Criteria Comparative Learning for Real-Scene Image Super-Resolution

Yukai Shi, Hao Li, Sen Zhang, Graduate Student Member, IEEE, Zhijing Yang, and Xiao Wang, Member, IEEE

Abstract—Real-scene image super-resolution aims to restore real-world low-resolution images into their high-quality versions. A typical RealSR framework usually includes the optimization of multiple criteria which are designed for different image properties, by making the implicit assumption that the ground-truth images can provide a good trade-off between different criteria. However, this assumption could be easily violated in practice due to the inherent contrastive relationship between different image properties. Contrastive learning (CL) provides a promising recipe to relieve this problem by learning discriminative features using the triplet contrastive losses. Though CL has achieved significant success in many computer vision tasks, it is non-trivial to introduce CL to RealSR due to the difficulty in defining valid positive image pairs in this case. Inspired by the observation that the contrastive relationship could also exist between the criteria, in this work, we propose a novel training paradigm for RealSR, named Criteria Comparative Learning (Cria-CL), by developing contrastive losses defined on criteria instead of image patches. In addition, a spatial projector is proposed to obtain a good view for Cria-CL in RealSR. Our experiments demonstrate that compared with the typical weighted regression strategy, our method achieves a significant improvement under similar parameter settings.

Index Terms—Comparative Learning, criteria, real-world scene, image super-resolution.

I. INTRODUCTION

REAL-WORLD image super-resolution [1]–[5] mainly refers to the restoration of real-scene low-resolution images into high-quality ones, which can be obtained by learning a projection function $F(\cdot)$:

$$I_o = F(I_{lf} | \theta),$$

where $I_o$ and $I_{lf}$ are the output high-resolution and the low-resolution images respectively, $\theta$ is the parameters of $F(\cdot)$. To obtain different resolution images in real-scene configuration, $I_{hf}$ and $I_{lf}$ are collected by different optical sensors [6]–[10] with various resolution settings, which is different from the traditional image super-resolution paradigm [11]–[16] that generates $I_{lf}$ using downsampling techniques. Therefore, compared with the traditional image super-resolution task, RealSR suffers a severer pixel displacement due to the difference between the camera settings to obtain $I_{hf}$ and $I_{lf}$. Although alignment-based methods have been developed to alleviate this problem [1], current RealSR datasets [4], [5] still fail to guarantee absolute alignment on pixel-level.

In the mainstream RealSR approaches [17]–[19], diverse losses or criteria have been integrated by using their weighted sum to achieve a trade-off between the perceptual- and pixel-similarities:

$$L_{asum} = \alpha C_{adv}(I_o, I_{hf}) + \beta C_{per}(I_o, I_{hf}) + \gamma C_{pix}(I_o, I_{hf}).$$

where $I_o$ is the output of projection function $F$, $[C_{adv}, C_{per}, C_{pix}]$ are the adversarial-, perceptual- and Euclidean- criteria, which focus on restoring different aspects of the images. And $[\alpha, \beta, \gamma]$ are the weights for each loss function, respectively. ESRGAN [20] uses $L_{asum}$ to pursue the trade-off between multiple criteria. SR-VAE [21] employs the KL loss to measure the divergence between latent vectors and standard Gaussian distribution. Similarly, DASR [22] employs generative adversarial nets [23] to learn the domain distance between the synthetic image and the real-scene image. Then, an Euclidean criterion is used to constrain the bottom image features. These methods implicitly make a strong assumption that the sole ground-truth images can provide a good trade-off between multiple criteria. However, is that a universal solution?

To answer this question, we re-examine the learning paradigm of typical RealSR pipelines. Accordingly, we found that the ground-truth images are beyond a trade-off between different image properties. For example, suppose we want to generate a realistic and rich texture, the Euclidean criterion plays a positive constraint in the adversarial learning paradigm by regularizing the generative model to preserve a stable and consistent structure. Nevertheless, when it comes to restoring a clear and sharp edge, this generative effect from the adversarial criterion for rich texture plays a negative role against obtaining a sharp edge. In previous works [20], [24], $L_{asum}$ is adopted by assuming all criteria are positively contributed...
to image enhancement. As illustrated in our visual analysis in Fig. 1 and Fig. 2, the usage of $L_{\text{ssim}}$ tends to achieve a trade-off between the perceptual- and pixel-similarities. Suppose a local region inherently has sharp edge, due to the adversarial criterion takes a considerable proportion, a weighted sum of perceptual- and pixel-criterion often restore a relatively blurry result. This bottleneck motivates us to investigate the contrastive effects among the criteria adaptively.

The contrastive learning (CL) paradigm [25], [26] provides a promising framework to account for the contrastive relationships, which focus on learning a good feature representation by constructing positive- and negative-training pairs. Specifically, CL attempts to make positive instances close to each other in the high-dimension feature space, while repelling away the negative instances. A basic CL contrastive loss function reads:

$$
\mathcal{L}_{\text{CL}} = \log \frac{\sum_{i=1}^{N} e^{\langle z_i, z_i^+ \rangle/\tau}}{\sum_{k=1}^{K} e^{\langle z_i, z_k^- \rangle/\tau}},
$$

where $z_i$, and $\{z_i^+, z_k^-\}$ are the hypersphere features of the input anchor sample, and its corresponding positive and negative samples, respectively. $\tau$ is a temperature parameter. Generally, the hypersphere projection of samples is implemented by a convolutional network [25]. In the ImageNet challenge [27], SimCLR [28] obtain the $z_i^+$ with data augmentation such as rotation, crop, cutout and resize. And $z_i^-$ is an arbitrary sample within the training mini-batch. In image processing tasks like de-raining, SID [29] captures the $z_i^+$ by searching the clean patch, and the $z_i^-$ is a patch which is full of raindrop.

Although CL has proven successful in many computer vision tasks, however, it remains non-trivial to introduce CL to RealSR, due to the difficulty in defining valid positive samples under the RealSR setting. Specifically, CL methods usually define the positive and negative relationships upon image patches, while in RealSR there are no trivial pixel-level positive samples other than the ground-truth images. Although a ground-truth image can be regarded as perfect positive samples, invalid gradients could occur during optimization when taking the derivative of the attached pixel loss: $\|I_{sc} - F(I_c)\|_2^2$. Moreover, since the ground-truth images have already been used as the labels in Eqn. (2), the repeated use of the ground-truth samples as the input when constructing the contrastive loss could make the network fail to learn the desired discriminative features. Therefore, the positive patches for RealSR are hard to be well defined.

In this work, we tackle this problem by proposing a novel CL training paradigm for RealSR, named Criteria Comparative Learning (Cria-CL). Inspired by the observation that the inherent contrastive relationship in RealSR also exists between the criteria, e.g., the contrastive effect between the Euclidean- and the adversarial-criterion working on preserving the clear structure and smooth texture simultaneously, Cria-CL attempts to explore such contrastive relationship between criteria by defining the contrastive loss directly on criteria instead of image patches. In addition, in contrast to simply repelling the negative criteria pairs, we formulate the negative contrastive loss using Gaussian potential kernel to introduce uniformity into Cria-CL and provide symmetric context [30], [31]. Furthermore, a spatial projector is developed to obtain a good view for multi-criteria learning to facilitate the training process and enhance the restoration performance.

The contributions are summarized as:

1. To explore a new training paradigm for RealSR with appropriate contrastive learning, we build our comparative learning framework upon image restoration criteria (e.g., Euclidean-, perceptual- and adversarial-criterion).

2. In contrast to repelling negative data pair simply, in this paper, we extend the uniformity assumption [30], [31] into criteria to provide fresh symmetric contexts for the multi-task paradigm.

3. To verify the generalization on out-of-distribution (OOD) data, we built a new RealSR-Zero dataset for evaluation, in which the poor-quality photos are shot by a iPhone4 device and only test image are provided.

4. Extensive experiments are conducted to verify that each proposed component is solid while the unified framework shows a clear improvement toward state-of-the-art methods.

II. RELATED WORK

A. Real-Scene Image Super-Resolution

Different from the traditional image SR that generally focuses on simple synthesized degradation [32]–[35], RealSR [36] needs to handle complicated and rough degradation in real-world scenes [6], [37]. The first attempt is initially to estimate the degradation process for given LR images, and then apply a paired data-based model for super-resolution. KernelGAN [17] proposed to generate blur kernel from label images via a kernel estimation GAN before applying the ZSSR [38] method. SSPG [39] apply k-nearest
neighbors (KNN) matching into neural architecture design. Then, a sample-discriminating learning mechanism based on the statistical descriptions of training samples is used by SSPG to enforce the generative model focus on creating realistic pictures. CDC [36] employs a modularized CNN to enhance different cases. SwinIR [40] investigates a transformer, which gives attractive performance on various image processing tasks. EMASRN [41] facilitates performance with limited parameter number by using an expectation-maximization attention mechanism. TSAN [42] also addresses the attention mechanism in image super-resolution by realizing a coarse-to-fine restoration framework. Wan et al. [43] applies real-world image restoration model into old photos to build up a practical enhancement framework. Impressionism [44], the winner of NTIRE 2020 challenge [4], proposed to estimate blur kernels and extract noise maps from source images and then apply the traditional degradation model to synthesize LR images. Real-ESRGAN [45] introduced a complicated degradation modeling process to better simulate real-world degradation and employ a U-Net discriminator to stabilize the training dynamics. Yet, these methods cannot give out satisfactory results for images with degradations not covered in their model. To remedy this, several methods try to implicitly grasp the underlying degradation model through learning with the external dataset. DASR [22] proposed a domain-gap aware training strategy to calculate the domain distance between generated LR images and real images that both are used to train the SR model. USR-DA [2] proposed an unpaired SR training framework based on feature distribution alignment and introduced several losses to force the aligned feature to locate around the target domain.

B. Contrastive Learning

Unsupervised visual representation learning recently achieves attractive success in natural language processing and high-level computer vision tasks [25], [46], [47]. Bert [46] uses masked-LM and next sentence prediction to implement the pre-trained model on a large-scale text dataset. This training strategy contributes to learning general knowledge representations and facilitates reasoning ability in downstream tasks. MoCo [25] revitalizes the self-supervised training for high-level computer vision tasks by proposing momentum contrast learning. Specifically, MoCo builds up positive/negative data queues for contrastive learning, and fills the gap between unsupervised and supervised representation learning.

C. Contrastive Learning for Image Processing

Many efforts are devoted to contrastive-based image processing tasks. Recently, [48] address the mutual information for various local samples with contrastive learning. Reference [49] proposes a novel contrastive feature loss by non-locally patch searching. Reference [29] further explore contrastive feature learning with border tasks by incorporating individual domain knowledge. However, the aforementioned methods still suffer inflexibility of the fixed sample size and a trade-off between different criteria. In this paper, we mainly investigate the feature contrastive learning under multi-task configuration.

III. CRITERIA COMPARATIVE LEARNING

As the keypoint of this paper is to address the contrast effect between criteria under real-world image setting, we use a simplified RRDB backbone [20] for feature extraction and multiple criteria are constructed to emulate a general RealSR framework. As shown in Fig. 3, given an input image $I_{lf}$, we apply a feature extractor $F_{\alpha}(\cdot)$ to produce two intermediate results $[I_{align}, I_{anif}]$ as:

$$[I_{align}, I_{anif}] = F_{\alpha}(I_{lf}|\theta_{\alpha}).$$

Typically, in general real-world scene frameworks for image restoration, multiple criteria are adopted as object function for optimization. To realize a criteria comparative learning, we first calculate losses according to each criterion:

$$L_{SR} = \begin{cases} C_{a} = L_{d}(I_{align}, I_{hf}) \\ C_{i}^{+} = L_{i}^{+}(I_{align}, I_{hf}) \\ \ldots \ldots \\ C_{k}^{-} = L_{k}^{-}(I_{anif}, I_{hf}) \end{cases}$$

where $C_{a}$ is the anchor criterion. $C_{i}^{+}$ and $C_{k}^{-}$ are positive and negative criteria toward $C_{a}$. Note that the data type
of calculated results $[C_a, C^+_a, C^-_a]$ in Equ. 5 is tensor, among which we can apply a secondary loss computation and backpropagation. Thus, we can utilize these tensors to realize a criteria comparative learning in RealSR to achieve feature disentanglement. First, we apply Equ. 3 into multitask configuration by replacing the positive/negative patches with criteria:

$$\tilde{L}_{Cria} = -\log \frac{\sum_{i=1}^{N} e^{\psi^+(C_a, C^+_a)}}{\sum_{k=1}^{K} e^{\psi^-(C_a, C^-_a)}}. \tag{6}$$

In addition, $\psi^+$ and $\psi^-$ are similarity measurement function for positive/negative criteria.

To enhance interpretability, we further factorize $\tilde{L}_{Cria}$ into positive and negative items:

$$\tilde{L}_{Cria} = -\log \left( \sum_{i=1}^{N} e^{\psi^+(C_a, C^+_a)} \right) + \log \left( \sum_{k=1}^{K} e^{\psi^-(C_a, C^-_a)} \right).$$

As the $C$ indicates a computation result of tensors (e.g., $C_{pix} = ||I_{align} - I_{hfs}||_2$), we directly minimizes the loss for positive pair as:

$$L_{align}^{Cria}(C_a, C^+_a; \eta) = \sum_{i=1}^{N} (C_a(I_{align}, I_{hfs}) - C^+_a(I_{align}, I_{hfs}))^\eta. \tag{8}$$

As shown in Fig. 3, we attempt to realize a uniform distribution on hyperspherical space to provide symmetric context. Instead of repelling negative criteria irreguarly, the criteria are enforced to reach a uniform distribution on the hypersphere. Different from the uniformity assumption in [31], we realize $\psi^-$ by proposing the following uniformity loss for negative criteria to provide symmetric context:

$$L_{unif}^{Cria}(C_a, C^-_a; \tau) = -\log \left( \sum_{k=1}^{K} e^{-\tau(C_a(I_{unif}, I_{hfs}) - C^-_a(I_{unif}, I_{hfs}))^2} \right). \tag{9}$$

A. Spatial-View Projection

A good viewpoint for criterion disentanglement is non-trivial for contrastive learning. In RealSR, an image contains rich texture, which often lead to $[C_a, C^-_a]$ not in the same distribution space diametrically. Hence, its unreasonable to take contrastive multi-task learning for the whole image without looking into special cases. Searching a local patch with a fixed size for spatial view has inherent inflexibility. Thus, we apply a non-local spatial-view search and projection:

$$I_{SR} = I_{align} * \tilde{S} + I_{unif} * \tilde{\tilde{S}}, \tag{10}$$

where $[S, \tilde{S}]$ are spatial masks for $I_{unif}$ and $I_{align}$, we obtain them by extracting multi-task oriented feature representations and the original image with $F_{\beta}$. Then, we apply $[S, \tilde{S}]$ by the class activation map (CAM) [50] fashion to realize spatial projection for positive and negative criterion, and present final output $I_{SR}$. As illustrated in Fig. 3, the pairwise criteria are jointly optimized with $[F_a, F_{\beta}]$ for a comparative learning.

B. Model Details

We apply RRDBs as the backbone [20] in feature extractor $F_a$. Specifically, we added two sub-branches at the end of RRDBs, each sub-branch consists of three Residual blocks [51]. Then, we send the intermediate output of RRDBs into two sub-branches, each sub-branch uses different loss (e.g. L1 loss and adversarial loss) for optimization, and produces $I_{align}$ and $I_{unif}$ respectively. In addition, we send the original image $I_{LR}$ into another feature extractor $F_{\beta}$, which consists of three Residual blocks, an upsampling operation and a softmax operation to obtain the spatial masks $[S, \tilde{S}]$. The upsampling operation is to keep the spatial masks $[S, \tilde{S}]$ with the same spatial sizes $(H \times W)$ as $[I_{align}, I_{unif}]$.

C. Anchor Selection

How to choose a fixed anchor criterion and corresponding negative/positive counterpart is a critical issue in our algorithm. With a limited criterion number, we successively pick up pixel-, adversarial- and perceptual- criterion as anchor to observe the experimental result. As depicted in Tab. VI, by adopting $C_{adv}$ as anchor criterion, our model shows a poor results. Since a pure adversarial loss often performs unsteadily during training, which causes all criteria to become positive.
counterparts. As shown in Tab.I, once we set any negative criterion for $C_{adv}$ as the negative item, the performance becomes poor. Literally, Euclidean criterion can find distinct positive/negative examples and presents a solid performance. We therefore use Euclidean criterion as positive/negative examples and presents a solid performance.

We therefore use Euclidean criterion as positive/negative examples and presents a solid performance. Its note that we have employed a spatial-view projection, thus only regional pixels rather than full image will be handle with this operation as the source domain images. We adopt the ×4 scale to evaluate our model.

Our experiments are implemented by Pytorch 1.4.0 with 4 NVIDIA Tesla V100 GPUs. We use Adam [55] as optimizer, where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batchsize and total iterations are set to 16 and $10^6$, respectively. The initial learning rate is $1 \times 10^{-4}$ and decayed by $2^x$ at every $2 \times 10^5$ iterations. We use flip and random rotation with angles of $90^\circ$, $180^\circ$ and $270^\circ$ for data augmentation. In evaluation protocols, we adopted PSNR, SSIM, LPIPS [56], and NIQE [54] to verify our model. Also, we evaluate the inference speed on an NVIDIA Tesla V100 GPU.

B. Qualitative and Quantitative Comparison

1) RealSR-Zero: To perform a comparison on RealSR-Zero, we use label-free measure index NIQE and mean opinion score (MOS) for evaluation. In Tab. I, Cria-CL outperforms Real-ESRGAN with 0.1666 over the NIQE index, which verify that our criteria comparative algorithm help to generates richer details with high-fidelity. We also conduct human perception study by recruiting 20 volunteers for subjective assesment on RealSR-Zero. More specific, 50 pairs of patch generated by each method were shown to volunteers for a side-by-side comparison. Its note that Cria-CL wins highest preference with 6.25% better MOS score than Real-ESRGAN. As shown in Fig. 6, the proposed model is able to avoid over-smooth and produce realistic texture. For instance, compared with Real-ESRGAN, our algorithm restores realistic texture on the green stone as well as maintains sharp edge on Fig. 5, which reveals that the spatial-view projection a appropriate view for feature disentanglement in criteria comparative learning.

2) RealSR: As depicted in Tab. II, we present a quantitative comparison. Compared with Real-ESRGAN, Cria-CL
Fig. 5. Super-resolution results on the RealSR-Zero dataset. We trained all methods on DF2K [52] and conduct evaluation on RealSR-Zero. Particularly, RealSR-Zero is an out-of-distribution (OOD) dataset that only includes images for testing.

Fig. 6. Super-resolution results on the RealSR-Zero dataset.

achieves a 1.38 dB gain. Our method obtains a 0.0296 LPIPS improvement over Real-ESRGAN. Compared with ADL, Cria-CL shows a 0.92 dB, which is clear improvement on RealSR task. Moreover, our algorithm restore a clear text on the second row of Fig. 7, which address that the criteria comparative algorithm learns richer feature for image restoration. Essentially, Real-ESRGAN and ADL are the newest state-of-the-art works, which are published in top-tier conferences and journals. This indicates that the effectiveness of Cria-CL and the contrastive relationship among criteria is worth to be fully addressed.

3) NTIRE2020 and CameraSR: As depicted in Tab. III, compared with USR-DA [2], Cria-CL achieves a significant improvement with 0.81 dB PSNR and 0.0323 LPIPS gain on NTIRE2020 challenge data. Compared with Real-ESRGAN, our model shows a improvement with 1.3 dB and 0.0324 LPIPS. As depicted in Tab. IV, our model outperform Real-ESRGAN with 0.963 dB and 0.002 LPIPS. As USR-DA and Real-ESRGAN are recently proposed RealSR frameworks and exhibited a high-fidelity image restoration in RealSR task. Our model still achieves a significant improvement over them, which fully address the effectiveness of the proposed criteria comparative algorithm. Apart from that, Cria-CL still achieves a good visual effect in CameraSR dataset. As show in the top row of Fig. 9, other methods restore blurry texture and edges in the building roof. By contrast, our model obtains smooth texture, clear boundary and fewer artifacts, which indeed justify the effectiveness of the criteria comparative algorithm and spatial-view projection.
C. Ablations

We conduct extensive ablations of our Cria-CL framework on NTIRE 2020 challenge data to verify the effectiveness of each component.

1) Criteria Comparative Algorithm: In Tab. V, a plain model achieves 0.3 dB gain by using $\mathcal{L}_{\text{align}}$ which verifies the effectiveness of alignment loss between positive losses. For a fair comparison, we apply $\mathcal{L}_{\text{Cria}}$ on the ‘Baseline’ model, which achieves a significant promotion with a 0.94 dB gain. We conduct this quantitative analysis on NTIRE2020 dataset, which shows that each proposed component is necessary for our model.
2) Visual Effects: We also illustrate the visual effect of each proposed component in Fig. 10. Obviously, the enhanced images by $L_{\text{Cria}}$ restore more correct details with fewer artifacts, indicating that the uniformity constraint can significantly improve the visual qualities under the multi-task paradigm.

3) Spatial-View Projection: We show the effect of spatial-view projection in Tab. V. With the spatial-view mechanism, our model obtains a 1.19 dB improvement. Without spatial-view projection, $L_{\text{Cria}}$ and $L_{\text{align}}$ exhibit limited performance improvement. This shows that in the Cria-CL framework, spatial-view projection is required for a good view of feature disentanglement among multi-criteria training conditions.

D. Efficiency

We conduct the efficiency analysis toward the state-of-the-art methods on the RealSR dataset with their official
implementation and equal hardware environment. As shown in Tab. VII, Cria-CL achieves competitive efficiency with attractive performance. Specifically, Cria-CL obtains a faster running efficiency with a 1.83 dB improvement over FSSR. Compared with ESRGAN, Cria-CL obtains promising running efficiency by reducing 10% inference cost. Compared with DASR, the proposed Cria-CL demonstrates superior efficiency and achieves significant improvements by 1.57 dB. This shows that the Cria-CL learns effective feature representations for RealSR with fewer parameters.

V. DISCUSSION

As the proposed Cria-CL shows promising results on the RealSR task, a few open problems still need to be further explored. Cria-CL sets the pixel loss as the anchor and achieves attractive performance. Nevertheless, when Cria-CL uses adversarial loss as the anchor for contrastive multi-task learning, the performance becomes worse. This suggests the positive counterpart toward the adversarial criterion required for further investigation.

Except for the RealSR task, Cria-CL has the potential to be applied to other real-world image processing tasks, such as de-razing, image enhancement, and de-hazing. We hope Cria-CL will bring diverse insight to the image processing tasks that include contrastive learning.

REFERENCES

[1] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” 2019, arXiv:1904.00523.

[2] W. Wang, H. Zhang, Z. Yuan, and C. Wang, “Unsupervised real-world super-resolution: A domain adaptation perspective,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 4318–4327.

[3] S. Son, J. Kim, W.-S. Lai, M.-H. Yang, and K. M. Lee, “Toward real-world super-resolution via adaptive downsampling models,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Aug. 24, 2021, doi: 10.1109/TPAMI.2021.3106790.

[4] A. Lugmayr et al., “NTIRE 2020 challenge on real-world image super-resolution: Methods and results,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2020, pp. 494–495.

[5] A. Lugmayr et al., “AIM 2019 challenge on real-world image super-resolution: Methods and results,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 3575–3583.

[6] X. Xiang, L. Zhu, J. Li, Y. Wang, T. Huang, and Y. Tian, “Learning super-resolution reconstruction for high temporal resolution spike stream,” IEEE Trans. Circuits Syst. Video Technol., early access, Nov. 23, 2021, doi: 10.1109/TCASVT.2021.3130147.

[7] Z. He, S. Tang, J. Yang, Y. Cao, M. Y. Yang, and Y. Cao, “Cascaded deep networks with multiple receptive fields for infrared image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2310–2322, Aug. 2018.

[8] Y. Mo, Y. Wang, C. Xiao, J. Yang, and W. An, “Dense dual-attention network for light field image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 7, pp. 4431–4443, Jul. 2022.

[9] Y. Zhou, W. Deng, T. Tong, and Q. Gao, “Guided frequency separation network for real-world super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2020, pp. 428–429.

[10] Y. Lin, S. Zhang, T. Chen, Y. Lu, G. Li, and Y. Shi, “Exploring negatives in contrastive learning for unpaired image-to-image translation,” 2022, arXiv:2204.11018.

[11] C. Xie, W. Zeng, and X. Lu, “Fast single-image super-resolution via deep network with component learning,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 12, pp. 3473–3486, Dec. 2019.

[12] Z. Zhang, J. Shao, Z. Liang, X. Liu, and H. T. Shen, “Multi-branch networks for video super-resolution with dynamic reconstruction strategy,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3954–3966, Oct. 2021.
A. Shocher, N. Cohen, and M. Irani, “Zero-shot’ super-resolution using deep internal learning,” in Proc. CVPR, 2018, pp. 3118–3126.

Y. Hu, J. Li, Y. Huang, and X. Gao, “Image super-resolution with self-similarity prior guided network and sample-discriminating learning,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 4, pp. 1966–1985, Apr. 2021.

J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2021, pp. 1833–1844.

Y. Hu, J. Li, Y. Huang, and X. Gao, “Image super-resolution with self-similarity prior guided network and sample-discriminating learning,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1273–1284, Mar. 2022.

J. Zhang et al., “A two-stage attentive network for single image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1020–1033, Mar. 2022.

Z. Wan et al., “Bringing old photos back to life,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 2747–2757.

X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, “Real-world super-resolution via kernel estimation and noise injection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2020, pp. 466–467.

X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 1905–1914.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

T. Chen, L. Lin, R. Chen, X. Hui, and H. Wu, “Knowledge-guided multi-label few-shot learning for general image recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1371–1384, Mar. 2022.

T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in Proc. ECCV, 2020, pp. 319–345.

A. Andonian, T. Park, B. Russell, P. Isola, J.-Y. Zhu, and R. Zhang, “Contrastive feature loss for image prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2021, pp. 1934–1943.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2921–2929.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 136–144.

C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu, “Camera lens super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1652–1660.

A. Mittal, R. Soudagarajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209–212, Mar. 2013.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015, pp. 1–15.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 586–595.

Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, “Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 701–710.

M. Fritsche, S. Gu, and R. Timofte, “Frequency separation for real-world super-resolution,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 3599–3608.

Yukai Shi received the Ph.D. degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2019. He is currently a Lecturer with the School of Information Engineering, Guangdong University of Technology, China. His research interests include computer vision and machine learning.

Hao Li received the B.E. degree from the School of Information Technology, Guangdong University of Technology, Guangzhou, China, where he is currently pursuing the M.S. degree. His research interests include computer vision and machine learning.

Sen Zhang (Graduate Student Member, IEEE) received the B.S. and Ph.D. degrees in mathematics and computing science from Sun Yat-sen University, Guangzhou, China, in 2003 and 2008, respectively. He was a Visiting Research Scholar at the School of Computing, Informatics and Media, University of Bradford, U.K., from July 2009 to December 2009. He was a Research Fellow at the School of Engineering, University of Lincoln, U.K., from January 2011 to January 2013. He is currently a Professor with the School of Information Engineering, Guangdong University of Technology, China. He has published over 60 peer-review journals and conference papers. His research interests include time-frequency analysis, image retrieval, machine learning, and pattern recognition.

Zhijing Yang received the B.S. and Ph.D. degrees in mathematics and computing science from Sun Yat-sen University, Guangzhou, China, in 2003 and 2008, respectively. He was a Visiting Research Scholar at the School of Computing, Informatics and Media, University of Bradford, U.K., from July 2009 to December 2009. He was a Research Fellow at the School of Engineering, University of Lincoln, U.K., from January 2011 to January 2013. He is currently a Professor with the School of Information Engineering, Guangdong University of Technology, China. He has published over 60 peer-review journals and conference papers. His research interests include time-frequency analysis, image retrieval, machine learning, and pattern recognition.

Xiao Wang (Member, IEEE) received the B.S. degree from West Anhui University, Lu’an, China, in 2013, and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 2019. From 2015 and 2016, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has visited the UBTECH Sydney Artificial Intelligence Centre, Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia, in 2019. He is currently a Post-Doctoral Researcher with the Peng Cheng Laboratory, Shenzhen, China. His current research interests include computer vision, machine learning, pattern recognition, and deep learning.