Modulated Contrast for Versatile Image Synthesis

Fangneng Zhan[1], Jiahui Zhang[2], Yingchen Yu[2], Rongliang Wu[2], Shijian Lu[*2]
[1] S-Lab, Nanyang Technological University [2] Nanyang Technological University

Abstract

Perceiving the similarity between images has been a long-standing and fundamental problem underlying various visual generation tasks. Predominant approaches measure the inter-image distance by computing pointwise absolute deviations, which tends to estimate the median of instance distributions and leads to blurs and artifacts in the generated images. This paper presents MoNCE, a versatile metric that introduces image contrast to learn a calibrated metric for the perception of multifaceted inter-image distances. Unlike vanilla contrast which indiscriminately pushes negative samples from the anchor regardless of their similarity, we propose to re-weight the pushing force of negative samples adaptively according to their similarity to the anchor, which facilitates the contrastive learning from informative negative samples. Since multiple patch-level contrastive objectives are involved in image distance measurement, we introduce optimal transport in MoNCE to modulate the pushing force of negative samples collaboratively across multiple contrastive objectives. Extensive experiments over multiple image translation tasks show that the proposed MoNCE outperforms various prevailing metrics substantially. The code is available at MoNCE.

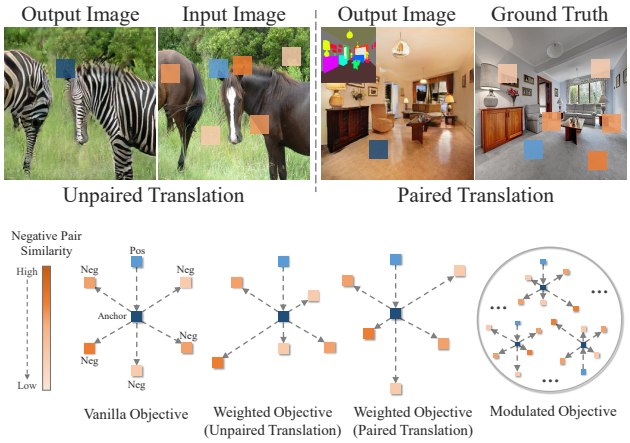

Figure 1. Comparison of different contrastive objectives: For the contrastive objective of a single image patch, vanilla contrastive objective repels all negative samples indiscriminately. The introduced weighted contrastive objective adaptively adjusts the weights of negative pairs according to the pair similarity. With inverse weighting strategies for unpaired and paired translation tasks, the weighted objective can be applied to improve the generation performance substantially. The modulated contrastive objective introduces optimal transport to modulate the learning objectives of all image patches as a whole.

1. Introduction

Multifarious image generation tasks [26, 27, 30, 45, 46, 50, 56, 57] often entail multifaceted metrics to measure the inter-image similarity with regard to different properties such as image structures, image semantics and image perceptual realism, etc. Defining generic metrics to fulfil multiple objectives is challenging as different visual properties are usually entangled in pixels and the notion of visual similarity is often subjective. Image similarity measurement remains a very open research challenge in visual generation tasks.

To measure and minimize the content variation in unpaired image translation, Zhu *et al.* [56] design a cycle-consistency loss to ensure that input images can be recovered from the output images. Different from unpaired image translation, paired image translation entails certain metrics to measure the perceptual similarity between output images and ground truth. Among various distance metrics [36, 37], perceptual loss [17] emerges as a powerful metric in line with human perception by leveraging the internal activation of pre-trained networks. However, above metrics are designed based on point-wise deviations, which undesirably minimize the average deviation to all possible instances. For example, a semantic map corresponds to numerous real images, minimizing the average deviation to all possible real images tends to produce blurred generation results.

Instead of minimizing the point-wise deviation, the prevailing contrastive learning [5, 14, 41] aims to pull positive samples towards an anchor and push negative samples far away from it. It has recently been adopted in image generation tasks for preserving image contents in unpaired im-

*Corresponding author, E-mail: shijian.lu@ntu.edu.sg

age translation [26], perceiving image similarity in paired image translation [2], or serving as a contrastive regularization term in image dehazing [38]. However, all these studies adopt the vanilla contrast that shares a critical constraint – negative samples are indiscriminately pushed away from the anchor regardless of their similarity to the anchor.

In this work, we formulate contrastive learning as a versatile metric for various image translation tasks as shown in Fig. 1. In unpaired image translation, contrastive learning allows to preserve image contents by maximizing the mutual information of corresponding patches [26]. In paired image translation, contrastive learning is employed to measure the perceptual similarity between images in line with human judgement, by leveraging pre-trained networks for feature extraction. However, vanilla contrastive objective repels all negative samples indiscriminately, which is apparently sub-optimal as negative samples usually have different similarity with the anchor. Certain weighting strategy is desired to formulate more effective contrast by adaptively adjusting the pushing force of negative samples.

Aiming to boost the translation performance, we comprehensively investigated different weighting strategies for negative samples and some non-trivial conclusions are drawn for the selection of weighting strategies in different scenarios. Intuitively, hard negative samples (i.e., with high similarity to the anchor) should be assigned higher weights (referred as *hard weighting*), complying with the rationale of hard negative sampling [29]. It is true for unpaired image translation where negative samples can be easily pushed apart as illustrated in the similarity distributions of negative & positive pairs in Fig. 2. However, for paired image translation, negative samples are hard to be pushed apart from the anchor (or positive pairs) as there is severe overlap for the similarity distribution of negative & positive pairs as in Fig. 2. In this scenario, we surprisingly find that the intuitive hard weighting strategy tends to impair the performance, and an inverse weighting strategy as shown in Fig. 1 allows to improve the performance. In addition, as in PatchNCE loss [26], contrastive learning for measuring image similarity involves several sub-objectives as each image patch is associated with a contrastive objective. Re-weighting each sub-objective separately without overall coordination tends to be sub-optimal. We propose a **Mo**dulated **N**oise **C**ontrastive **E**stimation (**MoNCE**) loss that employs optimal transport [28] to modulate the re-weighting of all negative samples collaboratively across the multiple objectives. With a cost matrix designed based on the similarity of negative pairs, optimal transport allows to retrieve an optimal transport plan which serves as the weights for negative samples to reach an overall optimal objective.

The contributions of this work can be summarized in three aspects. First, we formulate contrastive learning as a versatile metric in multifarious image translation tasks. Second, we extensively investigate the effect of negative pair weighting in contrastive learning and propose to adopt different weighting strategies according to the similarity distribution of negative pairs. Third, we propose a modulated contrast that exploits optimal transport to modulate the re-weighting of all negative pairs collaboratively across multiple contrastive objectives.

2. Related Work

Image Generation Loss Image generation tasks entail various losses to achieve dedicated purposes in image synthesis [23, 24, 32, 39, 40, 43, 44, 47–49]. For instance, unpaired image translation is usually associated with certain losses to encourage correlation between the input and output images. Specially, Zhu *et al.* [56] design a cycle-consistency loss to preserve the image content by ensuring the input image can be recovered from the translation result. However, cycle-consistency loss assumes the relationship between the two domains is a bijection which is often too restrictive for image translation tasks. Therefore, several works [1, 3, 12] aim to explore one-way translation and bypass the bijection constraint of cycle-consistency. At the other end, paired image translation entails certain metric to measure the perceptual similarity between images in line with human perception. By leveraging the internal activation of pre-trained neural networks, perceptual loss [11, 13, 17, 33] emerges as a powerful metric in image translation that coincides with human perception [53]. However, all above metrics are designed based on point-wise absolute deviation, which tends to estimate the median of all possible instances.

With the emergence of contrastive learning, a popular line of research introduces contrastive learning in image generation [10, 18, 42, 51, 52]. Specially, CUT [26] proposes to maximize the mutual information between corresponding patches via noise contrastive estimation [25] for preserving the contents in unpaired image translation. Andonian *et al.* [2] introduce contrastive learning to measure the inter-image similarity in paired image translation. AECR-Net [38] introduces a contrastive regularization for image dehazing by pulling the restored image closer to the clear image and push it far away from the hazy image in the representation space. NEGCUT [34] presents an instance-wise hard negative sample generation framework for contrastive learning in Unpaired image-to-image Translation. However, all previous losses are designed based on the vanilla contrastive learning which indiscriminately repels all negative samples regardless of their similarity to the anchor.

Contrastive Learning The contrastive learning [5, 14, 41] has recently become a prominent tool in unsupervised representation learning, leading to state-of-the-art results. The goal of contrastive learning (CL) is to learn a generic feature

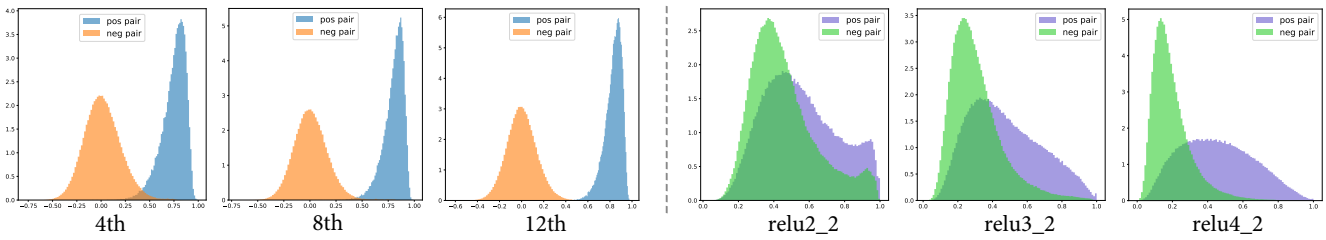

Figure 2. Histograms of positive & negative pair similarity in unpaired and paired image translation. The three plots on the left are the results of 4th, 8th, 12th layers of CUT model [26] for unpaired image translation (Horse to Zebra). The three plots on the right are the results of $relu2\_2, relu3\_2, relu4\_2$ layers of pre-trained VGG-19 in SPADE [27] for paired image translation (ADE20K). The very distinct similarity distribution leads to inverse weighting strategies for unpaired and paired image translation.

embedding by pulling positive points towards an anchor and push negative points far away from it. However, the objective of conventional contrastive learning is misleading as negative samples will be pushed apart indiscriminately with the same weights regardless of their similarity to the anchor. To alleviate the undesired repelling of similar pairs, a popular line of research explores to reweight the NCE loss by increasing the importance of positive pairs [6] or allocating different importance for negative pairs [29]. Besides, Chen *et al.* [4] propose large-margin contrastive learning (LMCL) to distinguish intra-cluster and inter-cluster pairs and only push away inter-cluster pair. However, all described methods explore to re-weight a single contrastive objective for feature representation, which is infeasible for the cases accompanied with multiple contrastive objectives and cannot generalize to the area of image generation.

3. Proposed Method

In this section, we first formulate the contrastive learning as a versatile metric for unpaired and paired image translation tasks. Then we establish the weighting strategies for unpaired and paired image translation according to the similarity distribution of positive and negative pairs. Finally, we derive our designed modulated noise contrastive estimation (MoNCE) loss which enables to coordinate the reweighting of negative pairs across multiple objectives.

3.1. Versatile Metric for Image Translation

Given images in two domains, image translation aims to translate images from the input domain to appear like images from the output domain. The datasets for training translation model could be unpaired (i.e., unpaired image translation) and paired (i.e., paired image translation), and different loss terms are entailed for image translations with different dataset setting. GAN loss is usually shared across unpaired and paired image translation to fight against artifacts in translated images, and other loss terms are usually designed specifically to fulfill various objectives, e.g., cycle loss [56] for content preservation, perceptual loss [17] for assessing human perceptual similarity. However, most metrics are designed by computing the absolute mean error which tends to minimize the average deviation to all possible instances and leads to blurs in the generated images.

In this work, we formulate contrastive loss as a versatile metric in various translation tasks, just by properly selecting the positive and negative pairs. For unpaired image translation, previously proposed PatchNCE [26] has validated the effectiveness of contrastive learning for the preservation of content. PatchNCE aims to maximize the mutual information between patches in the same spatial location from the generated image X and the ground truth Y as below:

$$\mathcal{L}(X,Y) = -\sum_{i=1}^{N} \log \frac{e^{x_i \cdot y_i/\tau}}{e^{x_i \cdot y_i/\tau} + \sum_{\substack{j=1 \\ j\neq i}}^{N} e^{x_i \cdot y_j/\tau}}, \qquad (1)$$

where $X = [x_1, x_2, \cdots, x_N]$ and $Y = [y_1, y_2, \cdots, y_N]$ are encoded image feature sets, τ is the temperature parameter, N is the number of feature patches. Normally, multi-layer features (1th, 4th, 8th, 12th and 16th layers of the encoder) are employed in PatchNCE, which is formulated as $\mathcal{L}^m(X,Y) = \sum_{l=1}^{L} \mathcal{L}(X_l, Y_l)$, where X_l and Y_l denote the corresponding feature sets in l layer of the encoder.

For paired image translation, we aim to measure the perceptual similarity between translated images and the ground truth in line with human perception. Consistent with the multi-layer setting in perceptual loss [17], we employ pre-trained VGG-19 network [31] to extract the same layer features $(relu1\_2, relu2\_2, relu3\_2, relu4\_2, relu5\_2)$ from translated images and ground truth to construct contrastive learning pairs. By treating feature patches in same spatial location from the translated image and the ground truth as positive pairs and feature patches in different location as negative pairs, Eq. (1) can be utilized to maximize the mutual information between translated images and the ground truth. According to the experimental results in Table 3, the PatchNCE with pre-trained VGG-19 for feature extraction rivals the well-known perceptual loss [17] in terms of paired image translation in various evaluation metrics.

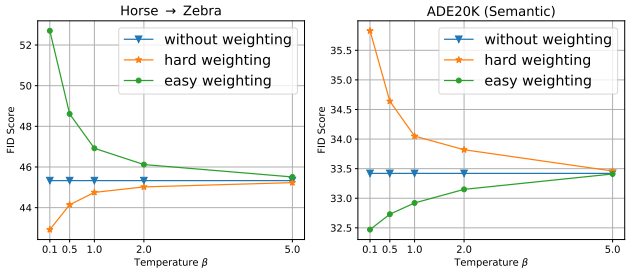

Figure 3. Image translation performance (FID) using different weighting strategies with varying temperature β. The two graphs are the results of unpaired image translation (Horse $\rightarrow$ Zebra using CUT model [26] with WeightNCE) and paired image translation (ADE20K using SPADE model [26] with WeightNCE), respectively.

Considering the superior performance of PatchNCE for unpaired and paired image translation, contrastive learning can serve as a versatile metric in various image translation tasks. However, the vanilla objective of PatchNCE will repel all negative samples indiscriminately regardless of their similarity to the anchor, which tends to be sub-optimal as the inherent information of negative samples is not equal.

3.2. Weighted Contrastive Objective

As each negative sample poses different similarity to anchor, the pushing force of each negative sample should be accordingly adjusted for better contrastive learning [35]. To adjust the pushing force of a negative samples, a simple yet feasible approach is to adjust its weight in the contrastive objective. According to Eq. (1), a higher weight of a negative pair (e.g., $e^{x_i \cdot y_j/\tau}$) indicates a higher importance in contrastive objective, i.e., enlarged pushing force for this negative pair. Thus, the weighted version of Eq. (1) (denoted by **WeightNCE**) can be formulated as:

$$-\sum_{i=1}^{N} \log \frac{e^{x_i \cdot y_i/\tau}}{e^{x_i \cdot y_i/\tau} + Q(N-1)\sum_{\substack{j=1 \\ j\neq i}}^{N} w_{ij} \cdot e^{x_i \cdot y_j/\tau}}, \quad (2)$$

where Q denotes the weight of negative terms ($Q = 1$ by default) in the denominator, w_{ij} ($j \neq i$) denotes the weight between sample y_j and anchor x_i and is subjected to $\sum_{\substack{j=1 \\ j\neq i}}^{N} w_{ij} = 1, i \in [1, N]$.

The weighting strategy could essentially boil down to two categories: assigning higher weights to hard negative samples (referred as **hard weighting** w_{ij}^{+}) and assigning higher weights to easy negative samples (referred as **easy weighting** w_{ij}^{-}). To determine the weighting strategy for unpaired and paired image translation, we illustrate the similarity histograms of positive and negative pairs in three middle layers (4th, 8th, 12th for unpaired image translation, $relu2\_2, relu3\_2, relu4\_2$ for paired image translation) after the contrastive learning is completed. As shown in Fig. 2, for unpaired image translation, there is few overlap between the similarity histograms of positive and negative pairs after contrastive learning, which indicates that positive and negative pairs can be easily pushed apart. In this end, the hard weighting strategy may help to boost the performance, as the model can focus on learning from more informative negative samples (hard negative samples) which has been proved to be beneficial for contrastive learning [29, 35]. However, for paired image translation, there is severe overlap for the similarity histogram of positive and negative pairs, which indicates many negative samples are hard to be distinguished from the positive samples. In this case, hard weighting may not make for contrastive learning as naively using too hard negative samples may degrade the contribution of moderate ones, yielding worse representation [16]. It is reasonable to conjecture that easy weighting may contribute to the contrastive learning in this scenario by assigning lower weights to these hard negative samples which reduces their effects in the contrastive objective.

We validate above conjecture by apply both hard weighting and easy weighting strategy to unpaired and paired image translation, respectively. For the contrastive objective of a single patch, hard weighting weights w_{ij}^{+} and easy weighting weights w_{ij}^{-} are determined with a positive and negative relation to the similarity between sample y_j and anchor x_i as below:

$$w_{ij}^{+} = \frac{e^{(x_i \cdot y_j)/\beta}}{\sum_{j=1}^{N} e^{(x_i \cdot y_j)/\beta}} \quad w_{ij}^{-} = \frac{e^{(1-x_i \cdot y_j)/\beta}}{\sum_{j=1}^{N} e^{(1-x_i \cdot y_j)/\beta}}, \quad (3)$$

where β denotes the weighting temperature parameter. We take the value of temperature β and the FID score of generated images as the abscissa and ordinate, respectively, as show in Fig. 3. Treating the generation performance without weighting as the baseline, we can observe that the performance of unpaired image translation benefits from hard weighting strategy, and presents a positive correlation with the decreasing of β. On the other hand, paired image translation performance benefits from easy weighting strategy, and also presents a positive correlation with the decreasing of β, which is consistent with our conjecture. Despite some previous work [29, 35] proves the effectiveness of hard negative samples for contrastive learning, we would clarify that the bad effect of excessively hard samples have overwhelmed their positive effect in the case of paired image translation.

In the above experiments, the weighting strategy in Eq. (3) is applied to each contrastive sub-objective separately. However, all contrastive sub-objectives are contributing to the final objective as in Eq. 2. Weighting each sub-objective independently without overall coordination may result in conflicts between different sub-objectives, and thus tends to be sub-optimal for the final objective.

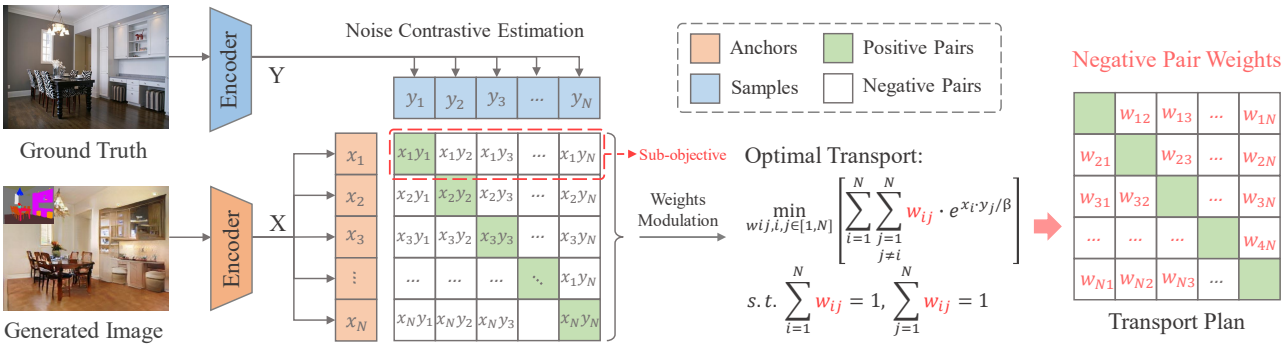

Figure 4. Framework of the proposed modulated contrast. There are multiple sub-objectives for the contrastive learning between feature set $X = [x_1, x_2, x_3, \cdots, x_N]$ and $Y = [y_1, y_2, y_3, \cdots, y_N]$. To modulate the weights of negative pairs across multiple sub-objectives, optimal transport with a cost matrix C (defined by $C_{ij} = e^{x_i \cdot y_j / \beta}$ for unpaired translation, $C_{ij} = e^{(1-x_i \cdot y_j)/\beta}$ for paired translation) is conducted between feature sets X and Y to minimize the total transport cost, yielding an optimal transport plan which serves as the weights of the corresponding negative pairs.

3.3. Modulated Contrastive Objective

As we are exploring **re-weighing** strategies, the total weight associated with a feature (x_i or y_j) is expected to be constant, thus yielding below constraints:

$$\sum_{i=1}^{N} w_{ij} = 1, \quad \sum_{j=1}^{N} w_{ij} = 1, \quad i, j \in [1, N]. \tag{4}$$

Considering the contrastive objective as illustrated in Fig. 4, a feature y_j serves as negative sample for multiple sub-objectives. As the total weights associated with y_j is constant (i.e., $\sum_{i=1}^{N} w_{ij} = 1$), there may be conflicts for the weighting strategies of y_j in different sub-objectives, e.g., several sub-objectives all expect a higher weight for y_j while the total weights of y_j is constrained. Therefor, we aims to modulate the assignment of weights w_{ij} ($i, j \in [1, N], i \neq j$) across multiple sub-objectives with the constraint of constant total weight.

Targeting to modulate the weights assignment for all negative pairs, a weight modulation goal shared across all contrastive sub-objectives should be determined. We take easy weighting strategy as an example to derive the final weight modulation goal. By assigning higher weights to negative pairs with low similarity, the easy weighting strategy for a contrastive sub-objective in Eq. (2) is equivalent to reducing the negative term $\sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} \cdot e^{x_i \cdot y_j / \tau}$. As the contrastive objectives of all image patches are summed to form the final objective, the shared modulation goal across multiple contrastive objectives can be regarded as reducing the total loss of negatives terms. To derive the expression mathematically, the objective of the modulation goal is formulated as *minimizing* the total loss of negative terms with regarding to $w_{ij}, i, j \in [1, N]$:

$$\min_{w_{ij}, i, j \in [1,N]} \left[\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} \cdot e^{x_i \cdot y_j / \tau} \right]. \tag{5}$$

The formation of Eq. (5) with constraint in Eq. (4) can be regarded as an optimal transport (OT) [28] problem between $[x_1, x_2, \cdots, x_N]$ and $[y_1, y_2, \cdots, y_N]$ with a cost matrix C defined by $C_{ij} = e^{x_i \cdot y_j / \beta}$ for $i \neq j$ and $C_{ij} = \inf$ for $i = j$. Similar to weighting temperature in Eq. (3), β in the cost matrix C serves as a cost temperature that indicates the smoothness of the optimal transport. A smaller β tends to assign higher weights for small cost entries C_{ij} and a large β tends to assign equal weights for all cost entries. Detailed parameter study of β can be found in the experiment part.

The optimal transport aims to retrieve a transport plan T which minimizes the total transport cost as formulated below:

$$\min_{T} \langle C, T \rangle, \quad s.t. \ (T\vec{1}) = 1, \ (T^\top \vec{1}) = 1, \tag{6}$$

where $\langle C, T \rangle$ denotes the inner product of C and T. Thus, solving the transport plan T is equivalent to solve the weight parameters as $w_{ij} = T_{ij}$. The Sinkhorn algorithm [8] can be applied to Eq. (6) for approximating optimal transport solution, yielding the desired optimal transport plan T. With the derived transport plan matrix T as the weights of negative pairs, the modulated objective for easy weighting strategy is accordingly determined. For hard weighting strategy, the modulated objective can be derived similarly, just redefining the cost matrix C in Eq. 6 as $C_{ij} = e^{(1-x_i y_j)/\beta}$ for $i \neq j$ and $C_{ij} = \inf$ for $i = j$.

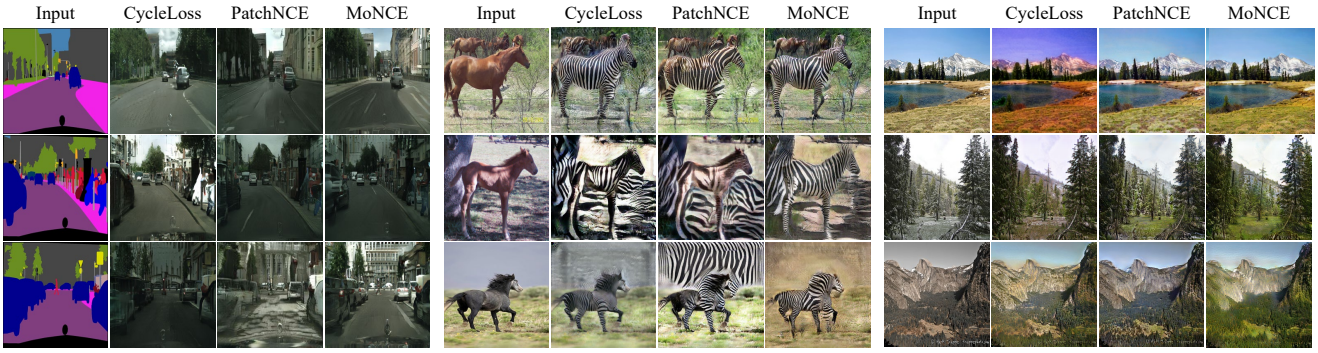

Figure 5. Qualitative comparison of different losses for unpaired image translation tasks including Cityscapes (Semantic → Image), Horse → Zebra, Winter → Summer. The structure of CUT [26] is employed for the translation.

CUT [26]	Cityscapes (Semantic → Image)				Horse → Zebra		Winter → Summer	
	FID ↓	mAP ↑	pixAcc ↑	classAcc ↑	FID ↓	SWD ↓	FID ↓	SWD ↓
Baseline (GAN Loss)	139.9	9.705	23.44	14.17	129.8	74.85	136.2	47.80
+Cycle Loss [56]	75.97	20.53	55.87	25.23	76.37	50.54	86.14	38.79
+PatchNCE [26]	57.16	24.29	78.22	30.67	45.33	32.02	80.25	36.92
+WeightNCE	55.94	24.98	77.92	31.96	42.92	31.58	79.32	36.39
+MoNCE	**54.67**	**25.61**	**78.41**	**33.02**	**41.86**	**30.80**	**78.18**	**35.95**

Table 1. Unpaired image translation performance on different tasks with CUT [26] as the model structure.

F/LSeSim [54]	Horse→Zebra		Winter→Summer	
	FID ↓	SWD ↓	FID ↓	SWD ↓
Random SeSim [20]	72.18	48.85	125.1	57.48
FSeSim	43.26	36.77	79.14	35.79
LSeSim+PatchNCE	40.12	34.77	78.30	34.47
LSeSim+WeightNCE	38.67	32.59	76.98	33.89
LSeSim+MoNCE	**37.21**	**32.12**	**76.04**	**33.10**

Table 2. Unpaired image translation with F/LSeSim [54] as the model structure.

4. Experiments

4.1. Experimental Settings

Datasets: For unpaired image translation, we conducted experiments on Cityscapes, Horse → Zebra, and Winter → Summer. For paired image translation, we conducted experiments on ADE20K, CelebA-HQ, and DeepFashion.

• Cityscapes [7] contains 2,975 training and 500 validation images captured on street. We conduct unpaired semantic-to-image translation on this dataset.

• Horse → Zebra [56] collects 1187 horse images and 1474 zebra images from ImageNet [9] for training and validation.

• Winter → Summer [56] contains 1,200 winter images and 1,540 summer images for training and validation.

• ADE20k [55] consists of 20k training images with 150-class segmentation masks. We conduct image generation by using its semantic segmentation as conditional inputs.

• CelebA-HQ [22] consists of 30,000 face images. We use its semantic map and edge maps for conditional generation.

• DeepFashion [21] contains 52,712 person images. We use its keypoints as conditional inputs in experiments.

Evaluation Metrics: Several evaluation metrics are adopted in our experiment to assess image translation performance. *Fréchet Inception Score (FID)* [15] and sliced *Wasserstein distance (SWD)* [19] are adopted to measure distribution discrepancy and statistical distances of low level patches between translated images and real images, respectively. For semantic image translation tasks, we employ pre-trained segmentation model to evaluate the segmentation accuracy, e.g., mean average precision (mAP) and pixel accuracy (Acc).

Implementation Details: All experiments are conducted with an image resolution of 256×256. For contrastive learning setting, we keep the same with CUT [26], e.g., 256 negative samples, temperature parameter $\tau = 0.07$. The default temperatures β and weight term weight Q in WeightNCE and MoNCE are 0.1 and 1, respectively, for all tasks. We re-train all compared methods following above setting to ensure fair comparison.

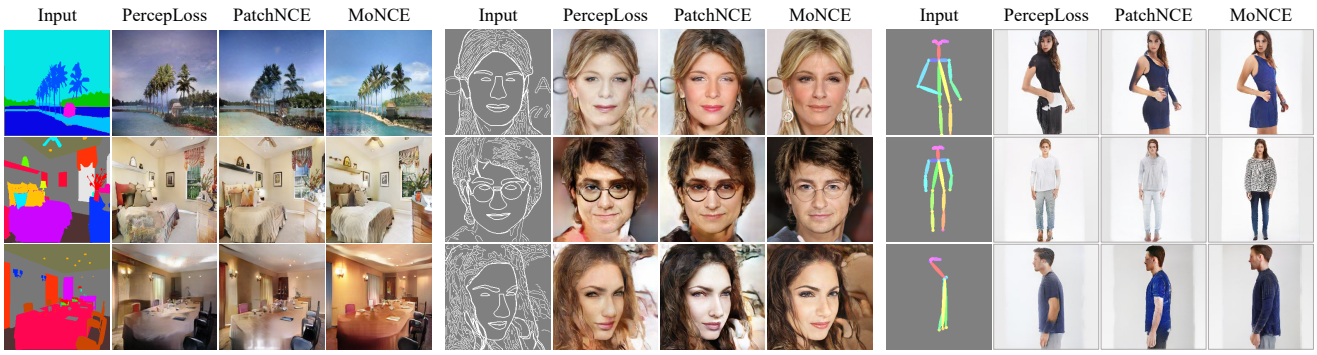

Figure 6. Qualitative comparison of different losses for paired image translation tasks including ADE20K (Semantic), CelebA-HQ (Edge), and DeepFashion (Keypoint). The structure of SPADE [27] is employed for the translation.

SPADE [27]	**ADE20K (Semantic)**			**CelebA-HQ (Semantic)**		**CelebA-HQ (Edge)**		**DeepFashion (Keypoint)**	
	FID ↓	mIoU ↑	Acc ↑	FID ↓	SWD ↓	FID ↓	SWD ↓	FID ↓	SWD ↓
Baseline (GAN Loss)	87.32	31.32	76.79	86.91	25.93	84.04	27.35	**28.57**	22.18
+PercepLoss [17]	33.68	42.23	81.96	36.54	**17.28**	31.53	18.25	35.74	24.03
+PatchNCE [26]	33.42	44.91	81.92	33.38	21.90	30.81	23.14	38.04	23.53
+WeightNCE	32.47	45.69	83.52	32.15	18.86	30.49	21.51	36.83	22.98
+MoNCE	**31.62**	**46.30**	**84.29**	**30.01**	17.39	**29.75**	**18.11**	33.96	**21.58**

Table 3. Paired image translation with different types of conditional input. The model structure of SPADE [27] is employed to compare the performance of different losses.

4.2. Unpaired Image Translation

We evaluate our proposed loss on the classical unpaired image translation task. We first adopt the model structure of CUT [26] to conduct comparison between CycleLoss [56], PatchNCE loss [26], and our proposed WeightNCE and MoNCE. Complying with the discussion in the Sec. 3.2, the weighting strategy of assigning higher weights to hard negative samples is adopted for unpaired image translation. As shown in Table 1, the model with GAN Loss only is used as the Baseline. The four different losses are further included into the Baseline, respectively, for comparisons. We can observe that the proposed WeightNCE and MoNCE both outperform the CycleLoss and PatchNCE consistently in all compared unpaired translation tasks. With an overall weight modulation across multiple contrastive objectives, the proposed MoNCE outperforms WeightNCE across all evaluation metrics. Fig. 5 shows qualitative comparisons on unpaired image translation. All compared methods adopt the same structure with CUT and the only variation comes from different losses.

Besides CUT model, we also compare the four losses with the F/LSeSim [54] model, which exploits the spatial patterns of self-similarity to preserve image structures in unpaired image translation. The content loss [20] using random sampled features for computing self-similarity is selected as the baseline (Random SeSim). F/LSeSim could employ a pre-trained VGG-16 [31] (namely FSeSim) or a PatchNCE (namely LSeSim) to learn spatially correlative maps. Here, we replace the PatchNCE with our WeightNCE and MoNCE to conduct the comparison. As shown in Table 2, the learnable self-similarity setting (LSeSim+PatchNCE) outperforms the fixed self-similarity setting with pre-trained VGG-16 [31]. Consistent with the results in CUT, replacing the PatchNCE with our WeightNCE and MoNCE also bring notable improvement in translation quality.

4.3. Paired Image Translation

For paired image translation, we adopt the structure of SPADE [27] to perform the comparison between PercepLoss [17], PatchNCE [26], and our proposed WeightNCE and MoNCE. The SPADE model with GAN loss only is selected as the baseline. Then the Baseline is combined with different losses to perform the comparisons. As shown in Table 3, PatchNCE with pre-trained VGG-19 for feature extraction could rival the well-known PercepLoss across all generation tasks in terms of generation quality. Considering the performance of vanilla PatchNCE in unpaired and paired image translation, contrastive learning has good potential to serve as a versatile metric for measuring image similarity. Besides, we can observe the MoNCE is advantageous to WeightNCE and both of them outperform the PatchNCE consistently, which verify the effectiveness of

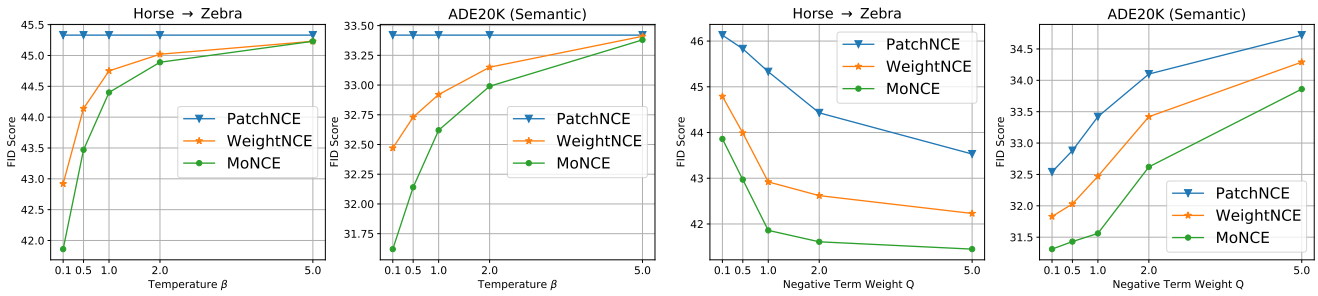

Figure 7. The effect of varying temperature parameter β and negative term weight Q on unpaired image translation (Horse $\rightarrow$ Zebra) and paired image translation (ADE20K (Semantic)).

MoNCE Variants			Performance	
Bidirectional	Pre-trained	Frozen	FID ↓	mIoU ↑
✓	✓	✓	31.31	**46.71**
✓	✓	✗	**30.94**	44.89
✓	✗	✗	40.47	38.22
✗	✗	✗	42.86	36.16
✗	✓	✗	31.37	45.44
✗	✓	✓	31.62	46.30

Table 4. Ablation study of MoNCE variants on paired image translation (ADE20K). The configuration at the grey row is the default setting of MoNCE.

our weighting strategies and modulation mechanism.

Fig. 6 shows qualitative experiments with SPADE with different losses. We can see that the images translated with PatchNCE tends to present less artifacts compared that with PercepLoss, as contrastive learning aims to maximize the mutual information of corresponding images instead of naively minimizing the point-wise absolute deviation. With an overall modulation of easy weighting stategies, our MoNCE outperforms PatchNCE clearly with more fine details in generated images.

4.4. Discussion

We conduct experiments on unpaired image translation (Horse $\rightarrow$ Zebra) and paired image translation (ADE20K (Semantic)) to examine the effect of the cost temperature β in Eq. (6). As show in Fig. 7, the generation performance (FID score) of unpaired and paired image translation improves consistently while decreasing the temperature β. However, we find the model training tends to be unstable and even fail with small temperature β, e.g., 0.01. We also ablate the effect of the negative terms weight Q in Eq. (2). As shown in Fig. 7, the performance of unpaired image translation and paired image translation presents positive correlation and negative correlation with the increasing of negative term weight Q, respectively. Although the FID is improved with a larger Q, we observe that the content preservation performance is actually degraded for unpaired image translation. Based on above observation, we set the temperature β as 0.1 and the negative term weight Q as 1 by default.

We also explore several variants of contrastive learning on paired image translation (ADE20K), including without pre-trained VGG-19 network, unfrozen pre-trained VGG-19 network, and bidirectional designing of contrastive learning (including the contrastive objective with ground truth patches as anchors) introduced in [2]. As shown in Table 4, learning the feature extractor from scratch without pre-training tends to impair the generation performance drastically. Including bidiretional design to the proposed MoNCE improve the generation performance slightly. Unfreezing the pre-trained VGG-19 network improves the FID, while it hurts the mAP score.

5. Conclusion

We have formulated contrastive learning as a versatile metric for various image translation tasks, which is on par with the prevailing losses designed in corresponding tasks. With a target to re-weighting negative pairs for performance gain, we explore and establish the weighting strategies for unpaired and paired image translation according to the similarity distribution of positive and negative pairs. To modulate the re-weighting of all negative pairs associated with the full image, we further derive a MoNCE which employs optimal transport to retrieve the optimal weights for negative pairs across multiple contrastive objectives. Our thorough and extensive analysis of negative pair weighting strategies lays a sound foundation for the exploration of contrastive learning in image generation.

6. Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8983–8992, 2019. 2

[2] Alex Andonian, Taesung Park, Bryan Russell, Phillip Isola, Jun-Yan Zhu, and Richard Zhang. Contrastive feature loss for image prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1934–1943, 2021. 2, 8

[3] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[4] Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, and Masashi Sugiyama. Large-margin contrastive learning with distance polarization regularizer. In *International Conference on Machine Learning*, pages 1673–1683. PMLR, 2021. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[6] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *arXiv preprint arXiv:2007.00224*, 2020. 3

[7] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020. 2

[11] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29:658–666, 2016. 2

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 2

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2

[15] 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*, pages 6626–6637, 2017. 6

[16] Sangryul Jeon, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Mining better samples for contrastive learning of temporal correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1034–1044, 2021. 4

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 2, 3, 7

[18] Minguk Kang and Jaesik Park. ContraGAN: Contrastive Learning for Conditional Image Generation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[20] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 6, 7

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6

[23] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conference on Computer Vision*, pages 427–443. Springer, 2018. 2

[24] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 2

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[26] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 1, 2, 3, 4, 6, 7

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1, 3, 7

[28] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2, 5

[29] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 2, 3, 4

[30] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 1

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 7

[32] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 2

[33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 2

[34] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14020–14029, 2021. 2

[35] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 4

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[37] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 1

[38] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 2

[39] Rongliang Wu and Shijian Lu. Leed: Label-free expression editing via disentanglement. In *European Conference on Computer Vision*, pages 781–798. Springer, 2020. 2

[40] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020. 2

[41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 1, 2

[42] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021. 2

[43] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14114–14123, 2021. 2

[44] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 2

[45] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[46] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Kaiwen Cui, Aoran Xiao, Shijian Lu, and Ling Shao. Bi-level feature alignment for semantic image translation & manipulation. *arXiv preprint*, 2021. 1

[47] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Changgong Zhang, Shijian Lu, Ling Shao, Feiying Ma, and Xuansong Xie. Gmlight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*, 2021. 2

[48] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. 2

[49] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3287–3295, 2021. 2

[50] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 1

[51] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021. 2

[52] Jiahui Zhang, Shijian Lu, Fangneng Zhan, and Yingchen Yu. Blind image super-resolution via contrastive representation learning. *arXiv preprint arXiv:2107.00708*, 2021. 2

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2

[54] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16407–16417, 2021. 6, 7

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 3, 6, 7

[57] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 1