)) computes the perceptual similarity between two images based on image patches. A lower LPIPS means more perceptual similarity between two images. We compute this metric between the translated outputs and corresponding ground-truth images to measure the image translation ability.

Image Quality is also adopted to measure the image generation quality based on paired data. We compute the peak signal-to-noise ratio (**PSNR**) to measure the quality of reconstruction of lossy compression. A higher PSNR score indicates better image translation performance. The structural similarity index (**SSIM**) is computed to evaluate the structural similarity. The higher SSIM, the better the translation performance.

Downstream vision tasks are conducted to evaluate the performance of the unpaired image-to-image translation. We adopted a Deeplab-v3 model[2] pre-trained on Cityscapes dataset (Cordts et al., 2016), performing semantic segmentation on the translated outputs based on the evaluation scripts.[3] We compute the Intersection-Over-Union (IoU) between the outputs and the ground truths. We report the mean IoU (mIoU) of all the categories.

[2] https://github.com/srihari-humbarwadi/DeepLabV3_Plus-Tensorflow2.0

[3] https://github.com/mcordts/cityscapesScripts

4.3. Dataset

Foggy Cityscapes (Sakaridis, Dai, Hecker, & Gool, 2018) is a recently proposed synthetic foggy dataset simulating fog on real scenes (Cordts et al., 2016) with three different levels of visibility: 150 m, 300 m and 600 m. For this dataset, we choose clear images and foggy images with 150 m visibility following the original Train/Val split. In all the experiments, 2975 clear and the corresponding foggy images are adopted for training and the other 500 clear and corresponding foggy images for performance evaluation.

IIIT-CFW (Mishra, Rai, Mishra, & Jawahar, 2016) is a caricature dataset, which has both the photo images of the celebrities and the corresponding cartoon faces in the wild. 8928 annotated cartoon faces of famous personalities in the world with varying professions. Also, it provides 1000 real faces of the public figure for cross-modal retrieval tasks. Due to the lack of the paired samples (the facial orientation and expression of the photo and caricature for the same identity vary a lot), it is not suitable to perform paired training on this dataset. For this dataset, we adopt the images from the first 800 public celebrities for training and others for evaluation.

Alderley is originally proposed for the SeqSLAM algorithm (Milford & Wyeth, 2012), which collected the images for the same route twice: once on a sunny day and another time on a stormy rainy night. Every frame in the dataset is GPS-tagged, thus each nighttime frame has a corresponding daytime frame. Due to the dynamic objects (pedestrians and cars), the daytime and nighttime images with frame correspondence are not paired. For this dataset, we inherit the frame correspondence to formulate our positive pool (the k consecutive images according to the frame correspondence) in our experiment. This dataset provides 14,607 frame matchings and we choose the first 12,000 daytime and nighttime images for training and the rest for evaluation.

4.4. Performance comparison

4.4.1. Defogging

In this section, we performed the low-level defogging task on a complex and diverse scene dataset. In detail, we performed the defogging task on the synthetic Foggy Cityscapes dataset (Sakaridis et al., 2018) following the official Train/Val split. We compared the proposed method with MUNIT (Huang et al., 2018), CycleGAN (Zhu et al., 2017), and ToDayGAN (Anoosheh et al., 2019), which were optimized by the unpaired training strategy.

Fig. 4. The visual comparison between different methods for defogging task on Foggy Cityscapes (Sakaridis et al., 2018) dataset.

Table 1
Quantitative comparison of defogging task on Foggy Cityscapes (Sakaridis et al., 2018) dataset. The symbol ↑ (↓) indicates that the larger (smaller) the value, the better the performance. The best three values in each metric are denoted into red, green and blue. Pix2pixHD* indicates that Pix2pixHD method is optimized in the supervised learning manner.

Method	SSIM↑	PSNR↑	LPIPS↓	FID↓	mIoU↑
CycleGAN (Zhu et al., 2017)	0.8789	25.91	0.3612	72.12	48.15
MUNIT (Huang et al., 2018)	0.8601	24.17	0.3731	88.12	43.12
ToDayGAN (Anoosheh et al., 2019)	0.8821	26.12	0.3515	67.43	51.92
Pix2pixHD* (Wang, Liu, Zhu, Tao, et al., 2018)	0.9051	27.18	0.3414	59.38	57.91
AnaGAN	0.8945	26.87	0.3324	63.23	56.14

Table 2
Quantitative comparison of photo-to-caricature translation task on IIIT-CFW (Mishra et al., 2016) dataset.

Method	FID↓
MUNIT (Huang et al., 2018)	178.7
CycleGAN (Zhu et al., 2017)	136.3
CariGAN (Cao et al., 2018)	110.4
DualPathGAN (Zheng et al., 2019)	86.45
AnaGAN	**47.27**

Table 3
Quantitative comparison of night-to-day translation task on Alderley dataset (Milford & Wyeth, 2012).

Method	FID↓
MUNIT (Huang et al., 2018)	141.5
CycleGAN (Zhu et al., 2017)	169.2
ToDayGAN (Anoosheh et al., 2019)	107.2
TSIT (Jiang et al., 2020)	93.81
AnaGAN	**75.14**

Due to the existence of the paired samples (the foggy images are synthetic from a physical model), we also performed the fully supervised Pix2pixHD method. To make a fair comparison, all the methods were optimized under the image resolution 1024×512. The qualitative results of applying different methods on the Foggy Cityscapes dataset (Sakaridis et al., 2018) were shown in Fig. 4. As illustrated, our AnaGAN could synthesize outputs with detailed content information while preserving the content and structural information of the original foggy images. The quantitative comparison was also reported in Table 1. Compared with other methods (MUNIT and CycleGAN can only generate fuzzy outputs while some detailed information has been lost after the image translation), our AnaGAN can effectively remove the fogginess of the input images and synthesize images with reasonable feature patterns. Besides the visual quality measurement, we also adopted the downstream semantic segmentation task as one objective evaluation metric to measure the visual translation performance. The mIoU scores of different methods are provided in Table 1. From the comparison, our method could still achieve the best semantic segmentation performance.

4.4.2. Photo-to-caricature

Besides the defogging task, we also performed the unpaired photo-to-caricature image translation based on a challenging IIIT-CFW (Mishra et al., 2016) dataset. Due to the highly abstracted characteristic of the caricature images, the translation from the photo images to the corresponding caricature counterparts is extremely challenging. Synthesizing a vivid and lifelike caricature output requires the model to extract and understand the implicit facial representations of the caricature images, and then exaggerate and magnify the learned representations. We compared our method with CariGAN (Cao et al., 2018), MUNIT (Huang et al., 2018), CycleGAN (Zhu et al., 2017), and DualPathGAN (Zheng et al., 2019). The qualitative comparison of different methods was shown in Fig. 5. As illustrated, our AnaGAN could generate caricature outputs according to the content representation of the input photo images. As for the quantitative comparison between various methods, we reported the FID scores of different methods in Table 2. Our method could also achieve the lowest FID score among all the methods, which indicates our method could align the photo and caricature domains well.

4.4.3. Night-to-day for autonomous driving

In autonomous driving, it is laborious and sometimes difficult to collect abundant data with clean and correct annotations for various computer vision tasks. Most of the available datasets contain images mostly from daytime driving. Models trained on those datasets are subject to performance degradation once they are tested on a different domain such as rainy night conditions. One possible solution is to perform night-to-day translation so that we can obtain a robust and accurate visual perception. Considering the dynamic objects and moving pedestrians, it is nearly impossible to collect paired daytime–nighttime images. The unpaired I2I is a popular choice for night-to-day image enhancement. Due to the refection and strong lights, it is extremely challenging to perform the translation from the rainy nighttime domain to the daytime domain. We compare the proposed method with current night-to-day translation methods such as ToDayGAN (Anoosheh et al., 2019) and TSIT (Jiang et al., 2020). Due to the lack of paired samples, we do not perform the paired methods for this task. The visual qualitative translation performance among different methods is illustrated in Fig. 6. The proposed method could better preserve the content information of the billboard, which is significantly important for autonomous driving. Besides, we also report the FID scores of different methods in Table 3 and

Fig. 5. The qualitative photo-to-caricature translation results of different methods on IIIT-CFW (Mishra et al., 2016) dataset.

Fig. 6. The qualitative comparison between different methods on night2day translation task for autonomous driving.

Table 4
Quantitative comparison of defogging task on Foggy Cityscapes (Sakaridis et al., 2018) dataset. The experiments are conducted under image resolution 1024 × 512.

Method	SSIM↑	PSNR↑	LPIPS↓	FID↓	mIoU↑
MUNIT (Huang et al., 2018)	0.8601	24.17	0.3731	88.12	43.12
Ana-MUNIT	**0.8872**	**25.26**	**0.3656**	**71.98**	**46.34**
UNIT (Liu et al., 2017)	0.8571	23.56	0.3894	108.4	41.92
Ana-UNIT	**0.8696**	**24.71**	**0.3792**	**89.54**	**44.76**

our AnaGAN achieves the best score. Our AnaGAN can perform reasonable night-to-day translation to enhance the recognition performance under adverse conditions.

4.5. Discussion and future work

General strategy. The proposed AnaGAN is also a general strategy to analogical learning for unpaired I2I tasks. We can also incorporate our method with other unsupervised learning methods such as MUNIT (Huang et al., 2018) and UNIT (Liu et al., 2017) (termed Ana-MUNIT and Ana-UNIT) and achieve a remarkable performance gain without introducing the extra parameters (see Table 4).

Influence of k. In this part, we first aim to demonstrate that the content-based image matching system Φ could pick up the effective samples with similar content representations under the cross-domain setting. To quantitatively measure the retrieval performance of the pre-trained Φ, we adopted the Recall@K protocol (Song, Xiang, Jegelka, & Savarese, 2016) as the main evaluation metric. We have reported the R@1, R@10, R@20, R@50 and R@100 precision in Table 5. To better illustrate the difference between the proposed metric-based sampling strategy and the previous random sampling strategy, we further investigate the *sampling efficiency* under various settings. Given one random x from $\mathbb{X}$, we assume that $\mathbb{Y}$ has one specific sample y, which matches x. For the random sampling strategy, the sampling efficiency is $\frac{1}{n}$ when the domain $\mathbb{Y}$ has n training samples. The sampling efficiency of the proposed metric-based sampling strategy is $\frac{1}{k} \times R@k$ since we only need to sample from the positive pool. The sampling efficiency of different settings on Foggy Cityscapes dataset is illustrated in Table 5. We observe that the proposed metric-based sampling strategy could heavily promote sampling efficiency.

We then explored the influences of choosing different k on the image translation performance. The quantitative comparison of using different values of k is shown in Table 6. As reported, the approximate number of k could lead to better translation performance without introducing additional parameters and particularly designed architectures. It is easy for the model to learn only from the selected samples when the positive pool is small (with only a few positive training samples) and thus leads to a performance drop. A large positive pool with redundant samples

Table 5
The cross-domain image retrieval results under the Foggy → Clear and Clear → Foggy settings. We adopt a pre-trained ResNet (He et al., 2016) model on ImageNet (Deng et al., 2009) to obtain the feature vectors after the global average pooling layer and compute the retrieval precision. Please note that 2975 training pairs are adopted for computing the retrieval precision and sampling efficiency. The sampling efficiency of the random sampling strategy is $0.0339e^{-2}$.

	Foggy → Clear					Clear → Foggy				
	R@1	R@10	R@20	R@50	R@100	R@1	R@10	R@20	R@50	R@100
R@K (%)	1.41	4.71	6.86	10.99	16.24	0.61	2.35	4.24	7.87	12.74
Sampling efficiency ($1e^{-2}$)	1.41	0.471	0.343	0.220	0.162	0.61	0.235	0.212	0.157	0.127

Table 6
Quantitative comparison of defogging task on Foggy Cityscapes (Sakaridis et al., 2018) dataset. The symbol ↑ (↓) indicates that the larger (smaller) the value, the better the performance. The best value is highlighted in bold.

Method	k	SSIM↑	PSNR↑	LPIPS↓	FID↓	mIoU↑
AnaGAN	1	0.8845	26.37	0.3401	73.45	53.14
AnaGAN	10	0.8901	26.73	0.3336	69.52	53.72
AnaGAN	20	0.8923	26.64	**0.3308**	72.45	56.04
AnaGAN	50	**0.8945**	**26.87**	0.3324	**63.23**	**56.14**
AnaGAN	100	0.8865	26.45	0.3363	67.98	55.03

Table 7
Quantitative comparison of defogging task on Foggy Cityscapes (Sakaridis et al., 2018) dataset. The symbol ↑ (↓) indicates that the larger (smaller) the value, the better the performance. The best value is highlighted in bold.

Method	α	SSIM↑	PSNR↑	LPIPS↓	FID↓	mIoU↑
AnaGAN	1.0	0.8797	26.35	0.3367	77.98	54.87
AnaGAN	1.5	**0.8967**	**26.92**	**0.3246**	68.24	54.32
AnaGAN	2.0	0.8945	26.87	0.3324	**63.23**	**56.14**
AnaGAN	5.0	0.8927	26.52	0.3354	67.93	54.36

Table 8
Quantitative FID comparison of using different values of α on photo-to-caricature and night-to-day translation tasks. Lower is better.

Method	α	Photo-to-caricature	Night-to-day
AnaGAN	1.0	61.76	84.21
AnaGAN	1.5	51.72	**73.67**
AnaGAN	2.0	**47.27**	75.14
AnaGAN	5.0	54.19	87.82

makes it very difficult for the model to learn reasonable and effective translation functions. There is a tradeoff between translation performance and sampling efficiency. Finally, we have to admit the selection of k is dependent on the vision translation task and the dataset. It requires some empirical priors to do the hyper-parameter selection.

Choice of α. We also explored the quantitative results of choosing different values of α on the above three vision translation tasks. The quantitative comparison of using different α is reported in Table 7 and Table 8, respectively. $\alpha = 2.0$ leads to the highest mIoU score in Table 7, which indicates the translated images could better serve for the downstream semantic segmentation task. In Table 8, we achieved the best translation performance when $\alpha = 2.0$ on photo-to-caricature task while $\alpha = 1.5$ on night-to-day task. An approximate α could promote the translation performance by enhancing the analogical learning ability. With the dual analogical adversarial loss, our method can better utilize the content similarity and relevant relationships between samples and alleviate the influence of the negative samples.

Influence of negative samples. In this paper, we argue that the samples with content mismatch could result in the performance degradation of the unpaired I2I task. We have designed corresponding defogging experiments on Foggy Cityscapes (Sakaridis et al., 2018) dataset. In detail, we choose CycleGAN as the baseline backbone and the original CycleGAN sample (x, y) randomly form

Table 9
Quantitative comparison of defogging task on Foggy Cityscapes (Sakaridis et al., 2018) dataset. The experiments are conducted under image resolution 1024×512.

Method	SSIM↑	PSNR↑	LPIPS↓	FID↓	mIoU↑
CycleGAN (Zhu et al., 2017)	**0.8789**	**25.91**	**0.3612**	**72.12**	**48.15**
CycleGAN-Neg	0.8745	25.56	0.3678	83.76	43.76

$\mathbb{X}$ and $\mathbb{Y}$, separately. Based on CycleGAN, for each x, we formulate a negative sample subset $\mathbb{Y}_n(y_n|x)$ with k samples $\{y_n^i|x\}_1^k$ from $\mathbb{Y}$ with most irrelevant content features. Similarly, for y, we formulate a negative sample subset $\mathbb{X}_n(x_n|y)$. (x, y_n) and (y, x_n) are adopted for unpaired training following the same experimental setup of CycleGAN. We denote this modified CycleGAN as CycleGAN-Neg. The quantitative results of the original CycleGAN and CycleGAN-Neg are reported in Table 9. As illustrated, when CycleGAN is only optimized by the sampled data with a content mismatch, there would be visible performance degradation.

Future Work: *End-to-end Manner*: in our paper, we adopt an offline pre-trained model to compute the content similarity to formulate the analogical learning. The metric learning network is not optimized with the unpaired I2I model through the training process. We target to train a learnable engine to select the efficient samples for the adversarial training and optimize the whole framework in an end-to-end manner, while the losses of different stages can be transferred to each other task. With the joint training, the analogical retrieval and the unpaired I2I could benefit mutually. We leave this as our future work.

5. Conclusion

In this paper, to address the content mismatch caused by the random sampling strategy used in previous unpaired I2I methods, we introduced a novel analogical learning-based generative adversarial network method termed "AnaGAN" to boost the performance of unpaired I2I translation. With an effective metric-based sampling strategy, the model could utilize the content relationships between training samples to achieve more effective adversarial training. Besides this, we have also designed a dual analogical adversarial loss to force the model to learn from the effective samples and alleviate the influence of negative irrelevant samples. Comprehensive experiments and ablation studies have been conducted to illustrate the superior performance of the proposed method. Our method is also a generic solution and can be extended to other methods can achieve a performance gain.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Number 62171419, the

finance science and technology Q19 project of 630 Hainan province of China under Grant Number ZDKJ202017 and the Natural Science Foundation of Shandong Province of China under Grant Number ZR2021LZH005.

References

Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., & Courville, A. C. (2018). Augmented cycleGAN: Learning many-to-many mappings from unpaired data. In *Proceedings of machine learning research*: *vol. 80*, *Proceedings of the international conference on machine learning* (pp. 195–204). PMLR.

Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., & Van Gool, L. (2019). Night-to-day image translation for retrieval-based localization. In *International conference on robotics and automation* (pp. 5958–5964). IEEE.

Bau, D., Zhu, J., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., et al. (2019). GAN dissection: Visualizing and understanding generative adversarial networks. In *International conference on learning representations*.

Bellet, A., Habrard, A., & Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. arXiv preprint arXiv:1306.6709.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *European conference on computer vision* (pp. 850–865). Springer.

Borji, A. (2019). Pros and cons of GAN evaluation measures. *Computer Vision And Image Understanding, CVIU*, *179*, 41–65.

Cambria, E., Gastaldo, P., Bisio, F., & Zunino, R. (2015). An ELM-based model for affective analogical reasoning. *Neurocomputing*, *149*, 443–455.

Campo, M., Espinoza, J., Rieger, J., & Taliyan, A. (2018). Collaborative metric learning recommendation system: Application to theatrical movie releases. arXiv preprint arXiv:1803.00202.

Cao, K., Liao, J., & Yuan, L. (2018). Carigans: Unpaired photo-to-caricature translation. arXiv preprint arXiv:1811.00222.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems* (pp. 2172–2180).

Chen, K., & Forbus, K. (2021). Visual relation detection using hybrid analogical learning. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 1* (pp. 801–808).

Chen, Y., Pan, Y., Yao, T., Tian, X., & Mei, T. (2019). Mocycle-GAN: Unpaired video-to-video translation. In *Proceedings of ACM international conference on multimedia* (pp. 647–655).

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., & Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE conference on computer vision and pattern recognition* (pp. 8789–8797). IEEE.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *IEEE conference on computer vision and pattern recognition* (pp. 3213–3223). IEEE.

Danielsson, P. E. (1980). Euclidean distance mapping. *Computer Graphics And Image Processing*, *14*(3), 227–248.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics And Intelligent Laboratory Systems*, *50*(1), 1–18.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Duan, Y., Zheng, W., Lin, X., Lu, J., & Zhou, J. (2018). Deep adversarial metric learning. In *IEEE conference on computer vision and pattern recognition* (pp. 2780–2789). IEEE.

Elgammal, A., Duraiswami, R., & Davis, L. S. (2003). Probabilistic tracking in joint feature-spatial spaces. In *IEEE conference on computer vision and pattern recognition, vol. 1* (p. I). IEEE.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Gentner, D. (2006). Analogical reasoning, psychology of. In *Encyclopedia of cognitive science*. Wiley Online Library.

Gentner, D., & Smith, L. A. (2013). Analogical learning and reasoning. In *The Oxford handbook of cognitive psychology* (pp. 668–681). NY: Oxford University Press New York.

Gomez, R., Liu, Y., Nadai, M. D., Karatzas, D., Lepri, B., & Sebe, N. (2020). *Proceedings of ACM international conference on multimedia*.

Gong, R., Dai, D., Chen, Y., Li, W., & Van Gool, L. (2020). Analogical image translation for fog generation. arXiv preprint arXiv:2006.15618.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems* (pp. 6626–6637).

Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition* (pp. 84–92). Springer.

Hu, J., Lu, J., & Tan, Y. (2014). Discriminative deep metric learning for face verification in the wild. In *IEEE conference on computer vision and pattern recognition* (pp. 1875–1882). IEEE.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *European conference on computer vision* ECCV, (pp. 172–189).

Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition* (pp. 5967–5976). IEEE.

Jiang, L., Zhang, C., Huang, M., Liu, C., Shi, J., & Loy, C. C. (2020). Tsit: A simple and versatile framework for image-to-image translation. In *European conference on computer vision* (pp. 206–222). Springer.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711). Springer.

Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry*, *11*(9), 1066.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of machine learning research*: *vol. 70*, *Proceedings of international conference on machine learning*. PMLR.

Kim, J., Kim, M., Kang, H., & Lee, K. (2020). U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International conference on learning representations*. ICLR.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *16*(7), 373–381.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE conference on computer vision and pattern recognition* (pp. 105–114). IEEE.

Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M., & Yang, M. H. (2018). Diverse image-to-image translation via disentangled representations. In *European conference on computer vision* (pp. 35–51).

Li, Z., & Tang, J. (2015). Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions On Multimedia, TMM*, *17*(11), 1989–1999.

Li, N., Zheng, Z., Zhang, S., Yu, Z., Zheng, H., & Zheng, B. (2018). The synthesis of unpaired underwater images using a multistyle generative adversarial network. *IEEE Access*, *6*, 54241–54257.

Liu, M., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In *Advances in neural information processing systems* (pp. 700–708).

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., et al. (2019). STGAN: a unified selective transfer network for arbitrary image attribute editing. In *IEEE conference on computer vision and pattern recognition* (pp. 3673–3682). IEEE.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are GANs created equal? A large-scale study. In *Advances in neural information processing systems* (pp. 698–707).

Luo, Y., Liu, P., Guan, T., Yu, J., & Yang, Y. (2020). Adversarial style mining for one-shot unsupervised domain adaptation. In *Advances in neural information processing systems, vol. 33* (pp. 20612–20623).

Milford, M. J., & Wyeth, G. F. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE international conference on robotics and automation* (pp. 1643–1649). IEEE.

Mishra, A., Rai, S. N., Mishra, A., & Jawahar, C. (2016). IIIT-CFW: A benchmark database of cartoon faces in the wild. In *European conference on computer vision workshops* (pp. 35–47).

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., et al. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal Of Cognitive Neuroscience*, *16*(2), 260–271.

Pizzati, F., Charette, R. d., Zaccaria, M., & Cerri, P. (2020). Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *IEEE conference on computer vision and pattern recognition* (pp. 2990–2998).

Prade, H., & Richard, G. (2017). Analogical proportions and analogical reasoning-an introduction. In *International conference on case-based reasoning* (pp. 16–32). Springer.

Sakaridis, C., Dai, D., Hecker, S., & Gool, L. V. (2018). Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *European conference on computer vision* (pp. 707–724).

Sanakoyeu, A., Kotovenko, D., Lang, S., & Ommer, B. (2018). A style-aware content loss for real-time hd style transfer. In *European conference on computer vision* (pp. 698–714).
Song, H. O., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *IEEE conference on computer vision and pattern recognition* (pp. 4004–4012). IEEE.
Su, J., Chu, H., & Huang, J. (2020). Instance-aware image colorization. In *IEEE conference on computer vision and pattern recognition* (pp. 7965–7974). IEEE.
Tang, H., Liu, H., Xu, D., Torr, P. H., & Sebe, N. (2021). Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions On Neural Networks And Learning Systems, TNNLS*.
Tao, X., Hong, X., Shi, W., Chang, X., & Gong, Y. (2020). Analogy-detail networks for object recognition. *IEEE Transactions On Neural Networks And Learning Systems, TNNLS*.
Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical reasoning in the classroom: Insights from cognitive science. *Mind, Brain, and Education*, *9*(2), 100–106.
Wang, Z., Du, B., & Guo, Y. (2019). Domain adaptation with neural embedding matching. *IEEE Transactions On Neural Networks And Learning Systems, TNNLS*, *31*(7), 2387–2397.
Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional GANs. In *IEEE conference on computer vision and pattern recognition* (pp. 8798–8807).
Wang, T., Liu, M., Zhu, J., Yakovenko, N., Tao, A., Kautz, J., et al. (2018). Video-to-video synthesis. In *Advances in neural information processing systems* (pp. 1152–1164).
Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z., & Shen, H. T. (2020). Universal weighting metric learning for cross-modal matching. In *IEEE conference on computer vision and pattern recognition* (pp. 13005–13014).
Yi, R., Liu, Y.-J., Lai, Y. K., & Rosin, P. L. (2019). Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *IEEE conference on computer vision and pattern recognition* (pp. 10743–10752).
Yi, Z., Zhang, H., Tan, P., & Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE international conference on computer vision* (pp. 2849–2857).
Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE conference on computer vision and pattern recognition* (pp. 586–595).
Zhang, Y., & Lu, H. (2018). Deep cross-modal projection learning for image-text matching. In *European conference on computer vision* (pp. 686–701).
Zhang, H., Sindagi, V., & Patel, V. M. (2017). Image de-raining using a conditional generative adversarial network. arXiv preprint arXiv:1701.05957.
Zheng, Z., Wang, C., Yu, Z., Wang, N., Zheng, H., & Zheng, B. (2019). Unpaired photo-to-caricature translation on faces in the wild. *Neurocomputing*, *355*, 71–81.
Zheng, Z., Wang, C., Yu, Z., Zheng, H., & Zheng, B. (2018). Instance map based image synthesis with a denoising generative adversarial network. *IEEE Access*, *6*, 33654–33665.
Zheng, Z., Wu, Y., Han, X., & Shi, J. (2020). Forkgan: Seeing into the rainy night. In *European conference on computer vision* (pp. 155–170).
Zheng, Z., Yu, Z., Wu, Y., Zheng, H., Zheng, B., & Lee, M. (2021). Generative adversarial network with multi-branch discriminator for imbalanced cross-species image-to-image translation. *Neural Networks*.
Zheng, Z., Yu, Z., Zheng, H., Yang, Y., & Shen, H. T. (2021). One-shot image-to-image translation via part-global learning with a multi-adversarial framework. *IEEE Transactions On Multimedia, TMM*.
Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision* (pp. 2242–2251). IEEE.