Domain Adaptive Image-to-image Translation

Ying-Cong Chen[1] Xiaogang Xu[1] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore
yingcong.ian.chen@gmail.com, {xgxu,leojia}@cse.cuhk.edu.hk

Abstract

Unpaired image-to-image translation (I2I) has achieved great success in various applications. However, its generalization capacity is still an open question. In this paper, we show that existing I2I models do not generalize well for samples outside the training domain. The cause is twofold. First, an I2I model may not work well when testing samples are beyond its valid input domain. Second, results could be unreliable if the expected output is far from what the model is trained. To deal with these issues, we propose the Domain Adaptive Image-To-Image translation (DAI2I) framework that adapts an I2I model for out-of-domain samples. Our framework introduces two sub-modules – one maps testing samples to the valid input domain of the I2I model, and the other transforms the output of I2I model to expected results. Extensive experiments manifest that our framework improves the capacity of existing I2I models, allowing them to handle samples that are distinctively different from their primary targets.

1. Introduction

In recent years, unpaired image-to-image translation (I2I) [44, 8, 21, 23] has attracted quite a lot of interest in computer vision, graphics, and machine learning. Given images of certain domain $\boldsymbol{A}^-$, it learns a mapping $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ to another domain $\boldsymbol{A}^+$ without requiring any paired information. It can serve a wide range of applications, including image attribute manipulation [8], style transfer [44], data augmentation [11], domain adaptation [14], to name a few.

Despite great success, these approaches could be less effective when the testing images are not in the same domain as the training set. Specifically, when a model $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ is trained with domain $\boldsymbol{A}$ ($\boldsymbol{A}=\boldsymbol{A}^-\cup\boldsymbol{A}^+$), it may not perform well when applied on another domain $\boldsymbol{B}$. Fig. 1 shows an example of applying a *neutral* $\mapsto$ *smile* model trained on human faces to a cat face. Intuitively, the process of *getting smile* should include raising the corner of a mouth and changing other smiling related muscles. Human can easily imagine how this happens on a cat face, even if he/she has never seen such a smiling cat before. However, as shown in Fig. 1(b), an I2I model does not have such capacity. When an image is out of the domain $\boldsymbol{A}^-$, the model cannot modify its attribute from "$-$" to "$+$" correctly. Accordingly, it generates artifacts and changes almost no target attribute.

(a) I2I (b) I2I (on cat) (c) DAI2I

Figure 1. Applying a *neutral* $\mapsto$ *smile* I2I model on human and cat faces. The I2I model is trained on human faces. The 1st and 2nd rows are input and output respectively. (a) Result on a human face. (b) Directly applying the model on a cat face. (c) Our result.

In this paper, we propose the Domain Adaptive Image-to-image translation (DAI2I) framework to enable I2I models to handle out-of-domain samples. The *out-of-domain* here has two meanings. First, the input samples are from a new domain $\boldsymbol{B}^-$ instead of $\boldsymbol{A}^-$. As $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ is trained with $\boldsymbol{A}^-$, it may not parse information of $\boldsymbol{B}^-$ correctly. Second, in practice the expected output domain $\boldsymbol{B}^+$ may not be available during training. Take Fig. 1 as an example. Capturing many smiling cat photos is not easy. As a result, we lack essential information to define the expected output domain. Different from existing I2I tasks, modeling the output domain with GAN is infeasible, because there exists *no* real data ($\boldsymbol{B}^+$) to train the discriminator.

From the discussion above, it is clear that out-of-domain image-to-image translation is still an open problem. In this paper, we adopt two assumptions to make it tractable. The first assumption is that $\boldsymbol{A}$ and $\boldsymbol{B}$ can be translated bidirectionally. This implies $\boldsymbol{A}$ and $\boldsymbol{B}$ are semantically related; otherwise the translation, such as a chair mapping to a cat,

would be either meaningless or visually implausible. The other assumption is that the relation between $\boldsymbol{A}^-$ and $\boldsymbol{A}^+$ can be generalized to $\boldsymbol{B}^-$. Thus, even if no $\boldsymbol{B}^+$ is presented during training, there exists an imaginable counterpart based on other samples.

Based on these assumptions, we introduce two mapping functions, $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ and $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$, to our DAI2I framework, which conduct translation between domains $\boldsymbol{A}$ and $\boldsymbol{B}$. $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ serves as an *adapter* that maps target images to the valid input domain of the base I2I model $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$. $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ works as a *reconstructor* that maps the output of I2I model $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ to the expected target domain.

Besides, we introduce a *perceptual analogy loss* that enables our model to leverage the relation between $\boldsymbol{A}^-$ and $\boldsymbol{A}^+$ to define the expected output domain $\boldsymbol{B}^+$. This allows training without any sample of $\boldsymbol{B}^+$. Finally, we propose a style feature extraction and adaptation scheme for the reconstructor to handle input images of highly diverse styles. Our total contribution is the following.

- We make the first attempt to address out-of-domain image-to-image translation.
- We propose the Domain Adaptive Image-To-Image translation (DAI2I) framework. Our model generalizes a base image-to-image translation model to handle images of significant different styles.
- We conduct extensive experiments to demonstrate the effectiveness of our model.

2. Related Work

Image-to-image Translation Unpaired Image-to-image translation (I2I) [44, 8, 21, 23] aims to translate images from domain $\boldsymbol{A}^-$ to $\boldsymbol{A}^+$. CycleGAN [44], DualGAN [42] and DiscoGAN [19] are pioneering methods. Following methods improved quality and flexibility, including addressing the domain scalability issue [8, 43], multi-modality issue [21, 17], discreteness issue [6, 31], etc. It is still difficult to explore generalization capacity. Almost all methods assume testing and training samples are in the same domain. Our framework is complementary to these methods by handling out-of-domain samples.

Recently, OST [2] and FUNIT [24] were proposed to address the generality issue in image-to-image translation. Specifically, OST [2] allows learning $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ when $\boldsymbol{A}^-$ contains very few samples. This is different from our approach since our objective is to learn $F_{\boldsymbol{B}^-\mapsto\boldsymbol{B}^+}(\cdot)$ that conducts translation in a new domain. FUNIT [24] learns a model that maps a source image to an unknown target class by presenting few target samples during testing. It learns $\boldsymbol{F}_{\boldsymbol{A}\mapsto\boldsymbol{B}^+}(\cdot)$ such that $\boldsymbol{F}_{\boldsymbol{A}\mapsto\boldsymbol{B}^+}(A; B_1^+, B_2^+ \cdots, B_n^+) \in \boldsymbol{B}^+$, where $B_1^+, B_2^+ \cdots, B_n^+$ are samples of $\boldsymbol{B}^+$ given during testing. This method is also inherently different from ours, since we do not assume that $\boldsymbol{B}^+$ is available during either training or testing.

Domain Adaptation Domain adaptation (DA) aims to transfer knowledge from a label-rich source domain to a label-scarce target of interest. A large amount of methods have been proposed, including instance re-weighting [15, 9], covariance alignment [35, 36], Maximum Mean Discrepancy [30, 26], pixel-level adaptation [14, 29], etc. Our method can be categorized in it since it adapts the model trained from a source domain $\boldsymbol{A}$ to a target domain $\boldsymbol{B}$. Yet it is different from existing approaches because it focuses on the *generation* task, while others took understanding tasks for image classification, segmentation, etc.

Image Analogy Given a pair of images A^- and A^+ along with a target B^-, image analogy [12, 22, 7, 32, 1, 40] aims to synthesize a new image B^+ such that B^+ relates to B^- in the same way as A^+ relates to A^-. This basic idea has motivated the perceptual analogy loss of our model. However, our work is fundamentally different from image analogy for two reasons. First, paired data is required for most existing approaches [12, 7, 32, 1, 40], while it is not needed in our model. Second, our work can handle high-level change, while most existing ones focus on low-level modification [12, 22, 7, 1, 40] in style transfer, image filtering, texture synthesis, etc.

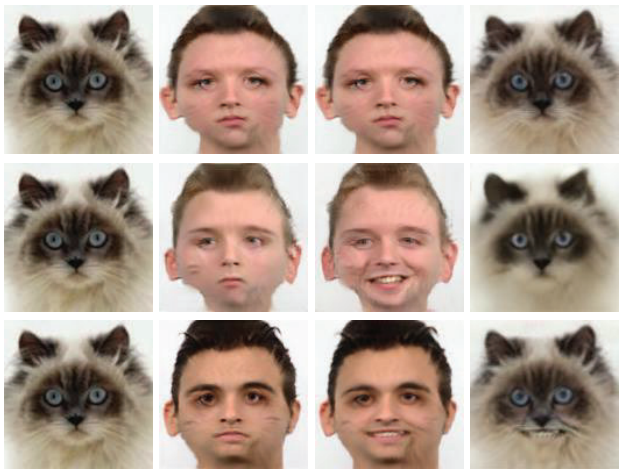
3. Proposed Method

Given a trained image-to-image translation model $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$, which modifies certain attribute (e.g., smiling) for domain $\boldsymbol{A}$ (e.g., human faces), our objective is to transform $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ to $F_{\boldsymbol{B}^-\mapsto\boldsymbol{B}^+}(\cdot)$ so as to handle samples of another domain $\boldsymbol{B}$ (e.g., cat faces). $F_{\boldsymbol{B}^-\mapsto\boldsymbol{B}^+}(\cdot)$ is expected to translate $\boldsymbol{B}^-$ (e.g., common cat faces) to $\boldsymbol{B}^+$ (e.g., smiling cat faces) without introducing other irrelevant changes. We assume that images of $\boldsymbol{A}^-$, $\boldsymbol{A}^+$ and $\boldsymbol{B}^-$ are available during training. $\boldsymbol{B}^+$ is not used for training, since in practice it could be hard to obtain.

In this paper, all mapping functions, including $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$, $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ and $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ are implemented by convolutional neural networks. We use bold font, such as $\boldsymbol{A}^-$ and $\boldsymbol{A}^+$, to denote image collection of certain domains. A^- and A^+ in normal font, contrarily, denote samples of the corresponding collections. The subscripts "-" and "+" refer to the attribute labels. For simplicity, we assumes the base I2I model $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ only changes one attribute in this section. As shown in the experiments, our model can handle multiple attributes simultaneously with multi-domain translation models like StarGAN [8].

3.1. Analysis

We take the task of turning a neutral cat face into a smiling one for illustration. The absence of $\boldsymbol{B}^+$



(a) $\boldsymbol{B}^-$ (b) $\boldsymbol{A}^-$ (c) $\boldsymbol{A}^+$ (d) $\boldsymbol{B}^+$

Figure 2. Illustration of the difficulties of training our DAI2I model. (a) shows the input images. (b) and (c) are the output of $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ and $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$. (d) are the final output. The 1st row illustrates the case that $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ fails to translate the label from $\boldsymbol{A}^-$ to $\boldsymbol{A}^+$. The 2nd row illustrates the case that $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ fails to translate $\boldsymbol{A}^+$ to $\boldsymbol{B}^+$. The 3nd row illustrates the results of our proposed method. It shows that the cat gets smiling after processed by $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$, $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ and $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$. Noted that we only care about the quality of the final results (d) rather than intermediate (b) and (c) since (b) and (c) are invisible to users.

(i.e., smiling cats) prevents us from directly learning $F_{\boldsymbol{B}^-\mapsto\boldsymbol{B}^+}(\cdot)$. Nevertheless, we learn a pair of mapping functions $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ and $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ that conduct translation between $\boldsymbol{A}$ (human faces) and $\boldsymbol{B}$ (cat faces). In this setting, we finally obtain $F_{\boldsymbol{B}^-\mapsto\boldsymbol{B}^+}(\cdot)$ by sequentially stacking $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$, $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ and $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$, i.e., $F_{\boldsymbol{B}^-\mapsto\boldsymbol{B}^+}(\cdot) = F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot))$.

This first turns a neutral cat face to a human one, then change its expression with the base model $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}$, and finally transform it back to a cat face. Here, $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ can be viewed as pixel-level adaptation like that of [14], which coverts *invalid* input samples to *valid* ones such that they can be processed by $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$. On the other hand, $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ transforms the *non-target* output produced by $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ to the *target* ones.

Training of $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ and $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ is not trivial. Our first attempt is to use CycleGAN [44] to train $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ and $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$. However, the result is unsatisfactory with several reasons. First, $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ may not always translate samples of $\boldsymbol{B}^-$ (neutral cat faces) to perfectly match $\boldsymbol{A}^-$ (neutral human faces). In this case, $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ does not work well given that $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(B)$ is out of its valid input domain. This is illustrated in Fig. 2 in the 1st row – the expression of human face is not changed.

Second, even when $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot))$ works correctly, we found that $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ always produces samples of $\boldsymbol{B}^-$ instead of $\boldsymbol{B}^+$. As shown in the Fig. 2 in the 2nd row, even if the expression of human face is modified, the expression of cat face is untouched.

Replacing CycleGAN with other I2I methods does not solve this problem. It is because the adversarial loss of CycleGAN [44] imposes $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ to generate samples that is indistinguishable from samples of $\boldsymbol{B}$. Note that $\boldsymbol{B}^+$ (smiling cat faces) are unreachable in our setting. This loss would encourage $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ to produce samples in $\boldsymbol{B}^-$ (neutral cat faces), which impedes the model from modifying the target attribute.

Removing the adversarial loss on $\boldsymbol{B}$ also does not address the issue. Generating novel images without seeing any examples of its kind is very difficult. In the following, we present our solution, which leads to intriguing and inspiring results as shown in Fig. 2 (3nd row).

3.2. Domain-Adaptive Image Translation Model

To address the aforementioned problems, we introduce several loss functions to train the adapter network $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ and the reconstructor network $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$. Note that we assume $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ is already trained, and its weight is kept fixed throughout the optimization process.

Adversarial Loss We use an adversarial loss to enforce $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ to translate images of domain $\boldsymbol{B}$ to domain $\boldsymbol{A}$. Here LSGAN [27] is adopted, which is formulated as

$$\min_{\boldsymbol{D}} \boldsymbol{\mathcal{L}}_{GAN_{\boldsymbol{D}}} = \mathbb{E}(||\boldsymbol{D}(\hat{A})||_2) + \mathbb{E}(||\boldsymbol{D}(A) - 1||_2), \quad (1)$$

$$\min_{F_{\boldsymbol{B}\mapsto\boldsymbol{A}}} \boldsymbol{\mathcal{L}}_{GAN} = \mathbb{E}(||\boldsymbol{D}(\hat{A}) - 1||_2), \quad (2)$$

where $A \in \boldsymbol{A}^-$, $B \in \boldsymbol{B}$, $\hat{A} = F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(B)$, $\mathbb{E}(\cdot)$ denotes computing the mean over a batch, $\boldsymbol{D}(\cdot)$ is the discriminator parameterized by a neural network. Spectral normalization [28] is adopted in $\boldsymbol{D}(\cdot)$. This turns $\hat{A}$ towards a valid input of $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$.

One may concern that another adversarial loss to train $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ is needed so that its output domain is constrained to $\boldsymbol{B}$. Although this is a common practice in bidirectional I2I models [44, 21, 23], we do not incorporate this loss because of the absence of $\boldsymbol{B}^+$. We found that this loss hinders our model from modifying the target attributes.

Adaptation Loss In practice, optimizing the adversarial loss (1) is not easy due to the $\min\max$ formulation. When $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ is not perfectly optimized, it may not produce valid input to $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$. To remedy this, we propose explicitly enforcing $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ to generate samples that can be effectively processed by $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$, formulated as

$$\min_{F_{\boldsymbol{B}\mapsto\boldsymbol{A}}} \boldsymbol{\mathcal{L}}_{ADA} = \mathbb{E}[-\log(\mathcal{C}(\hat{A}^+))], \quad (3)$$

where $\hat{A}^+ = F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(\hat{A}) = F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B))$, and $\mathcal{C}(\cdot)$ is a classification network that maps $\boldsymbol{A}^-$ to 0 and $\boldsymbol{A}^+$ to 1. Similar to $F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(\cdot)$, $\mathcal{C}(\cdot)$ is pretrained with $\boldsymbol{A}$ and keeps fixed during optimization. The idea here is to enforce $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\cdot)$ to produce samples whose target attribute can be successfully translated by $F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(\cdot)$.

Note that $\boldsymbol{\mathcal{L}}_{GAN}$ and $\boldsymbol{\mathcal{L}}_{ADA}$ work cooperatively to encourage $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\cdot)$ to map to the *valid* input domain of $F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(\cdot)$. $\boldsymbol{\mathcal{L}}_{GAN}$ guides training in the sample level, encouraging $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B)$ to be indistinguishable from samples of $\boldsymbol{A}^-$; while $\boldsymbol{\mathcal{L}}_{ADA}$ supervises $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B)$ in the model level, making it adaptively fit the pretrained network $F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(\cdot)$. Thus, $F_{\boldsymbol{A}^- \mapsto \boldsymbol{A}^+}(\cdot)$ can translate the attribute of $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B)$ from "$-$" to "$+$" as expected.

Reconstruction Loss Since $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\cdot)$ is expected to be the inverse function of $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\cdot)$, we incorporate the reconstruction loss as

$$\min_{F_{\boldsymbol{B} \mapsto \boldsymbol{A}}, F_{\boldsymbol{A} \mapsto \boldsymbol{B}}} \boldsymbol{\mathcal{L}}_{rec} = \mathbb{E}[||F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\hat{A}) - B||_1], \quad (4)$$

where $\hat{A} = F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B)$. This loss enforces $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B)$ to be invertible with $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\cdot)$, as required in our model. It also provides regularization of $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\cdot)$, making $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(B)$ semantically relevant to B.

Perceptual Analogy Loss Note that $\boldsymbol{\mathcal{L}}_{rec}$ alone is not sufficient to model the relation of $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\cdot)$ and $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\cdot)$. It only encourages $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\hat{A}) = B$, which does not imply $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\hat{A}^+) = B^+$, where B^+ denotes the expected translated version of B. Therefore, it is necessary to explicitly model the relation between $\hat{A}^+$ and B^+ to ensure that $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\hat{A}^+)$ leads to the correct result. This is challenging because $\boldsymbol{B}^+$ is not available during training.

Inspired by image analogies [12], we propose a perceptual analogy loss

$$\min_{F_{\boldsymbol{B} \mapsto \boldsymbol{A}}, F_{\boldsymbol{A} \mapsto \boldsymbol{B}}} \boldsymbol{\mathcal{L}}_{PA} = \mathbb{E}[||\mathcal{V}_{\boldsymbol{B}} - \alpha \mathcal{V}_{\boldsymbol{A}}||_1], \quad (5)$$

where $\mathcal{V}_{\boldsymbol{B}} = \Phi(\hat{B}^+) - \Phi(\hat{B})$, $\mathcal{V}_{\boldsymbol{A}} = \Phi(\hat{A}^+) - \Phi(\hat{A})$, and $\hat{B}^+ = F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\hat{A}^+)$. $\Phi(\cdot)$ is a latent space that encodes semantic information of images, and α is a scalar to amplify or reduce the scale of $(\Phi(\hat{A}^+) - \Phi(\hat{A}))$.

The rationale is the following. As $\hat{A}$ is semantically related to B, the relation between $\hat{B}^+$ and B is supposedly analogous with that between $\hat{A}^+$ and $\hat{A}$. Note that the relation here is represented as linear difference in the latent space $\Phi(\cdot)$. The underlying assumption is that $\Phi(\cdot)$ unfolds images to a flat manifold, where the change of target attribute becomes linear.

Choice of $\Phi(\cdot)$ The latent space $\Phi(\cdot)$ plays a key role in our model. To an extreme, if $\Phi(\cdot)$ is the RGB space, optimizing $\boldsymbol{\mathcal{L}}_{PA}$ encourages our model to simply copy $(\hat{A}^+ - \hat{A}^-)$ to B, which could look artificial, because most semantic attribute changes are actually nonlinear in this space.

Bengio *et al.* [3] showed that a well trained CNN could unfold natural images to a space where semantic changes become linear. Following work [5, 39] also indicates that high-level attribute changes can be achieved by linearly interpolating in the ImageNet pretrained deep feature space. This suggests that by *seeing* a large volume of images, a deep neural network could unfold natural images to a space where many semantic changes are linear and the assumption of Eq. (5) is mostly true.

We follow the setting of [5, 39] and use ReLU3_1, ReLU4_1 and ReLU5_1 features of VGG-19 [34] to form $\Phi(\cdot)$. It works well in our experiments. We also believe it is possible to find/learn other space for certain specific attributes, which will be explored in our future work.

Calibration of Domain Shift As $\boldsymbol{A}$ and $\boldsymbol{B}$ are of different domains, $\Phi(B)$ and $\Phi(\hat{A})$ may also suffer from distribution shift. We remedy this by introducing domain-specific batch normalization [4], i.e.,

$$\begin{aligned} \Phi(B)_i &= \frac{\Phi(B)_i}{\sigma_i^B}, \quad \Phi(\hat{B}^+)_i = \frac{\Phi(\hat{B}^+)_i}{\sigma_i^B}, \\ \Phi(\hat{A})_i &= \frac{\Phi(\hat{A})_i}{\sigma_i^A}, \quad \Phi(\hat{A}^+)_i = \frac{\Phi(\hat{A}^+)_i}{\sigma_i^A}, \end{aligned} \quad (6)$$

where $\Phi(\cdot)_i$ denotes the ith channel of $\Phi(\cdot)$. σ_i^A and σ_i^B are the standard variation of $\Phi(\hat{A})_i$ and $\Phi(\hat{B})_i$ respectively. They are computed by the moving average scheme. Note that we do not normalize the means because they are canceled out in Eq. (5). Despite its simplicity, this normalization scheme can largely improve the quality of our model. We have also tried other normalization including Batch Whitening [33] and CORAL [35]. They do not lead to significant improvement, albeit costing much computation.

The final loss functions of $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}$ and $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}$ are

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{F_{\boldsymbol{B} \mapsto \boldsymbol{A}}} &= \lambda_{GAN} \boldsymbol{\mathcal{L}}_{GAN} + \boldsymbol{\mathcal{L}}_{ADA} + \boldsymbol{\mathcal{L}}_{rec} + \boldsymbol{\mathcal{L}}_{PA}, \\ \boldsymbol{\mathcal{L}}_{F_{\boldsymbol{A} \mapsto \boldsymbol{B}}} &= \boldsymbol{\mathcal{L}}_{rec} + \boldsymbol{\mathcal{L}}_{PA}, \end{aligned} \quad (7)$$

where λ_{GAN} is set as 0.1 with cross-validation.

3.3. Handling Multiple Target Domains

Note that the above solution works best when target images come from one domain. If target images are from different domains with highly different appearance (e.g., oil painting, sketches and cats, as shown in Fig. 3(a)), the above model may fail. The reason is that $F_{\boldsymbol{B} \mapsto \boldsymbol{A}}(\cdot)$ could map target images of different styles to an unitary domain $\boldsymbol{A}$, which tends to suppress the original style information. This makes it hard to reconstruct $F_{\boldsymbol{A} \mapsto \boldsymbol{B}}(\cdot)$ since style information is needed in this phase.

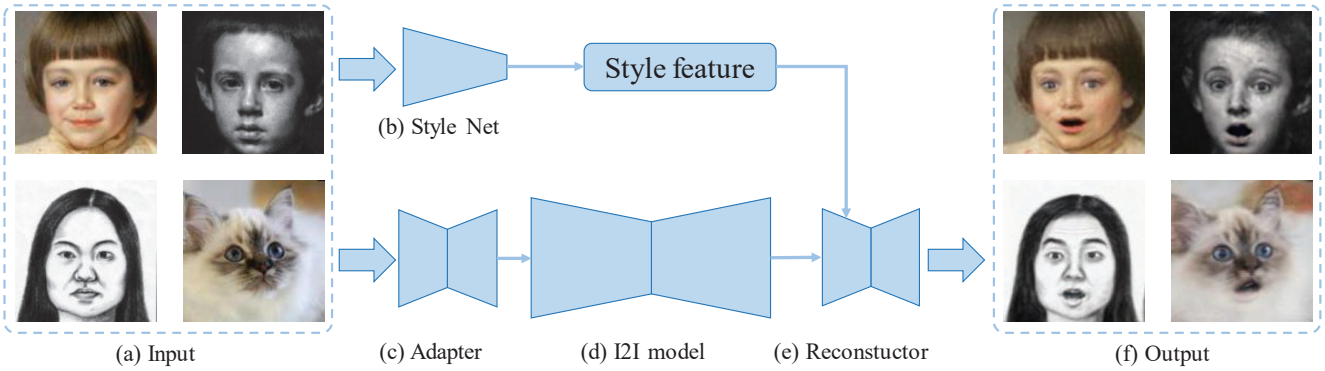


Figure 3. Illustration of our framework. (a) is the input image. (b) is the style network $S(\cdot)$. It extracts style feature, which controls affine parameters of AdaIN layers in the reconstructor network $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$. (c) is the adapter network $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$, which adapts a target image to the valid input domain of the base I2I model. (d) is the base I2I model $F_{\boldsymbol{A}-\mapsto\boldsymbol{A}+}(\cdot)$, which maps a neural expression image towards "surprise" expression. (e) is the reconstructor network $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$. (f) contains results by translating attribute of target images. Note that even if the four input images are of different styles/categories, the target attribute is still modified successfully.

To address this problem, we additionally incorporate style network $S(\cdot)$ that maps an input target image to a style feature, which is a $1\times1\times c$ vector. Then, we add an adaptive instance normalization (AdaIN) [16] layer after each convolutional layer of $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ (except for the output layers). The affine parameters of these AdaIN layers are controlled by this style feature. Specifically, for the i-th convolutional layer, the AdaIN layer works as

$$y_i = \gamma_{S(x_i)}\left(\frac{x_i - \mu(x_i)}{\sigma(x_i)}\right) + \beta_{S(x_i)}, \tag{8}$$

where x_i and y_i refer to the input and output of the AdaIN layer, $\mu(x_i)$ and $\sigma(x_i)$ denote the mean and variance of x_i across the spatial dimensions, $\gamma_{S(x_i)}$ and $\beta_{S(x_i)}$ are parameters of the AdaIN layers, which are implemented by linearly projecting $S(x_i)$ to match the channel number of x_i.

This style feature extraction-adaptation scheme provides a skip path for $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ to access the style information. Thus decent reconstruction can be achieved. In our supplementary material, we visualize the learned style feature, which suggests that it captures appearance information of the input image. The whole framework is shown as Fig. 3.

4. Experiments

4.1. Ablation Study

We first evaluate each component in our framework quantitatively. CelebA [25] contains 200K celebrity images, each with 40 attribute labels. We use these image to form domain $\boldsymbol{A}$. Each attribute can be used to divide $\boldsymbol{A}$ into $\boldsymbol{A}^-$ and $\boldsymbol{A}^+$. To form domain $\boldsymbol{B}$, we generate four stylized versions using the method of [18][1].

[1]We use the implementation from https://github.com/pytorch/examples/tree/master/fast_neural_style, which provides 4 pretrained models for different styles, including *candy, mosaic, udnie* and *rain-princess*.

Separation of training and testing sets follows the setting of [25]. For the stylized domain, *only* samples with negative labels are involved during training; while for the original image domain $\boldsymbol{A}$, all training samples are incorporated. Thus, the model cannot see any stylized images of positive labels during training. In our experiments, we use attributes 'Smiling', 'Smaller Eyes', 'Mustache' and 'Mouth Open' to evaluate our approach.

Evaluation Metrics We introduce the translation accuracy (ACC) to quantify how effective a model modifies the label of a target sample from "-" to "+", which is defined as

$$ACC = \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\boldsymbol{B}}(\hat{B}_i^+), \tag{9}$$

where N is the number of testing samples, and $\hat{B}_i^+$ is the ith generated sample. $\mathcal{C}_{\boldsymbol{B}}(\cdot)$ is a classifier trained on stylized images, which outputs 1 for "+" and 0 for "−".

Besides modifying the target attribute, the model should not introduce much disturbance to the input sample. Thus, we also use the Frchet Inception Distance (FID) [13] to measure the quality of the generated images. FID scores differentiate between generated and real samples. They are sensitive to various disturbance [13], such as noise, blurring, and swirling.

Effectiveness of Each Component We evaluate each term using the stylized CelebA dataset described above. The α in Eq. (5) is set to 1, and its influence will be discussed later. Table 1 compares ACC and FID by ablating each component in our model.

The 1st column ("Base I2I") presents a baseline that directly applies the base I2I model trained on $\boldsymbol{A}$ (original image) to $\boldsymbol{B}$ (stylized image). This does not yield good-quality results. Our approach ("**DAI2I**") consistently and significantly improves the performance of "I2I" and achieves

Attribute	Metrics	Base I2I	w/o $\mathcal{L}_{ADA}$	w/o $\mathcal{L}_{rec}$	w/o $\mathcal{L}_{PA}$	w/o BN	w/o AdaIN	**DAI2I**
Smiling	ACC(%)	15.3	0.2	87.9	15.3	92.4	95.3	96.1
	FID	56.6	10.8	41.9	56.6	41.4	18.5	14.9
Smaller Eyes	ACC(%)	53.4	2.6	77.2	77.2	47.7	80.5	80.1
	FID	130.7	7.3	72.8	50.1	7.9	11.3	8.6
Mustache	ACC(%)	12.1	1.6	75.9	68.4	52.5	96.2	96.9
	FID	178.9	16.7	26.4	88.1	14.6	15.2	14.2
Mouth Open	ACC(%)	88.3	1.8	31.7	37.4	51.0	90.1	90.3
	FID	65.7	8.4	25.9	49.9	6.0	7.4	6.2

Table 1. Evaluating our approach with stylized CelebA data. "Base I2I" means applying StarGAN to the stylized images, which is the baseline. The 2nd ("w/o $\mathcal{L}_{ADA}$") - 6th ("**DAI2I**") columns report the performance on variants of our framework on domain $\boldsymbol{B}$. "w/o $\mathcal{L}_{ADA}$", "w/o $\mathcal{L}_{rec}$" and "w/o $\mathcal{L}_{PA}$" denote ablating Eqs. (3), (4) and (5) respectively, while keeping other parts intact. "no BN" means removing distribution calibration in Eq. (6). "no AdaIN" means removing adaptive instance normalization in Section 3.3. Finally, **DAI2I** denotes our final full model. For each row, the best result is marked in red.

(a) input (b) $\alpha = 1$ (c) $\alpha = 2$ (d) $\alpha = 3$

Figure 4. Results of using different α on stylized CelebA data. Rows 1-4 correspond to Smiling, Smaller Eyes, Mustache and Mouth Open.

much higher ACC and lower FID scores.

Note that $\mathcal{L}_{ADA}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{PA}$, and distribution calibration (Eq. (6)) are all important in our DAI2I model; disabling each would cause performance drop. For example, without $\mathcal{L}_{ADA}$, ACC reduces significantly and the DAI2I model fails to change anything because $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$ does not map the input image to the valid set of the I2I model. Thus the target attribute cannot be translated successfully.

Removing $\mathcal{L}_{PA}$ causes both FID and ACC drop. This indicates that the perceptual analogy loss not only guides the model to modify the target attribute of $\boldsymbol{B}$, but also prevents false changes. We have also ablated the distribution calibration ("w/o BN"), which also causes degradation of performance. This suggests that perceptual analogy loss works better on well aligned deep features. $\mathcal{L}_{rec}$ is also useful in our model, as it provides useful regularization on $F_{\boldsymbol{A}\mapsto\boldsymbol{B}}(\cdot)$ and $F_{\boldsymbol{B}\mapsto\boldsymbol{A}}(\cdot)$. Discarding the AdaIN introduced in Section 3.3 degrades performance. Finally, removing $\mathcal{L}_{GAN}$ makes the model totally fail.

Influence of α In Eq. (5), α is used to control the scale of $(\Phi(\hat{A}^+) - \Phi(\hat{A}))$. A large α amplifies the difference of $\Phi(\hat{A}^+)$ and $\Phi(\hat{A})$, and makes the effect stronger. However, since $F_{\boldsymbol{A}^-\mapsto\boldsymbol{A}^+}(\cdot)$ may not be perfect, it may introduce subtle artifacts, which could be amplified when α increases. This is illustrated in Fig. 4. When α is too large, undesired structures may appear.

4.2. Comparison with Other Methods

4.2.1 Cross-domain Expression Manipulation

In this section, we demonstrate that our framework can handle cross-domain expression manipulation on diverse real-world data. RaFD [20] is a face dataset that contains 67 people displaying 8 expressions, including a "neutral" expression and 7 other emotional ones. This dataset serves as the source domain $\boldsymbol{A}$. The "neutral" expression forms $\boldsymbol{A}^-$, while others form $\boldsymbol{A}^+$. Three other datasets serve as domain $\boldsymbol{B}$, including a sketch dataset, an oil painting dataset, and a cat face dataset. The sketch dataset [41] contains 187 images (128 for training and 59 for testing). The oil painting dataset [21] contains 1,664 images (1,572 for training and 92 for testing). The cat face dataset [21] contains 870 images (770 for training and 100 for testing).

Expressions of these three target datasets are not biased, and are thus treated as "neutral". We use the StarGAN model [8] trained with RaFD [20] as our base I2I model. Then, we train a unified DAI2I that adapts the base I2I model for sketches, oil paintings and cat faces.

Results and Analysis We first compare our DAI2I with StarGAN, the base I2I model. As shown in Fig. 5, directly applying StarGAN trained on RaFD does not lead to satisfactory results on out-of-domain samples. In most cases, StarGAN cannot modify the target attribute correctly, and

Datasets	Methods	happy	angry	sad	contemptuous	disgusted	fearful	surprised	Overall
Sketch	Base I2I (%)	63.6	21.9	21.4	28.9	9.0	14.3	3.8	26.3
	DAI2I (%)	99.3	82.2	59.2	77.2	72.2	37.3	53.0	68.7
Painting	Base I2I (%)	48.7	26.5	30.1	30.3	21.9	15.5	34.5	30.7
	DAI2I (%)	93.5	33.8	54.0	55.0	49.6	31.2	65.4	55.1
Cat	Base I2I (%)	7.2	16.6	16.4	22.0	4.5	10.9	12.7	22.9
	DAI2I (%)	74.8	44.9	32.2	27.5	28.0	21.9	58.3	43.0

Table 2. Expression recognition test on each attribute. Each column corresponds to one target attribute. Each entry reports the percentage that the chosen attribute is consistent with the target one. The better one is marked in red.

Datasets	happy	angry	sad	contemptuous	disgusted	fearful	surprised	Overall
Sketch	92.9	94.4	88.5	86.3	89.6	88.3	87.5	89.8
Painting	84.3	76.8	82.6	77.6	78.3	80.3	76.7	79.7
Cat	69.9	86.5	92.4	84.0	74.1	82.0	79.3	81.0

Table 3. Quality comparison test on each attribute. Each column corresponds to one target attribute. Each entry reports the percentage that our method is preferred by subjects. All entries are larger than 50%, suggesting that our results are consistently preferred by subjects.

yet introduces strong artifacts. In comparison, our DAI2I model successfully modifies the target attributes without bringing much irrelevant change. More results are presented in our supplementary material.

In addition to visual comparison, we also conduct user study on the Amazon Mechanical Turk, including expression recognition and quality comparison tests. Each set of Tables 2-3 is computed by 2,500 comparisons. In the expression recognition test, given an edited image, subjects are asked to select the best-matched expression from 7 possible candidates. In Table 2, we report the percentages that the chosen expression is the same as expected.

In the quality comparison test, subjects are given an original image and two edited ones (ours vs. StarGAN) of the same identity and the same target expression, and are asked to pick one with better quality. Table 3 reports the percentages that our approach is chosen. It shows that our approach largely outperforms the base I2I model (StarGAN) and manifests the usefulness of our model in this challenging task.

4.2.2 Cross-Domain Novel View Synthesis

Given a single 2D image, the target of novel view synthesis is to generate images from other viewpoints. Recent work [37] shows that it can be formulated as an I2I problem. In this section, we show that our framework can also handle cross-domain samples.

Datasets and settings Multi-PIE [10] contains 337 persons under 13 horizontal camera poses with 15° intervals. This dataset is used as $\boldsymbol{A}$. We take the frontal view as $\boldsymbol{A}^-$, and the $-30°$, $-15°$, $15°$ and $30°$ views as $\boldsymbol{A}^+$. To evaluate the cross-domain performance, we use the sketch [41] and oil painting [21] datasets described above as $\boldsymbol{B}$. Note that only images in frontal view are used for training and testing. CRGAN [37] trained with Multi-PIE is used as our base I2I model. We compare our DAI2I with two related approaches, i.e., CRGAN [37] and DRGAN [38].

Results and Analysis As shown in Fig. 6 (rows 2 and 3), both CRGAN [37] and DRGAN [38] do not perform well when directly applied to sketch and oil painting images. Although they successfully synthesize face photos of the target view, the color, illumination and style are different from the input images. In contrast, our model synthesizes the sketch and oil photos without falsely changing other factors. This manifests the strong capacity of our approach in creating novel views in the form of sketches even without seeing any non-frontal sketch/oil painting images.

5. Limitations and Conclusion

We have stated early in the introduction that our framework is based on the assumption that images of the source and target domain can be transformed bidirectionally, and attribute changes in the source domain can be transferred to the target domain in certain latent space. Violating it may produce less satisfactory results. For example, to replace the sketch dataset with cat face dataset in Section 4.2.2 is not suggested, since viewpoint change of 2D human faces does not generalize well for cat faces.

Given an image-to-image translation model trained on a certain domain, this paper has presented a general framework to adapt it for a new domain. On the one hand, this extends the applicability of existing models, allowing for a lot of interesting applications. On the other hand, it also shows a way for a neural network to generate new images that do not look like training data. This is achieved by generalizing the relation of one domain to another, which simulates how human creates new arts through analogy. Extensive experiments manifest that our framework works with different I2I models, largely improving their performance on unseen target domains.

Figure 5. Results of cross-domain expression manipulation. The 1st row shows an intra-domain example that applies StarGAN on a RaFD image for reference. The 2nd-3rd, 4th-5th and 6th-7th present cross-domain expression manipulation on sketches [41], oil painting [21] and cat faces [21] respectively.

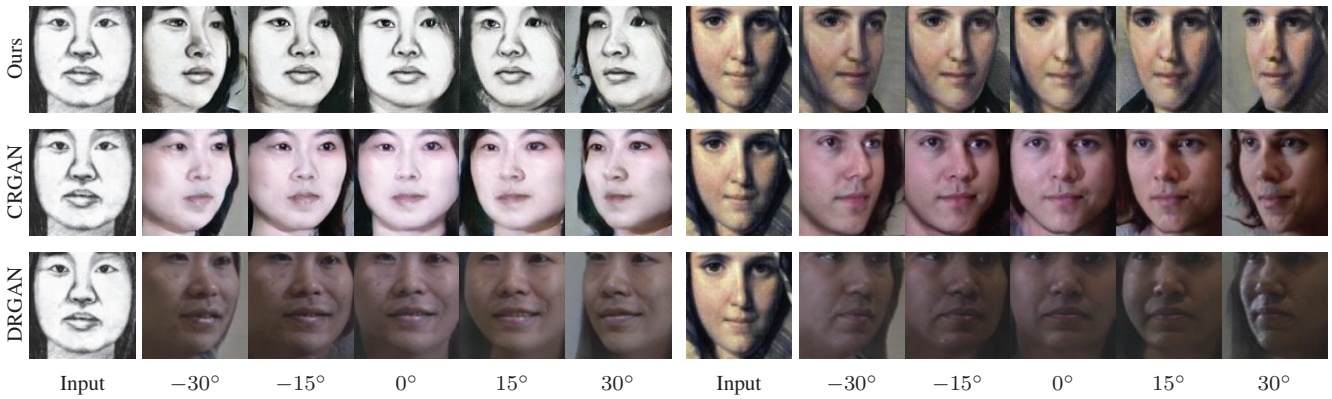


Figure 6. Results of cross-domain novel view synthesis on a sketch image (left) and an oil painting image (right). The first row is our results. The second and third rows are the results of directly applying CRGAN and DRGAN trained on Multi-PIE [10] to the target images.

References

[1] Connelly Barnes, Fang-Lue Zhang, Liming Lou, Xian Wu, and Shi-Min Hu. Patchtable: Efficient patch queries for large datasets and applications. *Siggraph*, 2015. 2

[2] Sagie Benaim and Lior Wolf. One-shot unsupervised cross domain translation. In *Advances in Neural Information Processing Systems*, 2018. 2

[3] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International Conference on Machine Learning*, 2013. 4

[4] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulò. Just dial: Domain alignment layers for unsupervised domain adaptation. In *International Conference on Image Analysis and Processing*, 2017. 4

[5] Ying-Cong Chen, Huaijia Lin, Ruiyu Li, Xin Tao, Michelle Shu, Yangang Ye, Xiaoyong Shen, and Jiaya Jia. Facelet-bank for fast portrait manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[6] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019. 2

[7] Li Cheng, SV N Vishwanathan, and Xinhua Zhang. Consistent image analogies using semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2

[8] Yunjey Choi, Minje Choi, and Munyoung Kim. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 6

[9] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2013. 2

[10] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 2010. 7, 8

[11] Anant Gupta, Srivas Venkatesh, Sumit Chopra, and Christian Ledig. Generative image translation for data augmentation of bone lesion pathology. *arXiv e-prints*, 2019. 1

[12] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 2, 4

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 5

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, 2018. 1, 2, 3

[15] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2007. 2

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, 2018. 2

[18] 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. 5

[19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, 2017. 2

[20] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 2010. 6

[21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018. 1, 2, 3, 6, 7, 8

[22] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics*, 2017. 2

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 1, 2, 3

[24] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. *arXiv e-prints*, 2019. 2

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015. 5

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015. 2

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 3

[29] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[30] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. 2011. 2

[31] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation:

Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision*, 2018. 2

[32] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, 2015. 2

[33] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. *arXiv e-prints*, 2019. 4

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*, 2014. 4

[35] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2016. 2, 4

[36] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016. 2

[37] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. Cr-gan: learning complete representations for multi-view generation. *arXiv e-prints*, 2018. 7

[38] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 7

[39] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4

[40] Guangyu Wang, Tien-Tsin Wong, and Pheng-Ann Heng. Deringing cartoons by image analogies. *ACM Transactions on Graphics*, 2006. 2

[41] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. 2008. 6, 7, 8

[42] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision*, 2017. 2

[43] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal. Modular generative adversarial networks. In *European Conference on Computer Vision*, 2018. 2

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 3