Best of Both Worlds: Learning Arbitrary-scale Blind Super-Resolution via Dual Degradation Representations and Cycle-Consistency

Shao-Yu Weng[1] Hsuan Yuan[1] Yu-Syuan Xu[2] Ching-Chun Huang[1] Wei-Chen Chiu[1]
[1] National Yang Ming Chiao Tung University [2] MediaTek Inc.
{vivian5190.ee10, yuan040686.cs10, chingchun, walon}@nycu.edu.tw, Yu-syuan.xu@mediatek.com

Abstract

Single image super-resolution (SISR) for reconstructing from a low-resolution (LR) input image its corresponding high-resolution (HR) output is a widely-studied research problem in the field of multimedia applications and computer vision. Despite the magic leap brought by recent development of deep neural networks for SISR, such problem is still considered to be quite challenging and non-scalable for the real-world data due to its ill-posed nature, where the degradations happened to the input LR images are usually complex and even unknown (in which the degradations in the test data could be unseen or different from the ones shown in the training dataset). To this end, two branches of SISR methods have emerged: blind super-resolution (blind-SR) and arbitrary-scale super-resolution (ASSR), where the former aims to reconstruct SR images under the unknown degradations, while the latter improves the scalability via learning to handle arbitrary up-sampling ratios. In this paper, we propose a holistic framework to take both blind-SR and ASSR tasks (accordingly named as arbitrary-scale blind-SR) into consideration with two main designs: 1) learning dual degradation representations where the implicit and explicit representations of degradation are sequentially extracted from the input LR image, and 2) modeling both upsampling (i.e. LR→HR) and downsampling (i.e. HR→LR) processes at the same time, where they utilize the implicit and explicit degradation representations respectively, in order to enable the cycle-consistency objective and further improve the training. We conduct extensive experiments on various datasets where the results well verify the effectiveness of our proposed framework in handling complex degradations as well as its superiority with respect to several state-of-the-art baselines.

1. Introduction

In recent years we have witnessed a large improvement for addressing the task of single-image super-resolution (denoted as SISR, which aims to recover a high-resolution image from a low-resolution input) with the help of deep learning techniques to increase the model capacity. Nevertheless, lots of early attempts have an assumption upon the degradation process being bicubic downsampling (i.e. the low-resolution image is made by applying bicubic downsampling on its high-resolution counterpart) [12, 15, 16, 19, 34, 35], which leads to the first practical challenge while being directly applied to real-world data: As real images often undergo the degradations which are different and more complex than the bicubic one, the mismatch upon degradation process hence results in the performance drop for super-resolution models. Another practical challenge comes from the fact that most of these methods are typically designed for performing super-resolution/upsampling with a specific scale (or a set of fixed scales, usually integer ones), thus limiting their applicability to arbitrary-scale upsampling.

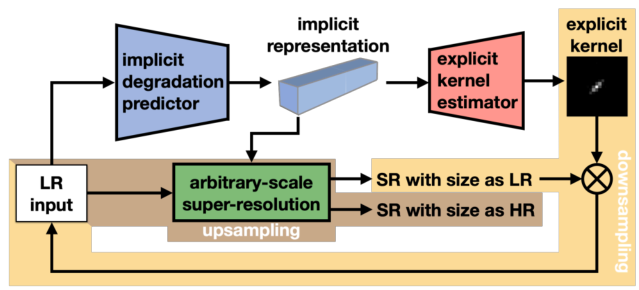

Figure 1. Conceptual illustration of our proposed method for tackling arbitrary-scale blind super-resolution problem, with highlighting our two main designs: 1) utilizing both implicit degradation representation and explicit degradation kernel, and 2) integrating both upsampling and downsampling processes (shaded by light brown color and light yellow color respectively) into a holistic framework for enabling the training objective of cycle-consistency.

Corresponding to the aforementioned two challenges, two research problems of SISR have been proposed: blind super-resolution (blind-SR) and arbitrary-scale super-resolution (ASSR). For the blind-SR, it aims to reconstruct high-resolution images under unknown degradations, in which the model typically puts no prior assumption on the type of degradations (mostly known as degradation/downsampling kernels) or can be adaptive to the degradation kernel of the low-resolution input image [2, 13, 18, 20, 26, 29, 36]. Without loss of generality, blind-SR methods typically consist of two stages: a predictor firstly estimates degradation kernels from the input LR images, followed by a super-resolution module to take the estimated kernel as prior information or a condition for performing the upsampling. Particularly, the predictor plays a crucial role and can be categorized into two categories according to the form of its outputs: estimation of explicit degradation kernels [2, 13, 20, 29] or estimation of implicit degradation representations [18, 26, 36]. The former provides a more direct way of using estimated degradation for the super-resolution module (e.g. explicit and physically-meaningful priors upon kernels can be easily applied for simpler utilization in super-resolution [2]), but it may suffer the performance drop when there is a mismatch between the estimated kernels and the actual ones [33]; The latter in turns learns to extract latent/implicit representation of degradations (instead of estimating the explicit kernels) in which different degradations in the representation space should be distinguishable. While the flexibility in the implicit representations helps to alleviate the kernel mismatch problem, but how to ensure the discriminativeness among various degradation representations and their integration with super-resolution module would be other concerned issues.

For arbitrary-scale super-resolution (ASSR), its main goal is to enable the super-resolution model to handle arbitrary upsampling scales (e.g. continuous scaling factors instead of just integer ones). The seminal works of ASSR (e.g. LIIF [6] and LTE [17]) typically learn to represent an image as a continuous function (e.g. implicit neural representation based on multi-layer perceptron) which can map the queries of continuous image coordinate to their corresponding RGB values, thus being able to produce outputs at arbitrary scales as the coordinates are continuous.

In this paper, we propose to tackle both blind-SR and ASSR problems at the same time (i.e. we name it **arbitrary-scale blind-SR task**), which is the first of its kind to the best of our knowledge. We argue that the direct integration over existing blind-SR and ASSR techniques is actually nontrivial, as most of the blind-SR works conduct their studies under the fixed scaling factors while most of ASSR works have prior assumption upon the degradation process (cf. Table 1 for the inferior performance of their direct combination, e.g. DCLS+LIIF or LTE). We take the two blind-SR categories as examples here: For blind-SR methods of estimating explicit kernels, given two LR images produced by applying the identical degradation kernel on the same HR image but undergoing different downsampling scales, the predictor should still output the same estimation for them. Hence, besides the difficulty stemming from the potential mismatch between estimated kernels and the actual ones (due to the ambiguity resulting from the downsampling process [20]), the input images from different downsampling scales bring another burden/ambiguity for the model learning; While for the blind-SR method of estimating the implicit degradation representations, the introduction of different scaling factors (in addition to various degradations) also further makes the learning of representations and ensuring discriminativeness more complicated.

Our proposed framework provides a feasible solution for addressing blind-SR and ASSR simultaneously, where there are several key design choices: 1) We adopt both implicit and explicit degradation representations in our framework for better leveraging their advantages, in which they are sequentially estimated from the input LR image (i.e. firstly predicting implicit representation from LR input then inferring the explicit kernel from the predicted implicit representation). In particular, the LR input is projected into the wavelet domain where the resultant high-frequency subbands are used to extract distinctive implicit representations via the predictor; 2) The predicted implicit degradation representation is incorporated into the super-resolution module/subnetwork which supports arbitrary-scale upsampling, for achieving the adaptive ability of blind-SR. In addition to the upsampling process, we convolve the super-resolution result with the previously estimated explicit degradation kernel for realizing the degradation/downsampling process and reverting to the LR image. Moreover, super-resolution result used in such downsampling process is actually with the same image size as the original LR input image, thus the ambiguity/uncertainty in terms of downsampling scale is alleviated. The upsampling and downsampling processes together form a closed-loop, which allows us to exploit the cycle-consistency objective for further driving the model optimization, and it is believed that jointly considering both upsampling and downsampling in the model training benefits the regularization against the ill-posed nature of the super-resolution problem [9]. By holistically tackling the arbitrary-scale blind-SR task, we demonstrate the effectiveness of our proposed method through extensive experiments and comparisons with respect to various baselines. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first work proposing to explicitly address both blind-SR and ASSR problems jointly.
- We utilize both implicit and explicit degradation representations in which they are respectively incorporated with the arbitrary-scale super-resolution module and the degradation process to form our holistic framework for the task of arbitrary-scale blind-SR.
- Both the upsampling and downsampling processes are modeled in our framework to construct a closed-

loop which is experimentally shown to benefit overall model training (with noting that the downsampling is achieved via inherent strengths of ASSR instead of adopting separate downsampling module or relying on bicubic downsampling as other approaches).

2. Related Works

Blind Super-resolution. Several pioneering works [7, 15, 16] have achieved promising results in SISR by using deep networks with predefined degradation, such as bicubic subsampling. However, these methods suffer from severe performance drops when being applied to the real-world data produced by unknown degradations (which are typically different from the ones in the training set). To address this issue, blind SISR methods have emerged, which aim to reconstruct high-resolution images from low-resolution images without knowing the degradation in advance. These methods can be categorized into two groups: those that explicitly estimate the degradation kernels [2, 13, 20, 29] and those that implicitly derive the degradation embeddings [18, 26, 36]. In the former group (i.e. estimating explicit degradation kernels), KernelGAN [2] estimates the kernel by taking advantage of the internal cross-scale recurrence property among images at different scales without requiring additional training data. However, this approach can be time-consuming due to the iterative kernel estimation process, and it may suffer from kernel mismatch issues when projecting the kernel from the low-resolution space to the high-resolution space. To mitigate these issues, KernelNet [29] and DCLS [20] ease the task into image deblurring. KernelNet [29] first estimates the coarse kernel in the low-resolution space and then refines it in the high-resolution space using self-convolution techniques. DCLS [20] reformulates the task as deblurring in the Fourier transform domain and derives a new low-resolution space degradation kernel. In the latter group (i.e. estimating implicit degradation representations), DASR [26] and CDSR [36] learn to distinguish different degradations in the feature space using contrastive learning. This strategy helps to avoid the kernel mismatch problem and enables adaptive use of the degradation representations in the SISR model. Unlike previous works that only utilize one of these strategies, our approach leverages both explicit and implicit degradations to address the blind SISR problem in a more holistic manner.

Arbitrary-scale Super-resolution. Previous SR research mostly focuses on a fixed-scale or only integer scales, while ASSR [6, 10, 17, 21, 24, 27, 30] is more realistic in real-world scenarios to consider arbitrary or continuous scales. MetaSR [10] is the first work to address ASSR by dynamically predicting the weights for the upscaling convolution modules. Inspired by the recent advance upon implicit neural representations (INR) for 3D shape reconstruction, LIIF [6] adopts a multi-layer perceptron (MLP) to learn a continuous representation for images which takes the continuous image coordinate as well as the image features around the coordinate as input and output the RGB value at the given coordinate. However, MLPs are known to struggle with learning high-frequency components [25]. LTE [17] addresses this issue by encoding image textures in Fourier space. While SRNO [27] introduces neural operator [14] to capture global relationships thus avoiding the point-wise limitation of MLP. ITSRN [30] and ITSRN++ [24] further propose implicit transformers that fully utilize the INR structure on screen image content. However, all these works are limited to single degradations (i.e. being less adaptive to other unseen/unknown degradations). In this work, we aim to advance ASSR under unknown degradations.

Cycle-consistency Loss. Cycle-consistency loss [8, 9, 22, 31, 37], which was originally introduced in CycleGAN [37], has not only been widely adopted in image translation tasks but also (conceptually) extended to the model designs for various applications. In the context of SR, this loss has been extended and applied in mapping relationships among different domains. For example, [22] uses cycle-consistency loss across the domains of clean LR and LR images, while DRN [9] frames it into a dual learning task among LR and HR images. And CinCGAN [31] utilizes both cycles as [22] and [9] into its optimization functions. Additionally, this strategy is also applied to zero-shot SR [8] that learns an image-specific mapping between LR and HR images. In our work, our proposed method models both the upsampling (i.e. LR→HR) and downsampling (i.e. HR→LR) processes thus the cycle-consistency loss is enabled. This helps to constrain the possible solution space of the arbitrary-scale blind super-resolution model to those that are consistent with the information provided by the input LR image, hence benefiting our model training.

3. Proposed Method

Without loss of generality, the single image super-resolution problem follows the following degradation model:

$$y = (x \otimes k_h)_{\downarrow_s} + n \tag{1}$$

where y represents the low-resolution (LR) image, x denotes the high-resolution (HR) image, k_h represents the degradation kernel applied to x, $\downarrow_s$ denotes the downsampling operation with a scaling factor s, and $\otimes$ is the convolution operation. The term n typically denotes the white Gaussian noise. In the subsequent sections, we conduct our investigation mainly on the noise-free scenario (i.e. $n = 0$) as following the common practice [2, 29], in which the Equation 1 can thus be further reformulated as convolving the HR image that is already downsampled to the LR space (i.e. $x_{\downarrow_s}$) with the degradation kernel k_l in the LR space as well [20]:

$$y = x_{\downarrow_s} \otimes k_l \tag{2}$$

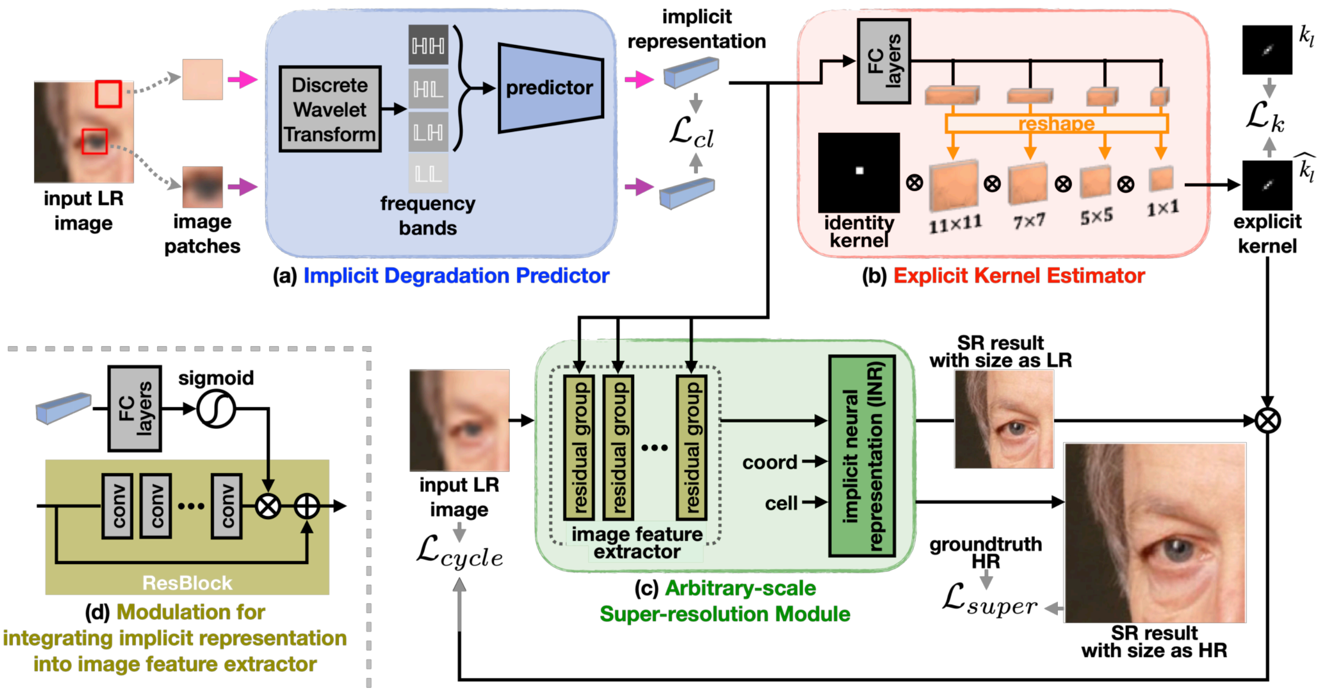

Figure 2. Overview of our proposed framework for the arbitrary-scale blind-SR task, in which it is composed of major components: (a) implicit degradation predictor (cf. Section 3.1), (b) explicit kernel estimator (cf. Section 3.2), and (c) arbitrary-scale super-resolution module (cf. Section 3.3). The input LR image is firstly gone through the implicit degradation predictor to derive the implicit degradation representation, then the implicit representation is not only adopted to estimate the explicit degradation kernel in LR space by using the explicit kernel estimator, but also taken as the condition for arbitrary-scale super-resolution module to output the super-resolution results, in which the manner of integrating implicit representation into the arbitrary-scale super-resolution module is based on (d) the modulation mechanism (noting that the residual groups in the image feature extractor of arbitrary-scale super-resolution module are built upon stacks of residual blocks). Moreover, the super-resolution result with the same size as LR image is further convolved with the estimated explicit kernel in LR space, where an upsampling-downsampling cycle (cf. Section 3.4) is therefore formed and it is experimentally shown to be beneficial for the overall model training.

3.1. Implicit Degradation Predictor

Our proposed framework starts with estimating implicit degradation representation from the input LR image. We are firstly inspired by the prior work from DASR [26] to learn the implicit representations for degradations, while we need to further take the variety in terms of scaling factors into account. As the additional scaling variety increases the overall complexity of learning implicit degradation representations in the original image space (noting that the original DASR only considers the images of the same/fixed scale), we step forward to adopt the lesson learned from [28] into our design for alleviating the complexity: it finds that the high-frequency parts of the LR image patches contain more important information of degradation (since the high-frequency parts are typically suppressed during blurring and downscaling, which results in different levels of degradation in different frequency bands for the same image), thus the wavelet transform is utilized to firstly project the input LR image into wavelet representations where the model is then trained upon. Accordingly, we leverage such insight to firstly employ the **wavelet transform** to the LR image patches (where the Haar transform is adopted as the wavelet basis) for deriving the low- and high-frequency components (denoted as y_L and y_H, respectively), then take the high-frequency subbands y_H as the input to train our predictor of implicit degradation representations.

The training of the predictor does follow the same practice as in DASR [26] and CDSR [36] to base on the **contrastive learning**. To be detailed, a training set composed of HR-LR image pairs is firstly synthesized by following Equation 1 being noise-free, where an HR image x is randomly cropped to obtain two patches, and the same degradation/blurring kernel and downsampling scale are applied to both patches to create two LR image patches y^1 and y^2 (in accordance with the common assumption that the patches from the same image ideally should have the same degradation). Subsequently, SimSiam [5] algorithm is adopted to perform contrastive learning upon the high-frequency subbands of y^1 and y^2, which are denoted as y^1_H and y^2_H respectively. The SimSiam model consists of an encoder E and a prediction head G, in which the contrastive objective $\mathcal{L}_{cl}$ is defined as follows:

$$\mathcal{L}_{cl} = \frac{1}{2}\mathcal{D}\left(E(y_H^1), G(E(y_H^2))\right) + \frac{1}{2}\mathcal{D}\left(E(y_H^2), G(E(y_H^1))\right) \tag{3}$$

where $\mathcal{D}(\cdot)$ computes the negative cosine similarity [5]. The resultant encoder E is then our predictor for estimating the implicit degradation representation of input LR image.

3.2. Explicit Kernel Estimator

Our framework proceeds to construct the explicit degradation kernel from the implicit representation estimated by the predictor E, via the help of an explicit kernel estimator M. The design of our explicit kernel estimator is similar to the one in DCLS [20] but being simpler (i.e. the subnetwork of the kernel estimator in DCLS) as our input for the estimator is already the implicit degradation representation while the estimator in DCLS needs to start from the LR input image. The detailed architecture of our explicit kernel estimator M is provided in the appendix A.2, so as the predictor E. Basically, the input implicit degradation representation is firstly projected to a lower dimension using two fully connected layers, and further processed through four separately fully connected layers before being reshaped into four corresponding convolution filters of size 11×11, 7×7, 5×5, and 1×1 respectively. Finally, these filters are convolved sequentially with an identity kernel (of size 41×41) to obtain the estimation of explicit degradation kernel (of size 21×21, the maximum kernel size used in our experiments). Such manner of deriving explicit kernel is named **degradation-dependent deep linear convolution** in our work, following the same naming rules as DCLS [20].

Please particularly note that, the explicit kernel estimated by our estimator M is actually aimed to be the degradation kernel in the LR space (cf. Equation 2), thus being denoted as $\widehat{k_l}$. The reason behind our performing kernel estimation in the LR space (i.e. refer to Equation 2 instead of Equation 1) is that the super-resolution of reconstructing x given y needs to estimate the degradation kernel k_h through the uncertainty of downsampling scale s (cf. Equation 1) hence leading to the potential mismatch of kernels. In contrast, the scenario described in Equation 2 puts all components (i.e. y, $x_{\downarrow_s}$, and k_l) in the same LR space, thus simplifying the kernel estimation process (as well as alleviating the issue of kernel mismatch).

Following the practice/derivation in DCLS [20], the objective $\mathcal{L}_k$ to drive the training of explicit kernel estimator is based on the L1 error between $\widehat{k_l}$ and its corresponding groundtruth k_l, i.e. $\left|\widehat{k_l} - k_l\right|_1$, in which k_l is defined as follows for the purpose of ensuring numerical stability during optimization:

$$k_l = \mathcal{F}^{-1}\left(\frac{\overline{\mathcal{F}(x_{\downarrow s})}}{\overline{\mathcal{F}(x_{\downarrow s})}\mathcal{F}(x_{\downarrow s}) + \epsilon}\mathcal{F}((x \otimes k_h)_{\downarrow s})\right) \tag{4}$$

where $\mathcal{F}$ denotes the Discrete Fourier Transform, $\mathcal{F}^{-1}$ is the inverse of $\mathcal{F}$, $\overline{\mathcal{F}(\cdot)}$ is the complex conjugate of $\mathcal{F}$, and ϵ is a small number to prevent the denominator from being zero. Note that the small values in k_l are zeroed out as in [2, 20] for better numerical stability.

3.3. Adaptive Arbitrary-scale SR Module

Once obtaining the degradation representation, our proposed framework learns to integrate the degradation representation into the arbitrary-scale super-resolution module in order to realize the arbitrary-scale blind-SR (as now the arbitrary-scale super-resolution module is adaptive to the estimated degradation information). The architecture of our arbitrary-scale super-resolution module is composed of the image feature extractor and the implicit neural representation (INR), where EDSR [19] is exploited for implementing the image feature extractor while the implementation of INR follows the LTE [17] fashion (where the input is composed of the continuous image coordinate, the cell size indicating the size/shape of the query pixel, and the extracted image features being projected into Fourier space, in which the output is the predicted RGB value at the given image coordinate) in our full model. And the implicit degradation representation provided by the predictor E is incorporated into the image feature extractor via a modulation mechanism (similar to the one used in [26]): the feature maps of the residual blocks in the image feature extractor are weighted in a channel-wise manner by the coefficients transformed from the degradation representation, where the transformation is done by two fully-connected layers and a sigmoid activation function. The training of such arbitrary-scale super-resolution module, which being adaptive to the implicit degradation representation, is driven by the L1 error between the groundtruth high-resolution image x and the output image $\widehat{x}$ composed of the predicted pixel values from our arbitrary-scale super-resolution module:

$$\mathcal{L}_{super} = \left|\widehat{x} - x\right|_1 . \tag{5}$$

3.4. Cycle-Consistency

As previously motivated in the introduction, with being inspired by several related works [9, 31] which have shown the benefit of modeling the dual paths between LR and HR (i.e. upsampling in terms of LR→HR and downsampling in terms of HR→LR) in a unified framework, here in our proposed method we also adopt such idea to better regularize the super-resolution output via enforcing its reverted/degraded version to be close to the original LR image y. In particular, as now our super-resolution module supports arbitrary upsampling scales, we can easily produce $\widehat{x_{\downarrow_s}}$ with letting the upsampling scale of our arbitrary-scale super-resolution module to be 1 (noting that the details of setting query image coordinate and the cell size as the input for LTE-based INR according to the desired upsampling scale are provided in the appendix A.3), which provides an estimation of $x_{\downarrow_s}$ (i.e. the HR image that is already downsampled to the LR space). Afterwards, with following the process described in Equation 2, we convolve $\widehat{x_{\downarrow_s}}$ with $\widehat{k_l}$ in which the result should be close to the original input LR image y thus leading to the cycle-consistency objective:

$$\mathcal{L}_{cycle} = \left| \widehat{x_{\downarrow_s}} \otimes \widehat{k_l} - y \right|_1 \tag{6}$$

The overall optimization loss for training our arbitrary-scale super-resolution module now becomes:

$$\mathcal{L}_{SR} = \mathcal{L}_{super} + \lambda \mathcal{L}_{cycle} \tag{7}$$

where we empirically set λ to 0.1 in all our experiments.

3.5. Training Procedure

The training of our entire framework follows a two-stage procedure, which comprises the degradation representation training stage and the arbitrary-scale super-resolution training stage. In the first stage, the implicit degradation predictor and explicit kernel estimator are jointly trained to improve the representative power of the degradation representations, where the optimization objective is the summation over the contrastive learning loss $\mathcal{L}_{cl}$ and the kernel estimation loss $\mathcal{L}_k$; While in the second stage, both the implicit degradation predictor and the explicit kernel estimator are fixed, and only the arbitrary-scale super-resolution module is optimized with $\mathcal{L}_{SR}$. During inference, we only require the implicit degradation predictor and the arbitrary-scale super-resolution module to achieve the task of arbitrary-scale blind-SR, in which our model supports upsampling any input image to arbitrary scales without the need of prior knowledge upon the degradation kernels. Please note that, as the explicit kernel estimator is not used during inference, the kernel mismatch problem is consequently prevented.

4. Experiments

Dataset. We adopt the DIV2K [1], which contains 800 high-resolution images, for training our proposed model as well as other baselines, in which the corresponding low-resolution images are synthesized according to Equation 1 with noise-free scenario. To be specific, in order to increase the variety of degradations seen in training dataset, we follow [26] to take the anisotropic Gaussian kernel as our degradation kernel where the filter size of the kernel is sampled from the odd numbers $\sim$ [7, 21], while the weights in a kernel are determined by two random eigenvalues λ_1, $\lambda_2 \sim \mathcal{U}(0.2, 4)$ in a covariance matrix and a random rotation angle $\theta \sim \mathcal{U}(0, \pi)$, in which $\mathcal{U}$ denotes the uniform distribution. Moreover, there are three downsampling operations being randomly selected to apply, i.e. bicubic, bilinear, and area interpolations, with the downscaling scale $s \sim \mathcal{U}(1, 4)$. The patch size of LR images (e.g. the contrastive objective for training our implicit predictor is based on image patches) is set to 48×48 for all the training settings. Five datasets are used for evaluation, i.e. Set5 [3], Set14 [32], BSD100 [23], Urban100 [11] and the DIV2K validation dataset, where each dataset is processed with unknown degradations as the same procedure as the training dataset generation.

Performance Evaluation. The evaluation metrics are PSNR and SSIM, where both PSNR and SSIM are larger the better. Both of them are evaluated under the Y channel of YCbCr space, following the setting in [20, 36].

4.1. Quantitative and Qualitative Results

We evaluate the performance of our model with four baseline arbitrary-scale super-resolution methods (including **MetaSR** [10], **LIIF** [6], **LTE** [17] and **SRNO** [27]) under unknown degradations. Please note that, as our proposed method aims to tackle the novel task of arbitrary-scale blind-SR, which is the first of its kind, we thus have to make comparison with either the ASSR methods or blind-SR methods. However, blind-SR methods can't generalize to various of scales. For fair comparison, we take the state-of-the-art blind-SR methods, DCLS [20] with combination to LIIF [6] and LTE [17] as the representative of blind-SR baselines. We also include a naive baseline which directly applies bicubic upsampling upon LR input to obtain the super-resolution results, denoted as **Bicubic**. To demonstrate the overall performances upon arbitrary upsampling scales, we show both integer scales and continuous scales in Table 1. Our method outperforms all the baselines on almost all benchmarks for both integer and continuous scales, with being the second-best on only a few metrics or datasets (i.e. in average our proposed method performs the best). These results clearly demonstrate the effectiveness of our method in handling unknown degradations under various scales, and highlight its advantages over existing methods.

From the qualitative comparison in Figure 3, we show the super-resolution images of all methods, under both integer and continuous scales. Our method outperforms all the others in terms of showing sharp edges, preserving structural patterns (e.g. streak), and having less distortion.

4.2. Analysis and Discussions

Study – using k_h instead of k_l for explicit kernel? Here we conduct an investigation on training our explicit kernel estimator to estimate the degradation kernel in HR space (cf. k_h in Equation 1) instead of the one in LR space (cf. k_l in Equation 2). In Figure 4, there are images being blurred with the same k_h, which we set (λ_1, λ_2, θ) = (1.1, 2.5, 65),

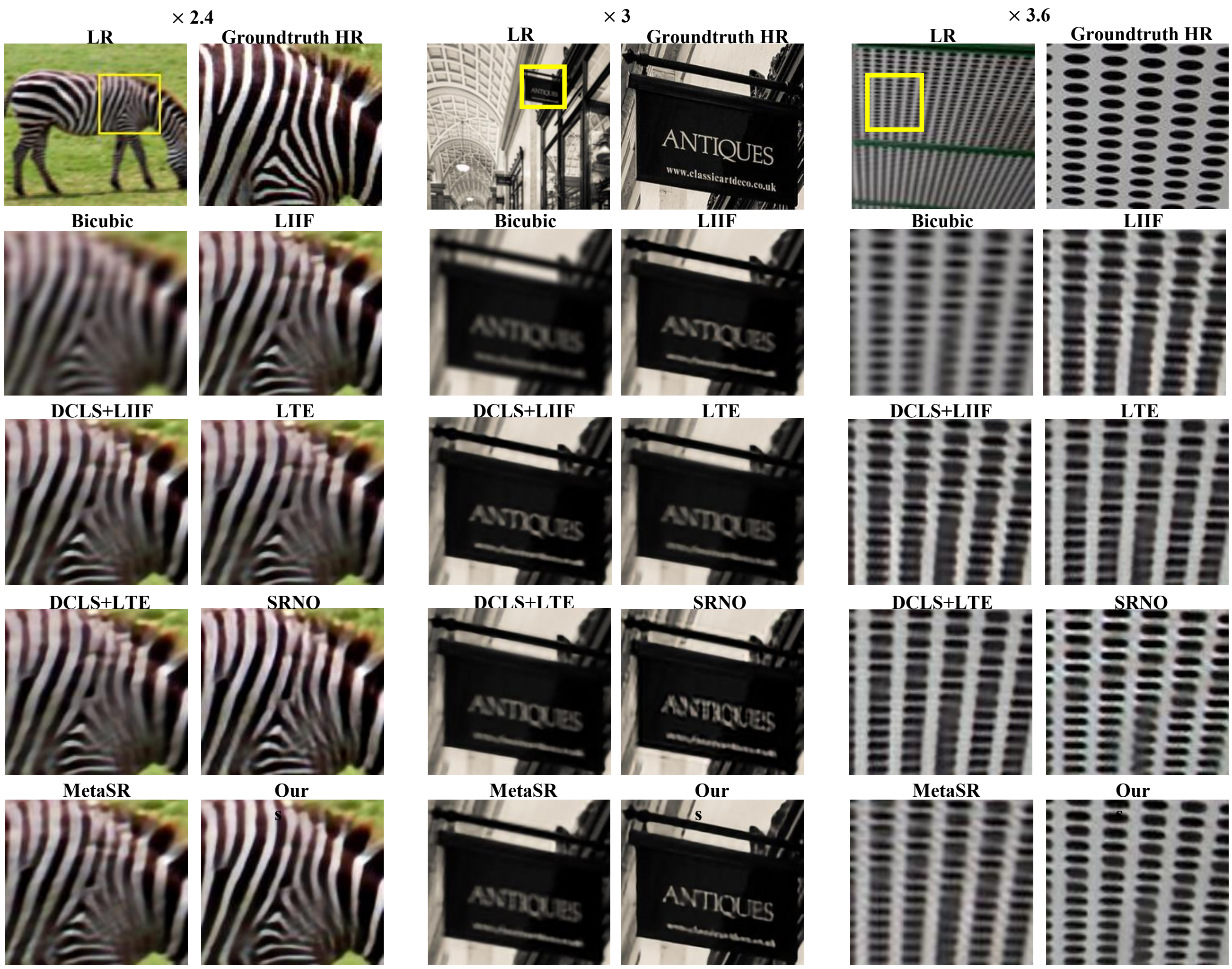

Figure 3. Qualitative comparison conducted on various upsampling scales. The image in the top row is from Set14, while the image in the second and the last rows are from Urban100. The magnified areas are indicated with yellow boxes.

but downsampled with different scales, which are 2.2, 1.1, 3.0, 4.2, 6, 10. The top row indicates groundtruth k_h while the bottom row shows the estimated k_h. We can observe that the estimated kernels are highly affected by the scales and deviate from the the groundtruth k_h, which conclude the infeasibility for the estimator to produce the same k_h under different scales.

Study – model designs. To verify the contribution of our model designs, we conduct ablation studies on two vital features: 1) arbitrary-scale super-resolution module with implicit degradation representation and 2) a closed-loop with the estimated explicit kernel in LR space (denoted as Implicit degradation and Cycle respectively in Table 2). Note that without implicit representation integrated to image feature extractor, it becomes the original EDSR [19]. Table 2 shows the results on DIV2K validation dataset with scale set to 3. It shows that both designs improve performance individually and perform the best when being combined.

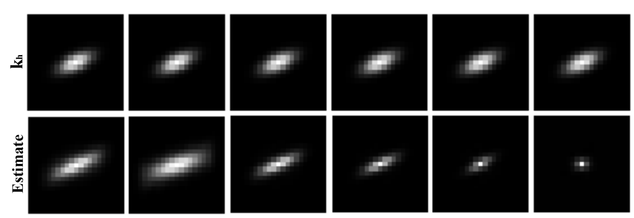

Figure 4. Experiments with the model variant of training explicit kernel estimator to predict k_h instead of k_l. The top row is the ground truth degradation kernels k_h, and the bottom row is the corresponding estimations produced by the estimator. It is shown that when the same blur kernel is used but with different downsampling scales [2.2, 1.1, 3.0, 4.2, 6, 10], the estimated kernels are inconsistent with the groundtruth.

Performance on real-world degradations. Here in Figure 5 we provide a qualitative comparison upon real-world images obtained from the RealSR version-3 [4] dataset to validate the robustness of our proposed method against real-

Table 1. Quantitative comparison on various datasets with various upsampling scales. The best performance is in red while the second best is in blue. All models are trained with continuous scales randomly sampled from $\mathcal{U}(1, 4)$.

Dataset		Set5	Set14	BSD100	Urban100
Scale	Method	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM
×3	Bicubic	27.6694 / 0.7655	24.8790 / 0.6265	25.4217 / 0.6134	22.3680 / 0.6059
	DCLS [20]+LIIF [6]	27.4310 / 0.8231	24.2159 / 0.7032	24.7017 / 0.6798	22.1304 / 0.6924
	DCLS [20]+LTE [17]	27.4674 / 0.8222	24.3413 / 0.7056	24.8001 / 0.6812	22.2755 / 0.6948
	MetaSR [10]	29.1847 / 0.8528	25.9156 / 0.7272	26.5609 / 0.7065	23.3855 / 0.6980
	LIIF [6]	29.3787 / 0.8630	25.9210 / 0.7325	26.6672 / 0.7140	23.3933 / 0.7049
	LTE [17]	29.4247 / 0.8644	25.8921 / 0.7327	26.6800 / 0.7145	23.3823 / 0.7057
	SRNO [27]	29.2663 / 0.8635	25.8981 / 0.7391	26.5954 / 0.7196	23.2896 / 0.7078
	Ours	29.5681 / 0.8696	25.9791 / 0.7423	26.8306 / 0.7242	23.5128 / 0.7131
×3.6	Bicubic	27.0250 / 0.7263	24.7365 / 0.6700	24.7794 / 0.5804	21.7952 / 0.5680
	DCLS [20]+LIIF [6]	27.0487 / 0.8141	24.6793 / 0.7101	23.8700 / 0.6448	21.2705 / 0.6439
	DCLS [20]+LTE [17]	26.9578 / 0.8129	24.6589 / 0.7066	23.8323 / 0.6425	21.3129 / 0.6437
	MetaSR [10]	25.8354 / 0.7130	28.6290 / 0.8266	25.5168 / 0.6636	22.4664 / 0.6490
	LIIF [6]	28.7843 / 0.8395	25.9306 / 0.7212	25.5267 / 0.6702	22.3859 / 0.6549
	LTE [17]	28.8638 / 0.8435	25.9306 / 0.7219	25.5356 / 0.6707	22.3626 / 0.6551
	SRNO [27]	28.4700 / 0.8371	25.8687 / 0.7245	25.4639 / 0.6751	22.3067 / 0.6564
	Ours	29.1043 / 0.8475	25.9760 / 0.7275	25.5867 / 0.6763	22.4304 / 0.6614
×4	Bicubic	26.7935 / 0.7173	25.0350 / 0.6174	24.9124 / 0.5868	22.0827 / 0.5763
	DCLS [20]+LIIF [6]	26.4664 / 0.7894	24.3381 / 0.6823	23.9654 / 0.6458	21.9636 / 0.6673
	DCLS [20]+LTE [17]	26.3301 / 0.7857	24.2402 / 0.6789	23.8380 / 0.6406	21.9944 / 0.6675
	MetaSR [10]	28.5168 / 0.8147	26.0609 / 0.7059	25.7275 / 0.6666	22.9445 / 0.6577
	LIIF [6]	29.0035 / 0.8346	26.2131 / 0.7143	25.7439 / 0.6732	23.1839 / 0.6703
	LTE [17]	29.2171 / 0.8402	26.3126 / 0.7173	25.7677 0/ .6741	23.0556 / 0.6765
	SRNO [27]	28.8063 / 0.8329	26.1791 / 0.7193	25.6898 / 0.6774	23.0131 / 0.6720
	Ours	29.3732 / 0.8440	26.4834 / 0.7243	25.8560 / 0.6811	23.2178 / 0.6804

Table 2. Ablation study based on DIV2K for our designs.

Implicit degradation	Cycle	PSNR / SSIM
✓	✓	29.8102 / 0.8312
-	✓	29.6253 / 0.8293
✓	-	29.6870 / 0.8279
-	-	29.5949 / 0.8246

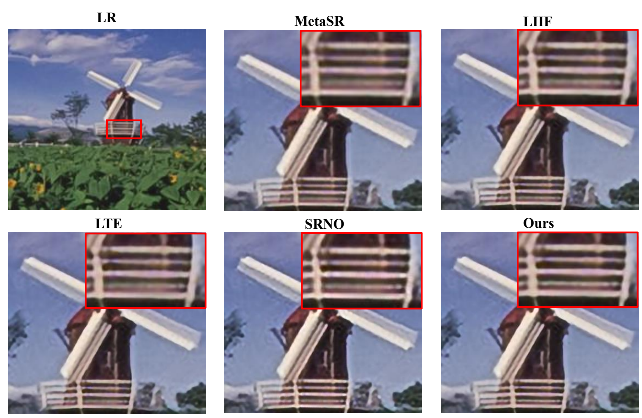

Figure 5. Qualitative results upon real-world degradations, where the images are obtained from RealSR version-3 [4] dataset and the upscaling factor is ×4.

world degradations (more are provided in the appendix). Compared to the other ASSR baselines, our method clearly generates the fences, which demonstrates that our proposed scheme is better suited for capturing the details.

5. Conclusion

We introduce a holistic framework that addresses the arbitrary-scale blind-SR problem. We propose to take advantage from both implicit and explicit degradations representations as well as optimize the overall framework with introducing the cycle formed by both upsampling and downsampling processes. In particular, the implicit degradation is adaptively integrated with the arbitrary-scale super-resolution module while the explicit degradation kernel is convolved with the super-resolution results to regularize the model optimization. With comparable or even better quantitative and qualitative performance at arbitrary-scale blind-SR compared to several baselines, we show the effectiveness of our method in terms of both supporting arbitrary upsampling scales and handling unknown degradations.

Acknowledgement. This work is mainly funded by MediaTek, National Science and Technology Council 111-2628-E-A49-018-MY4 and 112-2221-E-A49-087-MY3, Higher Education Sprout Project of the National Yang Ming Chiao Tung University, as well as Ministry of Education, Taiwan.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 6

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3, 5

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

[4] Jianrui Cai, Shuhang Gu, Radu Timofte, and Lei Zhang. Ntire 2019 challenge on real image super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 7, 8

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4, 5

[6] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 8

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[8] Mohammad Emad, Maurice Peemen, and Henk Corporaal. Dualsr: Zero-shot dual learning for real-world super-resolution. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3

[9] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 5

[10] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 6, 8

[11] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[13] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[14] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021. 3

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[17] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 6, 8

[18] Fengjun Li, Xin Feng, Fanglin Chen, Guangming Lu, and Wenjie Pei. Learning generalizable latent representations for novel degradations in super-resolution. In *ACM Conference on Multimedia (MM)*, 2022. 2, 3

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1, 5, 7

[20] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 6, 8

[21] Cheng Ma, Peiqi Yu, Jiwen Lu, and Jie Zhou. Recovering realistic details for magnification-arbitrary image super-resolution. In *IEEE Transactions on Image Processing (TIP)*, 2022. 3

[22] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[23] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2001. 6

[24] Sheng Shen, Huanjing Yue, Jingyu Yang, and Kun Li. Itsrn++: Stronger and better implicit transformer network for continuous screen content image super-resolution. In *ArXiv:2210.08812*, 2022. 3

[25] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[26] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 5, 6

[27] Min Wei and Xuesong Zhang. Super-resolution neural operator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 6, 8

[28] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[29] Mehmet Yamac, Baran Ataman, and Aakif Nawaz. Kernelnet: A blind super-resolution kernel estimation network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[30] Jingyu Yang, Sheng Shen, Huanjing Yue, and Kun Li. Implicit transformer network for screen content image continuous super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[31] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 3, 5

[32] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, 2012. 6

[33] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[35] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[36] Yifeng Zhou, Chuming Lin, Donghao Luo, Yong Liu, Ying Tai, Chengjie Wang, and Mingang Chen. Joint learning content and degradation aware feature for blind super-resolution. In *ACM Conference on Multimedia (MM)*, 2022. 2, 3, 4, 6

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3