Instance-wise Hard Negative Example Generation for Contrastive Learning in Unpaired Image-to-Image Translation

Weilun Wang [1], Wengang Zhou [1,2]*, Jianmin Bao [3], Dong Chen [3], Houqiang Li [1,2†]
[1] CAS Key Laboratory of GIPAS, EEIS Department, University of Science and Technology of China (USTC)
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3] Microsoft Research Asia
wwlustc@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn, {jianbao, doch}@microsoft.com

Abstract

Contrastive learning shows great potential in unpaired image-to-image translation, but sometimes the translated results are in poor quality and the contents are not preserved consistently. In this paper, we uncover that the negative examples play a critical role in the performance of contrastive learning for image translation. The negative examples in previous methods are randomly sampled from the patches of different positions in the source image, which are not effective to push the positive examples close to the query examples. To address this issue, we present instance-wise hard Negative Example Generation for Contrastive learning in Unpaired image-to-image Translation (NEGCUT). Specifically, we train a generator to produce negative examples online. The generator is novel from two perspectives: 1) it is instance-wise which means that the generated examples are based on the input image, and 2) it can generate hard negative examples since it is trained with an adversarial loss. With the generator, the performance of unpaired image-to-image translation is significantly improved. Experiments on three benchmark datasets demonstrate that the proposed NEGCUT framework achieves state-of-the-art performance compared to previous methods.

1. Introduction

Image-to-image translation aims to transfer images from the source domain to the target domain with the content information preserved, which is of significant importance on various applications such as style transfer [13, 24, 29, 37], domain adaption [4, 19, 20, 33, 51] and image colorization [2, 48, 58, 60]. Due to the inconvenience of collecting paired training data, recent methods are usually based on the unpaired setting. In that case, cycle-consistency loss has been widely used to preserve the consistency between the source images and generated images, for instance, CycleGAN [61], StarGAN [7], UNIT [34] and MUNIT [22].

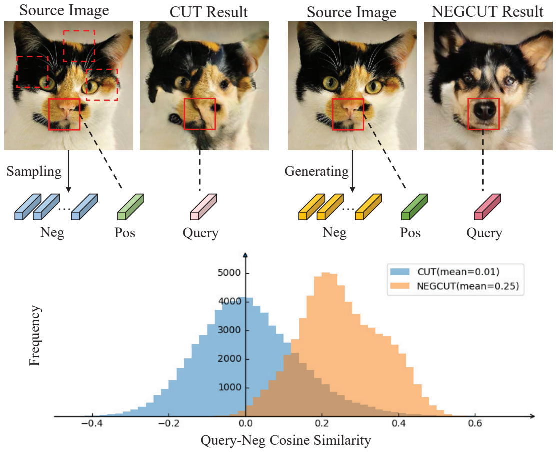

Figure 1. We visualize the generated images along with the distribution of cosine similarity between query and negative samples in CUT [41] and our method. The blue histogram refers to the distribution in CUT while the orange histogram refers to the distribution in our method.

The recently proposed method CUT [41] introduces contrastive learning in unpaired image-to-image translation and achieves better performance over methods [31, 34, 61] that use cycle-consistency loss. In this paper, we aim to further improve the performance of contrastive learning for unpaired image-to-image translation. We uncover that the performance of contrastive learning relies heavily on the hardness of negative samples. As shown in Figure 1, the negative samples in the method [41] are randomly sampled from the patches of different positions in the image, which sometimes leads to the translated results in poor quality and the contents not preserved consistently. Also we calculate the cosine similarities between the query patches and negative patches, and we can find that their cosine similarities are around 0. In other words, these negative patches are not challenging enough to push the positive examples close to the query examples, which will result in the framework not taking full advantage of contrastive learning.

*Corresponding author: Wengang Zhou and Houqiang Li.

To address the above issue, we present instance-wise hard Negative Example Generation for Contrastive learning in Unpaired image-to-image Translation (NEGCUT) in this paper. More precisely, we propose a novel negative generator to excavate hard negative examples. For a source image, we first extract its features on different layers of image generator encoder and embed them into feature vectors. Based on the embedded features from source images, the negative generator produces instance-wise negative examples related to the source image. Moreover, the negative samples should be diverse enough to push the query patch closer to the positive patch. To this end, we add the noise as an extra input for the generator. However, the noise input can probably be ignored for the generator, thus the generator can generate similar examples for different input noises. This is also called the mode collapse issue [45]. Inspired by the mode seeking loss in MSGAN [38], we introduce diversity loss to the generator to encourage the generator to produce diverse hard negative samples for different input noise.

To generate challenging negative samples for contrastive learning, the main idea is to train the negative generator against the encoder network in an adversarial manner. Two components in the framework, *i.e.*, the encoder network and negative generator, are updated alternatively to play a min-max game. On one hand, the encoder network narrows the distance between query and positive samples against hard negative samples to minimize contrastive loss. On the other hand, the negative generator produces hard negative samples close to the positive samples to maximize contrastive loss. Intuitively, the framework will reach an equilibrium where the encoder learns detailed and distinguishing representation to discriminate the positive samples from generated hard negative samples. In Figure 1, we visualize the generated images along with the distribution of cosine similarity between the query and negative samples in the CUT and NEGCUT. It is observed that the negative samples produced by negative generator are harder than those sampled in the method [41], which push the encoder network to learn distinguishing representation and finally results in fine-grained correspondence of structures and textures.

Our contributions are summarized as follows,

- We identify that instance-wise negative examples that increase hardness as training process play a critical role in the performance of contrastive learning for unpaired image-to-image translation.
- We propose a novel framework NEGCUT to mine instance-wise hard negative examples for contrastive learning in unpaired image-to-image translation.
- Extensive experiments on three benchmark datasets demonstrate the superiority of our method, which achieves new state-of-the-art performance. The generated images of our method are of better visual performance with consistent detailed correspondence.

2. Related Work

In this section, we briefly introduce the related topics, including contrastive learning, image-to-image translation and hard negative mining.

2.1. Image-to-Image Translation

Image-to-image translation (I2I) [30, 35, 44, 50, 52, 54, 59, 61, 62] aims to transfer images from source to target domain with the content information preserved. Earlier methods [5, 23, 42, 52] apply an adversarial loss [14], along with a reconstruction loss to train their model based on the paired training data. However, due to the difficulty of collecting a large amount of paired data, recent methods are usually based on the unpaired setting. In that cases, cycle-consistency loss has been widely used to preserve the consistency between the source images and generated images instead, for instance, CycleGAN [61], DiscoGAN [27], DualGAN [55] and U-GAT-IT [26]. Based on the assumption that the generated result should be translated back by an inverse mapping, cycle-consistency learns the mapping from target to source domain and check whether the source images are reconstructed. However, the assumption is overly strict compared to the actual situation, where the images between the two domains are not one-to-one mapping. In view of this, CUT [41] involves contrastive learning in unpaired image-to-image translation to learn the correspondence between source and generated images, which outperforms previous methods using cycle-consistency loss.

2.2. Contrastive Learning

Contrastive learning is a framework that learns representation by comparing similar and dissimilar pairs. Recent methods [1, 6, 16, 17, 18, 40] based on the theory of maximizing mutual information have achieved wide success on unsupervised representation learning. These methods take full advantage of noise-contrastive estimation [15], mapping the images into an embedding space where associated samples are brought together in contrast with unrelated samples. For a single query sample, the associated samples are referred to as positive samples while the unrelated samples are referred to as negative samples. With similarity measured by dot production, a form of a contrastive loss, called InfoNCE, is proposed as a representative loss function for noise-contrastive estimation.

2.3. Hard Example Mining

Hard example mining is a classic method to solve the problem of sample imbalance in several areas, *i.e.*, object detection and unsupervised representation learning. In earlier methods, hard example mining are used to optimize SVMs [11], shallow neural networks [43] and boosted decision trees [10]. Recent work [21, 25, 32, 36, 46, 47, 53]

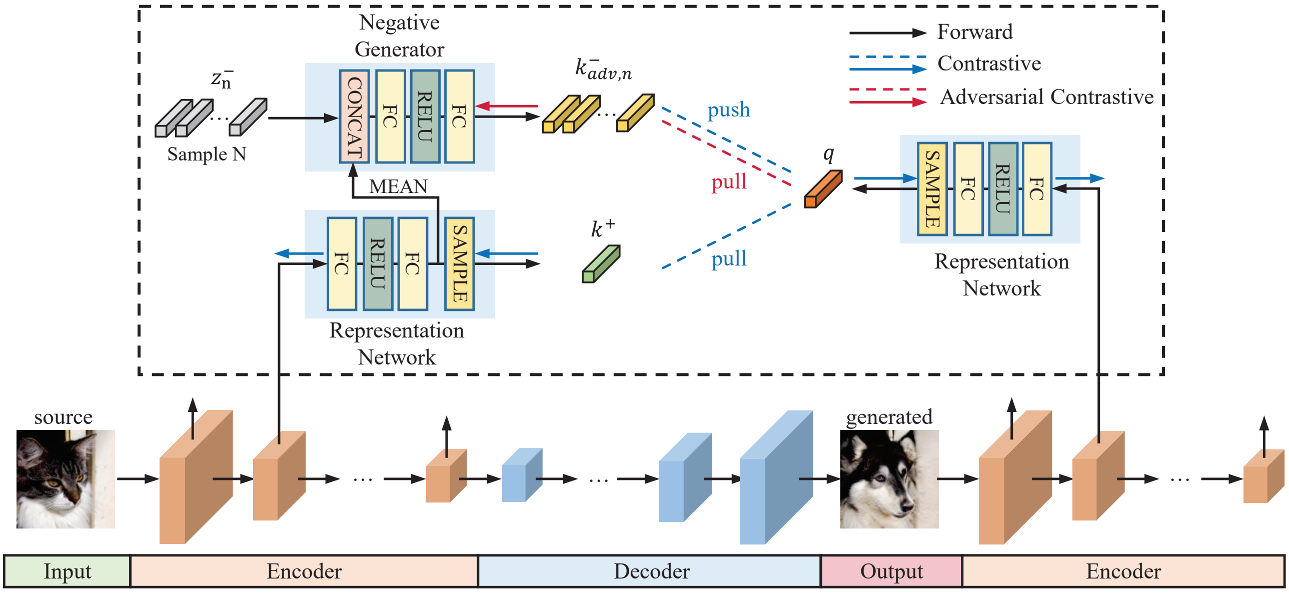

Figure 2. The overview of our NEGCUT framework. We perform hard negative example generation for adversarial contrastive learning on multiple layers of the image generator encoder. The black arrows show the forward propagation of our framework while the blue and red arrows show the backward propagation of contrastive loss and adversarial contrastive loss, respectively. On each layer, the representation network randomly samples the source and translated features at the spatial dimension, and produces the query and positive samples. The negative generator produces challenging negative samples by the mean vector of features from the representation network. The query, positive and generated negative samples are involved for contrastive learning in an adversarial manner.

selects hard examples for training deep networks. In [47], an image descriptor is learned to independently select the hard positive and negative samples from a large set. In [36] and [46], online hard examples selection is investigated on image classification and object detection, respectively. Lin *et .al* design a novel focal loss [32] to focus training on a sparse set of hard examples, which addresses the imbalance between different classes in object detection. In unsupervised representation learning, a triple loss is used [53] to mine the hard negative samples from a large set. In [21], adversarial learning is involved to generate challenging negative samples for unsupervised representation learning.

3. Methods

In this paper, we present a novel framework NEGCUT to mine instance-wise hard negative examples for contrastive learning in unpaired image-to-image translation. Different from previous work which randomly samples negative examples from the patches in the image, our method generates instance-wise hard negative examples through adversarial learning. With the produced hard negative examples, our framework can generate images with detailed and fine-grained correspondence on structures and textures. The rest of this section is organized as follows: We begin with reviewing the related method in previous work in Sec. 3.1. In Sec. 3.2, we outline the NEGCUT framework and introduce the details of hard negative example generation through adversarial learning. Finally, we discuss the objective function utilized in our framework in Sec. 3.3.

3.1. Preliminaries and Motivation

We first briefly review the method leveraging contrastive learning in unpaired image-to-image translation developed in CUT [41]. To generate images of target domain with the content information maintained, the main idea is to learn the correspondence between the source and generated images. Compared with previous methods using cycle-consistency loss, CUT applies contrastive loss to learn the correspondence instead, which directly maximizes the mutual information between the source and generated images. The contrastive loss is formulated as follows,

$$l(\mathbf{q},\ \mathbf{k}^+,\ \mathbf{k}^-) = \\ -\log[\frac{\exp(\mathbf{q}\cdot\mathbf{k}^+/\tau)}{\exp(\mathbf{q}\cdot\mathbf{k}^+/\tau)+\sum_{n=1}^{N}\exp(\mathbf{q}\cdot\mathbf{k}_n^-/\tau)}], \quad (1)$$

where $\mathbf{q}$ is the query samples from the generated image, $\mathbf{k}^+$ is the positive samples from the corresponding position of the query in the source image, $\mathbf{k}_n^-$ is the negative samples from the other positions in the source images, and τ is the temperature factor.

CUT develops the contrastive learning in a multi-layer patch-wise manner, which is formulated as follows,

$$\mathcal{L}_{\text{PatchNCE}}(G,H,\mathbf{X}) = \mathbb{E}_{x\sim X}\sum_{l=1}^{L}\sum_{s=1}^{S_l} l(\mathbf{q}_{l,s},\mathbf{k}_{l,s}^+,\mathbf{k}_{l,s}^-), \quad (2)$$

where $\mathbf{q}_{l,s}$, $\mathbf{k}^+_{l,s}$ and $\mathbf{k}^-_{l,s}$ are extracted from the features of source image $\mathbf{X}$ and generated image $\mathbf{Y}$ at different intermediate layers l of generator encoder. With the contrastive loss, the generator learns to narrow the distance between the query and positive samples against negative samples at different layers, which is equivalent to maximizing the mutual information between the source and generated images.

By replacing the cycle-consistency loss with the contrastive loss, CUT generates more realistic and corresponding images compared with previous methods. However, the randomly-sampled negative examples in CUT cannot take full advantage of contrastive learning. The approach to estimating the negative examples plays a critical role in the performance of contrastive learning. Negative examples in CUT are not challenging enough to push the encoder network to learn distinguishing representation, which leads to the translated results in poor quality and the contents not preserved consistently. Different from these, we propose a novel framework NEGCUT to mine instance-wise hard negative samples for unpaired image-to-image translation through adversarial learning.

3.2. NEGCUT

3.2.1 Framework Architecture

Figure 2 gives an overview of our framework, which consists of *Image Generator*, *Representation Network* and *Negative Generator*. Image generator G takes the source image $\mathbf{X}$ as input and generates the translated image $\mathbf{Y}$. Regarding two variants of a single image, *i.e.*, the source image $\mathbf{X}$ and the generated image $\mathbf{Y}$, we conduct multi-layer patchwise contrastive learning to learn the correspondence between these two images. On a certain layer of the image generator encoder, the query and positive samples are produced by the representation network through embedding the spatially sampled feature vectors into high-dimensional representation space.

To increase the similarity between the query and positive samples, the negative generator mines instance-wise hard negative samples against positive samples. Based on the embedded features of the source image, diverse challenging negative examples are generated by taking various randomly-sampled noise vectors as input. In our framework, the encoder network (*i.e.*, the image generator and representation network) and the negative generator are alternately updated with the adversarial contrastive loss. With more challenging negative examples produced by the negative generator, the encoder network will learn distinguishing representation to discriminate positive samples from the challenging negative samples, which leads to fine-grained and robust correspondence between the source and generated images. Additionally, a discriminator is applied to ensure the domain and realness of the generated image.

3.2.2 Hard Negative Example Generation

In this section, we formally present hard negative example generation for contrastive learning in unpaired image-to-image translation. As shown in Figure 2, we perform contrastive learning on multiple layers of the image generator encoder. For a certain layer, we employ a representation network $H^i(\cdot)$ to embed the feature of different patches. The representation network is a 2-layer MLP network independently mapping the feature vector at each pixel from the source and translated images to a M-dimension vector. Based on the feature after mapping, we randomly sample S positions in the spatial dimension and take the normed vectors as query and positive samples for contrastive learning, which is formulated as follow,

$$q = \frac{H^i_s(\mathbf{F}^{\mathbf{Y}}_i)}{\|H^i_s(\mathbf{F}^{\mathbf{Y}}_i)\|_2}, k^+ = \frac{H^i_s(\mathbf{F}^{\mathbf{X}}_i)}{\|H^i_s(\mathbf{F}^{\mathbf{X}}_i)\|_2}, \tag{3}$$

where $\mathbf{F}^{\mathbf{X}}_i$ and $\mathbf{F}^{\mathbf{Y}}_i$ are the source features and the translated features at the i-th layer of image generator encoder, respectively. $H^i_s(\mathbf{F}^{\mathbf{X}}_i)$ and $H^i_s(\mathbf{F}^{\mathbf{Y}}_i)$ refers to the s-th positive and query examples sampled, respectively.

To push the positive sample close to the query sample, we generate challenging negative samples with a carefully designed multi-layer negative generator $\{N^0, N^1, \cdots, N^l\}$. Base on the spatially-average features from representation network $\overline{H^i(\mathbf{F}^{\mathbf{X}}_i)}$, the negative generator produces hard negative samples with noise vector z_n, which is formulated as follows,

$$k^-_{\mathrm{adv},n} = \frac{N^i(\overline{H^i(\mathbf{F}^{\mathbf{X}}_i)}; z_n)}{\|N^i(\overline{H^i(\mathbf{F}^{\mathbf{X}}_i)}; z_n)\|_2}. \tag{4}$$

For a positive sample, we generate multiple negative examples through sampling various noise vectors from standard Gaussian distribution.

To generate challenging negative samples for contrastive learning, the main idea is to train the negative generator against the encoder network in an adversarial manner, which is formulated as follows,

$$\begin{aligned} &\min_{\theta_{\mathcal{H}}, \theta_G} \max_{\theta_{\mathcal{N}}} l(\mathbf{q}, \mathbf{k}^+, \mathbf{k}^-_{adv}) = \\ &-\log\Big[\frac{\exp(\mathbf{q}\cdot\mathbf{k}^+/\tau)}{\exp(\mathbf{q}\cdot\mathbf{k}^+/\tau) + \sum_{n=1}^{N}\exp(\mathbf{q}\cdot\mathbf{k}^-_{\mathrm{adv},n}/\tau)}\Big]. \end{aligned} \tag{5}$$

From Equation (5), it is observed that the encoder network (*i.e.*, the representation network $\mathcal{H} = \{H^0, H^1, \cdots, H^l\}$ and the image generator G) narrows the distance between the query samples and positive samples against the negative samples to minimize contrastive loss. On the contrary, the negative generator $\mathcal{N} = \{N^0, N^1, \cdots, N^l\}$ produces challenging negative examples to maximize the contrastive loss. Intuitively, the encoder network and the negative generator will reach an equilibrium by alternate training, where the negative generator produces challenging negative samples and the encoder network learns distinguishing representation to discern the positive samples from the negative samples.

In Figure 2, we further illustrate how the negative generator, representation network and image generator are updated. The negative generator is first updated with negative contrastive loss, which is formulated as follows,

$$\theta_{N^i} \leftarrow \theta_{N^i} + \eta_{\mathcal{N}} \frac{\partial l(\mathbf{q}, \mathbf{k}^+, \mathbf{k}^-_{adv})}{\partial \theta_{N^i}}. \tag{6}$$

The backpropagation of the negative contrastive loss is cut off before the representation network and does not affect the weights of the representation network and image generator. After that, the representation network is updated with positive contrastive loss, which is formulated as follow,

$$\theta_{H^i} \leftarrow \theta_{H^i} - \eta_{\mathcal{H}} \frac{\partial l(\mathbf{q}, \mathbf{k}^+, \mathbf{k}^-_{adv})}{\partial \theta_{H^i}}. \tag{7}$$

Since the contrastive learning is developed in a multi-layer manner, the total adversarial contrastive loss for the negative generator and representation network is formulated as follows,

$$\mathcal{L}_{AdCont} = \mathbb{E}_{x \sim X} \sum_{l=1}^{L} \sum_{s=1}^{S_l} l(\mathbf{q}_{l,s}, \mathbf{k}^+_{l,s}, \mathbf{k}^-_{adv,l,s}). \tag{8}$$

The image generator is trained along with the representation network. Through the back-propagation of adversarial contrastive loss, the image generator receives the gradient at different layers of the encoder. The image generator is updated with the summation of these gradient, which is formulated as follows,

$$\begin{aligned} \theta_G \leftarrow \theta_G - \eta_G \sum_{i=0}^{l} (& \frac{\partial l(\mathbf{q}, \mathbf{k}^+, \mathbf{k}^-_{adv})}{\partial \mathbf{F}_i^{\mathbf{X}}} \frac{\partial \mathbf{F}_i^{\mathbf{X}}}{\partial \theta_G} \\ & + \frac{\partial l(\mathbf{q}, \mathbf{k}^+, \mathbf{k}^-_{adv})}{\partial \mathbf{F}_i^{\mathbf{Y}}} \frac{\partial \mathbf{F}_i^{\mathbf{Y}}}{\partial \theta_G}), \end{aligned} \tag{9}$$

where $\mathbf{F}_i^{\mathbf{X}}$ and $\mathbf{F}_i^{\mathbf{Y}}$ are the features of the source and translated images at the i-th layer of encoder, respectively.

However, when the adversarial contrastive loss is the only function used to update the negative generator, it is observed that the generated negative examples lose diversity and collapse to one negative example. This is because the adversarial contrastive loss focuses on generating hard negative samples rather than diverse negative samples, though diversity is helpful for the performance. To this end, we introduce the diversity loss to generate diverse challenging negative samples with different input noise. The diversity loss encourages the generation of distinctive results when different noise vectors are brought in, which is formulated as follows,

$$\mathcal{L}_{div} = -\|N^i(\overline{H^i(\mathbf{X}_i)}, z_1) - N^i(\overline{H^i(\mathbf{X}_i)}, z_2)\|_1, \tag{10}$$

where z_1 and z_2 are two different input noise randomly sampled from standard Gaussian distribution.

3.3. Other Objectives

Besides the adversarial contrastive loss and diversity loss mentioned above, our framework is also optimized by generative adversarial loss.

Generative Adversarial Loss. Since the ground-truth images are unavailable in unpaired image-to-image translation, we develop adversarial learning [14, 57] to constrain the realness and domain of the generated images. For the image generator $G(\cdot)$ and the discriminator $D(\cdot)$, we unitize the LSGAN$_{100}$ loss [39], which is formulated as follows,

$$\begin{aligned} \mathcal{L}^D_{gan} &= \mathbb{E}_{x_r}[(1 - \mathbf{D}(x_r))^2] + \mathbb{E}_{x_f}[\mathbf{D}(x_f)^2], \\ \mathcal{L}^G_{gan} &= \mathbb{E}_{x_f}[(1 - \mathbf{D}(x_f)^2], \end{aligned} \tag{11}$$

where x_r and x_f indicates the real image distribution and the generated images distribution, respectively.

Overall Loss. The overall loss for the negative generator and encoder network is the weighted summation of above losses, which is formulated as follows,

$$\begin{aligned} \mathcal{L}_{\mathcal{H}} &= \mathcal{L}_{AdCont}, \\ \mathcal{L}_G &= \mathcal{L}_{AdCont} + \lambda_1 \mathcal{L}^G_{gan}, \\ \mathcal{L}_{\mathcal{N}} &= -\mathcal{L}_{AdCont} + \lambda_2 \mathcal{L}_{div}, \end{aligned} \tag{12}$$

where λ_1 and λ_2 are the trade-off parameters balancing different losses. In our experiments, λ_1 and λ_2 are set to 1 and 1, respectively.

4. Experiment

4.1. Experiment Setup

Datasets. To demonstrate the superiority of our method, we train and test our method on three benchmark datasets, *i.e.*, *Cityscapes* [9], *Cat→Dog* [8] and *Horse→Zebra* [61] datasets with translation between various different domains. The *Cityscapes* dataset contains a diverse set of images recorded in the street scenes with high-quality pixel-level annotations. The *Cat→Dog* dataset is a dataset of 10,000 high-quality cat and dog face images extracted from the *AFHQ* dataset. The *Horse→Zebra* dataset consists of about 2,500 images of horse and zebra in different scenes. We learn the translation from semantic masks to real images, from cat images to dog images and from horse images to zebra images on three datasets, respectively. For all the datasets, we resize images to the same resolution of 256 × 256 to train our network.

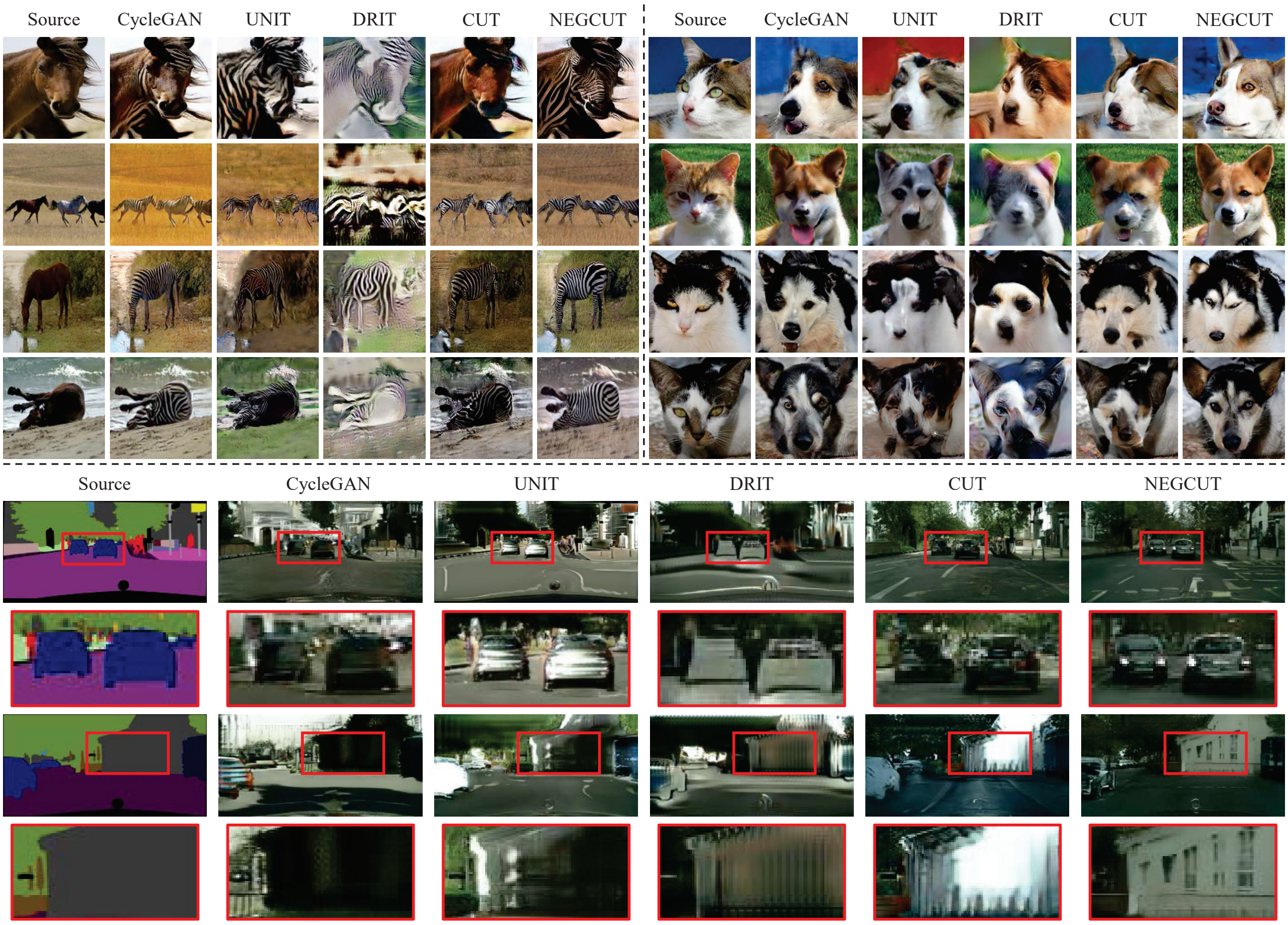

Figure 3. Qualitative results with the four challenging methods, *i.e.*, CycleGAN [61], UNIT [34] , DRIT [31], CUT [41], on three benchmark datasets. Compared with previous methods, the generated images of our method show superior performance with correct correspondence between the source and generated image.

Implementation Details. To make a fair comparison, we set the hyperparameters consistent with previous methods [41]. We conduct our adversarial contrastive learning on the 1-st, 5-th, 9-th, 13-th, 17-th layers of the generator encoder. The number of negative samples for contrastive learning is set to 256 in our framework. The dimension of the query, positive and negative samples is set to 256. For the whole framework, we utilize Adam optimizer [28]. The training lasts 400 epochs in total. The learning rate is set to 2e-4 and linearly reduces after 200 epochs. The whole framework is implemented by Pytorch and we perform experiments on NVIDIA RTX 3090Ti.

Evaluation Metrics. We evaluate the realness of generated images by the FID metric. FID measures the distance between two sets of images. To calculate the FID metric, we first embed the generated images and ground-truth images into the feature space with an Inception model [49]. The FID metric is computed by the mean value and covariance of the generated image set $(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{Y}})$ and the ground-truth image set $(\mu_{\hat{\mathbf{Y}}}, \Sigma_{\hat{\mathbf{Y}}})$:

$$\mathrm{FID}(\mathbf{Y}, \hat{\mathbf{Y}}) = \|\mu_{\mathbf{Y}} - \mu_{\hat{\mathbf{Y}}}\|_2^2 + \mathrm{Tr}(\Sigma_{\mathbf{Y}} + \Sigma_{\hat{\mathbf{Y}}} - 2(\Sigma_{\mathbf{Y}}\Sigma_{\hat{\mathbf{Y}}})^{\frac{1}{2}}). \tag{13}$$

In addition, to evaluate the relevance between source images and generated images, we apply several metrics different from FID on the *Cityscapes* dataset. With a pre-trained segmentation model [56], we calculate the mAP, pixel accuracy (pAcc) and class accuracy (cAcc) metrics on the source semantic labels and generated real images. The higher mAP, pAcc and cAcc represent that the generated images are more relevant to source semantic labels.

4.2. Comparison with the State-of-the-art Methods

We compare our method with several state-of-the-art methods of unpaired image-to-image translation, *i.e.*, CUT [41], CycleGAN [61] and DRIT [31]. The quantitative result on three benchmark datasets is shown in Table 1. From the table, it is observed that our method achieves new state-of-the-art performance on three datasets. Compared

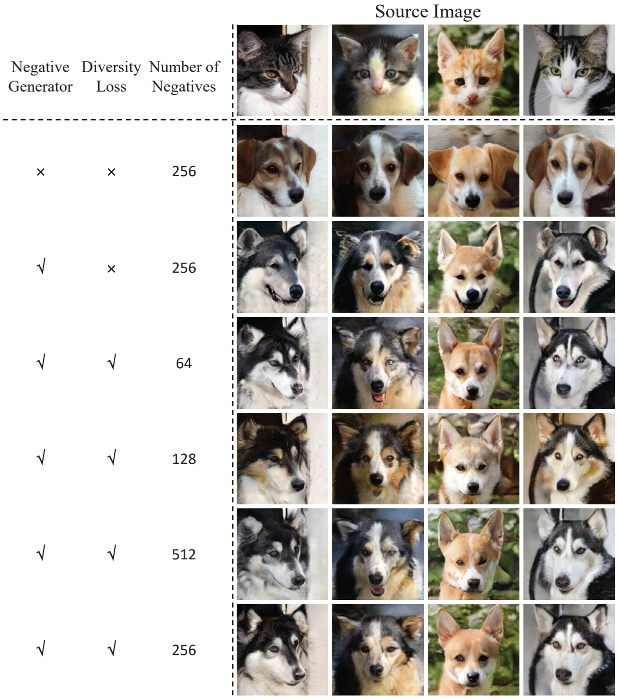

Figure 4. Qualitative comparison among different designs of negative generator. When the negative generator and diversity loss are employed and the number of negatives is set to 256, the generated images have the best visual quality and most correct correspondence.

with the most challenging method, *i.e.* CUT, our method outperforms it 14.0%, 26.6% and 13.0% relatively on FID metric on three datasets. Additionally, since only the image generator is used at inference time, NEGCUT does not introduce extra test time consumption compared with CUT.

Furthermore, we make a qualitative evaluation on three datasets with several competitive methods, *i.e.* CUT, CycleGAN, UNIT and DRIT. From Figure 3, it is observed that the images generated by our method have better visual performance compared with previous methods. Especially, our generated images keep a better correspondence with the source images compared with the most challenging method CUT. This benefits from the challenging negative samples generated by the negative generator. The negative examples sampled randomly in CUT help the network learn the correspondence between source images and generated images at the beginning, but become less and less effective as the training process proceeds. In contrast, our negative samples produced by negative generator keep challenging via adversarial learning, which forces the image generator and representation network to learn the fine-grained correspondence. Due to this reason, the images generated by our method have better correspondence with the source images in details, *i.e.*, textures and postures.

Method	**Cityscapes**				**Cat→Dog**	**H→Z**
	mAP↑	**pAcc↑**	**cAcc↑**	**FID↓**	**FID↓**	**FID↓**
CycleGAN [61]	20.4	55.9	25.4	76.3	85.9	77.2
UNIT [34]	16.9	56.5	22.5	91.4	104.4	133.8
DRIT [31]	17.0	58.7	22.2	155.3	123.4	140.0
Distance [3]	8.4	42.2	12.6	81.8	155.3	72.0
SelfDistance [3]	15.3	56.9	20.6	78.8	144.4	80.8
GCGAN [12]	21.2	63.2	26.6	105.2	96.6	86.7
CUT [41]	24.7	68.8	30.7	56.4	76.2	45.5
FastCUT [41]	19.1	59.9	24.3	68.8	94.0	73.4
NEGCUT	**27.6**	**71.4**	**35.0**	**48.5**	**55.9**	**39.6**

Table 1. Comparison with state-of-the-art methods on unpaired image translation, *i.e.* CycleGAN, UNIT, DRIT, CUT, *etc.* H→Z refers to the *Horse→Zebra* dataset. ↑ indicates the higher the better, while ↓ indicates the lower the better. It is notable that our method outperforms previous methods on various metrics.

Settings			**Cityscapes**				**Cat→Dog**	**H→Z**
Negative Generator	**Diversity Loss**	**Number of Neg.**	**mAP↑**	**pAcc↑**	**cAcc↑**	**FID↓**	**FID↓**	**FID↓**
×	×	256	27.3	**71.9**	34.5	49.7	110.9	59.6
✓	×	256	27.0	71.1	33.7	91.5	83.0	72.1
✓	✓	64	26.9	71.1	33.7	49.7	59.3	59.3
✓	✓	128	27.2	71.4	33.9	49.8	86.9	51.8
✓	✓	512	27.3	71.3	34.3	51.2	62.8	44.0
✓	✓	256	**27.6**	71.4	**35.0**	**48.5**	**55.8**	**39.6**

Table 2. Ablation study for several different designs, *i.e.*, negative generator, diversity loss and the number of negative samples. H→Z refers to the *Horse→Zebra* dataset. ↑ indicates the higher the better, while ↓ indicates the lower the better. Without negative generator or the diversity loss, the produced negative examples are not challenging enough, which leads to inferior performance under most of the indicators on three datasets.

4.3. Ablation Study

We perform several ablation experiments to verify the effectiveness of several designs in our framework, *i.e.* negative generator, diversity loss and number of negative samples. We report the quantitative results in Table 2.

To evaluate the necessity of generating instance-wise negative samples with negative generator, we design a variant without negative generator. As an alternative, we directly update the negative samples in the feature vector space in this variant. In such case, the learned negative samples are widely distributed in the feature space and unrelated to the source instance. From Table 2, it is observed that, without the negative generator, the framework obtains an inferior performance under most of the indicators, which verifies the effectiveness of generating *instance-wise* negative samples. After that, we conduct an ablation study on the diversity loss by comparing the framework with and without the diversity loss. In Table 2, it demonstrates that the framework with diversity loss outperforms that without diversity loss. This is because, in the variant without diversity loss, the produced negative samples lose diversity in the early stage of training and maintain less diverse during training. Under this situation, the negative generator fails to produce challenging negative samples, which leads to the poor performance. Additionally, we analyze the visual results under these different settings. In Figure 4, it can be seen that, without the negative generator or diversity loss, the generated image has a far inferior visual quality on fidelity and correspondence between the source and generated image.

Based on the framework with negative generator and diversity loss, we further make an ablation study on the number of negative samples. From Table 2, it can be seen that the performance reaches the top when the number of negative samples equal to 256. When the number of negative samples is more than 256, the main challenging negative samples have been contained in the 256 negative samples. The extra generated negative samples may contain some irrelevant disturbance resulting in inferior performance, and increase the computation cost notably. Nevertheless, too few negative samples may result in the ineffectiveness to push positive samples closer to the query. Comprehensively considering the performance and computational consumption, the best number of negative samples is 256 in our framework. Furthermore, we compare the generated images under different numbers of negative examples. In Figure 4, it is observed that, when the number of negative samples is set to 256, the generated images have the best visual quality and most correct correspondence.

4.4. Visualization of Hard Negative Examples.

To further demonstrate the effect of hard negative examples, we visualize hard negative examples by retrieving regions based on generated features. In Figure 5, we first retrieve 8 hard negative examples based on each query feature. After that, we visualize these hard negative examples by retrieving the most related patches in the image. It is observe that the retrieved hard negative examples share similar semantic meanings with the query patch in structure and texture. This indicates that the generated hard examples can encourage the model to generate content consistent results.

Furthermore, we compare the learned similarity by representation network in CUT and NEGCUT. For each query q, we calculate the similarity maps through computing $\exp(\mathbf{q} \cdot \mathbf{k}^{+}/\boldsymbol{\tau})$ on all the pixels of the image. From Figure 6, it is observed that, in the similarity maps of CUT, the corresponding areas are scattered over the entire image and several unrelated areas are also associated. Additionally, when the query point is sampled from a part of the foreground, *i.e.* head of the horse, the whole foreground is associated in the similarity maps of CUT, which demonstrates that the representation network in CUT has difficulty on discriminating different parts of the foreground. Different from that, the corresponding areas in the similarity maps of NEGCUT are concentrated on the neighborhood of query points or areas with the same semantic, which verifies that the representation network in NEGCUT learns more distinguishing representation and accuracy correspondence under the help of instance-wise hard negative samples.

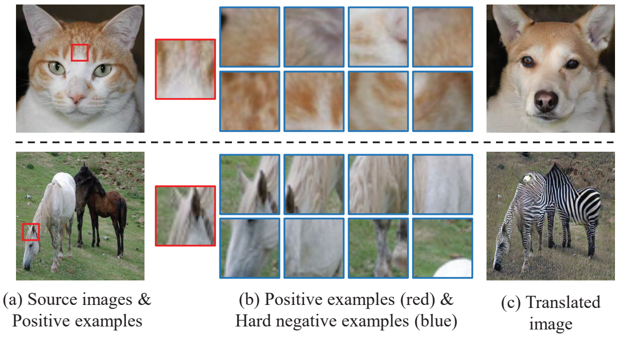

Figure 5. Visualization of negative examples by retrieving regions based on generated features. We visualize 8 hard negative examples by retrieving the most related patches in the image. It is observed that the retrieved patches share similar semantic meanings with the query patch in structure and texture.

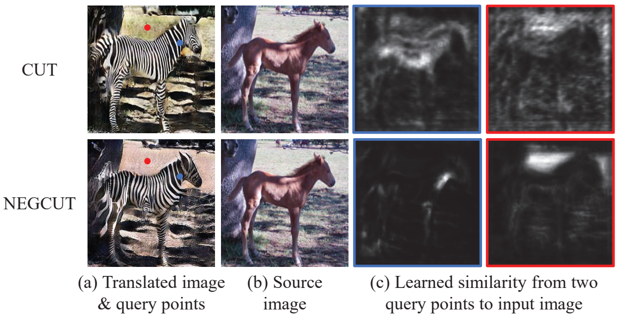

Figure 6. Visualization of the learned similarity by representation network in CUT and NEGCUT. Two similarity maps are learned from two query points sampled from the foreground (blue) and background (red). Compared with the similarity learned by CUT, our similarity maps are more concentrated on the neighbourhood of query points, which verifies that our method learns distinguishing representation with the help of hard negative samples.

5. Conclusion

In this paper, we propose a novel framework called NEGCUT to mine challenging negative samples for contrastive learning in unpaired image-to-image translation. Specifically, we design a negative generator trained against the encoder network in an adversarial manner. The two components in our framework, *i.e.*, the encoder network and the negative generator, are updated alternately to learn distinguishing representation to discriminate positive samples against generated hard negative samples. Extensive experiments on three benchmark datasets demonstrate the superiority of our method. Our method achieves state-of-the-art performance and shows a better correspondence between source images and generated images compared with previous methods.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Contract 61836011, 61822208, and 62021001, and in part by the Youth Innovation Promotion Association CAS under Grant 2018497. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019.

[2] Federico Baldassarre, Diego González Morín, and Lucas Rodés-Guirao. Deep koalarization: Image colorization using cnns and inception-resnet-v2. *arXiv preprint arXiv:1712.03400*, 2017.

[3] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *NeurIPS*, 2017.

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, pages 3722–3731, 2017.

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, pages 1511–1520, 2017.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[10] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *BMVC*, 2009.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2009.

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. In *CVPR*, 2019.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[15] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[17] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, pages 4182–4192. PMLR, 2020.

[18] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[19] 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 *ICML*, pages 1989–1998. PMLR, 2018.

[20] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[21] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. *arXiv preprint arXiv:2011.08435*, 2020.

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[25] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, pages 21798–21809, 2020.

[26] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

[27] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pages 1857–1865. PMLR, 2017.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[29] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *CVPR*, pages 10051–10060, 2019.

[30] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representation. In *ECCV*, 2018.

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[33] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, pages 7202–7211, 2019.

[34] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, pages 701–709, 2017.

[35] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019.

[36] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

[37] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, pages 4990–4998, 2017.

[38] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019.

[39] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[41] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, pages 319–345. Springer, 2020.

[42] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019.

[43] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *TPAMI*, 20(1):23–38, 1998.

[44] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Moressi, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. *arXiv preprint arXiv:1711.05139*, 2017.

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

[46] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *ICCV*, pages 761–769, 2016.

[47] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, and Francesc Moreno-Noguer. Fracking deep convolutional image descriptors. *arXiv preprint arXiv:1412.6537*, 2014.

[48] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, pages 7968–7977, 2020.

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[50] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*, 2019.

[51] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018.

[53] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015.

[54] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019.

[55] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017.

[56] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, pages 472–480, 2017.

[57] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu. Adversarial attribute-text embedding for person search with natural language query. *TMM*, 22(7):1836–1846, 2020.

[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.

[59] Rui Zhang, Tomas Pfister, and Jia Li. Harmonic unpaired image-to-image translation. In *ICLR*, 2019.

[60] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *TOG*, 9(4), 2017.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[62] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017.