# LightBSR: Towards Lightweight Blind Super-Resolution via Discriminative Implicit Degradation Representation Learning

Jiang Yuan[1,2,4]*, Ji Ma[1,2]*, Bo Wang[1]†, Guanzhou Ke[3], Weiming Hu[1]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences,
[2]North China Electric Power University, [3]Beijing Jiaotong University, [4]YunQue AGI, Hangzhou, China
{yuanj, maji}@ncepu.edu.cn, wangbo@ia.ac.cn, guanzhouk@gmail.com, wmhu@nlpr.ia.ac.cn

## Abstract

*Implicit degradation estimation-based blind super-resolution (IDE-BSR) hinges on extracting the implicit degradation representation (IDR) of the LR image and adapting it to LR image features to guide HR detail restoration. Although IDE-BSR has shown potential in dealing with noise interference and complex degradations, existing methods ignore the importance of IDR discriminability for BSR and instead over-complicate the adaptation process to improve effect, resulting in a significant increase in the model's parameters and computations. In this paper, we focus on the discriminability optimization of IDR and propose a new powerful and lightweight BSR model termed LightBSR. Specifically, we employ a knowledge distillation-based learning framework. We first introduce a well-designed degradation-prior-constrained contrastive learning technique during teacher stage to make the model more focused on distinguishing different degradation types. Then we utilize a feature alignment technique to transfer the degradation-related knowledge acquired by the teacher to the student for practical inferencing. Extensive experiments demonstrate the effectiveness of IDR discriminability-driven BSR model design. The proposed LightBSR can achieve outstanding performance with minimal complexity across a range of blind SR tasks. Our code is accessible at: https://github.com/MJ-NCEPU/LightBSR.*

## 1. Introduction

The goal of blind super-resolution (BSR) [7, 15, 16, 22, 25, 32, 33, 35–37, 39, 44] is to restore the detail-rich high-resolution (HR) image from its low-resolution (LR) counterpart with unknown and complex degradations. The key to BSR methods lies in accurately estimating the degradation information of the LR image and embedding it into the SR module effectively to guide image upsampling. As a more practical LR image restoration technique, BSR has attracted a lot of attention and developed two research directions: explicit degradation estimation-based (EDE-) BSR and implicit degradation estimation-based (IDE-) BSR.

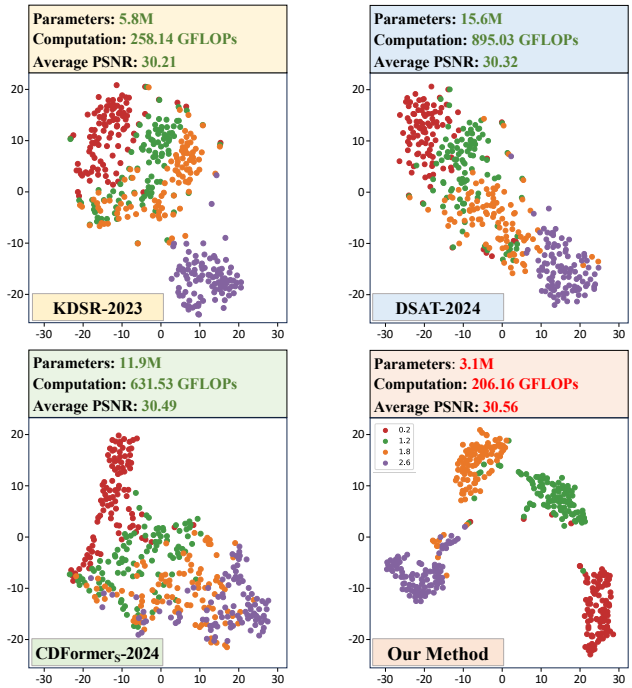



Figure 1. Comparison of different methods in terms of model complexity, SR performance and IDR distribution across four blur kernels on the DIV2K [2] val-set.

The EDE-BSR paradigm [3, 7, 13, 17, 25, 42] requires an explicit mapping between degradation parameters (e.g., blur kernel, noise level) and the estimator. Only when the degradation type of the LR image falls within the predefined degradation scope can the EDE-BSR method achieve satisfactory image restoration results. However, real-world degradation scenarios are nearly infinite, making it impractical to annotate all possible types of degradation in advance. Additionally, precise kernel estimation often necessitates more parameters and iterations [7, 13], significantly increas-

*Authors contributed equally. Work done during the internship at CASIA.
†Corresponding author.

ing the model and time complexity of such methods.

The IDE-BSR paradigm [21, 22, 32, 34, 36, 37, 43] transforms the problem of estimating degradation information into a degradation-related representation learning problem. The core objective is to construct a latent representation space that can generalize to various degradation domains, including unseen types during training, and effectively adapt the extracted implicit degradation representation (IDR) to LR image features to guide HR image restoration. Benefiting from its more comprehensive understanding of the intrinsic characteristics of intricate degradation patterns and simpler training process, IDE-BSR paradigm has shown significant potential in real-world scenarios with noise interference and complex degradation combinations.

The previous works have explored the application of various representation learning techniques in IDE-BSR task, such as contrastive learning [8, 21, 32], meta learning [6, 36], distillation learning [10, 37], and diffusion model [11, 22]. Despite some progress, these methods did not particularly emphasize the discriminative power of the learned IDR space. As shown in the t-SNE plots of Fig.1, even the latest methods [21, 22] struggle to distinguish different degradation types effectively. This means that these models do not accurately understand different degradation patterns, and the extracted IDR contains degradation-independent noise. As a result, they have to increase the complexity of the adaptation process to solve the issue of noise interference for improving SR effect. For example, although CDFormer$_S$ [22] and DSAT [21] achieved high average PSNR values on the DIV2K val-set [2], the cost was an increase of more than twice the number of parameters and computational costs compared to KDSR [37]. However, the increasing complexity of models undoubtedly hinders the practicability of BSR technology. This phenomenon raises an important question: *is it possible to enhance SR effect by strengthening IDR discriminability, thereby circumventing complex adaptation processes and achieving a lightweight BSR model?*

Based on the above observations and reflections, we attempt to focus on enhancing the discriminability of IDRs rather than stacking adaptation networks in this paper, aiming to form a new powerful and parameter-efficient BSR model design. The proposed method is termed **LightBSR**. The overall architecture is shown in Fig.2, where the core modules are an IDR estimation module (IDR-EM) and an IDR adaptation module (IDR-AM). The former consists of an IDR-estimator and an IDR-converter, used to extract channel- and spatial-wise IDRs from the input LR image. The latter consists of a series of hierarchical adaptation components, used to effectively adapt IDRs to LR image features. Regarding model learning, as shown in Fig.3, we follow the distillation learning framework that has a natural advantage in lightweight, allowing us to perform very complex implicit degradation modeling in the teacher stage without affecting the model complexity and inference efficiency in the student stage. During teacher training stage, we introduce a well-designed degradation-prior-constrained contrastive learning technique, which uses specific degradation parameters as degradation reference priors (DRP) to reinforce the teacher's learning of discriminative degradation-related information in LR images. In the student training stage, we utilize a feature alignment-based distillation technique to transfer the degradation-related knowledge acquired by the teacher to a simplified student model that only takes the LR image as input to meet the task requirement.

Extensive experiments on both synthetic and real-world images demonstrate the effectiveness of our novel BSR model design. Thanks to the high discriminability of IDRs, even with a low complexity degradation adaptation module, it is easy to build a high-quality mapping from LR to HR with the generalization to any degradation scenarios. Compared to existing advanced EDE- and IDE-BSR methods under the same settings, our method achieves outstanding performance across various blind SR tasks with minimal parameters and computational costs.

## 2. Related Work

The BSR aims to extract degradation information from LR images using a learnable estimator instead of manually set parameters [38, 41], to guide reconstruction. **EDE-BSR.** Early methods explicitly estimate degradation. IKC [7] iteratively refines the estimator with generated SR results. DAN [13] jointly predicts blur kernels and SR images in a dual-branch design. DCLS [25] introduces constrained least-square filtering to produce deblurred features. These approaches often involve costly iterative estimation. **IDE-BSR.** Recent works model degradation implicitly by learning latent representations integrated with LR features. DASR [32] first applied contrastive learning [8] to distinguish degradations. IDMBSR [43] uses kernel width and noise level as weak supervision. MRDA [36] adopts meta-learning [6] in a multi-stage setup. KDSR [37] introduces knowledge distillation [1, 10, 14], transferring knowledge from HR-supervised teachers to students for degradation estimation. CDFormer [22] and DSAT [21] stack large Transformer blocks [5, 19] for performance gains, but their complexity limits practical use. **Our Goal.** Instead of enlarging models, we enhance the discriminability of the IDR space by combining contrastive learning and distillation, enabling lightweight yet effective BSR. Further analysis is in the supplementary material.

## 3. Method

### 3.1. Architecture Overview

The architecture of the proposed LightBSR is shown in Fig.2, which mainly consists of an LR Feature Extractor (a 3×3 Conv layer), an IDR Estimation Module (IDR-EM), an IDR Adaptation Module (IDR-AM) and an Upscaler [29].

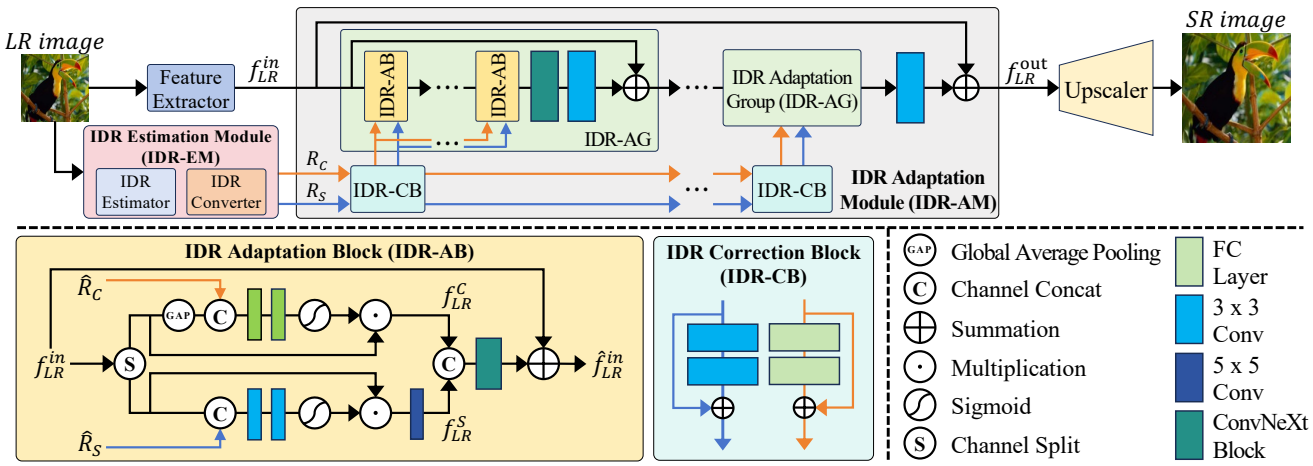



Figure 2. The architecture of our designed implicit degradation estimation-based lightweight blind SR network.

The LR image is first fed into both the low-level LR image feature extractor and IDR-EM simultaneously. Secondly, outputs are fed into IDR-AM for a series of modulation and fusion. Thirdly, the adapted result is fed into the Upscaler to generate the SR image. More design details are below.

#### 3.1.1. IDR Estimation Module (IDR-EM)

IDR-EM consists of an IDR-estimator and an IDR-converter. The former is used to extract the original IDR, and its structure is a six-layer Conv network like KDSR [37], but halves channels of each layer to reduce parameters. The latter includes a spatial transformation branch (consisting of a pixel-shuffle layer and a Conv layer) and a channel transformation branch (consisting of a GAP operator and a FC layer) for further generating spatial-wise IDR and channel-wise IDR based on original IDR, respectively.

#### 3.1.2. IDR Adaptation Module (IDR-AM)

IDR-AM adopts a hierarchical adaptation process, consisting of a series of IDR Correction Blocks, IDR Adaptation Blocks and IDR Adaptation Groups. Details are as follows.

**1) IDR Correction Block (IDR-CB).** The IDR-CB is designed to further optimize channel-wise IDR $R_C$ and spatial-wise IDR $R_S$. The channel-level correction includes two FC layers with GELU [9] and a residual connection, while the spatial-level correction includes two Conv layers with GELU and a residual connection. Sec.4.2 shows that this module can enhance the robustness of adaptation module to IDR estimation errors to a certain extent.

**2) IDR Adaptation Block (IDR-AB).** IDR-AB is the basic adaptation unit. As shown in Fig.2, taking LR features $f_{LR}^{in}$ and corrected IDRs as inputs, IDR-AB modulates IDRs from channel and spatial perspectives, and integrates IDRs into LR features. Given its intricate design, we will elaborate on the internal structure of IDR-AB in detail in Sec.3.2.1 by combining the specific calculation process.

**3) IDR Adaptation Group (IDR-AG).** The IDR-AG is composed of eight IDR-ABs, followed by a ConvNeXt block [23] and a 3×3 Conv layer for further feature fusion. And a residual link is added between the input and output of an IDR-AG for restoring low-level details.

Finally, the IDR-AM consists of eight IDR-AGs and a 3×3 Conv layer. A residual link is also applied to preserve the original image details and semantic information.

### 3.2. KD-based Model Learning

As shown in Fig.3, the training of proposed LightBSR is divided into two stages: For teacher training, contrastive learning (CL) and degradation reference prior (DRP) are introduced into the training process to reinforce the discriminability of the learned IDR latent space. For student training, the degradation knowledge learned by the teacher is transferred to the student through feature alignment.

#### 3.2.1. Teacher Training

The CL-based teacher training framework mainly consists of a DRP generator, a principal branch, a momentum branch and a negative sample queue $\mathcal{Q}$. Each branch includes an IDR-estimator and a projector (a two-layer FC network). The training is divided into two stages, of which the first stage pre-trains the IDR-estimator by only updating two CL branches, and the second stage fine-tunes the whole model parameters. Note that during teacher training, the IDR-estimator and the IDR-converter are separate.

**1) Data Preparation.** Given a batch of $B$ LR images, we randomly crop $D$ patches (size is $3 \times H \times W$) for each LR image to form $B$ positive sample sets. For DRP generation, we refer to dimensionality stretching strategy [41]. Specifically, first, the blur kernel of size $k \times k$ is vectorized and projected into a $t$-dimensional space via PCA, resulting in a $t \times 1$ vector. Next, the noise level value $\sigma$ is replicated 3 times and concatenated with the blur kernel vector, yielding a $(t+3) \times 1$ vector. Then, this vector is stretched to match the LR patch dimensions, resulting in the tensor representation of DRP $F_{drp} \in \mathbb{R}^{(t+3)\times H\times W}$. Finally, the input for teacher training is the concatenation of each patch and its cor-

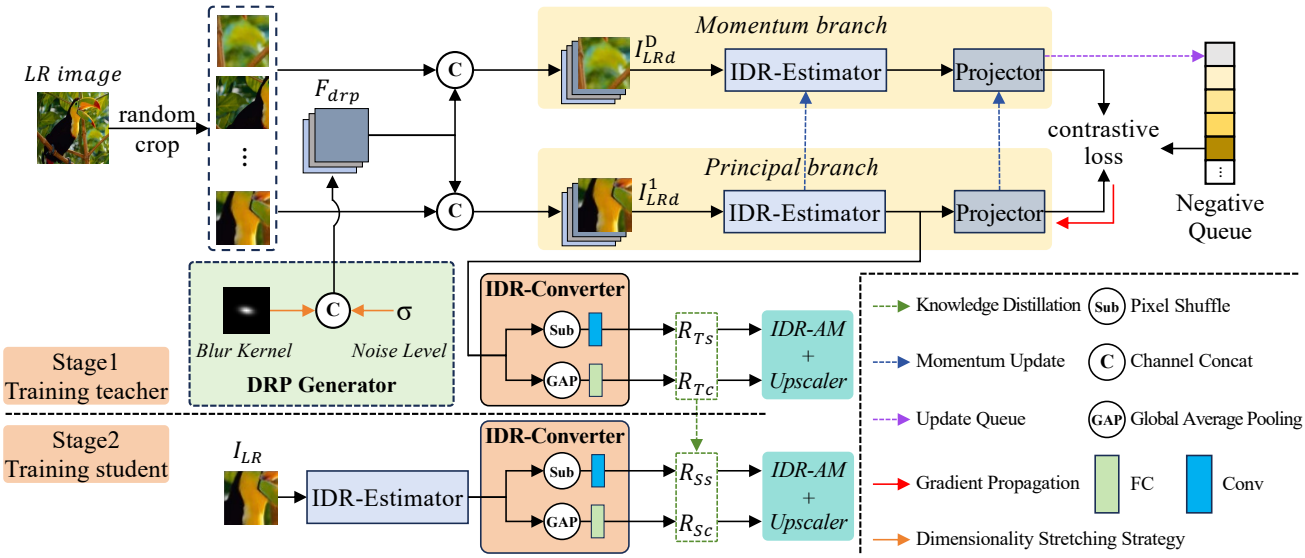



Figure 3. The training illustration of our LightBSR, in which the teacher stage uses complex structures to enhance IDR discriminability.

responding $F_{drp}$, defined as $I_{LRd} \in \mathbb{R}^{B\times D\times(t+6)\times H\times W}$.

**2) IDR-Estimator Pre-training.** The first stage can be seen as pre-training the IDR-Estimator. $I_{LRd}$ is fed simultaneously into the principal and momentum branches, producing $\mathcal{P}, \mathcal{M} \in \mathbb{R}^{B\times D\times 128}$, respectively. Then, the contrastive loss is computed based on $\mathcal{M}$, $\mathcal{P}$, and $\mathcal{Q}$, as follows:

$$L_{CL} = \frac{1}{D(D-1)} \sum_{i=1}^{D} \sum_{j=1}^{D} Sim(\mathcal{P}_i, \mathcal{M}_j), (i \neq j), \quad (1)$$

where the operator $Sim(\cdot)$ refers to [27], as follows:

$$Sim(\mathcal{P}_i, \mathcal{M}_j) = -\frac{1}{B} \sum_{k=1}^{B} log \frac{exp(\mathcal{P}_i^k \cdot \mathcal{M}_j^k/\tau)}{exp(\mathcal{P}_i^k \cdot \mathcal{M}_j^k/\tau) + \sum_{l=1}^{N} exp(\mathcal{P}_i^k \cdot \mathcal{Q}^l/\tau)}, \quad (2)$$

where $\tau$ denotes the temperature coefficient, and $N$ is the length of $\mathcal{Q}$. For the momentum branch, the parameters are updated using the momentum update strategy:

$$\theta^{\mathcal{M}} \leftarrow \alpha\theta^{\mathcal{M}} + (1-\alpha)\theta^{\mathcal{P}}, \quad (3)$$

where $\theta^{\mathcal{M}}$ and $\theta^{\mathcal{P}}$ are the parameters of the momentum and principal branch, $\alpha \in [0, 1]$ is the momentum coefficient.

**3) Overall Model Fine tuning.** The second stage expanded the parameter training scope based on the first stage. Specifically, after obtaining the output of the IDR-estimator of the principal branch, it is further fed into the IDR-converter for spatial- and channel-domain transformation, outputting a spatial-wise IDR $R_{Ts} \in \mathbb{R}^{8\times H\times W}$ and a channel-wise IDR $R_{Tc} \in \mathbb{R}^{48}$, respectively. These two IDRs, along with the LR image features, are then fed into the IDR-AM and Upscaler to generate the HR image. The L1 loss is used to measure the reconstruction error, as follows:

$$L_{SR} = \|I_{SR} - I_{HR}\|_1, \quad (4)$$

where $I_{SR}$ represents the output of the BSR network, $I_{HR}$ represents the HR version corresponding to the input LR image. The sum of $L_{SR}$ and $L_{CL}$ is defined as the total loss in the second stage for training the teacher network.

**4) IDR Adaptation.** As detailed in Sec.3.1, the adaptation process in IDR-AM is sequential, consisting of a series of IDR-AGs and a Conv layer, where each IDR-AG is composed of a series of IDR-ABs and two Conv blocks. As shown in Fig.2, taking output IDRs of IDR-EM and LR features as input, IDR-CB first optimizes IDRs to obtain $\widehat{R}_C$ and $\widehat{R}_S$ and sends them to the corresponding IDR-AG and next IDR-CB. In IDR-AG, the IDR-AB receives optimized IDRs and LR features as inputs. IDR-AB first divides the input LR features $f_{LR}^{in}$ into two parts along the channel dimension in a 1:3 ratio to fully use the redundancy between different channels and reduce computational complexity. For spatial-wise modulation, $\frac{1}{4}f_{LR}^{in}$ is concatenated with $\widehat{R}_S$ along the channel dimension, followed by modulation through two 3×3 Conv layers and a Sigmoid function. Next, the modulated spatial-wise IDR is multiplied with the original $\frac{1}{4}f_{LR}^{in}$, followed by a 5×5 Conv layer for further fusion to obtain the spatial-wise degradation-adapted LR image features $f_{LR}^{S}$. For channel-wise modulation, the remaining $\frac{3}{4}f_{LR}^{in}$ is first processed by a GAP operator, and then the pooled result is concatenated to $\widehat{R}_C$ along the channel dimension, followed by modulation through two FC layers and a Sigmoid function. Next, the modulated channel-wise IDR is multiplied with the original $\frac{3}{4}f_{LR}^{in}$ to obtain the channel-wise degradation-adapted LR features $f_{LR}^{C}$. Then $f_{LR}^{S}$ and $f_{LR}^{C}$ are concatenated and passed through a ConvNeXt block [23] for information complementary processing. Finally, a residual link is applied to integrate the original $f_{LR}^{in}$ into the processed result to enhance the image details, resulting in

| Method | DRP | CL | Set5 | Set14 | B100 | Urban100 |
|---|---|---|---|---|---|---|
| T1 | | | 31.60 | 28.25 | 27.40 | 25.69 |
| T2 | ✓ | | 31.96 | 28.45 | 27.47 | 25.94 |
| T3 | | ✓ | 31.75 | 28.34 | 27.43 | 25.77 |
| T4(Ours) | ✓ | ✓ | **32.00** | **28.49** | **27.50** | **26.02** |

Table 1. The effect of different training framework designs on the average PSNR under various kernel widths $\{1.2, 2.4, 3.6\}$.

| Method | $\widehat{R}_S$ | $\widehat{R}_C$ | IDR-CB | Set5 | Set14 | B100 | Urban100 |
|---|---|---|---|---|---|---|---|
| M1 | | | | 29.25 | 27.01 | 26.48 | 24.21 |
| M2 | ✓ | | ✓ | 31.86 | 28.43 | 27.46 | 25.87 |
| M3 | | ✓ | ✓ | 31.92 | 28.44 | 27.47 | 25.91 |
| M4(Ours) | ✓ | ✓ | ✓ | **32.00** | **28.49** | **27.50** | **26.02** |
| M5 | ✓ | ✓ | | 31.94 | 28.42 | 27.47 | 25.93 |
| M4- | M4 with error IDRs | | | 28.15 | 25.95 | 25.79 | 23.37 |
| M5- | M5 with error IDRs | | | 28.05 | 25.89 | 25.69 | 23.26 |

Table 2. The effect of different IDR-AM designs on the average PSNR under various kernel widths $\{1.2, 2.4, 3.6\}$.

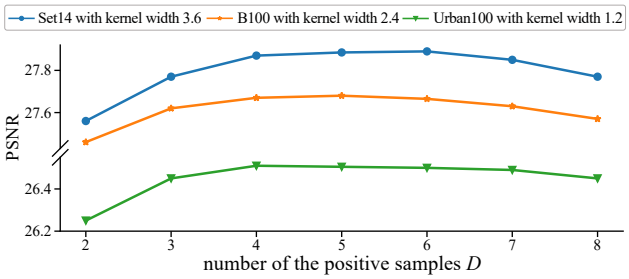



Figure 4. The effect of different numbers of positive samples.

the final output $\hat{f}_{LR}^{in}$ of current IDR-AB. After a series of hierarchical adaptations in IDR-AM, LR image features that fully integrate degradation information are obtained, denoted as $f_{LR}^{out}$, which would be used as input for the Upscaler.

#### 3.2.2. Knowledge Transfer

The student network mainly consists of an IDR-estimator IDR-E$_S$ and an IDR-converter IDR-C$_S$, with each component having the same structure as in the teacher network. In terms of input, the student network only takes the LR image as input, fulfilling the inference requirements. The spatial- and channel-wise IDRs output by IDR-C$_S$ are denoted as $R_{Ss}$ and $R_{Sc}$, respectively. After training the teacher network, the IDR-estimator of the principal branch and IDR-converter are exported as the teacher, defined as IDR-E$_T$ and IDR-C$_T$, respectively, and the learned knowledge is transferred to the student network using the distillation learning technique. Specifically, knowledge distillation is achieved by aligning the output between teacher and student converters, and the training is also divided into two stages, with the first stage only updating IDR-E$_S$ and IDR-C$_S$, and the second stage optimizing the whole model parameters.

In the first stage, for spatial-domain distillation, $L_2$ loss is used to perform pixel-level knowledge transfer from $R_{Ts}$ to $R_{Ss}$, which can be expressed as:

$$L_2 = \|R_{Ts} - R_{Ss}\|_2 . \tag{5}$$

For channel-domain representations $R_{Tc}$ and $R_{Sc}$, the Kullback–Leibler (KL) divergence loss is employed to match their distributions and the L1 loss is used to minimize their absolute differences, as follows:

$$L_{kl} = \sum_{i=1}^{C} R_{Tc}^{norm}[i] log(\frac{R_{Tc}^{norm}[i]}{R_{Sc}^{norm}[i]}), \tag{6}$$

$$L_1 = \|R_{Tc} - R_{Sc}\|_1 , \tag{7}$$

where $C$ is the number of channels, and $R_{Tc}^{norm}$ and $R_{Sc}^{norm}$ are softmax-normalized outputs of $R_{Tc}$ and $R_{Sc}$, respectively. The total loss of knowledge distillation is:

$$L_{DL} = L_2 + L_{kl} + \beta L_1, \tag{8}$$

where $\beta$ denotes the balance factor. In the second stage, $R_{Ss}$ and $R_{Sc}$ are further fed into IDR-AM and Upscaler for SR reconstruction, consistent with the teacher training. The sum of $L_{SR}$ and $L_{DL}$ is defined as the loss of the second stage for training the student network.

## 4. Experiments

### 4.1. Experimental Setup

**Datasets preparation.** For training, we use 800 images from DIV2K [2] and 2650 images from Flickr2K [30], consistent with previous works [32, 37]. For evaluation, we choose four standard benchmarks: Set5 [4], Set14 [40], B100 [26] and Urban100 [12]. The LR-HR sample pairs required for training follow the classical degradation model:

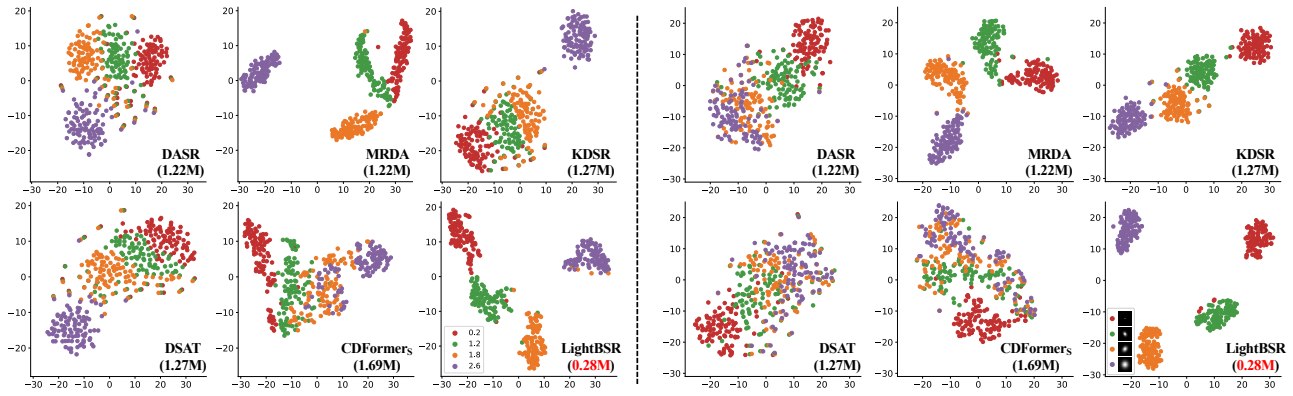
$$I_{LR} = (I_{HR} \otimes k) \downarrow_s + n, \tag{9}$$

where $I_{HR}$ is the original HR image, $I_{LR}$ is the degraded LR image, $\otimes$ denotes convolution operation, $k$ denotes the blur kernel, $\downarrow_s$ denotes the bicubic downsampling with scale factor $s$, and $n$ refers to additive Gaussian white noise. In experiments, two degradation settings are employed:

**Setting 1:** only contains isotropic Gaussian blur kernels. The kernel size is fixed to 21 $\times$ 21, and the range of kernel width is set to [0.2, 4.0] for $\times$4 SR.

**Setting 2:** considers anisotropic Gaussian blur kernels and Gaussian white noise for $\times$4 SR. The anisotropic Gaussian blur kernel characterized by a Gaussian probability density function $N(0, \Sigma)$, where covariance matrix $\Sigma$ is determined by two eigenvalues $\lambda_1, \lambda_2 \sim$ U(0.2, 4) and a rotation angle $\theta \sim$ U(0, $\pi$). The kernel size is fixed to 21 $\times$ 21 and the range of Gaussian white noise is set to [0, 25].

**Implementation Details.** For preprocessing, HR images are augmented via random rotation and flipping and the LR patch size is set to 64 $\times$ 64. For training, the batchsize $B$ is set to 64 and the number of the positive sample $D$ is set to 4. During IDR modeling, the PCA dimension $t$ is set to 15,




(a) Setting 1 (b) Setting 2

Figure 5. The t-SNE [31] plots of IDR distributions on the B100 [26] benchmark. a) 4 different isotropic blur kernels are selected under *Degradation Setting 1*. b) 4 different anisotropic blur kernels are chosen under *Degradation Setting 2*, with the noise level set to 4.

| Method | Param. | GFLOPs | Dataset | Set5 | | | Set14 | | | B100 | | | Urban100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Kernel width | 1.2 | 2.4 | 3.6 | 1.2 | 2.4 | 3.6 | 1.2 | 2.4 | 3.6 | 1.2 | 2.4 | 3.6 |
| **Bicubic** | — | — | — | 27.69 | 25.99 | 24.45 | 25.60 | 24.39 | 23.25 | 25.58 | 24.67 | 23.80 | 22.73 | 21.76 | 20.84 |
| **IKC** | 5.3M | 2876.30 | CNN | 31.76 | 30.35 | 30.26 | 28.44 | 28.17 | 26.63 | 27.43 | 27.28 | 26.41 | 25.63 | 25.02 | 24.07 |
| **DAN** | 4.3M | 1250.38 | CNN | 32.22 | 31.96 | 30.94 | 28.65 | 28.54 | 27.68 | 27.65 | 27.58 | 26.95 | 26.20 | 25.96 | 25.08 |
| **DCLS** | 13.6M | 496.72 | CNN | 32.35 | **32.19** | **31.14** | 28.66 | 28.57 | 27.78 | 27.73 | 27.65 | 27.02 | **26.50** | **26.24** | **25.34** |
| **DASR** | 5.8M | 208.80 | CNN | 31.92 | 31.75 | 30.59 | 28.44 | 28.28 | 27.45 | 27.52 | 27.43 | 26.83 | 25.69 | 25.44 | 24.66 |
| **IDMBSR** | 4.2M | — | CNN | 31.90 | 31.78 | 30.68 | 28.50 | 28.36 | 27.60 | 27.58 | 27.51 | 26.90 | 25.91 | 25.68 | 24.91 |
| **MRDA** | 5.8M | 386.25 | CNN | 32.36 | 32.11 | 30.89 | 28.67 | 28.57 | 27.62 | 27.67 | 27.58 | 26.91 | 26.26 | 26.02 | 25.08 |
| **KDSR** | 5.8M | 258.14 | CNN | 32.34 | 32.13 | 31.02 | 28.66 | 28.55 | **27.81** | 27.67 | 27.59 | 26.97 | 26.29 | 26.05 | 25.20 |
| **DSAT** | 15.6M | 895.03 | Transformer | **32.51** | 32.00 | 30.31 | 28.77 | 28.50 | 27.51 | 27.76 | 27.66 | 27.02 | 26.43 | 25.95 | 24.89 |
| **CDFormerS** | 11.9M | 631.53 | Transformer | 32.36 | 32.14 | 31.07 | 28.77 | 28.60 | 27.67 | 27.74 | 27.64 | 27.00 | 26.39 | 26.09 | 25.19 |
| **LightBSR** | **3.1M** | **206.16** | CNN | 32.47 | 32.17 | 31.06 | **28.82** | **28.65** | 27.79 | **27.77** | **27.69** | **27.04** | **26.50** | 26.20 | 25.31 |

Table 3. Quantitative comparison of PSNR metric for different methods under *Degradation Setting 1* on $\times 4$ SR. The best and second-best results are highlighted in **red** and blue, respectively. GFLOPs are calculated based on a 256 $\times$ 256 image size.

the temperature coefficient $\tau$ is set to 0.07, the momentum coefficient $\alpha$ is set to 0.999, and the balance coefficient $\beta$ in $L_{DL}$ is set to 0.1. For both teacher and student training, in the first stage, learning rate is set to 2e-4 for 100 epochs. In the second stage, cosine annealing strategy [24] is used to gradually decrease the learning rate from 2e-4 to 1e-6 over 600 epochs. Adam optimizer [18] is used for training.

## 4.2. Ablation Study

We conduct a detailed analysis of the effects of each component in the proposed method under *Degradation Setting 1*. Additional experimental results under *Degradation Setting 2* are provided in the supplementary material.

### 4.2.1. Core Components for Training

**1) Effect of DRP and CL.** These two are used to enhance the discriminability of IDRs. The ablation results are shown in Table 1. Using naive model T1 without any DRP or CL as the baseline, adding DRP (T2) or introducing CL (T3) both effectively improve performance, with average Peak signal-to-noise ratio (PSNR) on four benchmarks exceeding that of T1 by 0.22 dB and 0.09 dB, respectively. Moreover, the combination of DRP and CL, i.e. T4 (ours), can further improve performance, demonstrating the effectiveness and compatibility of the two components.

**2) Impact of Positive Sample Number.** To verify the impact of the number of positive samples, we gradually increase the value of $D$ from 2 to 8, with the results shown in Fig.4. When $D$ increased from 2 to 4, the performance on various benchmarks improves significantly, indicating that more positive samples can provide additional beneficial reference prior. However, when $D$ exceeds 4, the performance reaches saturation or even declines, possibly due to excessive redundant information impairing the estimator's robustness against degradation distribution shifts [28].

### 4.2.2. Core Components of IDR-AM

We perform detailed ablation studies on the core components of IDR-AM, namely IDR-AB and IDR-CB, and the ablation results are shown in Table 2.

**1) Effect of IDR-AB.** There are two degradation modulation branches in IDR-AB, corresponding to channel and spatial perspectives. Compared to the baseline (M1) without any modulation, employing only spatial-domain modulation (M2) or channel-domain modulation (M3) both improves SR performance, with the average PSNR increasing by 1.67 dB

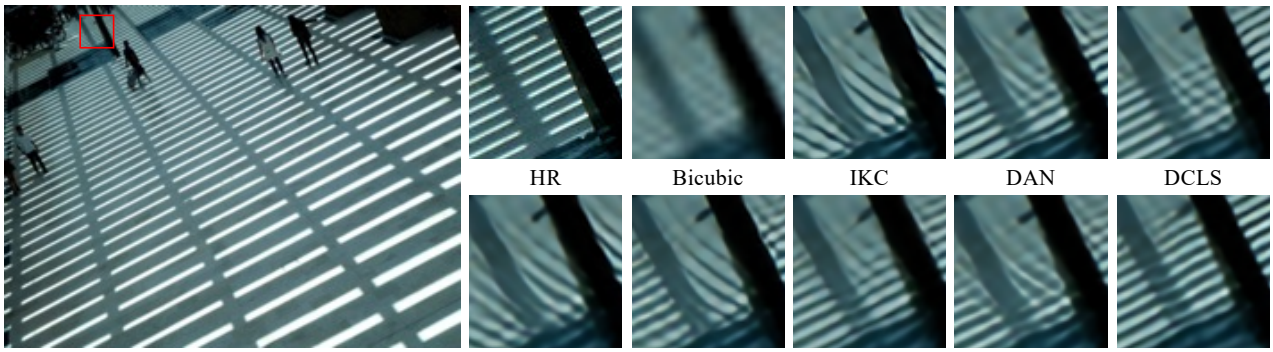



Figure 6. $\times$4 SR results for isotropic Gaussian blur kernel width 1.2 on the "img_093" of Urban100 benchmark.

| Noise | Method | Param. |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | **DnCNN+DAN** | 650K+4.33M | 26.18 | 26.14 | 24.89 | 24.64 | 24.61 | 24.52 | 24.37 | 24.01 | 23.63 |
| | **DnCNN+DCLS** | 650K+13.63M | 26.15 | 26.12 | 24.89 | 24.66 | 24.62 | 24.52 | 24.39 | 24.03 | 23.65 |
| | **DASR** | 5.84M | 26.60 | 26.52 | 25.87 | 25.69 | 25.64 | 25.60 | 25.48 | 25.23 | 24.85 |
| | **MRDA** | 5.84M | 26.73 | 26.68 | 25.98 | 25.80 | 25.77 | 25.71 | 25.59 | 25.29 | 24.82 |
| | **KDSR** | 5.80M | 26.73 | 26.70 | 26.02 | 25.87 | 25.83 | 25.76 | 25.65 | 25.37 | 24.94 |
| | **DSAT** | 15.64M | 26.71 | 26.66 | 25.95 | 25.80 | 25.74 | 25.66 | 25.55 | 25.30 | 24.85 |
| | **CDFormer$_S$** | 11.86M | 26.77 | 26.69 | 26.00 | 25.85 | 25.82 | 25.75 | 25.64 | 25.34 | 24.97 |
| | **LightBSR** | **3.09M** | **26.82** | **26.76** | **26.06** | **25.91** | **25.87** | **25.81** | **25.70** | **25.41** | **25.05** |
| 10 | **DnCNN+DAN** | 650K+4.33M | 25.69 | 25.63 | 24.62 | 24.42 | 24.39 | 24.29 | 24.18 | 23.89 | 23.54 |
| | **DnCNN+DCLS** | 650K+13.63M | 25.67 | 25.62 | 24.62 | 24.43 | 24.40 | 24.29 | 24.19 | 23.91 | 23.56 |
| | **DASR** | 5.84M | 26.02 | 25.94 | 25.27 | 25.11 | 25.08 | 25.03 | 24.93 | 24.69 | 24.33 |
| | **MRDA** | 5.84M | 26.12 | 26.05 | 25.36 | 25.20 | 25.18 | 25.12 | 25.01 | 24.76 | 24.36 |
| | **KDSR** | 5.80M | 26.11 | 26.05 | 25.38 | 25.24 | 25.21 | 25.15 | 25.05 | 24.81 | 24.46 |
| | **DSAT** | 15.64M | 26.07 | 26.04 | 25.29 | 25.18 | 25.14 | 25.09 | 24.97 | 24.72 | 24.36 |
| | **CDFormer$_S$** | 11.86M | 26.14 | 26.06 | 25.37 | 25.22 | 25.20 | 25.15 | 25.04 | 24.78 | 24.45 |
| | **LightBSR** | **3.09M** | **26.18** | **26.10** | **25.41** | **25.27** | **25.24** | **25.19** | **25.08** | **24.83** | **24.53** |

Table 4. Quantitative comparison of PSNR metric for different methods on the B100 benchmark under *Degradation Setting 2* on $\times$4 SR. The best and second-best results are highlighted in **red** and blue, respectively.

and 1.7 dB, respectively, validating the necessity of modulation from both perspectives. And the combination of the two branches, i.e. M4 (ours), further improves the PSNR, showing the complementarity of the two.

**2) Effect of IDR-CB.** The variant M5 without IDR-CB shows a 0.06 dB decrease in average PSNR compared to M4 (ours), preliminarily showing the effectiveness of this component. To further verify the effect of IDR-CB, we add random values ranging from 0 to 1 to the output of the IDR-Estimator, simulating incorrect IDRs and the results are shown at the bottom of Table 2. It can be seen that when receiving error IDRs, M5- (without IDR-CB) has an average PSNR 0.09 dB lower than M4- (with IDR-CB), showing that IDR-CB can improve the robustness of SR network to error degradation estimation to a certain extent.

## 4.3. Comparison with Previous Methods

We compare our approach under two degradation settings against the following BSR methods: IKC [7], DAN [13], DCLS [25], DASR [32], IDMBSR [43], MRDA [36], KDSR [37], DSAT [21] and CDFormer$_S$ [22]. Among these methods, IKC [7], DAN [13] and DCLS [25] follow EDE-BSR paradigm, while the others follow IDE-BSR paradigm. The out-of-distribution (OOD) degradation evaluations are provided in the supplementary material.

### 4.3.1. Experiments in Degradation Setting 1

**Quality of IDRs.** We apply four different blur kernel widths to generate LR images for the B100 benchmark [26], feeding them into the degradation estimators of DASR [32], MRDA [36], KDSR [37], DSAT [21], CDFormer$_S$ [22] and LightBSR. The t-SNE plots of IDRs output by each method are shown in Fig.5(a). Our method achieves better clustering of the same degradation and separation of different degradations, which shows the superiority of our method in discriminative implicit degradation estimation.

**Quantitative evaluation.** We use three different blur kernel widths $\{1.2, 2.4, 3.6\}$ to evaluate the SR performance of various methods across four classic benchmarks. The comparisons in Table 3 show that: (1) Compared to the best EDE-based BSR method [25], our approach achieves superior or matching performance under 8 out of 12 degradation conditions, while reducing the number of parameters by 78%

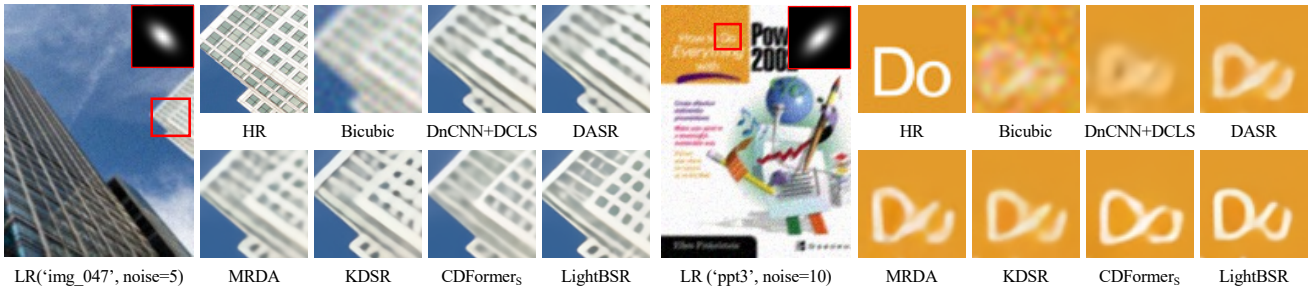



Figure 7. $\times$4 SR results for different anisotropic blur kernels and noise levels on the "img_047" of Urban100 and the "ppt3" of Set14.

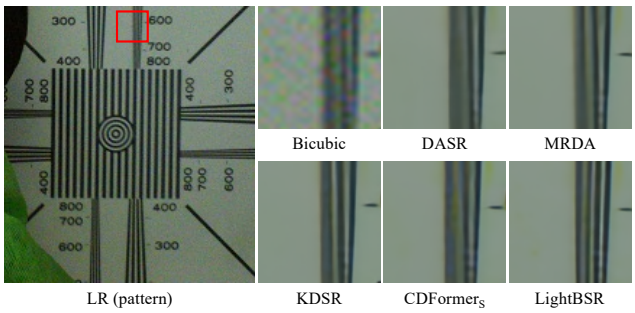



Figure 8. $\times$4 SR results on the "pattern" image of the RealWorld38 for various methods under *Degradation Setting 2*.

and computation by 59%, showing the great potential of the IDE-based BSR paradigm. (2) Compared with existing IDE-BSR methods, LightBSR achieves SOTA performance with fewer parameters and computation, showing the superiority of our method. For example, compared to CDFormer$_S$ [22], our method, which is based on CNN architecture, achieves superior results with only 26% of the parameters and 32% of the computational cost used by CDFormer$_S$, demonstrating that CNNs remain highly competitive in the BSR task.

**Qualitative evaluation.** The SR results of different methods under *Degradation Setting 1* are shown in Fig.6. Our method recovers the overall structure of the image more comprehensively and restore complex architectural texture details more clearly and completely. This shows that the discriminability of IDRs is crucial for image restoration.

#### 4.3.2. Experiments in Degradation Setting 2

**Quality of IDRs.** We use four different anisotropic blur kernels and set the noise level to 4 to generate LR images for the B100 benchmark [26]. The IDRs generation and t-SNE visualization follow *Degradation Setting 1*, and the results are shown in Fig.5(b). We can see that LightBSR also achieves the best discriminative IDR modeling.

**Quantitative evaluation.** Here we select nine different anisotropic blur kernels and two noise levels to evaluate various methods on B100 benckmark [26], with results shown in Table 4. It is evident that our LightBSR achieves SOTA performance under all combinations of blur kernels and noise levels. Compared to the EDE-based DCLS [25], LightBSR exceeds its average PSNR under all degradation conditions by 0.97 dB, corroborating the superiority of the IDE strategy in complex BSR tasks with noise interference. Compared to existing IDE-based methods, our method also improves performance with fewer parameters. For instance, LightBSR outperforms CDFormer$_S$ [22] by 0.05 dB in average PSNR across all blur kernel and noise combinations, while the number of parameters is only 26% of CDFormer$_S$.

**Qualitative evaluation.** The SR results of different methods under *Degradation Setting 2* are shown in Fig.7. Our LightBSR is able to restore rich details clearly and accurately, even under severe degradation conditions. In contrast, the SR images generated by DASR [32], MRDA [36], KDSR [37] and CDFormer$_S$ [22] are either highly blurred or difficult to recover some complex texture details.

#### 4.3.3. Experiments in Real Degradation Scenario

We use the "pattern" image from the RealWorld38 [20] to evaluate the generalizability of all methods to complex, unknown real degradation. Since no corresponding HR images are available, we use directly various models trained with *Degradation Setting 2* to perform qualitative comparisons. As shown in Fig.8, even if high-frequency details are lost, our method can still clearly restore line textures. More comparisons are provided in the supplementary material.

## 5. Conclusion

In this work, we demonstrate the importance of discriminative IDRs for the BSR task and reveal a novel lightweight design idea of BSR model. The proposed LightBSR follows the KD-based training framework, in which the teacher stage uses a well-designed degradation-prior-constrained contrastive learning technique to enhance the discriminability of the learned IDR latent space, and the student stage uses a feature alignment technique to transfer the learned degradation-related knowledge to the structurally simple student for practical reasoning. Extensive experiments show the effectiveness of our method, achieving excellent quantitative and qualitative results with minimal parameters and computation. Additionally, our work also shows that the CNN architecture still has great potential in the blind SR task.

**Acknowledgement.** This work was supported by the Natural Science Foundation of China (No. 62192782, 62036011), Beijing Natural Science Foundation (L223003), the Project of Beijing Science and Technology Committee (No. Z231100005923046).

## References


[1] Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and B Yoshua. Fitnets: Hints for thin deep nets. *ICLR*, 2:3, 2015. 2

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. 1, 2, 5

[3] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *NeurIPS*, 2019. 1

[4] 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. 5

[5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023. 2

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2

[7] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 1, 2, 7

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

[9] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[12] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 5

[13] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. In *NeurIPS*, 2020. 1, 2, 7

[14] Guanzhou Ke, Bo Wang, Xiaoli Wang, and Shengfeng He. Rethinking multi-view representation learning via distilled disentangling. In *CVPR*, pages 26774–26783, 2024. 2

[15] Asif Hussain Khan, Christian Micheloni, and Niki Martinel. Idenet: Implicit degradation estimation network for efficient blind super resolution. In *CVPR Workshops*, 2024. 1

[16] Asif Hussain Khan, Christian Micheloni, and Niki Martinel. Lightweight prompt learning implicit degradation estimation network for blind super resolution. *IEEE Trans. Image Process.*, 2024. 1

[17] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *CVPR*, 2021. 1

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

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 2

[20] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, 2022. 8

[21] Qingguo Liu, Pan Gao, Kang Han, Ningzhong Liu, and Wei Xiang. Degradation-aware self-attention based transformer for blind image super-resolution. *TMM*, 2024. 2, 7

[22] Qingguo Liu, Chenyi Zhuang, Pan Gao, and Jie Qin. Cd-former: When degradation prediction embraces diffusion model for blind image super-resolution. In *CVPR*, 2024. 1, 2, 7, 8

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3, 4

[24] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[25] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *CVPR*, 2022. 1, 2, 7, 8

[26] David Martin, Charless 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. 5, 6, 7, 8

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[28] Guillermo Ortiz Jimenez, Apostolos Modas, Seyed Mohsen Moosavi Dezfooli, and Pascal Frossard. Redundant features can hurt robustness to distributions shift. In *ICML*, 2020. 6

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 2

[30] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR Workshops*, 2017. 5

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[32] 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. 1, 2, 5, 7, 8

[33] Yue Wang, Jiawen Ming, Xu Jia, James H Elder, and Huchuan Lu. Blind image super-resolution with degradation-aware adaptation. In *ACCV*, 2022. 1

[34] Shao-Yu Weng, Hsuan Yuan, Yu-Syuan Xu, Ching-Chun Huang, and Wei-Chen Chiu. Best of both worlds: Learning arbitrary-scale blind super-resolution via dual degradation representations and cycle-consistency. In *WACV*, 2024. 2

[35] Shixiang Wu, Chao Dong, and Yu Qiao. Blind image restoration based on cycle-consistent network. *TMM*, 25:1111–1124, 2022. 1

[36] Bin Xia, Yapeng Tian, Yulun Zhang, Yucheng Hang, Wenming Yang, and Qingmin Liao. Meta-learning-based degradation representation for blind super-resolution. *IEEE Trans. Image Process.*, 32:3383–3396, 2023. 2, 7, 8

[37] Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Knowledge distillation based degradation estimation for blind super-resolution. In *ICLR*, 2023. 1, 2, 3, 5, 7, 8

[38] Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, and Yi-Min Tsai. Unified dynamic convolutional network for super-resolution with variational degradations. In *CVPR*, 2020. 2

[39] Zhixiong Yang, Jingyuan Xia, Shengxi Li, Xinghua Huang, Shuanghui Zhang, Zhen Liu, Yaowen Fu, and Yongxiang Liu. A dynamic kernel prior model for unsupervised blind image super-resolution. In *CVPR*, 2024. 1

[40] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 5

[41] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 2, 3

[42] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 1

[43] Yongfei Zhang, Ling Dong, Hong Yang, Linbo Qing, Xiaohai He, and Honggang Chen. Weakly-supervised contrastive learning-based implicit degradation modeling for blind image super-resolution. *KBS*, 249:108984, 2022. 2, 7

[44] Hongyang Zhou, Xiaobin Zhu, Jianqing Zhu, Zheng Han, Shi-Xue Zhang, Jingyan Qin, and Xu-Cheng Yin. Learning correction filter via degradation-adaptive regression for blind single image super-resolution. In *ICCV*, 2023. 1