A Practical Contrastive Learning Framework for Single Image Super-Resolution

Gang Wu[1] Junjun Jiang[1,2*] Xianming Liu[1,2] Jiayi Ma[3]
[1]Harbin Institute of Technology [2]Peng Cheng Laboratory [3]Wuhan University

Abstract

*Contrastive learning has achieved remarkable success on various high-level tasks, but there are fewer methods proposed for low-level tasks. It is challenging to adopt vanilla contrastive learning technologies proposed for high-level visual tasks straight to low-level visual tasks since the acquired global visual representations are insufficient for low-level tasks requiring rich texture and context information. In this paper, we propose a novel contrastive learning framework for single image super-resolution (SISR). We investigate the contrastive learning-based SISR from two perspectives: sample construction and feature embedding. The existing methods propose some naive sample construction approaches (*e.g.*, considering the low-quality input as a negative sample and the ground truth as a positive sample) and they adopt a prior model (*e.g.*, pre-trained VGG [48] model) to obtain the feature embedding instead of exploring a task-friendly one. To this end, we propose a practical contrastive learning framework for SISR that involves the generation of many informative positive and hard negative samples in frequency space. Instead of utilizing an additional pre-trained network, we design a simple but effective embedding network inherited from the discriminator network and can be iteratively optimized with the primary SR network making it task-generalizable. Finally, we conduct an extensive experimental evaluation of our method compared with benchmark methods and show remarkable gains of up to **0.21 dB** over the current state-of-the-art approaches for SISR.*

1. Introduction

Contrastive learning has been an effective paradigm for unsupervised representation learning. The approaches based on the pretext task of instance discrimination learn visual representations by making views from the same instance similar and that from different instances dissimilar [6–8, 22, 60]. The learned visual representations are beneficial for a wide variety of downstream tasks, particularly high-level tasks, and achieve promising results, *e.g.*, supervised image classification [29], image clustering [34], fine-gained image classification [3], and knowledge distillation [15, 52].

When it comes to low-level image processing tasks, there are some challenges to directly applying contrastive learning approaches. Firstly, the learned global visual representations are inadequate for low-level tasks that call for rich texture and context information. Secondly, a series of data augmentations have been proposed to generate positive and negative samples for high-level downstream tasks [4, 6, 22, 53]. However, except for some simple geometric augmentations (*e.g.*, rotation augmentation in widely used self-ensemble scheme), most of the complicated data augmentations cannot maintain the dense pixel correspondences and thus are not suitable for low-level tasks. Thirdly, a meaningful latent space (or embedding space) is required for contrastive loss. In contrast to high-level tasks that try to obtain the best semantic representations, low-level tasks aim at reconstructing restored results at data space. It is of great importance to explore a proper and meaningful embedding space where the contrastive loss can be effectively defined.

The current contrastive learning-based methods for low-level tasks mainly focus on exploiting negative samples, while let the ground truth image as the positive sample. For example, [59] treats the degraded image (the input hazy image) as a negative sample and presents a novel image dehazing method with a contrastive regularization. In [57] and [19], they take other examples in the dataset as negative samples for contrastive image super-resolution and underwater image restoration. These methods make a preliminary attempt and have proved the effectiveness of incorporating the contrastive constraint to the low-level tasks. Another line of research is to model the statistical characteristics of the image by contrastive learning. Dong *et al*. [14] assume that two patches from the same sample have similar noise distributions and two these from two different samples have two different noise distributions, and develop a residual contrastive loss for joint demosaicking and denoising. In a similar vein, Chen *et al*. [9] propose an unpaired image deraining method based on rain space contrastive

*Corresponding author: jiangjunjun@hit.edu.cn

learning, which can better help rain removal when compared with the original image space. In [63], Zhang *et al*. present a contrastive learning strategy in the feature channel space to obtain resolution-invariant features. They take feature maps of different channels as samples and postulate that the corresponding channel of low-resolution (LR) and high-resolution (HR) feature maps are positive while these from different are negative. Wang *et al* [55] apply the contrastive loss to pre-train a kernel estimation model which aims at separating different degradations and obtaining the degradation-aware representation.

Table 1 summarizes the characteristics of current contrastive learning-based image restoration methods, which are presented most recently. The positive samples are defined as the ground truth, while the negative samples are simply defined as the degraded images or other images in the dataset [19, 57, 59]. These negative samples are dissimilar to the reconstructed image and easily distinguished, *i.e.*, they are too distant to contribute to the contrastive loss. According to the specific image restoration tasks, another line of research [9, 14, 55, 63] all try to generate some invariant (global) features of image, which are immune to the noise, rain, resolution, and blur, based on constrictive learning. They overlooked the ingredient of constructing effective positive and negative pairs for the reconstructed image. In addition, since the constrictive losses of these methods are defined on some specific embedding space and cannot well generalize to other methods.

In this paper, we investigate contrastive learning for single image super-resolution (SISR) and propose a practical contrastive learning framework for SISR (PCL-SISR), where we simultaneously construct multiple negative samples and multiple positive samples. As revealed by recent studies, the super-resolved results of current deep learning methods are smooth and look unnatural and implausible (may be averaged from all possible outputs of the SR network). Based on these observations and hard negative mining studies [10, 46], we propose to generate multiple hard negative samples by applying some slight blurry to the ground truth and generate multiple informative positive samples by simply sharpening the ground truth, resulting in obtaining the informative positive and hard negative pairs for the super-resolved image. In this way, we believe that more hard negative samples will encourage the super-resolved image far from smooth results, while more positive samples will force the network to draw in more detailed information.

Furthermore, different from the existing methods that need an additional pre-trained feature embedding network (*e.g*., pre-trained VGG [48] model), we propose to leverage a cheap and task-friendly feature embedding network, the discriminator of our SR network, to embed positive/negative/anchor samples into a feature space where the contrastive loss will be effectively defined. For an anchor sample, the contrastive loss will push it away from negative samples and pull it close to positive ones. Since our feature embedding network is inherited from the discriminator of our SR network which distinguishes super-resolved images from high-resolution (HR) images, the embedded features will be sensitive to degradation. Therefore, the super-resolved image can be well separated from negative samples while remaining near to positive ones.

Our contributions are summarized as follows: (i) We propose a practical contrastive learning framework for the SISR task. We propose a valid and task-specific data augmentation strategy to generate multiple informative positive and hard negative samples. (ii) We rethink and propose a novel way to obtain a task-friendly or task-generalizable feature embedding where contrastive loss works efficiently by reusing the discriminator of our SR network. The proposed loss is generic and can be applied to any existing SISR and other image restoration frameworks. (iii) Extensive experiments show that our method, dubbed *PCL-SISR*, outperforms several representative SISR methods in terms of quantitative and qualitative results. In addition, ablation studies are conducted to verify the effectiveness of different parts of the proposed method.

2. Related work

2.1. Contrastive Learning

Contrastive learning is an effective paradigm in unsupervised representation learning by maximizing mutual information, and has been widely studied in recent years [6,22,23,43,60]. For a given anchor sample, the contrastive loss, similar to previous works [18, 47, 50] in deep metric learning, aims at pushing it away from negative samples and pulling it close to positive samples in latent space. On the one hand, the selection of the negative and positive samples is of great significance and it depends on specific downstream tasks. [6, 22, 60] take randomly augmented samples from the anchor sample as positive samples and samples from others are negative samples, and [53] analyses how to build an optimal augmentation strategy by minimizing mutual information between two augmented views. [4] generates positive and negative samples by feature space augmentation strategy and [10, 46, 49] improve final results using hard negative mining strategy. On the other hand, contrastive loss is conducted and worked in latent space, and the learned representation contains global semantic information which is instrumental in many high-level tasks.

2.2. Constrictive Learning for Image Restoration

For single image dehazing task, the proposal of [59] utilizes a pre-trained VGG [48] model to obtain the latent embeddings. For the restored image, it takes ground truth

Table 1. Comparison of our proposed contrastive learning framework with current contrastive learning approaches for low-level tasks.

Ref.	Task	Positive/Negative samples	Feature Embedding Space	Generalization
CVPR'21 [59]	Dehazing	Ground Truth/Hazy Image	Additional VGG Embedding	✓
IJCAI'21 [57]	SISR	Ground Truth/Other Images	Additional VGG Embedding	✓
IGARSS'21 [19]	Underwater Image Restoration	Ground Truth/Degraded Image	Additional VGG Embedding	✓
arXiv'21 [63]	Blind SR	Same Channel Feature/Different Channel Features	Task-specific Embedding	✗
CVPR'21 [55]	Blind SR	Same Blur Style/Different Blur Styles	Task-specific Embedding	✗
arXiv'21 [9]	Unpaired Deraing	Same Rain Style/Different Rain Styles	Task-specific Embedding	✗
arXiv'21 [14]	Demosaicking and Denoising	Same Noise Style/Different Noise Styles	Task-specific Embedding	✗
Ours	SISR	GT and Multiple Sharpen Images/Multiple Slightly Blury Images	Task-Generalizable Embedding	✓

image as positive sample and corresponding hazy image as negative sample respectively. Then contrastive loss is conducted with intermediate feature maps extracted from VGG model. For blind super-resolution (BSR) task, Wang *et al* [55] apply contrastive loss to pre-train a kernel estimation model which is trained to distinguish different degradations and obtains the degradation-aware representation. They assume patches from the same image are under the same degradation and that from different images are not the same so that positive and negative samples are selected respectively. The work in [63] builds a two-stage BSR model. Firstly, an image encoder is pre-trained with LR and HR pair images to learn resolution-invariant features by contrastive loss. Then, another contrastive loss is adopted to refine super-resolved images by drawing super-resolved image (anchor sample) closer to the ground truth image (positive sample) and away from the corresponding LR image (negative sample) in latent space. Zhang *et al* [63] introduce a new way to obtain latent space by an additional encoder. However, the encoder is trained with contrastive loss simultaneously and it cannot avoid trivial solution. When the encoder is just degradation-aware, the negative pair distance will large enough and the positive pair distance will be very small so that the contrastive loss will be useless. The way to obtain a proper latent space for SISR is our concern.

2.3. Constrictive Learning for I2I Translation

CUT [44] is the pioneer work that applies contrastive loss to the unpaired image-to-image (I2I) translation by proposed patch-based contrastive loss to learn better structure preserved and style transferred features in latent space. For the given patch of input image, positive and negative samples are the corresponding transferred patch and other random patches from the same domain. The recent work DCLGAN [20], based on CUT, takes the advantages of the CycleGAN [69] and adopts a bidirectional patch-based contrastive loss between source domain and target domain. It is reasonable that they apply contrastive loss between source and target domain in a common latent space based on CycleGAN framework. In this paper, we focus on single image super-resolution task where there is only an SR network working with LR input and we cannot obtain a common latent space for both LR and HR domains.

2.4. Single Image Super-Resolution

Deep-learning-based methods have dominated the single image super-resolution field in recent years. Dong *et al*. proposed the first CNN-based SR method, named SRCNN. Since that, various efficient and deeper architectures have been proposed for SR and the performance on benchmark datasets has been continuously improved by newly developed network architectures [11, 21, 26, 37–39, 42, 67, 68]. Most recently, transformer-based architectures are explored for SR [5, 36]. Besides investigating more powerful network architectures, some perceptual-driven approaches also have been proposed utilizing perceptual loss functions to achieve better visual quality [28, 33, 56, 66]. Compared with the aforementioned SISR methods, our proposed contrastive learning approach is a unified training pipeline and it can work with any exiting SR network to well balance the distortion and perception.

3. Method

In this section, we introduce our proposed practical contrastive learning framework for SISR in detail. We first introduce preliminaries of contrastive learning and then we describe our positive and negative sample generation strategy and the training of our feature encoder. At last we present our main framework which employs contrastive learning to further improve the performance of existing SISR works.

3.1. Preliminaries

Contrastive learning is one of the most powerful approaches for representation learning. It aims at pulling the anchor sample close to the positive samples and pushing it far away from negative samples in latent space [6, 22, 60]. For the image dataset $\mathcal{I}$, the representation learning model E is trained to extract representations $\mathcal{R} = \{r_i | r_i = E(I_i), I_i \in \mathcal{I}\}$ with InfoNCE loss [23, 43]. The loss $\mathcal{L}_{InfoNCE}$ is based on a softmax formulation and for the i-th sample the loss $\mathcal{L}_i$ is as formulated as follows:

$$\mathcal{L}_i = -\log \frac{\exp(r_i^T \cdot r_i^+ / \tau)}{\exp(r_i^T \cdot r_i^+ / \tau) + \sum_{j=1}^{K} \exp(r_i^T \cdot r_j^- / \tau)}, \tag{1}$$

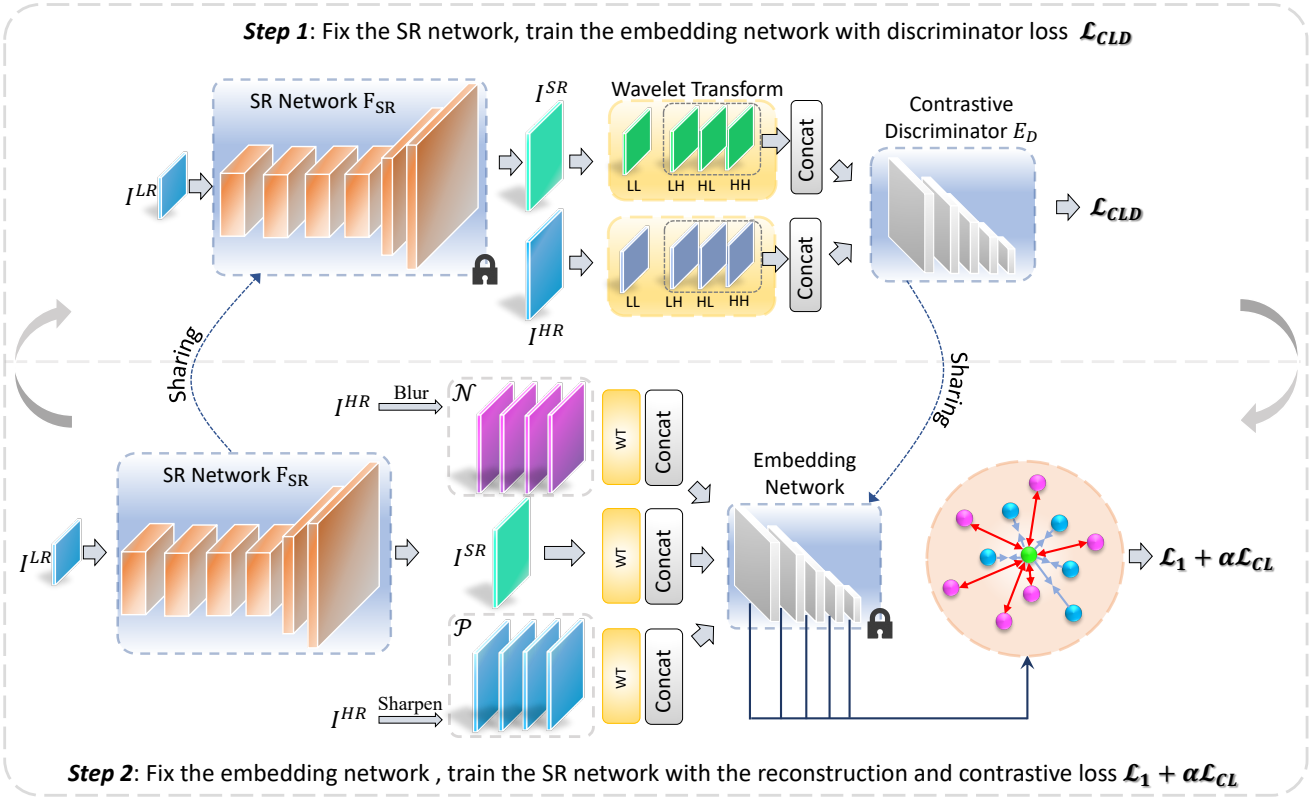

Figure 1. Overview of our proposed contrastive learning framework for SISR. We adopt the GAN-like training processing that updating our embedding network E_D and target SR network F_{SR} iteratively. We train our embedding network to learn degradation-aware features. Then E_D is frozen and SR network is trained with the pixel-wise construction loss $\mathcal{L}_P$ and our contrastive loss $\mathcal{L}_{CL}$.

where τ is the temperature hyper-parameter. r_i^+ means the representation of the positive sample usually generated by random data augmentations from the same sample I_i. K is the number of negative samples and $\{r_j^-\}_{j=1}^K$ is the set of negative representations from negative samples $\{I_j | I_j \in \mathcal{I}, j \neq i\}_{j=1}^K$ that are random selected other images from dataset. Finally, total contrastive loss is as follows:

$$\mathcal{L}_{InfoNCE} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i. \tag{2}$$

In addition, the work in [29] modified and applied contrastive loss to the supervised classification task where there are more than one positive samples. For the i-th image, this supervised contrastive loss is as follows:

$$\mathcal{L}_i = -\frac{1}{P}\sum_{p=1}^{P}\log\frac{\exp(r_i^T \cdot r_p^+/\tau)}{\exp(r_i^T \cdot r_i^+/\tau) + \sum_{j=1}^{K}\exp(r_i^T \cdot r_j^-/\tau)}, \tag{3}$$

where P is the number of the positive set, noted as $\{r_p^+\}_{p=1}^P$. Then the total supervised contrastive loss $\mathcal{L}_{SupCL}$ is:

$$\mathcal{L}_{SupCL} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i. \tag{4}$$

Contrastive learning is beneficial for various downstream tasks and achieves promising performance. As we described in Sec. 1, we can find that to employ contrastive learning methods, careful designed sample selection and construction strategy and a task-related latent space needs to be explored. Next, we will describe our sample generation strategy from the perspective of frequency domain and how to train a task-friendly embedding network in SISR instead of a pre-trained model, such as VGG network.

3.2. Positive and Negative Sample Generations

The SISR task aims to transform low-resolution images (noted as $\mathcal{I}^{LR}$) into sharp, realistic, and high-resolution images (noted as $\mathcal{I}^{HR}$). As LR images are formulated from an image degradation process and contain only the low-frequency information, SR model focuses on learning a reverse translation to recover the lost high-frequency components (*e.g.*, edge and texture information). With this in

mind, we take valid and task-specific data augmentations to generate our positive and negative samples.

Informative positive sample generation. In addition to the only HR ground truth, we further generate K_P sharpened images as positive set $\mathcal{P}_i$ by applying different high-pass kernels on HR image. For the i-th image, we denote its positive set as follows:

$$\mathcal{P}_i = \{P_j | P_j = \text{Sharpen}(I_i^{HR})\}_{j=1}^{K_P}, \tag{5}$$

where K_P is the number of positive samples. Sharpen presents a random sharpness function. This is different from the existing contrastive learning-based image restoration methods, which consider only the ground truth as the positive sample [59, 63]. It should be noted that in order to generate more informative positive samples, we apply different high-pass kernels to the HR ground truth. This positive sample generation strategy is designed following two observations: (i) the object of the SISR task is to obtain detailed results. We can use some informative positive samples to induce more high-frequency details for the reconstruction results. (ii) SISR is an ill-posed problem, and the mapping between the LR and HR images is "one-to-many". That is, the number of ground truth should not only be one, and there are many possible HR samples except the given ground truth [27]. Our proposed positive sample generation method can be seen as a very coarse one.

Hard negative sample generation. In order to introduce contrastive learning to the low-level image restoration problems, recent works [19,57,59] simply task the degraded images (*e.g.* the input hazy image or LR image) or other images in the dataset. When compared with the target reconstructed image, these negative samples are dissimilar to them and easy to be distinguished. Therefore, they may provide a loss constraint on the solution. A natural idea is whether we can feed some difficult hard examples, which are very similar to the ground truth, into the contrastive learning model. Therefore, in this paper we adopt multiple negative samples here which can narrow down the solution space and can further improve the performance of the SR network with contrastive loss. Specifically, inspired from the hard negative samples mining and adversarial training methods [10, 24, 46], here we generate slight blurry images from the ground truth as our hard negative sample set $\mathcal{N}_i$ because they are close to the ground truth, thus forcing the reconstructed SR image become closer to the ground truth. For the i-th image, we denote its negative set as follows:

$$\mathcal{N}_i = \{N_j | N_j = \text{Blur}(I_i^{HR})\}_{j=1}^{K_N}, \tag{6}$$

where K_N is the number of negative samples and we use $K_N = K_P = 4$ as default. Blur is our blur function with random Gaussian kernel.

3.3. Feature Embedding Network

In this section, we introduce a simple but efficient way to obtain a task-friendly embedding network. As described in Sec. 1, VGG based perceptual loss is widely adopted [33, 56] and recent work in [59] designs a contrastive loss based on the pre-trained VGG model. We believe that a task-friendly embedding network is better because the features obtained by VGG tend to be the high-level semantic information for the classification task. In addition, compared with training the SISR task, pre-training on the ImageNet is a very heavy task. Furthermore, a good embedding network for SISR should be degradation-aware so that contrastive loss can work even the super-resolved results are very close to the ground truth. In other words, a good embedding network should be able to distinguish changes in details.

Inspired by adversarial learning approaches in [17, 24], we find that the discriminator learned in the vanilla GAN framework is degradation-aware because it can correctly classify whether the input image is fake or not. With this in mind, we employ a GAN-like framework to obtain a task-friendly embedding network by forcing it to distinguish SR and HR images as illustrated in Fig. 1. Notably, to enhance the high-frequency components learning, we separate the image to low- and high-frequency parts, and our embedding network is trained with only the frequency components, which has been verified to be effective in real-world SR problem [16]. Here we use Haar wavelet transform to extract the informative high-frequency components. The four sub-bands decomposed by Haar wavelet transform, noted as LL, LH, HL, and HH. Then we stack the three high-frequency-related components (LH, HL, and HH) as the input and feed them to discriminator network E_D. Instead of training with conventional real or fake binary classification network, we adopt a contrastive discriminator loss proposed in [61] to train our E_D. It is a one-against-a-batch classification in the softmax cross-entropy formulation, and thus the loss function of the E_D can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{CLD} = &\mathbb{E}_{I^{HR}}\left[\log \frac{e^{E_D(\mathcal{H}_w(I^{HR}))}}{e^{E_D(I^{HR})} + \sum_{I^{LR}} e^{E_D(\mathcal{H}_w(I^{SR}))}}\right] \\ &+ \mathbb{E}_{I^{LR}}\left[\log \frac{e^{-E_D(\mathcal{H}_w(I^{SR}))}}{e^{-E_D(I^{SR})} + \sum_{I^{HR}} e^{-E_D(\mathcal{H}_w(I^{HR}))}}\right],\end{aligned} \tag{7}$$

where $I^{SR} = F_{SR}(I^{LR})$ represents super-resolved image, and $\mathcal{H}_w$ is the operator extracting LH, HL and HH sub-bands and concatenating them.

Here, we employ a GAN-like training strategy to learn a discriminator, *i.e.*, the task-friendly embedding network. It is a common way to train the embedding network with

any network structure and this is general for many low-level downstream tasks (*e.g.*, image denoising, compression artifacts removal and so on).

3.4. Contrastive Loss

As described in Sec. 3.2 and Sec. 3.3, we employ a task-friendly embedding network to obtain a frequency-based latent space, and we also generate sharpness and blurry images as our multiple positive and negative samples. To fully utilize these samples, we conduct our contrastive loss with multi-intermediate features from embedding network E_D and propose our contrastive loss based on the Eq. (3). For the target super-resolved image I_i^{SR}, its generated positive and negative sets are noted as $\mathcal{P}_i$ and $\mathcal{N}_i$ respectively. The feature representations for the super-resolved image, positive sample and negative sample are noted as f, p and n, respectively. We use superscript l is the layer index in E_D. Our contrastive loss for the i-th sample on the l-th layer is defined as follows:

$$\mathcal{L}_{i,l} = \frac{1}{K_P}\sum_{j=1}^{K_P} -\log \frac{\exp(s(f_i^l, p_j^l)/\tau)}{\exp(s(f_i^l, p_j^l)/\tau) + \sum_{k=1}^{K_N}\exp(s(f_i^l, n_k^l)/\tau)}, \tag{8}$$

where s is the similarity function. Let's note the shape of feature map as $C \times H \times W$. We adopt the mean value of pixel-wise cosine similarity as the similarity between feature maps. s is defined as follows:

$$s(f^x, f^y) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\frac{f_{hw}^x f_{hw}^y}{\|f_{hw}^x\|\|f_{hw}^y\|}. \tag{9}$$

Then, our total contrastive loss $\mathcal{L}_{CL}$ is as follows:

$$\mathcal{L}_{CL} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}\mathcal{L}_{i,l}, \tag{10}$$

where N is the number of the training images and L is the number of feature layers we used. We take the first 4 intermediate layers to calculate our contrastive loss $\mathcal{L}_{CL}$ and the L is 4 as default.

3.5. Training and Implementation Details

We train our SR network F_{SR} using L_1 loss and the proposed contrastive loss in Eq. (10). The total loss function $\mathcal{L}_{SR}$ is defined as follows:

$$\mathcal{L}_{SR} = \mathcal{L}_1 + \alpha\mathcal{L}_{CL}, \tag{11}$$

where α is a scaling parameter and we take $\alpha = 1$ as default.

In this paper, we propose a practical contrastive learning framework for SISR as illustrated in Fig. 1. The discriminator and the SR network are trained alternately. When training the discriminator, we fix the SR network, whose parameters are shared from the previous step. When training the SR network, the embedding network is frozen and its parameters are the same as the discriminator of the previous step. We generate positive and negative samples to make SR network learn sharper results. From the training process of our framework, we can learn that our proposed sample generation strategy and feature embedding network is general and can be seamlessly integrated with many other low-level image restoration networks, whose purpose is to recover a high-quality and visually pleasant result.

For training, we crop patches of size 48×48 from LR image with the corresponding HR patches. We augment the training data with random horizontal flips and 90 rotations. Our model is trained by ADAM optimizer [31] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We train for 200,000 iterations (200 epochs) and the batch size we used is 16.

4. Experiments

4.1. Experiment Setup

Datasets and Metrics. Following [11,37,67,68], we use DIV2K dataset [54] which contains 800 images for training and 100 images for evaluation. Datasets for testing include Set5 [1], Set14 [62], B100 [40], Manga109 [41], and Urban100 [25] with the up-scaling factor: $\times 4$. For comparison, we measure PSNR and SSIM [58] on the Y channel of transformed YCbCr space.

Comparison Methods. Our proposed contrastive learning framework for SISR task is generic and it can be applied to any existing method. We compare our method with the state-of-the-art (SoTA) methods. To evaluate the proposed method, we re-train EDSR[1] [37], RCAN[2] [67], and HAN[3] [42] with our contrastive learning approach. In addition, we also add some representative methods as the comparison methods: SRCNN [12], FSRCNN [13], VDSR [30], LapSRN [32], MemNet [51], SRMDNF [64], D-DBPN [21], RDN [68], SRFBN [35], and SAN [11]. We use Pytorch [45] to implement our proposed approach and for fair comparison, we re-train [37,42,67] with their official implementations.

4.2. Main Results

In Tab. 2, we tabulate the quantitative results of different methods. The EDSR-S has 16 residual blocks and 64 channels, while EDSR-L is the large version and has 32 residual blocks and 256 channels. Compared to existing methods, all of our re-trained models surpass original results on all the benchmark datasets. Specifically, when we compare the reconstruction results on the Manga109 dataset, re-trained RCAN by our proposed contrastive learning framework advances **0.21 dB** in terms of PSNR than the original RCAN model. Notably, our re-trained RCAN can achieve comparable performance against HAN, which is the SoTA method.

[1] https://github.com/sanghyun-son/EDSR-PyTorch
[2] https://github.com/yulunzhang/RCAN
[3] https://github.com/wwlCape/HAN

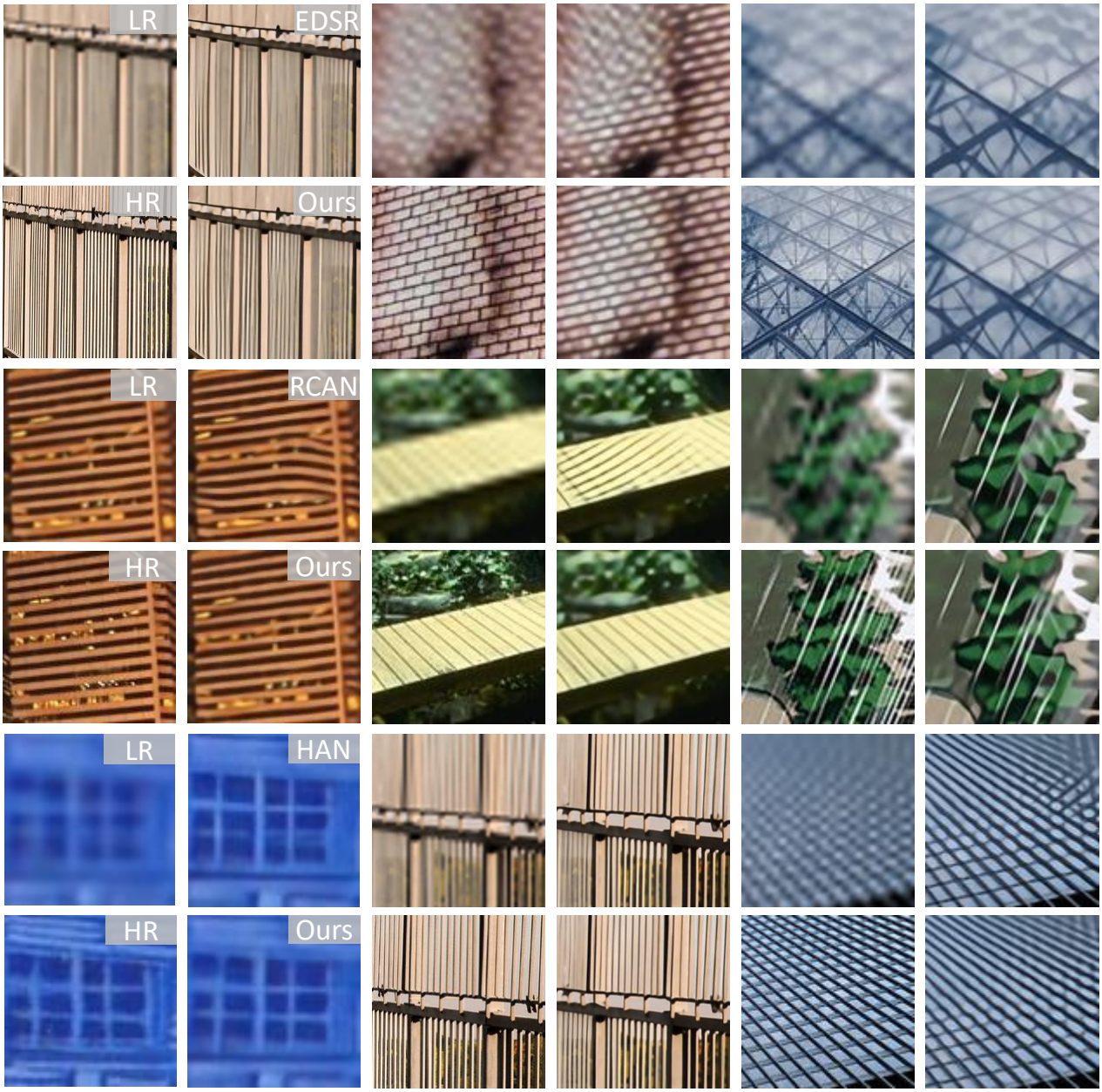

Figure 2. Visual comparison between the baselines and our improved methods: the first two rows: EDSR vs. Ours; the second two rows: RCAN vs. Ours; and the last two rows: HAN vs. Ours.

To verify the effectiveness of our method, here we only show the visual comparisons between the baseline method and its improved one. As shown in Fig. 2, by introducing the proposed contrastive constraints, the super-resolved results of our models contain clearer and sharper textures. Particularly, when the original methods cannot obtain reasonable edges, the re-trained counterparts can well recover clear and accurate textures. We also report the quantitative comparisons on benchmark datasets in terms of NIQE and LPIPS in Tab. 3. Our methods surpass the original methods in most cases. This is beyond our expectations because a lot of previous work could not well balance between distortion and perception [2].

4.3. Ablation Study

We conduct ablation studies with different training configurations to show the effectiveness of our proposed PCL-SISR. We calculate the metrics in terms of PSNR and LPIPS [65] on the Set5 dataset to compare the quantitative and qualitative performances as reported in Sec. 4.3. Config.

Table 2. Quantitative comparison with SoTA methods on benchmark datasets. The improvements of our method are in green.

Method	Set5		Set14		B100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRCNN	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221	27.58	0.8555
FSRCNN	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280	27.90	0.8610
VDSR	31.35	0.8830	28.02	0.7680	27.29	0.0726	25.18	0.7540	28.83	0.8870
LapSRN	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
SRMDNF	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
D-DBPN	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
RDN	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
SRFBN	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
SAN	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
EDSR-S	32.09	0.8938	28.58	0.7813	27.57	0.7357	26.04	0.7849	30.35	0.9067
+PCL	32.17	0.8948	28.61	0.7825	27.58	0.7365	26.07	0.7863	30.45	0.9078
Gains	0.08	0.0010	0.03	0.0012	0.01	0.0008	0.03	0.0014	0.10	0.0011
EDSR-L	32.48	0.8988	28.81	0.7879	27.73	0.7422	26.65	0.8037	31.04	0.9158
+PCL	32.52	0.8993	28.84	0.7881	27.74	0.7428	26.71	0.8056	31.20	0.9171
Gains	0.06	0.0005	0.03	0.0002	0.01	0.0006	0.06	0.0019	0.16	0.0013
RCAN	32.64	0.9002	28.85	0.7885	27.75	0.7432	26.75	0.8066	31.20	0.9170
+PCL	32.70	0.9005	28.89	0.7889	27.78	0.7437	26.84	0.8083	31.41	0.9185
Gains	0.06	0.0003	0.04	0.0004	0.03	0.0005	0.09	0.0017	0.21	0.0005
HAN	32.60	0.8997	28.88	0.7887	27.78	0.7437	26.78	0.8073	31.42	0.9174
+PCL	32.68	0.9006	28.91	0.7894	27.79	0.7441	26.88	0.8095	31.43	0.9188
Gains	0.08	0.0009	0.03	0.0007	0.01	0.0004	0.10	0.0022	0.01	0.0014

Table 3. Qaulitative comparison on benchmark datasets.

Method	Set14		Set5		B100		Urban100		Manga109	
	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS
EDSR-S	**6.166**	0.3058	**6.917**	0.2153	**6.349**	0.3419	**5.633**	0.2837	**5.018**	0.1698
+PCL	6.193	**0.3039**	7.043	**0.2139**	6.392	**0.3411**	5.658	**0.2825**	5.095	**0.1676**
EDSR-L	6.308	0.2957	7.155	0.2085	6.320	0.3343	5.461	0.2630	5.138	0.1616
+PCL	**6.166**	**0.2944**	**7.048**	**0.2080**	**6.251**	**0.3335**	**5.443**	**0.2618**	**4.998**	**0.1590**
RCAN	6.215	0.2958	7.275	0.2102	**6.314**	0.3351	5.521	0.2604	5.225	0.1608
+PCL	**6.214**	**0.2938**	**7.008**	**0.2074**	6.372	**0.3331**	**5.448**	**0.2577**	**5.004**	**0.1555**
HAN	**6.060**	**0.2912**	**6.796**	**0.2061**	6.406	**0.3302**	**5.408**	**0.2537**	**4.879**	**0.1501**
+PCL	6.211	0.2931	7.017	0.2079	**6.374**	0.3324	5.515	0.2571	5.068	0.1556

Table 4. Ablation study results. $\mathcal{L}_{CL}$ and $\mathcal{L}_{VCL}$ denote our contrastive loss and that utilizing pre-trained VGG model, respectively. $\mathcal{H}_w$ is wavelet transformation. $\mathcal{N}$ represents the negative set. LR, Rand and Gen represent utilizing only LR input, random selected other instances in mini-batch and our generated hard negative samples respectively. We calculate PSNR and LPIPS on Set5 dataset and the best results are **highlighted**.

Config	$\mathcal{L}_1$	$\mathcal{L}_{PCL}$	$\mathcal{L}_{VCL}$	$\mathcal{H}_w$	$\mathcal{N}$	↑PSNR	↓LPIPS
1	✓					32.639	0.2102
2	✓		✓		Gen	32.672	0.2098
3	✓	✓		✓	LR	32.692	0.2086
4	✓	✓			Gen	32.652	0.2083
5	✓	✓		✓	Rand	32.666	0.2086
6	✓	✓		✓	Gen	**32.697**	**0.2074**

6 is our proposed approach and it achieves the best results on both PSNR and LPIPS. Comparisons among Configs. 3, 5, and 6 clearly show the effectiveness of the generated negative samples. In addition, we can find that results of using LR input are better than randomly selected samples in mini-batch. This is reasonable because the LR input is 'harder' than selected other samples. Comparison between Configs. 2 and 4 demonstrates that our proposed $\mathcal{L}_{PCL}$ can well work on RGB space and it remains comparable to the VGG pre-trained model. Through our method obtaining lower PSNR (-0.02 dB), it achieves better LPIPS (-0.0015). To enhance the high-frequency components learning, we introduce wavelet transformation $\mathcal{H}_w$ and the embedding network is trained with only the frequency components. By comparing Configs. 4 and 6, we learn that this strategy is effective.

5. Conclusion and Discussion

In this paper, we propose a practical contrastive learning framework to further improve the performance of existing SISR approaches. We investigate contrastive learning approaches from two perspectives: sample selection and feature embedding. To make SR network better reconstruct losing high-frequency information, valid data augmentations are designed to generate our informative positive samples and hard negative samples in frequency space. In addition, we introduce a simple but efficient way to obtain a task-friendly embedding network inspired by adversarial learning. Then we propose a general framework to apply contrastive learning to SISR which can work with many other low-level tasks as well. At last, extensive experiments show that our proposed approach can achieve superior performance based on existing benchmark methods.

Limitations. The negative and positive samples are generated by simply blurring and sharpening operators. It may be more effective to learn the negative and position sample generation. In experiments, we take simple hyper-parameters as default. The impact of more hyper-parameters (*e.g.*, temperature factor τ, value range for randomly Sharpen and Blur, and numbers of intermediate layers L) can be studied to help us better understand the effect of different components and achieve even better results. Since we adopt a GAN-like framework, more training iterations are needed. A feasible solution is to utilize a pre-trained discriminator as the embedding network then the SR network can be trained with our contrastive loss directly.

Board Impact. In the future, we will explore contrastive learning approaches for more low-level tasks. Different from the traditional GAN-based perceptual methods, which only focus on whether the reconstructed image is real or not, we leverage the contrastive loss (push the result away from the smooth solution, and at the same time induce it to predict more details) to improve the perceptual quality. Therefore, it can not only constrain the solution space but also enhance the perceptual quality of the solution. This may be a direction for future SISR and other low-level tasks.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012. 6

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*, 2018. 7

[3] Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogério Feris, Raja Giryes, and Leonid Karlinsky. Fine-grained angular contrastive learning with coarse labels. In *CVPR*, 2021. 1

[4] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. In *NeurIPS*, 2020. 1, 2

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119, 2020. 1, 2, 3

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 1

[9] Xiang Chen, Jinshan Pan, Kui Jiang, Yufeng Huang, Caihua Kong, Longgang Dai, and Yufeng Li. Unpaired adversarial learning for single image deraining with rain-space contrastive constraints. *arXiv preprint arXiv:2109.02973*, 2021. 1, 2, 3

[10] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, 2020. 2, 5

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 3, 6

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2), 2016. 6

[13] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, 2016. 6

[14] Nanqing Dong, Matteo Maggioni, Yongxin Yang, Eduardo Pérez-Pellitero, Ales Leonardis, and Steven McDonagh. Residual contrastive learning for joint demosaicking and denoising. *arXiv preprint arXiv:2106.10070*, 2021. 1, 2, 3

[15] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. SEED: self-supervised distillation for visual representation. In *ICLR*, 2021. 1

[16] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, 2019. 5

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 5

[18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR (2)*, 2006. 2

[19] Junlin Han, Mehrdad Shoeiby, Tim Malthus, Elizabeth Botha, Janet Anstee, Saeed Anwar, Ran Wei, Lars Petersson, and Mohammad Ali Armin. Single underwater image restoration by contrastive learning. In *IGARSS*, 2021. 1, 2, 3, 5

[20] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *CVPRW*, 2021. 3

[21] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 3, 6

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 3

[23] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 2, 3

[24] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *CVPR*, 2021. 5

[25] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 6

[26] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM Multimedia*, 2019. 3

[27] Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Tackling the ill-posedness of super-resolution through adaptive target generation. In *CVPR*, 2021. 5

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, volume 9906, 2016. 3

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 1, 4

[30] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 6

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 6

[32] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 6

[33] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 3, 5

[34] Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, 2021. 1

[35] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, 2019. 6

[36] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCVW*, 2021. 3

[37] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CPVR*, 2017. 3, 6

[38] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCVW*, 2020. 3

[39] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 3

[40] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 6

[41] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multim. Tools Appl.*, 76(20), 2017. 6

[42] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 3, 6

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3

[44] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, volume 12354, 2020. 3

[45] Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[46] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021. 2, 5

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1, 2

[49] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *ICLR*, 2021. 2

[50] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 2

[51] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, 2017. 6

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 1

[53] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 1, 2

[54] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 6

[55] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 2021. 2, 3

[56] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 3, 5

[57] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. In *IJCAI*, 2021. 1, 2, 3, 5

[58] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13, 2004. 6

[59] 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 *CVPR*, 2021. 1, 2, 3, 5

[60] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1, 2, 3

[61] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *CVPR*, 2021. 5

[62] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, 2010. 6

[63] Jiahui Zhang, Shijian Lu, Fangneng Zhan, and Yingchen Yu. Blind image super-resolution via contrastive representation learning. *arXiv preprint arXiv:2107.00708*, 2021. 2, 3, 5

[64] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 6

[65] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[66] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, 2019. 3

[67] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, volume 11211, 2018. 3, 6

[68] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 3, 6

[69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3