Evaluating the Generalization Ability of Super-Resolution Networks

Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, Senior Member, IEEE, and Chao Dong

Abstract—Performance and generalization ability are two important aspects to evaluate the deep learning models. However, research on the generalization ability of Super-Resolution (SR) networks is currently absent. Assessing the generalization ability of deep models not only helps us to understand their intrinsic mechanisms, but also allows us to quantitatively measure their applicability boundaries, which is important for unrestricted real-world applications. To this end, we make the first attempt to propose a Generalization Assessment Index for SR networks, namely SRGA. SRGA exploits the statistical characteristics of the internal features of deep networks to measure the generalization ability. Specially, it is a non-parametric and non-learning metric. To better validate our method, we collect a patch-based image evaluation set (PIES) that includes both synthetic and real-world images, covering a wide range of degradations. With SRGA and PIES dataset, we benchmark existing SR models on the generalization ability. This work provides insights and tools for future research on model generalization in low-level vision.

Index Terms—Model generalization ability, super-resolution networks.

I. INTRODUCTION

Deep learning has achieved great success in constrained environment, and we have steadily moved our attention to its generalization ability. Generalization determines whether an algorithm can work well on unseen data. However, due to the data-driven nature, deep learning can easily overfit the training data, leading to unpredictable generalization behavior. In this article, we address the problem of generalization in the context of image super-resolution (SR) and restoration, which are classic low-level vision problems. Conventional SR models are typically trained under known degradation types and downsampling kernels, limiting their performance severely on real-world images [1], [2]. Improving the generalization ability of SR models is crucial for developing future methods. Notably, in this work, we extend the concept of the SR network beyond the classical bicubic super-resolution task to include simultaneous restoration. This broader scope allows us to investigate the generalization capability of models that perform both restoration and SR tasks, which are more relevant to real-world scenarios.

The problem of super-resolving images with unknown degradations is also called blind SR or real SR. According to a recent blind SR survey [10], existing methods can be categorized into four classes, including degradation specific methods [3], [4], kernel estimation methods [2], [5], unsupervised methods [6], [7], and internal statistical methods [8], [9]. Remarkably, recent works based on synthetic data have made significant progress. For example, BSRGAN [4] and Real-ESRGAN [3] show that when we can synthesize abundant degradations, the model can be applied in a wide range of real-world scenarios. This phenomenon is also observed in blind face restoration task, like GFP-GAN [11] and GPEN [12]. They all show impressive results on some real-world images. Recently, Dropout [13], [14] is also introduced into SR networks to improve the generalization performance.

However, here comes the problem: how to compare their generalization performance? We may also wonder: what are the failure cases? Whether extending the degradation types is a correct direction? All these require us to objectively evaluate the generalization ability. Existing works can only show some visual examples, but do not provide any feasible evaluation strategies. Without a standard evaluation metric and dataset, we cannot fairly compare different models, restricting the progress of their development.

Nevertheless, the evaluation of generalization is by no means easy! There is no specific assessment for generalization ability. In high-level vision tasks, like classification, they usually use the prediction accuracy on an unseen dataset or category as the generalization measure [15], [16], [17], [18], [19]. But in low-level vision tasks, like image restoration, there are no appropriate strategies. Can we use image quality assessment (IQA), such as PSNR and NIQE [20], to take the place of generalization assessment (GA)? The answer is NO, and there are
three main reasons. First, IQA is designed to evaluate the image quality, yet image quality is not equal to generalization ability. For example, traditional interpolation or filtering methods get lower IQA values than deep models in most cases, but they have a stable performance (good generalization ability) on all kinds of data [21]. Generalization should be a relative notion that is correlated with the method itself, not only the output. Second, IQA is highly sensitive to image content, thus will have different absolute values on different images. While in image restoration, we need to give a stable measurement on unseen degradations but not specific datasets. Third, IQA itself is not perfect. The reference-based IQA, like PSNR and SSIM, cannot be used in real-world images without ground truth, while the non-reference IQA, like NIQE and PI [22], cannot accurately evaluate the image quality. Gu et.al [23] have proved that existing IQA methods all have low correlation rates with human subjective scores on PIPAL dataset [24]. The above three issues have stopped IQA from being a qualified GA. More detailed explanations and experiments about IQA can be found in Section VI-C.

Additionally, we should note that GA is not proposed to take the place of IQA. They are two evaluation aspects, and both have great values. In general, we can first adopt IQAs to evaluate the model performance. If a model has much inferior performance than others, it is of little significance to extraly evaluate its generalization. For models with similar IQA performance, we can exploit GA to evaluate their generalization ability. This helps us to comprehensively appraise the models in a multi-dimensional way. Hence, IQA and GA are different but complementary with each other. They each describe a different aspect of the model. Nevertheless, the reasearch on model generalization in low-level vision is rarely discussed.

To fill in the gap, we make the first attempt to propose a complete evaluation protocol, including a GA index, a series of test datasets, and a benchmark, which could provide a comprehensive evaluation of the generalization ability. Before presenting our method, we clarify the basic definitions of generalization in SR task, and provide some general principles for the new index design. Afterwards, we introduce the first Generalization Assessment index for SR task – SRGA. SRGA is based on the statistical characteristics of internal features of the model, not output images. It is calculated on the test dataset, but is not sensitive to the data selection. More interestingly, it is a non-parametric and non-learning metric, which is guaranteed to have a good generalization ability itself. Moreover, the proposed SRGA does not require paired ground-truth images. Therefore, it can be used to evaluate the model generalization ability on real-world data, not just on synthetic data. To better validate our method, we collect a patch-based image evaluation set (PIES) that includes both synthetic data and real-world images, covering a wide range of degradations. On this basis, we can benchmark existing SR models on generalization ability. The benchmarking results are mostly consistent with our common sense. For instance, BSRGAN and Real-ESRGAN have a good generalization performance on most datasets, and are superior to other blind SR methods. We also have some surprising discoveries, like SwinIR [25] generalizes better on heavier noisy degradation, which could provide further insights on these methods. We hope our SRGA and datasets can help promote the development of blind SR methods, as well as other low-level vision problems.

II. PRELIMINARIES

Designing a generalization assessment (GA) method is a completely new task for low-level vision, thus we need to reach an agreement on some basic definitions and general principles. We achieve this goal by answering common questions, which could lay the foundation of the follow-up methodology. Actually, if you find this section tedious, it is also OK to read the Experiments Section first. There you may raise some questions, and you can find the answers here.

1) How to Define the Generalization Ability in SR Task? As a commonly-used definition, generalization is the ability of your model, after being trained to digest new data and make accurate predictions [26, 27]. There are three key components – model, data and prediction, which all have specific meanings in SR task. 1) Model in this article refers to the SR network. Generalization should be a property of the network, NOT output images. 2) Data refers to the input images with specific degradations. Generalization should be evaluated on degradations, NOT individual images. 3) Prediction refers to the processing effect of the output images. Generalization should measure the consistency of the processing effects across different input degradations, NOT absolute quality evaluation values. For example, an appropriate description would be: if an SR-net trained with data-A (well-performing domain) can achieve similar processing performance on data-B, then SR-net generalizes well on data-B.

2) How to Measure the Generalization Ability in SR Task? Generally, we use the generalization error to measure the generalization ability [26, 27]. The generalization error is easy to define in high-level vision tasks, which can be calculated by the distance between ground-truth (GT) labels and the model predictions. However, in low-level vision, there are no GT images in most cases, and one input image can correspond to multiple GT images, due to the ill-posed nature. It is hard to directly measure the restoration accuracy. It is acknowledged that deep model could achieve the best performance on in-distribution data (same distribution as the training data), but can hardly generalize to out-distribution data (different distribution from the training data). Thus, it is reasonable to define the generalization error in SR as the performance gap between the in-distribution and out-distribution data. As shown in Fig. 1, the in-/out-distribution inputs are distributed in two separate image spaces. After processing, they are expected to lie in the same space, indicating a better generalization ability. The gap between these two output spaces is the generalization error, and GA index is exactly the measurement of such a generalization error.

3) What are the General Principles for GA Design? From the above descriptions, we can summarize four unique characteristics of GA and introduce the corresponding design principles. 1) GA index should be correlated with both the model and the test data. Thus, it is better to take advantage of the model, not only the output image. 2) GA index should be a relative measure that quantifies the distance between the reference dataset (e.g., in-distribution data) and the test dataset (e.g., out-of-distribution data).
data). Therefore, it is preferred to select an appropriate reference dataset beforehand. However, in certain situations, identifying a suitable in-distribution dataset can be a daunting task or even impossible. In such cases, utilizing a pairwise metric between all test data may be more suitable for constructing the GA index. This approach offers a flexible and robust way of evaluating the generalization relationship across multiple datasets. 3) GA index in low-level vision should be sensitive to degradation, but insensitive to image content. Thus, it is important to disentangle image degradation from content. Thus, it is preferable to devise a non-learning (not rely on training data) and non-parametric (not rely on human settings) metric.

The rest of the paper is organized as follows. In Section III, we give the formulation of the classic image degradation model, and discuss the difference between evaluating model performance and generalization ability. The proposed SRGA index is presented in Section IV. Then, in Section V, we describe the collected PIES dataset. In Section VI, we demonstrate the effectiveness of the proposed SRGA and measure the generalization ability of several representative SR models.

III. FORMULATION

Model Performance: Given a trained SR model $G$ and a set of test input images $S_D = \{I_n\}_{n=1}^N$ with degradation $D$, the predicted SR results are obtained by: $I_{SR}^n = G(I_n)$. To evaluate the performance of model $G$, we can quantify the distance between the predicted output and the ground truth (GT) image $I_{HR}^n$

$$\text{Perf}(G, S_D) = \sum_{n=1}^N \text{Dist}(I_{SR}^n, I_{HR}^n),$$

where $\text{Dist}(\cdot, \cdot)$ is a distance or similarity function, such as $L_2$ error, PSNR, SSIM [28], LPIPS [29] or other image quality evaluation metrics. The (1) describes the average performance of model $G$ on the test set $S_D$. Note that the model performance is actually affected by image content and degradation simultaneously. That is, datasets with the same degradation but different image content will result in different performance scores.

Generalization Ability: Unlike model performance, generalization ability should characterize the consistency of the model’s processing effects across different types of input data, rather than absolute performance values. A model with good generalization ability should have similar processing effects for different types of inputs.

Formally, given two different input sets $S_{D_1}$ and $S_{D_2}$, generalization measures the difference between the processing effects of $G$ on $S_{D_1}$ and $S_{D_2}$

$$\text{Gen}(G, S_{D_1}, S_{D_2}) = \text{Dist}(\mathcal{F}(G, S_{D_1}), \mathcal{F}(G, S_{D_2})), \quad (2)$$

where $\mathcal{F}(G, S_D)$ is a function that represents the processing effect of model $G$ on dataset $S_D$. It should concentrate more on the image degradation $D$ instead of image content. In this article, we explore the intrinsic statistics of deep features of SR networks, and propose a statistical method for evaluating the generalization ability.

Blind Super-Resolution: Blind super-resolution aims at recovering and super-resolving an input low-resolution (LR) image with unknown degradation to a high-resolution (HR) version. Formally, a basic image degradation model is formulated as follows:

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

where $I_{HR}$ is the HR image, $I_{LR}$ is the degraded LR image, $\otimes$ denotes the convolution operation. There are mainly three types of degradation in this model, i.e., blur kernel $k$, additive noise $n$ and the downsampling operation $\downarrow_s$. For blind SR setting, the degradation information is unavailable during the training process. Generally, most current methods follow this degradation model or its variants [2], [3], [5]. In this article, we also follow the basic degradation formulation as (3) to make the PIES dataset and validate the proposed generalization index.

IV. STATISTICAL MODELING OF DEEP SR NETWORKS

A. Overview of the Proposed SRGA

The overview of the proposed generalization ability metric SRGA is depicted in Fig. 2. SRGA is built upon the statistical modeling of deep features of SR networks. It is a relative measurement that computes the distance of the feature distributions between the reference dataset and the candidate test dataset. The reference dataset used in SRGA is typically the one on which the model performs well, and is usually within the training distribution. We first obtain the corresponding deep features of the input datasets. These features are then compressed using principal component analysis (PCA), and the resulting projected feature sets are modeled using a generalized Gaussian distribution. Finally, the generalization error is measured by the Kullback-Leibler divergence (KLD) between the two probability distributions, leading to the proposed generalization index SRGA.

When a model performs well on a specific dataset, we can use it as a reference and compare the distance between other candidate datasets and the reference dataset using SRGA. However, if an appropriate reference dataset cannot be identified, we can treat all candidate datasets equally and calculate the SRGA.
score between each pair to obtain a generalization matrix. By using SRGA in this way, we can assess a model’s generalization performance beyond its training set and make informed decisions about its suitability for real-world applications. In the following, we will elaborate on the details of the adopted statistical methodology.

B. Revisit Natural Image Statistics

For decades, lots of efforts have been made to explore the statistics of natural images [30], [31], [32], [33]. For example, Mallat [34] discovered that coefficients of multi-scale and orthonormal wavelet decompositions of images could be described by the generalized Gaussian model. Moulin and Liu [35] analyzed the multi-resolution (wavelet) image denoising schemes using generalized Gaussian and complexity priors. Among these works, Mittal et al. [36] discovered that the image mean subtracted contrast normalized (MSCN) coefficients in the spatial domain have characteristic statistical properties that are strongly correlated to the image distortion. They adopted the generalized Gaussian distribution to model the MSCN coefficients, and successfully proposed the widely-used no-reference image quality assessment (NR-IQA) BRISQUE [36] and NIQE [20]. A recent work [21] further discovered the deep degradation representations (DDR) hidden in the SR networks: the deep features of SR networks are spontaneously discriminative to image degradations rather than image contents.

Inspired by these works, we present a new perspective for studying the generalization ability of deep models. Specifically, we explore the statistics of deep features of SR networks and build a statistical model to quantitatively measure the generalization ability.

C. Statistics of Deep Features of SR Networks

Given a set of $N$ input images $\{I_n\}_{n=1}^N$ with degradation $D$, for arbitrary super-resolution model $G$, we can obtain its corresponding deep features $\{F_n\} = \{G_F(I_n)\}$, where $F_n \in \mathbb{R}^{H \times W \times C}$, $H$, $W$ and $C$ are the height, width and depth, respectively. $G_F$ denotes the model containing all the layers before the last output layer, i.e., $F_n$ is the extracted deepest feature map before the output layer.

First, we apply PCA [37] to reduce the dimension of $\text{vec}(F_n)$ from $HWC$ to $D$, to alleviate the calculation cost and compress information. Specifically, $Y = [\text{vec}(F_1); \text{vec}(F_2); \ldots; \text{vec}(F_N)] \in \mathbb{R}^{N \times HWC}$. The PCA method will find a projection matrix $P \in \mathbb{R}^{HWC \times D}$, resulting in $X = YP$, $X \in \mathbb{R}^{N \times D}$. Denote set($X$) = $\{X[j,k] | j \in \{1, 2, \ldots, N\}, k \in \{1, 2, \ldots, D\}\}$. The cardinal of set($X$) is $N \times D$. Our finding is that the elements $x \in \text{set}(X)$ have the statistical properties that are determined by the model $G$ and the degradation type $D$. Formally, $x$ obeys probability distribution $p_{G,D}: x \sim p_{G,D}(x)$.

Notably, for all the SR models, we extract the output features of the penultimate layer to calculate the SRGA index. A simplified illustration is depicted in Fig. 4. This deepest layer contains all the processed information to produce the final results. There is no more processing layer afterwards. Thus, it is reasonable to use this layer to compute SRGA. To make an analogy, IQA adopts the three-channel output images to measure the model performance (image quality), while we utilize the penultimate multi-channel output feature maps to evaluate the generalization ability of deep models. Another reason is that different models

---

1Flatten the spatial feature map into a one-dimensional vector.

2In our experiments, $D$ is set to 300. Other reasonable values (> 300) also work fine.
have different architectures, it is convenient and universal to adopt the last layer.

In order to visualize how the distributions \( p_{G,D} \) vary as a function of the model \( G \) and input degradation \( D \), Fig. 3 plots the histograms of \( x \in \text{set}(X) \) with different \( G \) and \( D \) on PIEs dataset (described in Section V). From the statistical results, we can draw two important observations: 1) For a given model \( G \), different input degradation types \( D_1 \) and \( D_2 \) will lead to different probability distributions \( p_{G,D_1} \) and \( p_{G,D_2} \). By quantifying the distance between these two resulting probability distributions, it is possible to measure the difference in the processing effect on different input distributions. 2) For degradation type \( D \), different model \( G \) will produce different distributions, implying that each model has created its own feature manifold, which determines the generalization ability.

D. Generalized Gaussian Distribution

To model and analyze the statistics of the deep features of SR networks, we explore to fit the empirical distributions \( p_{G,D} \) with an appropriate statistical distribution. The Generalized Gaussian Distribution (GGD) is widely used for many relevant tasks, since its form is general and covers many common distributions like Gaussian distributions, Laplace distributions and uniform distributions. More importantly, extensive practice has demonstrated that GGD can fit natural image statistics, and the deep features extracted by SR networks inherit such properties. We discover that the statistical properties of deep features are changed by the presence of distortion. Thus, quantifying these changes can make it possible to measure the processing effect of the model on different input degradations. The GGD with zero mean is formulated as

\[
GGD(x; \alpha, \sigma^2) = \frac{\alpha}{2 \beta \Gamma(1/\alpha)} \exp \left( -\frac{|x|}{\beta} \right)^\alpha , \tag{4}
\]

where

\[
\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}, \tag{5}
\]

and \( \Gamma(.) \) is the gamma function

\[
\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \quad (z > 0). \tag{6}
\]

There are two parameters \( \alpha \) and \( \sigma \) in GGD. The parameter \( \alpha \) controls the “shape” of the distribution. \( \alpha = 1 \) and \( \alpha = 2 \) yield the Laplacian and the Gaussian density function, respectively.

Smaller values of the shape parameter lead to more peaked distributions. Another parameter \( \sigma \) controls the variance. Intuitively, the natural image contents are presented by the distribution, while the degradation changes the distribution parameters.

We try to match the empirical histograms of the feature sample values \( x \in X \) with the best possible GGD pdf. Specifically, we adopt the moment-matching based method proposed in [38] to estimate the parameters \( \alpha \) and \( \sigma \) of GGD. Table I summarizes the statistics of several representative models with different input degraded images. The complete results are in the supplementary materials, available online.

E. Measuring Model Generalization Ability

In the previous sections, we have revealed that different input degradation types will lead to different feature distributions. Ideally, if model \( G \) is robust enough and has perfect generalization ability, the outputs of different inputs should be of the same distribution or very close to each other. Hence, by quantifying the distance between these two probability distributions, we can measure the difference of processing effects between different input distributions, which reflects the model generalization performance.

In practice, given two sets of input images with different degradation types \( \{I_1\}_{D_1} \) and \( \{I_2\}_{D_2} \), we first obtain their corresponding deep feature sets \( \text{set}(X_{D_1}) \) and \( \text{set}(X_{D_2}) \). Then, we adopt GGD to fit the datapoints to model the feature statistics: \( p_{G,D_1}(x) = GGD(x; \alpha_1, \sigma_1^2) \), \( p_{G,D_2}(x) = GGD(x; \alpha_2, \sigma_2^2) \). Once the distribution parameters have been estimated with the deep features, we can capture the distribution change by calculating the Kullback-Leibler divergence (KLD) between distributions. Fortunately, the KLD of two zero-mean Generalized Gaussian Distribution \( GGD(x; \alpha_1, \sigma_1^2) \) and \( GGD(x; \alpha_2, \sigma_2^2) \) has the closed-form solution [40]

\[
D_{KL} = \ln \frac{\alpha_1 \sigma_2 \Gamma(1/\alpha_2) \sqrt{\Gamma(3/\alpha_1) \Gamma(3/\alpha_2)}}{\alpha_2 \sigma_1 \Gamma(1/\alpha_1) \sqrt{\Gamma(3/\alpha_1) \Gamma(3/\alpha_2)}} + \left( \frac{\sigma_1 \sqrt{\Gamma(1/\alpha_1) \Gamma(3/\alpha_2)}}{\sigma_2 \sqrt{\Gamma(1/\alpha_2) \Gamma(3/\alpha_1)}} \right)^{\alpha_2} \Gamma(\alpha_2/\alpha_1 + 1/\alpha_1) \Gamma(1/\alpha_1). \tag{7}
\]

Now, we have an analytical approach to quantitatively measure the difference in the output feature distributions caused by different input images. Specifically, to evaluate the generalization ability of model \( G \), a reference dataset \( S_{ref} \) could be selected (i.e., known in-distribution data that the model can perform well), then we compute the feature distribution distance
(FDD) between the reference dataset \( S_{D_{\text{ref}}} \) and the test dataset \( S_{D_{\text{test}}} \)

\[
FDD = D_{kl} \left( P (G, S_{D_{\text{ref}}}) , P (G, S_{D_{\text{test}}}) \right),
\]

\[
SRGA = \log_{10} \left( FDD + 10^{-\delta} \right) + \delta,
\]

where \( P(G, S_f) \) represents the generalized Gaussian distribution fitted by the deep features of model \( G \) with the input dataset \( S_f \). Smaller FDD means that \( G \) has similar processing effects on these two different input degradations. \( \delta \) is introduced to avoid zero point in logarithmic function and the min value is shifted to 0 by adding \( \delta \) (we set \( \delta = 5 \)). The proposed SRGA is a plausible representation of \( \text{Dist} \) and \( \mathcal{F} \) in (2).

In addition, if there are \( N \) test datasets \( \{ S_{D_{\text{test}}}^i \} \), we can calculate the mean value of the SRGA

\[
mSRGA = \frac{1}{N} \sum_{i} SRGA \left( P \left( G, S_{D_{\text{ref}}} \right) , P \left( G, S_{D_{\text{test}}}^i \right) \right).
\]

Notably, mSRGA describes the averaged generalization ability across multiple degradation datasets. It is actually averaging different degradations not the contents of the datasets. For the same degradation but different contents, SRGA is not sensitive; for different degradations, mSRGA can give a mean value. It is worth noting that such a measurement method does not require any paired ground truth (GT) image or even the final output image. Further, even if the contents of the two sets of input images are completely different, we can still measure the difference between them. Because the probability distribution is derived from the statistics of the network deep features and is almost irrelevant to specific image content (see Section VI-C).

In summary, SRGA satisfies all four properties for GA design in Section II. SRGA is correlated with both the model itself and the test data. It is a relative indicator that measures the difference of processing effects. It is not sensitive to specific image content but the degradation. Also, SRGA is a statistical approach without relying on learning process.

**Generalization Matrix:** Identifying a reference dataset that accurately reflects a model’s performance can be a tricky task. Fortunately, the use of SRGA offers a solution to this challenge by enabling us to evaluate a model’s generalization ability without relying on a specific reference dataset. By calculating the SRGA score between candidate datasets, we can obtain an SRGA matrix that effectively illustrates the processing effects of the model on different datasets. This approach allows us to compare the generalization abilities of different datasets by examining their SRGA distances. Datasets with similar SRGA distances demonstrate similar generalization abilities, and the model is expected to perform consistently on such datasets. By leveraging the SRGA matrix, we can evaluate a model’s performance in a more comprehensive manner and make informed decisions when selecting the most suitable model for a specific task. Therefore, SRGA is a valuable tool for assessing a model’s generalization ability and can aid in selecting the best model for a given task.

### V. Patch-Based Image Evaluation Set

For SR task, classical test datasets include Set5 [41], Set14 [42], BSD100 [43], Urban100 [44] and DIV2K-vali100 [45]. Most methods evaluate their performance on these public datasets. However, these datasets are mainly designed for evaluating the absolute model performance. As illustrated in Fig. 13, these datasets have severe bias on image contents, which cannot well reflect the model generalization on degradations. Further, these datasets do not provide sufficient fine-grained and continuous degradation types, making it difficult to precisely evaluate the generalization performance. Moreover, test images with large resolution require tremendous computational cost and storage to estimate the feature distributions. Therefore, we propose a new fine-grained Patch-based Image Evaluation Set (PIES). It contains a variety of test images with different degradation types and degrees, including both synthetic and real-world degradations. PIES dataset has the following characteristics: i) Patch-based. Instead of evaluating a whole image with large resolution, we focus on image patches with relatively small resolution (128 × 128 for HR and 32 × 32 for LR, i.e. × 4 SR). The degradation type and degree in one patch can be considered homogeneous and spatially invariant, which can facilitate analysis. ii) Fine-grained degradation types. PIES dataset contains different types of common degradation and covers a wide range of degradation degrees. The descriptions for PIES are summarized in Table II. Each subset contains 800 patches. With PIES dataset, we can both evaluate the model performance and generalization ability on a unified platform.

Notably, researchers can also define their own test dataset and reference dataset according to the practical needs. The proposed SRGA index is not restricted to specific datasets and does not require paired GT images. Thus, it can be applied in real-world images.

### VI. Experiments

In this article, we select representative SR methods to benchmark their generalization ability with PIES dataset, including SRResNet [1], IKC [2], DAN [5], DASR [46], RealESRGAN [3], BSRGAN [4] and SwinIR [25]. We train SRResNet with different training data as baselines. For other methods, we directly adopt their released pre-trained models. IKC, DAN and DASR methods mainly focus on blur degradation, thus their released model cannot deal with noisy images. RealESRGAN, BSRGAN and SwinIR-GAN (GAN version of SwinIR) are trained with multiple synthetic complex degraded data. We conduct experiments on × 4 SR. Since the blind SR methods are supposed to perform well on clean input images, we select PIES-Clean dataset as the reference dataset, and then calculate the SRGA index between the reference and other test datasets. Note again that SRGA is not limited to synthetic datasets, but is applicable to real-world datasets. A smaller SRGA value suggests that the model can well generalize to the test dataset.

\[\text{Note that model like SRResNet (train: blur2) performs well on blur2 images, thus, the reference dataset should be selected as PIES-Blur2 instead of PIES-Clean.}\]
TABLE II

| Datasets     | Has Ref? | Syn/Real? | Description                                                                 |
|--------------|----------|-----------|-----------------------------------------------------------------------------|
| PIES-Clean   | ✓        | Syn       | Image patches collected from DIV2K-val100, BSDS100, Urban100 and General100 datasets. The corresponding LR patches are downsampling using matlabbicubic function. |
| PIES-Blur    | ✓        | Syn       | Additionally apply Gaussian blur on PIES-clean dataset. The blur kernel is sampled in (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0) (16 subsets in total). |
| PIES-AnisoBlur | ✓      | Syn       | Additionally apply anisotropic Gaussian blur on PIES-clean dataset. The kernel size is 21, the kernel width is uniformly sampled in [0.6, 5] and the rotation is uniformly sampled in [0, π]. |
| PIES-Noise   | ✓        | Syn       | Add Gaussian noise on PIES-clean dataset. The noise level is sampled in (5, 10, 15, 20, 25, 30, 35, 40, 45, 50) (10 subsets in total). |
| PIES-BlurNoise | ✓      | Syn       | Apply both Gaussian blur and Gaussian noise. The blur kernel width is sampled in (1, 2, 4, 6), and the noise level is sampled in (10, 20, 30) (12 subsets in total). |
| PIES-RealCam | ✓        | Real      | Image patches selected from RealSR dataset [39]. RealSR dataset consists of LR-HR image pairs obtained by adjusting the lens of two digital single-lens reflex (DSLR) cameras (Nikon D810 and Canon 5DS). |
| PIES-RealLQ  | ×        | Real      | Image patches collected from the Internet, containing real-world images of various distortion types and degrees. |

It includes both synthetic and real-world images with fine-grained degradation types.

---

**Fig. 5.** PSNR curves (a) and (d), SRGA curves (b) and (e) and NIQE curves (c) and (f) of different models on blur and noise degradations, respectively. SRGA successfully quantifies the model generalization ability. Further, it can reflect more precise information about the model generalization ability. The detailed quantitative values are listed in the supplementary file, available online.

---

In this section, we first conduct validation experiments to verify the correctness of the proposed SRGA index, and benchmark the generalization ability of existing models. Then, we compare SRGA with IQA as generalization ability measurements. Due to the space limit, the complete quantitative and qualitative results are included in the supplementary file, available online.

**A. Sanity Check for SRGA**

As there is no ground truth or human label for generalization ability, it is hard to evaluate the effectiveness of GA index directly. Nevertheless, we can construct some cases for sanity check. **Case 1:** The baseline model SRResNet trained only with clean LR data is supposed to have poor generalization ability to handle blur images. The PSNR value of this model reaches the highest on the clean input data, and then decreases rapidly as the blur degree increases (Fig. 5(a)). After involving blur LR data into training, SRResNet (train: blur0-4) is eligible to restore blur images: the PSNR value decreases slowly especially in the range of blur level [0, 3]. In Fig. 5(b), we observe that the SRGA curves have the same trend with the PSNR curves in such cases. For example, the SRGA curve of SRResNet (train: clean) is steadily increasing, while the SRGA curve of SRResNet (train: blur0-4) grows slowly and its values are all lower than that of SRResNet (train: clean). Especially in blur range [0, 3], the SRGA values of SRResNet (train: blur0-4) only change a little within a small range. The SRGA curves successfully depict the model generalization ability in a quantitative manner. Similar results are also observed in baseline model SRResNet (train: clean) and SRResNet (train: noise0-20) (see Fig. 5(d) and (e)). **Case 2:** In another extreme case, we train a SRResNet model with only blur2 data. The SRGA curve can also successfully reflect the
trend of generalization. Case 3: SRResNet (train: clean) and SRResNet (train: noise0-20) have similar poor generalization trend on blur degradation; but SRResNet (train: noise0-20) has better generalization than SRResNet (train: clean) on noise degradation (see Table III). The SRGA results are consistent with our common sense. In principal, SRGA is theoretically and experimentally sound. It provides us with a tool to quantitatively measure the generalization, which is more reliable than intuitive judgment.

B. Evaluating the Model Generalization Ability

Model generalization ability measures the difference of processing effects on different types of input data. In the following, we analyze the model generalization results on blur, noise and real-world degraded input data. Note that the generalization ability is not equal to the output image quality. The benchmarking results are summarized in Table III. The detailed quantitative results are described in the supplementary file, available online.

Blur Degradation: Fig. 5(b) summarizes the distances of the resulting feature distributions between clean LR input dataset and degraded input datasets with various isotropic Gaussian blur levels. Smaller value means the model can well generalize to the test dataset. Larger SRGA values imply the model processes the input degradation more differently than the chosen reference dataset. As can be seen, the distances of SRResNet (train: blur0-4) are continuously smaller than SRResNet (train: clean), which is intuitive and reasonable. The results can successfully describe the model performance change in a quantitative and analytical manner. Interestingly, for mild blur degradation (blur [0, 3]), DAN [5] has better model performance and generalization ability than SRResNet (train: blur0-2) and IKC. However, when the blur level becomes larger, its performance (PSNR, SSIM) and generalization (SRGA) deteriorate dramatically, indicating that DAN can only well generalized within mild degradation region. The proposed SRGA metric successfully describes such phenomenon. Another important observation is that methods like Real-ESRGAN [3], BSRGAN [4] and SwinIR [25] have relatively more stable and better generalization ability, especially for severe degraded data. Particularly, when the blur level is larger than 5, SRResNet, IKC [2], DAN [5] and DASR [46] all fail to reconstruct the sharp images, while Real-ESRGAN, BSRGAN and SwinIR-GAN can still generate realistic images with sharp textures (see Fig. 7). It shows the superiority of such newly-proposed methods. This suggests that models trained with more degradations are likely to have better generalization despite that the absolute performance could deteriorate. For anisotropic blur degradation, as shown in Table III, SRGA can also successfully embody the generalization ability of different models. Methods like BSRGAN, Real-ESRGAN and SwinIR still have good generalization on anisotropic blur degradation.

Pairwise SRGA Matrix for Blur Degradation: One of the key advantages of SRGA is its ability to eliminate the need for ground-truth images, making it a practical tool for real-world scenarios where ground-truth datasets are unavailable. This allows all candidate test datasets to be treated equally. Concretely, we can compute the SRGA score for each pair of datasets and obtain an SRGA matrix that provides a comprehensive view of a model’s generalization ability. The resulting SRGA matrix provides a valuable tool for comparing and selecting models across a range of datasets, facilitating the identification of models with the best overall performance. Fig. 6 presents the detailed SRGA matrices of different models under fine-grained blur levels. From Fig. 6(a), we can observe that, except for the diagonal elements, the SRGA values at other positions are considerably high, indicating that the model’s performance in handling different degraded data significantly varies. This observation is consistent with our intuition and analysis that models trained only on clean data struggle with degraded data, leading to low generalization ability. Comparing Fig. 6(a) and 6(b), when adding blur data into training, the SRGA scores in low-level blur regions are relatively low, indicating that the model has similar processing effects on these datasets. The SRGA matrix provides a useful tool for describing the generalization distribution over different degradation data.

| Methods             | PIES-Blur | PIES-AnisoBlur | PIES-Noise | PIES-BlurNoise | PIES-RealCam | PIES-RealLQ |
|---------------------|-----------|----------------|------------|----------------|--------------|-------------|
| SRResNet (train: clean) | 3.639 (13) | 3.615 (13) | 3.727 (9) | 3.454 (10) | 3.563 (11) | 3.825 (8) |
| SRResNet (train: blur0-4) | 2.967 (7)  | 2.880 (6)  | -          | 3.366 (9)   | 3.125 (6)  | 3.811 (6)  |
| SRResNet (train: noise0-20) | 3.623 (12) | 3.605 (12) | 2.325 (8) | 3.167 (8) | 3.547 (10) | 3.827 (9) |
| SRResNet (train: blur0-4&noise0-20) | 2.978 (8) | 2.820 (7) | 1.574 (2) | 2.901 (6) | 2.566 (1) | 3.736 (3) |
| IKC | 3.416 (10) | 3.375 (11) | -          | -             | 3.432 (9)  | 3.845 (10) |
| DAN | 3.534 (11) | 3.310 (10) | -          | -             | 3.937 (13) | 4.070 (12) |
| DASR | 3.248 (9) | 3.157 (9) | -          | -             | 3.770 (12) | 4.033 (11) |
| Real-ESRGAN | 2.480 (4)  | 2.500 (6)  | 1.791 (4) | 2.692 (3) | 3.301 (7)  | 3.823 (7)  |
| Real-ESRNet | 2.336 (3)  | 2.456 (5)  | 1.566 (1) | 2.900 (5) | 2.787 (3)  | 3.770 (4)  |
| BSRGAN | 2.592 (5)  | 2.397 (4)  | 1.946 (5) | 2.850 (4) | 2.872 (5)  | 3.796 (5)  |
| BSRNet | 2.598 (6)  | 2.339 (3)  | 2.254 (7) | 2.964 (7) | 2.686 (2)  | 4.345 (13) |
| SwinIR-GAN | 1.639 (1)  | 1.852 (2)  | 1.996 (6) | 2.435 (2) | 3.379 (8)  | 3.662 (2)  |
| SwinIR-PSNR | 1.668 (2)  | 1.727 (1)  | 1.685 (3) | 2.297 (1) | 2.826 (4)  | 3.655 (1)  |

The null value indicates that the method cannot handle this type of degradation. The reference dataset is excluded when calculating the average. PIES-Blur, PIES-Noise and PIES-BlurNoise each contain multiple subdatasets. Since models may perform inconsistently on different degrees of degraded data, the average value may not reflect the actual situation. We recommend to refer to the detailed results in the supplementary file, available online. (0 ~ 2: well generalized; 2 ~ 3: mediocre; > 3: poor).
Fig. 6. SRGA matrices on blur degradation of (a) SRResNet (train: clean), (b) SRResNet (train: blur0-4), (c) DAN, (d) DASR, (e) SwinIR-PSNR, (f) SwinIR-GAN.
Fig. 7. Visual results of different models on PIES-Blur dataset.
Fig. 8. Visual results of different models on PIES-Noise dataset. Note that PSNR and NIQE evaluate the individual image quality, while SRGA measures the model generalization (the consistency of processing effect) across different degraded datasets.

**Noise Degradation:** As shown in Fig. 5(e), as the input noise level increases, the feature distributions become more divergent from that of the input clean data. When the noise level exceeds the range that the model can handle (denoise), the SRGA values increase dramatically. When noise level is extremely severe (e.g., noise level is larger than 30), most methods are unable to produce visually pleasing images. The results of SRResNet contains many noise residues and artifacts. The results of Real-ESRGAN (Real-ESRNet) are over-smoothed. The results of BSRGAN (BSRNet) are unsatisfactory with severe color distortion and artifacts. It is noteworthy that SwinIR-GAN shows peculiar characteristics when the noise level is large. The distance of the feature distributions between the clean input and severe noise input becomes smaller. The visual results of SwinIR-GAN can give a reasonable explanation and can well embody the superiority of the proposed distribution-based GA metric (see Fig. 8). SwinIR-GAN can produce relatively realistic output images by generating sharpened textures. Although the generated images are not consistent with the ground truth image, they look realistic with few noise residue and artifacts. SwinIR-GAN seems to treat the severe noisy input as a kind of texture and generate more textured details according to the noise distribution. This demonstrates SwinIR’s powerful learning and generation capabilities. However, if we use the IQAs (especially reference-based IQA) to deduce the model generalization ability, such important phenomenon and observation cannot be perceived (see navy blur curves in Fig. 5(d) and (e)). More discussions are in Section VI-C. SRGA successfully helps us discover such an interesting fact. Specifically, the PSNR values continuously decrease when the noise level become larger, since it measures the pixel-wise distance between the SR results and the GT images. However, SRGA can help us find that the processing effect of SwinIR-GAN on severe noisy images is not continuously deteriorating. For severe noisy images, SwinIR-GAN can still produce visually plausible realistic results.

**Real-World Degradation:** Another advantage of the proposed SRGA is that it does not require any paired reference images as ground-truth (GT). Thus, it can be applied to evaluate the model generalization ability on real-world images. As shown in Table III, SwinIR [25], BSRGAN [4] and Real-ESRGAN [3] generalize well on PIES-RealCam dataset. PIES-RealCam dataset is collected from [39], which captures paired images by zooming the camera’s focal length. The LR images mainly contain modest blur and noise degradations. Hence, it is reasonable that SRResNet (train: blur0-4&noise0-20) has good generalization performance on this dataset as well. This suggests that models trained with synthetic blur and noise data are able to generalize to LR images simply caused by camera focal length. The visual results of PIES-RealCam dataset are shown in Fig. 9. For PIES-RealLQ dataset, the LR images are collected from the Internet, which contain various degradations including blur, noise, compression, physical damage, etc. The degree of degradation also

---

**Table III:**

| Noise Level | PSNR | NIQE | SRGA |
|-------------|------|------|------|
| Clean       | 24.55| 7.37 | 6.00 |
| Noce5       | 23.87| 8.40 | 6.00 |
| Noise10     | 21.88| 10.31| 5.99 |
| Noise15     | 19.89| 12.01| 5.37 |
| Noise20     | 18.88| 13.27| 3.30 |
| Noise25     | 17.88| 14.82| 1.11 |
| Noise30     | 16.88| 16.48| 1.23 |
| Noise40     | 15.88| 18.21| 1.40 |

---

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 9. Visual results of different models on PIES-RealCam dataset.

ranges from mild to severe. As can be seen, IKC [2], DAN [5] and DASR [46] have quite different processing effects on this dataset, compared with the chosen reference PIES-Clean dataset. This implies that although these models have good performance on clean LR inputs, they do not generalize well on the collected real-world images. On the contrary, Real-ESRGAN [3] and SwinIR [25] have a relatively good generalization ability on this dataset. This again suggests that exploiting abundant synthetic data into training is helpful for covering more realistic cases, especially for large models. The visual results of PIES-RealLQ dataset are shown in Fig. 10.

Note again, the generalization does not equal to model performance. It just measures the difference in the processing effect of the model on different inputs. To fully evaluate the merits of a model, we need to assess it comprehensively, including model performance, generalization ability, number of parameters, computational cost, etc. SRGA only gives a new perspective for evaluating the generalization ability.

C. Comparing SRGA With IQA

Notably, GA and IQA are conceptually different. They are two different but complementary facets of evaluating the model. We are not proposing GA to supplant IQA. Previously, however, since there is not any viable metric for evaluating the model generalization, people tend to utilize IQA to measure the performance and to deduce the generalization. Notwithstanding, there are several limitations of IQA for measuring model generalization ability. In the following, we illustrate the advantages of SRGA and limitations of IQA in five aspects.

SRGA versus PSNR in Sensitivity of Image Content: IQA is highly sensitive to image content, thus will have different values on different images. Different datasets will have different absolute IQA values, even though they share the same degradation. While in image restoration, we need to give a stable measurement on unseen degradations but not specific datasets. For different datasets containing the same degradation type, the generalization ability of the model should be the same. An illustration is depicted in Figs. 12 and 13. The proposed SRGA metric stems from the statistical characteristics of deep features of SR model: for a given model, different input degradation types will lead to different feature probability distributions, and such distributions are less sensitive to the image content but have higher response to image degradation, as illustrated in Section IV-C. Similar property is also observed in previous literatures [20], [21], [36], [47]. We further validate this property.

We select different numbers of image patches from PIES-Blur dataset to fit the corresponding distribution parameters (σ and α) using SRRResNet(train:clean) model. Thus, the contents of the selected image sets with different numbers of patches are inconsistent. The patches are randomly selected and each setting is repeated three times. As shown in Fig. 11, with the increase of patch number, the estimated parameters gradually converge. Moreover, we select different datasets at different blur levels, so
that the contents of the images are not aligned. For example, the data for clean, blur1, and blur2 are from the PIES dataset, and the data for blur0.5, blur1.5, and blur2.5 are from the REDS [48] dataset. Fig. 12 plots the corresponding PSNR and SRGA curves. The PSNR values are largely affected by the image content, leading to a fluctuating PSNR curve (not monotonic). But the SRGA curve can still depict the generalization trend of the model, despite of the fact that we use different datasets for calculation.⁴

⁴In this experiment, PIES-Clean is selected as the reference dataset.

**Fig. 10.** Visual results of different models on PIES-RealLQ dataset.

**Fig. 11.** As the number of patches increases, the estimated distribution parameters gradually converge.

**Fig. 12.** Limitation of IQAs for measuring generalization ability. If the image contents are not aligned across degradations, PSNR values will fluctuate since it is heavily sensitive with the image contents while SRGA curves are more stable, focusing more on degradations. "PIES-REDS Interspersed" means that the data in each blur level are interspersed by PIES and REDS.

**SRGA versus PSNR in Luminance Jitter:** Another advantageous property of SRGA is that it is more robust to image luminance jitter, which is a common photographic disturbance in real camera imaging. To validate the robustness of SRGA, we manually shift the luminance of the test PIES-Blur images and maintain the original luminance of the reference PIES-Clean.
IQA is highly sensitive to image content. Different datasets will have different absolute IQA values, even though they share the same degradation. For example, under the same degradation conditions, the average PSNR value of Urban100 dataset is consistently lower than Set14 and BSDS100 datasets. More counterintuitively, the PSNR value of Urban100 with blur level 2 is even worse than those of Set14 and BSDS100 with blur level 4 (marked with red dashed lines and arrows). If we use IQA to indicate the model generalization ability, we may come to a conclusion that the model has better generalization ability on the degradation of blur level 4 rather than blur level 2. However, the visual results do not support such a claim. For image restoration task, the model generalization ability should focus more on the input degradations rather than the specific contents. Concretely, for different datasets containing the same degradation type, the generalization ability of the model should be the same, i.e., models should have the same generalization on Urban100 with blur level 2 and BSDS100 with blur level 2. IQA fails to satisfy such a property thus is not appropriate for evaluating the model generalization ability.

When the global brightness between the test and reference images is not aligned (luminance jitter), SR models may still produce good SR results, but IQA will be severely affected, especially for the most commonly-used PSNR. Nevertheless, SRGA is more robust to such disturbances. As revealed in Fig. 14, when the global luminance jitter is within [−10, +20], SRGA is basically unaffected and maintains good stability. However, PSNR is sensitive to such a disturbance, since it measures the absolute distance between paired pixels. This shows the great superiority of SRGA over IQA. SRGA can deal with such disturbances since the PCA operation has subtracted the mean of the features beforehand. If the brightness changes too much, SRGA will also change, because the input distribution has altered a lot. This is equivalent to adding a new degradation to the input image, which is out of the scope of this discussion.

**SRGA versus NR-IQA in Performance Stability:** For no-reference IQA, like NIQE, it is widely acknowledged that it is unstable and less accurate [4], [23]. As shown in Fig. 5(c), the NIQE curves of SRResNet (train: blur0-4) and SRResNet (train: blur0-4&noise0-20) on PIES-Blur dataset are counterintuitive and abnormal. The NIQE value reaches the maximum at Blur4 data, and then drops suddenly. To better illustrate the abnormal behavior of NIQE, we show some visual results of SRResNet (train: blur0-4) in Fig. 15. The visual examples reveal that the NIQE values cannot always reflect the actual image quality. In the second row, the NIQE value of Blur3.5 data is even higher than that of Blur4.0 and Blur4.5. However, the image quality of Blur3.5 is clearly better than Blur4.0 and Blur4.5 with sharper textures. Similar phenomena are observed in the third and fourth rows. This problem occurs not only in a few images, but also in a large amount of images, especially for images with blur level around 4.0. In addition, as shown in Fig. 5(f), NIQE values also fluctuate greatly on PIES-Noise dataset, compared with PSNR. In Fig. 8, the results of SwinIR-GAN on Noise>20 look visually similar, but their NIQE values differ a lot. This can be attributed to the fact that NIQE is simply determined by limited handcrafted features. Besides NIQE, Gu et al. [23], [24] have demonstrated that other existing NR-IQAs are also far from satisfactory. Inaccuracy and instability make NR-IQA unsuitable as a good GA.

**SRGA versus IQA in Real Data Without GT:** Full-reference IQA, like PSNR, SSIM and LPIPS, requires paired ground-truth (GT) images, which are impractical in real scenarios. No-reference IQA is inaccurate and instable as mentioned before. However, SRGA does not rely on paired reference images, since it considers the distance between statistical feature distributions. As shown in Table III, SRGA can be successfully applied on real-world datasets.
VII. Prospects and Limitations

Although the SRGA metric is designed specifically for blind super-resolution tasks, it can also be applied to other low-level vision tasks that face similar generalization challenges. For example, denoising, deblurring, and dehazing models are trained on a set of degraded images and expected to perform well on a wide range of in/out-distribution degraded images. The SRGA metric can be used to evaluate the generalization ability of these models across various types and levels of degradations. By using the SRGA metric in these tasks, we can obtain a reliable measure of a model’s generalization ability to handle real-world scenarios. In the future work, we will explore the possibility of extending the SRGA metric to other image restoration tasks and investigating its effectiveness in evaluating the generalization performance of other more models.

Since we are the first to propose a generalization assessment metric for SR networks, our work has several limitations. First, SRGA requires a relatively large amount of computation and storage. Specifically, to estimate the GGD parameters of a dataset with 200 output features and dimensions of 128 × 128 × 64 (with input dimensions of 32 × 32 × 3), the time cost is approximately 120 seconds using two physical Intel(R) Xeon(R) Gold 6258R CPUs with 32 GB RAM. Second, unlike IQA, it is hard to subjectively evaluate GA by user study, as GA requires to measure the processing consistency between degraded datasets, not individual images. Third, SRGA is a statistical method, which cannot be directly used as a loss function to optimize the network. But we can mimic its behavior to guide the design of future models.

VIII. Conclusion

In this article, we present a novel perspective for measuring the SR model’s generalization ability. The proposed generalization assessment index, named SRGA, utilizes the statistics of the internal deep features of SR networks. Specifically, we adopt generalized Gaussian distribution to model the deep features and then calculate the distance between different degraded input sets. Interestingly, SRGA does not require paired reference images or any learning process. The proposed SRGA and PIES datasets can help promote the development of blind SR methods, as well as other low-level vision problems.

REFERENCES

[1] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4681–4690.
[2] J. Gu, H. Lu, W. Zuo, and C. Dong, “Blind super-resolution with iterative kernel correction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1604–1613.
[3] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 1905–1914.
[4] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for deep blind image super-resolution,” 2021, arXiv:2103.14006.
[5] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, “Unfolding the alternating optimization for blind super resolution,” 2020, arXiv:2010.02631.
[6] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, “Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2018, pp. 701–710.
[7] S. Maeda, “Unpaired image super-resolution using pseudo-supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 291–300.
[8] A. Shocher, N. Cohen, and M. Irani, “Zero-shot super-resolution using deep internal learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3118–3126.
[9] S. Bell-Kligler, A. Shocher, and M. Irani, “Blind super-resolution kernel estimation using an internal-GAN,” 2019, arXiv:1909.06581.
[10] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, “Blind image super-resolution: A survey and beyond,” 2021, arXiv:2107.03055.
[11] X. Wang, Y. Li, H. Zhang, and Y. Shan, “Towards real-world blind face restoration with generative facial prior,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 9168–9178.
Yihao Liu received the BS and PhD degrees from the University of Chinese Academy of Sciences, in 2018 and 2023, respectively. During his PhD studies, he was affiliated with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. Currently, he holds a research position with the Shanghai Artificial Intelligence Laboratory. His research interests include computer vision and image processing, with a distinct focus on image and video restoration and enhancement.

Hengyuan Zhao is currently working toward the PhD degree with the National University of Singapore supervised by Mike Zheng Shou. He worked as a research intern with Multimedia Laboratory, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. He was supervised by Prof. Yu Qiao and Prof. Chao Dong. His research interests include computer vision and image/video processing.

Jinjin Gu received the bachelor’s degree in computer science and engineering from the Chinese University of Hong Kong (Shenzhen), in 2020, and is currently working toward the doctoral degree with the School of Electrical and Information Engineering, University of Sydney, Australia. He has published more than 30 papers in well-known conferences and journals such as CVPR, ECCV, NeurIPS. IEEE Transactions on Pattern Analysis and Machine Intelligence. He has received honors such as SenseTime Scholarship, Huawei Camera Academic Star, and the 2023 Yifan Award.
Yu Qiao (Senior Member, IEEE) is a professor with Shanghai AI Laboratory and the Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences. He has published more than 600 articles in international journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, IEEE Transactions on Image Processing, IEEE Transactions on Signal Processing, CVPR, and ICCV. His research interests include computer vision, deep learning, and bioinformation. He received the First Prize of the Guangdong Technological Invention Award, and the Jiuxi Lv Young Researcher Award from the Chinese Academy of Sciences. He is a recipient of the distinguished paper award in AAAI 2021. His group achieved the first runner-up with the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition, and the winner at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification. He served as the program chair of IEEE ICIST 2014.

Chao Dong is a professor with the Shenzhen Institute of Advanced Technology, Chinese Academy of Science (SIAT), and Shanghai AI Laboratory. In 2014, he first introduced deep learning method – SRCNN into the super-resolution field. This seminal work was chosen as one of the top ten “Most Popular Articles” of TPAMI, in 2016. His team has won several championships in international challenges “NTIRE2018, PIRM2018, NTIRE2019, NTIRE2020 AIM2020 and NTIRE2022. He worked in SenseTime from 2016 to 2018, as the team leader of Super-Resolution Group. In 2021, he was chosen as one of the World’s Top 2% Scientists. In 2022, he was recognized as the AI 2000 Most Influential Scholar Honorable Mention in computer vision. His current research interest focuses on low-level vision problems, such as image/video super-resolution, denoising, and enhancement.