CONTRIBUTED PAPER

Deep-Learning-Empowered Super Resolution: A Comprehensive Survey and Future Prospects

This survey offers systematic reviews and unifies four major super-resolution modalities under a consistent backbone-based taxonomy, thus providing a unique and contemporary perspective.

By Le Zhang, *Member IEEE*, Ao Li, Qibin Hou, *Member IEEE*, Ce Zhu, *Fellow IEEE*, and Yonina C. Eldar, *Fellow IEEE*

ABSTRACT | Super resolution (SR) has garnered significant attention within the computer vision community, driven by advances in deep learning (DL) techniques and the growing demand for high-quality visual applications. With the expansion of this field, numerous surveys have emerged. Most existing surveys focus on specific domains, lacking a comprehensive overview of this field. Here, we present an in-depth review of diverse SR methods, encompassing single-image SR (SISR), video SR (VSR), stereo SR (SSR), and light field SR (LFSR). We extensively cover over 150 SISR methods, nearly 70 VSR approaches, and approximately 30 techniques for SSR and LFSR. We analyze methodologies, datasets, evaluation protocols, empirical results, and complexity. In addition, we conducted a taxonomy based on each backbone structure according to the diverse purposes. We also explore valuable yet understudied open issues in the field. We believe that this work will serve as a valuable resource and offer guidance to researchers in this domain. To facilitate access to related work, we created a dedicated repository available at https://github.com/AVC2-UESTC/Holistic-Super-Resolution-Review

KEYWORDS | Convolutional neural network (CNN); diffusion model; generative adversarial network (GAN); super resolution (SR); Transformer.

I. INTRODUCTION

Super resolution (SR) represents a highly promising field within computer vision, aiming to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts. Its applications span across various domains, including surveillance video [1], [2], medical image enhancement [3], [4], reconstruction of old images [5], [6], efficient image transmission [7], and specialized areas such as SR microscopy [8], [9], [10] and ultrasound [11]. Recent studies also highlight advancements in single-image SR and model-based artificial intelligence (AI) methods for imaging [12], [13]. However, the reconstruction process encounters difficulties because of the inherent ambiguity arising from multiple HR patches that can degrade to the same LR patch, thereby making the problem ill-posed.

To address this problem, several traditional methods have been proposed [14], [15], [16]. Yang et al. [14] introduced a sparse coding-based approach to obtain HR

Received 29 May 2025; accepted 12 September 2025. Date of publication 8 October 2025; date of current version 17 October 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62020106011 and in part by the National Key Research and Development Program of China under Grant 2024YFE0100700. *(Corresponding author: Ce Zhu.)*

Le Zhang, **Ao Li**, and **Ce Zhu** are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China (e-mail: lezhang@uestc.edu.cn; aoli@std.uestc.edu.cn; eczhu@uestc.edu.cn).

Qibin Hou is with Tianjin Key Laboratory of Visual Computing and Intelligent Perception, School of Computer Science, Nankai University, Tianjin 300071, China (e-mail: houqb@nankai.edu.cn).

Yonina C. Eldar is with the Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

Digital Object Identifier 10.1109/JPROC.2025.3613233

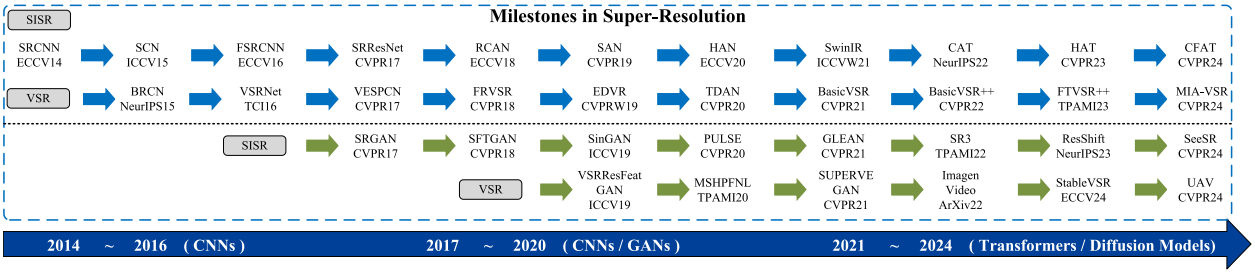

Fig. 1. ***DL-based SR has witnessed remarkable advancements in recent years. In 2014, the integration of CNN-based techniques brought about a revolution in SR, surpassing traditional methods. Subsequently, GAN-based methods were introduced to enhance perceptual quality in 2017. More recently, Transformer-based models have emerged, demonstrating promising performance in SR tasks. SR approaches can generally be categorized into two main branches: generative models (indicated in dark green arrows) and regression-based models (indicated in blue arrows). Generative models, such as GANs [6], [35], [36] and diffusion models [37], [38], [39], focus on generating perceptually realistic images, while regression models aim for pixelwise accuracy. Within this framework, SISR initially dominated the field, with significant advancements in resolution and detail generation. Later, VSR emerged, leveraging temporal information to further enhance video frame quality.***

feature representations and reconstruct arbitrary images. Chang et al. [15] utilized the neighbor embedding technique with a locally linear embedding technique to reconstruct HR images. Timofte et al. [16] proposed an anchored neighborhood regression method to improve SR quality. However, these traditional methods exhibit three primary drawbacks. First, they necessitate the manual design of the mapping function from LR to HR space, which is prone to suboptimal performance. Second, they entail high computational costs, rendering them impractical in many real-time applications. Finally, the quality of the reconstructed SR images, as indicated by metrics such as lower peak signal-to-noise ratio (PSNR) [17] and higher learned perceptual image patch similarity (LPIPS) [18], often does not meet the high expectations required for exceptional clarity and detail.

In contrast, deep learning (DL) methods have shown remarkable potential in SR by harnessing the expressive power of neural networks. DL-based approaches automatically learn a mapping function between LR and HR images through weight adjustments using gradient descent and backpropagation [19]. These methods effectively overcome the manual design requirement associated with traditional approaches, addressing the first drawback [20], [21], [22]. Deep learning techniques benefit from efficient structural designs [23], [24], [25], [26] and often require only a single pass to generate super-resolved images, enabling faster processing. This efficiently mitigates the second major drawback of traditional approaches. Furthermore, DL-based methods excel in extracting complex patterns [5], [27], [28], [29] that bridge the gap between LR and HR spaces, leading to promising advancements in both quantitative and qualitative aspects, thereby tackling the third drawback of traditional techniques.

Motivated by the advantages mentioned above, substantial efforts have been devoted to DL-based SR approaches, as depicted in Fig. 1. Pioneered by SRCNN [20], subsequent research has explored deeper networks for single-image SR (SISR) [5], [28], [30], [31], introduced spatiotemporal subpixel convolution networks for video SR (VSR) [32], employed parallax prior learning for stereo SR (SSR) [33], and utilized convolutional neural networks (CNNs) for light field SR (LFSR) [34]. With the evolution of SR technology, numerous surveys have been published to provide valuable guidance to researchers in this domain. To better illustrate the landscape of existing SR surveys and highlight the specific contributions of this work, we present a comparative overview in Table 1. From this table, it is evident that while numerous surveys exist, they often focus on a single modality (e.g., image SR as seen in [40] and [41], or VSR as in [1]), a specific technique (e.g., diffusion models for image SR by Moser et al. [42]), or a particular application domain (e.g., remote sensing SR in [43]). Furthermore, many cover literature up to earlier timeframes. This article distinguishes itself by offering a uniquely comprehensive and contemporary perspective. Specifically, we are the first to systematically review and unify four major SR modalities—image SR, VSR, SSR, and LFSR—under a consistent backbone-based taxonomy (i.e., CNN, Transformer, generative adversarial network (GAN), and diffusion models). This approach not only provides a holistic understanding of the architectural evolution and methodological trends across these diverse SR tasks, but also covers the most recent advancements up to early 2025, thereby addressing a significant gap in the existing literature for a truly integrative and up-to-date resource.

Although existing surveys have significantly contributed to addressing various SR challenges, a comprehensive review that spans all related domains is still lacking. To fill this gap, we present an in-depth review of various SR problems. These include SISR, which exclusively utilizes intraframe information to enhance the resolution of individual static images. In contrast, other SR methods such as VSR, SSR, and LFSR employ various types of interframe information, including temporal and angular data, to improve image quality and resolution. As shown

in Fig. 2, we first categorize SR methods within these domains based on two primary approaches: generative models and regression models. Generative models are further divided into GAN-based [44] and diffusion-based methods [45], [46], while regression models are classified according to their network backbone, either CNN-based [47] or Transformer-based [48], [49]. For each backbone type, we also perform a detailed classification based on their structural characteristics, providing a comprehensive overview of the landscape.

CNNs excel at capturing local features and delivering strong performance, leveraging their inherent characteristics of translation invariance and locality. Transformers, on the other hand, have gained prominence in visual tasks by effectively capturing long-range spatial dependencies, enabling larger perceptual fields, and yielding improved performance. Commonly used loss functions in SR, such as $\mathcal{L}_1$ or $\mathcal{L}_2$, tend to encourage the network to minimize pixelwise errors, which, in the context of an ill-posed problem, can lead to overly smooth results and suboptimal perceptual quality. To address this, GAN-based methods employ perceptual loss to enhance perceptual quality and provide a closer approximation to the human visual system (HVS) [6], [50].

The remaining sections of this article are organized as follows. In Section II, we revisit the definition of the SR task, explore evaluation metrics, elaborate on different domains, and discuss corresponding optimization objectives and network backbones. Subsequently, Section III provides a comprehensive review of SISR methods, while Section IV focuses on VSR techniques. We, then, delve into SSR approaches in Section V, followed by an in-depth review of LFSR methods in Section VI. Section VII addresses existing challenges and outlines potential future directions in the SR field. Finally, in Section VIII, we present concluding remarks.

II. PRELIMINARIES

A. SR Definition

In general, we define SR in the context of 2-D natural images. We assume that the HR image X_{hr} and the LR image X_{lr} obey the following degradation process:

$$X_{\text{lr}} = f_{\text{de}}\left(X_{\text{hr}}, \Theta\right) \downarrow_s + \epsilon \tag{1}$$

where $f_{\text{de}}(\cdot)$ represents the degradation process, Θ refers to the parameters of the degradation kernel, $\downarrow_s$ indicates a downsampling operation, and the variable ϵ denotes the additive white Gaussian noise (AWGN) with a standard deviation of σ (noise level).

There are two main approaches to model this reconstruction process.

1) Mapping-Based SR Model: In this regression approach, the objective is to learn a mapping function $\mathcal{F}$ by minimizing the distance between the predicted HR image $\mathcal{F}(X_{\text{lr}}, \theta)$ and the ground-truth HR image X_{hr}. This can be formalized as

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} E\left(\mathcal{F}\left(X_{\text{lr}}^i, \theta\right), X_{\text{hr}}^i\right) \tag{2}$$

where $E(\cdot)$ denotes a distance function (e.g., mean squared error), and N is the number of training samples $(X_{\text{lr}}^i, X_{\text{hr}}^i)$.

2) Degradation Matching SR Model: This approach focuses on reconstructing the HR image $\hat{X}_{\text{hr}}$ such that it best matches the given LR image X_{lr} after undergoing a specific degradation process. The relationship can be expressed as

$$\hat{X}_{\text{hr}} = \arg\min_{X_{\text{hr}}} ||X_{\text{lr}} - f_{\text{de}}\left(X_{\text{hr}}, \Theta\right)|| \tag{3}$$

where $f_{\text{de}}(\cdot)$ is the degradation function as defined in (1). Recent advancements in this model involve leveraging generative models to explore the latent space for HR images that mimic the degraded versions of the LR input [63], [64], [65].

For each specific sample $(X_{\text{lr}}^i, X_{\text{hr}}^i)$, $i \in 1, \ldots, N$, the details are defined according to the specific domains.

1) In SISR, only intraview information is used, where the LR image $X_{\text{lr}}^i \in \mathbb{R}^{H\times W\times C}$ is equivalent to the HR image $X_{\text{hr}}^i \in \mathbb{R}^{sH\times sW\times C}$ degraded by a magnification factor s. Here, H and W represent the height and width of the LR image, and C is the number of color channels.
2) For SSR, cross-view information can be utilized to provide more disparity correlation. The dimension is defined as: $X_{\text{lr}}^i \in \mathbb{R}^{M\times H\times W\times C}$ and $X_{\text{hr}}^i \in \mathbb{R}^{M\times sH\times sW\times C}$, where M indicates different perceptive views.
3) VSR introduces temporal information, and the dimension is defined as: $X_{\text{lr}}^i \in \mathbb{R}^{T\times H\times W\times C}$ and $X_{\text{hr}}^i \in \mathbb{R}^{T\times sH\times sW\times C}$, where T indicates the time serial index.
4) LFSR incorporates angular information, enabling the model to utilize multiple views from different angles for generating higher quality super-resolved images. This is exemplified by a light field camera that captures scenes from varied perspectives, encoding these as angular dimensions within the dataset. Specifically, the dimensions of an LFSR sample are defined as: $X_{\text{lr}}^i \in \mathbb{R}^{H\times W\times C\times U\times V}$ and $X_{\text{hr}}^i \in \mathbb{R}^{sH\times sW\times C\times rU\times rV}$, where U and V represent the angular dimensions and r signifies the angular magnification factor.

In practical scenarios, simulating all degradation processes is challenging. Therefore, to simplify the SR task, the bicubic kernel and Gaussian blur kernel are commonly used to simulate the degradation process. Unless explicitly stated, the reviewed methods in this article are

Table 1 Comparison of Existing SR Survey Articles

Survey	Venue	Primary Scope	Timeframe	Description
Yue *et al.* [51] (2016)	IEEE Signal Processing	Image SR	Up to 2015	Reviews regularized frameworks for multi/single-frame SR; discusses priors and optimization.
Yang *et al.* [40] (2019)	IEEE T-MM	Image SR	Up to 2018	Details CNN architectures and optimization strategies for SISR.
Wang *et al.* [41] (2021)	IEEE T-PAMI	Image SR	Up to 2019	Broadly surveys DL methods (CNN/GAN), learning paradigms, metrics, and applications for image SR.
Anwar *et al.* [52] (2020)	ACM Computing Surveys	Image SR	Up to 2019	Taxonomizes DL architectures for SISR; evaluates models and discusses real-world SR challenges.
Li *et al.* [53] (2020)	IET Image Processing	Image SR	Up to 2019	Surveys classical and DL (CNN/GAN) SISR; discusses loss functions, architectures and metrics.
Li *et al.* [54] (2021)	Irbm	Image SR (Medical Imaging)	Up to 2021	Reviews DL for Medical Image SR; covers architectures, loss functions and modality applications.
Liu *et al.* [1] (2022)	Artificial Intelligence Review	Video SR	Up to 2021	Systematically categorizes VSR methods by inter-frame information usage; discusses metrics and challenges.
Chen *et al.* [55] (2022)	Information Fusion	Image SR (Real-World SISR)	Up to 2021	Categorizes Real-World SISR methods; discusses datasets, metrics, and challenges.
Wang *et al.* [43] (2022)	Earth-Science Reviews	Image SR (Remote Sensing)	Up to 2022	Reviews DL for Remote Sensing SR; categorizes methods (CNN/GAN/Attention) and discusses data.
Bashir *et al.* [56] (2022)	PeerJ Computer Science	Image SR	Up to 2021	Reviews DL for SISR by learning paradigms; details metrics, datasets and applications.
Wang *et al.* [57] (2022)	Remote sensing	Image SR (Remote Sensing)	Up to 2022	Reviews DL for Remote Sensing SR; discusses architectures, learning strategies and challenges.
Liu *et al.* [58] (2022)	Annual Review of Biophysics	Super-Resolution Microscopy (SRM)	Up to 2022	Explores SMLM and MINFLUX for cell biology; discusses principles, advancements and data analysis.
Lepcha *et al.* [59] (2023)	Information Fusion	Image SR	Up to 2022	Surveys traditional and DL (CNN/GAN/Transformer) SR; discusses metrics and applications.
Prakash *et al.* [60] (2022)	Phil. Trans. R. Soc. A.	Super-Resolution Microscopy (SRM)	Up to 2022	Overviews SRM techniques (STED/SIM/SMLM); discusses principles, applications and tools.
Baniya *et al.* [61] (2024)	IEEE T-ETCI	Video SR	Up to 2023	Surveys VSR via key components and taxonomy; covers applications, challenges, and trends.
Moser *et al.* [42] (2024)	IEEE T-NNLS	Image SR (Diffusion Models)	Up to 2024	Reviews Diffusion Models for Image SR; covers theory, applications, and future challenges.
Li *et al.* [62] (2024)	ACM Computing Surveys	Image SR	Up to 2023	Systematically reviews DL for SISR by simulation, real-world, and domain-specific categories; notes challenges.
Ours	N/A	Image/Video/Stereo/Light Field SR	Up to 2025	Comprehensively reviews Image/Video/Stereo/LF SR; offers backbone-based taxonomy and discusses open issues.

designed under the condition of these two degradation approaches.

B. Metrics

1) Distortion Metrics: In SR, two commonly used metrics for evaluating distortion performance are the PSNR and the structural similarity index measure (SSIM) [17].

PSNR evaluates the quality of the reconstructed SR image X_{sr} by comparing it to the ground-truth HR image X_{hr}, and it is calculated as

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{mse}}\right) \tag{4}$$

where MAX represents the maximum possible image value. Choosing the correct value for MAX is critical, especially in imaging fields like MRI [66], [67], [68], where images may be stored in floating-point format. In such cases, using a fixed value like 255 may not be appropriate, and alternative normalization strategies should be considered. The mean square error (mse) is defined as

$$\text{mse} = \frac{1}{\text{CHW}} \sum_{k=0}^{C-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(X_{\text{sr}}(k,i,j) - X_{\text{hr}}(k,i,j)\right)^2 \tag{5}$$

where C, H, and W represent the channel, height, and width of the evaluated images, respectively.

SSIM offers a more perceptually motivated evaluation by modeling distortion through brightness, contrast, and structure changes. SSIM is computed on small windows, typically 8×8 or 11×11 pixels, of the image. The SSIM index between two image patches X_{sr} and X_{hr} is calculated as

$$\text{SSIM}(X_{\text{sr}}, X_{\text{hr}}) = \frac{\left(2\mu_{X_{\text{sr}}}\mu_{X_{\text{hr}}} + c_1\right)\left(2\sigma_{X_{\text{sr}}X_{\text{hr}}} + c_2\right)}{\left(\mu_{X_{\text{sr}}}^2 + \mu_{X_{\text{hr}}}^2 + c_1\right)\left(\sigma_{X_{\text{sr}}}^2 + \sigma_{X_{\text{hr}}}^2 + c_2\right)} \tag{6}$$

where $\mu_{X_{\text{sr}}}$ and $\sigma_{X_{\text{sr}}}^2$, and $\mu_{X_{\text{hr}}}$ and $\sigma_{X_{\text{hr}}}^2$ represent the mean and variance of X_{sr} and X_{hr}, respectively. The $\sigma_{X_{\text{sr}}X_{\text{hr}}}$ indicates the covariance between X_{sr} and X_{hr}, and c_1 and c_2 are the constant values used for numerical stability.

2) Perceptual Metrics: Beyond traditional distortion metrics, perceptual quality is often better captured by advanced methods such as the Frechet inception distance (FID) [69] and the LPIPS [18]. These metrics evaluate the similarity of high-level features rather than focusing purely on pixel differences.

FID measures the similarity between the distributions of real and generated images in a feature space and is defined as

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \tag{7}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the means and covariances of the feature vectors from real and generated images, respectively. FID is particularly useful for assessing how well the generated images mimic the overall distribution of real images, making it valuable in tasks involving natural image generation. In practice, a lower FID score indicates better perceptual similarity between real and generated images.

LPIPS evaluates the similarity between deep feature representations of two images and is computed as

$$\text{LPIPS}(x,y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot \left(\phi_l(x)_{\text{hw}} - \phi_l(y)_{\text{hw}}\right)||_2^2 \tag{8}$$

where ϕ_l denotes the deep features from layer l, and H_l and W_l are the height and width of the feature map at layer l, respectively. Note that using $w_l = 1$ for all l is equivalent to computing a cosine distance. In practice, LPIPS is widely adopted in tasks that prioritize perceptual similarity, as it leverages deep neural network activations to compare how similarly the networks perceive two images.

3) Complexity: In general, model complexity is indicated by two metrics: parameters and multiadd operations. The number of learnable parameters in a model reflects its size, while the number of multiplication and addition operations approximates the inference time on hardware

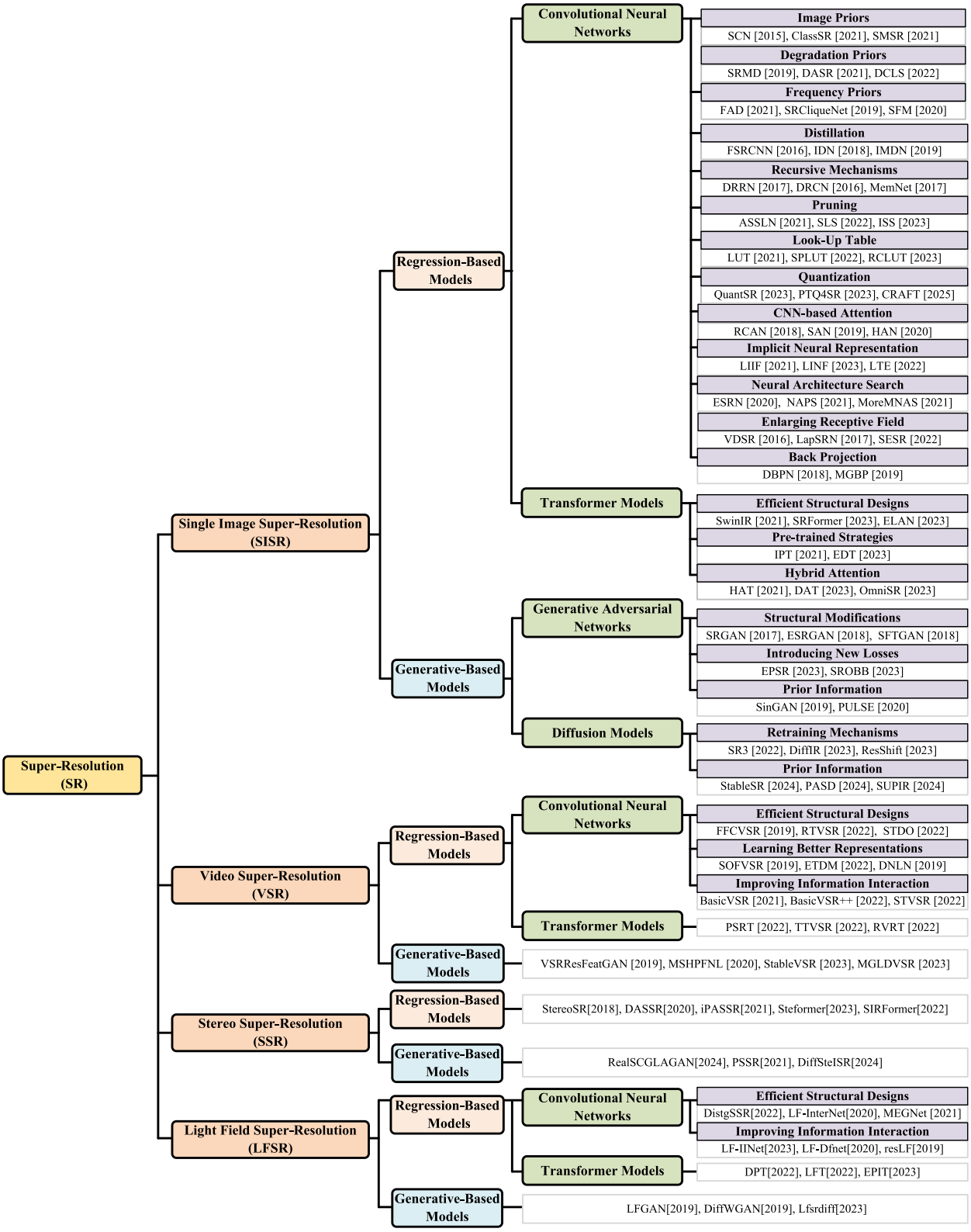

Fig. 2. ***Taxonomy of SR techniques in this article. This figure presents a hierarchical categorization of SR methods, initially divided into four major branches: SISR, VSR, SSR, and LFSR. Each branch is further divided into two primary pathways: regressive models (shown in yellow) and generative models (shown in blue). Within these pathways, methods are, then, categorized based on the network backbone, including CNNs, Transformer models, GANs, and diffusion models. Additionally, each model class is further subdivided based on distinct design purposes such as efficient structural designs, hybrid attention mechanisms, and retraining techniques.***

platforms. However, since most works only report the number of parameters, we focus on parameters as the complexity metric in this article.

C. Domain Elaboration and Corresponding Optimization Objective

1) Single-Image Super Resolution: Deep learning-based SISR techniques leverage information from a single input image to generate an HR output. These methods primarily focus on extracting intraview features, and a larger receptive field often leads to higher quality results [12], [13], [70].

a) Loss functions: Regarding the choice of loss function, pixelwise losses are commonly used in SISR. The most widely adopted loss function is the $\mathcal{L}_1$ loss

$$\mathcal{L}_1 = \frac{1}{\mathrm{CHW}} \sum_{k=0}^{C-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \|X_{\mathrm{sr}}(k,i,j) - X_{\mathrm{hr}}(k,i,j)\|_1 \quad (9)$$

where H and W represent the height and width of the evaluated images, respectively, C denotes the number of channels, and X_{sr} and X_{hr} represent the super-resolved and HR images, respectively.

To address sensitivity to outliers, some approaches employ a smooth approximation of the $\mathcal{L}_1$ loss, known as the Charbonnier loss [71]. It can be expressed as

$$\mathcal{L}_{\mathrm{Char.}} = \frac{1}{\mathrm{CHW}} \sum_{k=0}^{C-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \sqrt{(X_{\mathrm{sr}}(k,i,j) - X_{\mathrm{hr}}(k,i,j))^2 + \epsilon^2} \quad (10)$$

where ϵ is a small constant to ensure the function remains differentiable at zero. This constant helps to avoid the nondifferentiability issue of the $\mathcal{L}_1$ loss at zero, providing a robust, smooth loss function that is less sensitive to outliers than the squared error loss. Alternatively, some methods employ the $\mathcal{L}_2$ loss as their training objective

$$\mathcal{L}_2 = \frac{1}{\mathrm{CHW}} \sum_{k=0}^{C-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} (X_{\mathrm{sr}}(k,i,j) - X_{\mathrm{hr}}(k,i,j))^2 . \quad (11)$$

Since minimizing the above losses commonly leads to smooth results, GAN-based models introduced perceptual loss to obtain better quality super-resolved images

$$\mathcal{L}_{\mathrm{Perc.}} = \sum_i \|\varphi_i(X_{\mathrm{sr}}) - \varphi_i(X_{\mathrm{hr}})\|_2^2 \quad (12)$$

where $\varphi(\cdot)$ indicates the high-level feature extractor, commonly adopting the visual geometry group (VGG) network [72]. Consequently, $\varphi_i(X_{\mathrm{sr}})$ and $\varphi_i(X_{\mathrm{hr}})$ represent the feature maps of the SR and HR images, which are extracted from the ith layer of $\varphi(\cdot)$. Additionally, adversarial loss is introduced to train the GAN-based model, defined as

$$\mathcal{L}_{\mathrm{Adv.}} = -\log D(G(X_{\mathrm{lr}})) \quad (13)$$

where the discriminator is trained to minimize

$$\mathcal{L}_{\mathrm{Dis.}} = -\log(D(X_{\mathrm{hr}})) - \log(1 - D(G(X_{\mathrm{lr}}))) . \quad (14)$$

Here, $G(\cdot)$ represents the generator, and $D(G(\cdot))$ represents the probability of the discriminator over the training samples.

2) Video Super Resolution: VSR extends the concept of SISR to video sequences, leveraging the temporal coherence between consecutive frames. By exploiting the temporal information, VSR methods produce sharper and more temporally consistent HR video frames. VSR techniques involve multiple stages, including motion estimation, frame alignment, and temporal fusion. Motion estimation estimates the motion between consecutive frames to compensate for interframe misalignment. Frame alignment techniques align the LR frames to a common reference frame, enabling effective fusion and aggregation of information across frames. Temporal fusion methods combine the aligned frames to generate an HR video sequence.

a) Loss functions: In addition to the commonly used loss functions described in SISR, some works utilize optical flow to estimate the motion information and apply motion compensation loss, which can be formulated as follows:

$$\mathcal{L}_{\mathrm{mc}} = \sum_{i=-T}^{T} \|X_{\mathrm{lr}}^i - \tilde{X}_{\mathrm{lr}}^{0\to i}\|_1 + \alpha \|\Delta F_{i\to 0}\|_1 \quad (15)$$

where T is the size of the temporal span, X_{lr}^i represents the ith frame, and $\tilde{X}_{\mathrm{lr}}^{0\to i}$ indicates the backward warped X_{lr}^0 according to the estimated flow $F_{i\to 0}$. Additionally, $\Delta F_{i\to 0}$ is the total variation term. This motion compensation loss helps enhance the temporal coherence of the super-resolved video frames, leading to improved video quality and stability.

3) Stereo Super Resolution: SSR aims to enhance the resolution and quality of depth or disparity maps obtained from multiple LR stereo image pairs. By exploiting the inherent correlation between stereo images, SSR generates HR disparity maps that accurately represent the scene geometry. The SSR process involves three main stages: disparity estimation, disparity refinement, and disparity fusion. Disparity estimation techniques estimate the disparity map from each LR stereo image pair, while disparity refinement methods enhance the disparity maps by exploiting interview and intraview constraints. Disparity fusion techniques combine disparity maps from multiple viewpoints to generate an HR disparity map.

a) Loss functions: In addition to the commonly used $\mathcal{L}1$ loss, other loss functions are employed in SSR. Consider the left–right consistency, which can be formulated as

$$\begin{cases} X_{\text{lr}}^{\text{left}} = \mathbf{M}_{\text{right}\to\text{left}} \otimes X_{\text{lr}}^{\text{right}} \\ X_{\text{lr}}^{\text{right}} = \mathbf{M}_{\text{left}\to\text{right}} \otimes X_{\text{lr}}^{\text{left}} \end{cases} \tag{16}$$

where $\mathbf{M}_{\text{right}\to\text{left}}$ and $\mathbf{M}_{\text{left}\to\text{right}}$ represent the two parallax-attention maps generated by PAM [73] and $\otimes$ denotes the batchwise matrix multiplication. The photometric loss is defined as

$$\begin{aligned} \mathcal{L}_{\text{pho.}} = & \sum_{p\in\mathbf{V}_{\text{left}\to\text{right}}} \left\| X_{\text{lr}}^{\text{left}}(p) - \left(\mathbf{M}_{\text{right}\to\text{left}} \otimes X_{\text{lr}}^{\text{right}}(p)\right) \right\|_1 \\ & + \sum_{p\in\mathbf{V}_{\text{right}\to\text{left}}} \left\| X_{\text{lr}}^{\text{right}}(p) - \left(\mathbf{M}_{\text{left}\to\text{right}} \otimes X_{\text{lr}}^{\text{left}}(p)\right) \right\|_1 \end{aligned} \tag{17}$$

where p represents the pixel with a valid mask value, and $\mathbf{V}$ is defined as

$$\mathbf{V}_{\text{left}\to\text{right}}(i,j) = \begin{cases} 1, & \text{if } \sum_{k\in[1,W]} \mathbf{M}_{\text{left}\to\text{right}}(i,j,k) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where τ is a threshold and W is the width of stereo images. $\mathbf{M}_{\text{left}\to\text{right}}(i,j,k)$ indicates the contribution of pixel (i,j) in the left image to pixel (i,k) in the right image.

The smoothness loss is also utilized often and can be formulated as

$$\begin{aligned} \mathcal{L}_{\text{smo.}} = \sum_{\mathbf{M}} \sum_{i,j,k} & \left(\|\mathbf{M}(i,j,k) - \mathbf{M}(i+1,j,k)\|_1 \right. \\ & \left. + \|\mathbf{M}(i,j,k) - \mathbf{M}(i,j+1,k)\|_1 \right) \end{aligned} \tag{19}$$

where $\mathbf{M} \in \mathbf{M}_{\text{left}\to\text{right}}, \mathbf{M}_{\text{right}\to\text{left}}$. The cycle loss is also commonly used in SSR

$$\begin{aligned} \mathcal{L}_{\text{cyc.}} = & \sum_{p\in\mathbf{V}_{\text{left}\to\text{right}}} \|\mathbf{M}_{\text{left}\to\text{right}\to\text{left}}(p) - I(p)\|_1 \\ & + \sum_{p\in\mathbf{V}_{\text{right}\to\text{left}}} \|\mathbf{M}_{\text{right}\to\text{left}\to\text{right}}(p) - I(p)\|_1 \end{aligned} \tag{20}$$

where $I \in \mathbb{R}^{H\times W\times W}$ is a stack of H identity matrices. $\mathbf{M}_{\text{right}\to\text{left}\to\text{right}}$ and $\mathbf{M}_{\text{right}\to\text{left}\to\text{right}}$ can be defined as

$$\begin{cases} \mathbf{M}_{\text{right}\to\text{left}\to\text{right}} = \mathbf{M}_{\text{right}\to\text{left}} \otimes \mathbf{M}_{\text{left}\to\text{right}} \\ \mathbf{M}_{\text{right}\to\text{left}\to\text{right}} = \mathbf{M}_{\text{left}\to\text{right}} \otimes \mathbf{M}_{\text{right}\to\text{left}}. \end{cases} \tag{21}$$

4) Light Field Super Resolution: LFSR focuses on enhancing the spatial resolution and quality of light field data, which captures both spatial and angular information of a scene. LFSR typically involves two main steps: subaperture image SR and angular upsampling. Subaperture image SR enhances the resolution of each subaperture image within the light field. Angular upsampling increases the angular resolution of the light field by generating new views between existing views. By harnessing this angular information, LFSR enhances computational models' capability to perceive depth and texture in a manner akin to human vision, making it particularly valuable in applications such as depth estimation [74], [75], semantic segmentation [76], [77], and saliency detection [78].

a) Loss functions: In general, LFSR methods utilize the $\mathcal{L}_1$ loss as the main term of total loss. Considering the structural consistency, some works introduced the epipolar-plane image (EPI) gradient loss, which can be formulated as

$$\begin{aligned} \mathcal{L}_e = & \|\Delta_x \hat{E}_{y,v} - \Delta_x E_{y,v}\|_1 + \|\Delta_u \hat{E}_{y,v} - \Delta_u E_{y,v}\|_1 \\ & + \|\Delta_y \hat{E}_{x,u} - \Delta_y E_{x,u}\|_1 + \|\Delta_v \hat{E}_{x,u} - \Delta_v E_{x,u}\|_1 \end{aligned} \tag{22}$$

where $\hat{E}_{y,v}$ and $\hat{E}_{x,u}$ indicate the EPIs of the super-resolved light field (LF) images, and $E_{y,v}$ and $E_{x,u}$ represent the EPIs of the groud-truth LF images. In additon, (x,y) and (u,v) denote the spatial plane and angular plane, respectively, and the gradients are calculated along both spatial and angular dimensions with both horizontal and vertical EPIs.

D. Network Backbone

1) Convolutional Neural Network: CNN [79], [80], [81], [82], [83], [84], [85] is a DL model widely used for image analysis and computer vision tasks. It consists of multiple layers, each performing specific operations to extract and learn features from input images. The operations in a CNN can be described as follows.

a) Convolution operation: The convolution operation applies a filter to the input image using elementwise multiplication and summation. It can be represented as

$$\text{Conv}(i,j) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s,t)\, x(i-s, j-t) \tag{23}$$

where (i,j) is the coordinates of the input image, w is the filter (also known as the kernel), and the size of the filter is $(2a+1)\times(2b+1)$.

b) Activation function: After the convolution operation, an activation function is applied elementwise to introduce nonlinearity. One commonly used activation function is the rectified linear unit (ReLU) [86], defined as

$$\text{ReLU}(x) = \max(0, x). \tag{24}$$

This function sets negative values to zero and keeps positive values unchanged.

c) Pooling operation: Pooling reduces the spatial dimensions of the feature maps obtained from the convolutional layers. For example, max pooling selects the maximum value within a local region

$$\text{MaxPooling}\left(x\right)=\max_{k\in S}x\left(k\right) \tag{25}$$

where S indicates the region of calculation.

d) Fully connected layers: Fully connected layers are responsible for the final classification or regression tasks. Each neuron in these layers is connected to all neurons in the previous layer. The output of a fully connected layer is computed as

$$\text{FC}\left(x\right)=\sigma\left(Wx+b\right) \tag{26}$$

where x is the input vector, W is the weight matrix, b is the bias vector, and σ represents the activation function.

2) Transformer Network: The Transformer network [48] is a highly effective architecture widely applied to various computer vision tasks [49], [87], [88], [89], [90], due to its ability to capture long-range dependencies. The main components are described as follows.

a) Input embedding: The input image is divided into patches, which are linearly projected into higher dimensional embeddings. Let $\mathbf{X}=[x_1,x_2,\ldots,x_N]$ represent the input sequence of embeddings, where x_i denotes the embedding for the ith patch.

b) Positional encoding: To incorporate positional information, a positional encoding matrix $\mathbf{P}\in\mathbb{R}^{N\times d}$ is added to the input embeddings [48], [87], [91], [92], [93], where N is the sequence length, and d is the dimension of the embeddings.

c) Self-Attention: For the embedding sequence $\mathbf{X}$, the self-attention mechanism computes the attention weights $\mathbf{W}$ using the following definition:

$$\mathbf{Q}=\mathbf{X}\mathbf{W}_Q,\quad \mathbf{K}=\mathbf{X}\mathbf{W}_K,\ \mathbf{V}=\mathbf{X}\mathbf{W}_V \tag{27}$$

where $\mathbf{Q}\in\mathbb{R}^{N\times d_k}$, $\mathbf{K}\in\mathbb{R}^{N\times d_k}$, and $\mathbf{V}\in\mathbb{R}^{N\times d_k}$ are the linear projections of $\mathbf{X}$ into query, key, and value vectors, respectively. Here, $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are the learnable weight parameters, and d_k is the dimension of the key vectors. The self-attention operation can be formulated as

$$\mathbf{W}=\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{28}$$

d) Feed-forward networks: After the self-attention step, a positionwise feed-forward network (FFN) is applied to each embedding independently. It consists of two linear transformations followed by a nonlinear activation function:

$$\text{FFN}=\text{ReLU}\left(xW_1+b_1\right)W_2+b_2 \tag{29}$$

where x represents the input features, W_1 and W_2 are the learnable weight parameters, and b_1 and b_2 are the bias terms.

3) Generative Adversarial Network: A GAN [44] is a powerful DL model that comprises two key components: a generator G and a discriminator D. GANs are commonly used for generating synthetic data that closely resemble real data. The main modules of GAN can be described as follows.

a) Generator: The generator takes random noise as input and produces synthetic data samples. It can be represented as a function G that maps a noise vector z to a generated sample x

$$x=G\left(z\right). \tag{30}$$

In some applications, instead of random noise z, an LR image X_{lr} is used as input to generate the super-resolved image X_{sr}.

b) Discriminator: The discriminator is used to differentiate between real and generated data samples. It is represented by a function D that outputs a probability indicating the likelihood that the input sample is real

$$D\left(x\right)=P\left(\text{real}|x\right). \tag{31}$$

4) Diffusion Models: Diffusion models [42], [94] refer to a class of generative models that simulate a diffusion process, where data are transformed into noise and then denoised to recover the original distribution. These models can be divided into three main categories: 1) denoising diffusion probabilistic models (DDPMs); 2) score-based generative models; and 3) variational diffusion models. In this section, we focus on DDPMs, a prominent approach within this class [45]. The forward diffusion process gradually transforms data into Gaussian noise, and the reverse process models the denoising to recover the original data distribution, as shown in Fig. 3.

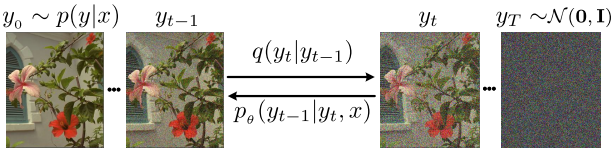

Fig. 3. *In the forward diffusion process q (left to right), Gaussian noise is incrementally added to the target image. Conversely, in the reverse inference process p (right to left), the target image is iteratively denoised under the guidance of a source image x.*

a) Forward process: This process gradually corrupts a data sample $\mathbf{x}_0$ into a sample $\mathbf{x}_T$ that is close to Gaussian noise, using a series of small Gaussian transitions

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{32}$$

where the hyperparameters $\beta_{1:T}$ subject to $0<\beta_t<1$, which indicate the noise level added at each step, and $\mathbf{I}$ is the identity matrix. Each step of the diffusion process generates a noisy data $\mathbf{x}_t$, and the entire diffusion process forms a Markov chain

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)=\prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right). \tag{33}$$

A key characteristic of the diffusion process is that we can directly sample $\mathbf{x}_t$ at any step t based on the original data $\mathbf{x}_0$

$$\mathbf{x}_t \sim q\left(\mathbf{x}_t \mid \mathbf{x}_0\right). \tag{34}$$

Here, we define $\alpha_t=1-\beta_t$ and $\overline{\alpha_t}=\prod_{i=1}^{t} \alpha_i$. By employing the reparameterization trick, we have

$$\mathbf{x}_t=\sqrt{\overline{\alpha_t}} \mathbf{x}_0+\sqrt{1-\overline{\alpha_t}} \boldsymbol{\epsilon}. \tag{35}$$

b) Reverse process: The reverse process is defined by learning to reverse the forward diffusion process. Starting from $\mathbf{x}_T$ assumed to be drawn from a normal distribution, the process is modeled as

$$p\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right). \tag{36}$$

The denoising network predicts the mean $\mu(\mathbf{x}_t, t)$ and covariance $\Sigma(\mathbf{x}_t, t)$ in this process. Importantly, the mean and covariance parameters are crucial for accurately simulating the reverse diffusion process and have significant impacts on the quality of the generated samples. In the SR domain, the conditional formulation $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{z})$, conditioned on $\mathbf{z}$ (e.g., an LR image), uses $\mu_\theta(\mathbf{x}_t, \mathbf{z}, t)$ and $\Sigma_\theta(\mathbf{x}_t, \mathbf{z}, t)$ instead.

c) Optimization objective: If we consider the intermediate variables as latent variables, then the diffusion model is actually a latent variable model containing T latent variables. It may be considered a special case of hierarchical variational autoencoders (VAEs) [95]. We can, then, use variational inference to derive the variational lower bound (VLB, also known as ELBO) as the maximization objective, which leads to

$$\begin{aligned}\log p_\theta\left(\mathbf{x}_0\right) &= \log \int p_\theta\left(\mathbf{x}_0 : \mathbf{x}_T\right) d\mathbf{x}_{1:T} \\ &= \log \int \frac{p_\theta\left(\mathbf{x}_0 : \mathbf{x}_T\right) q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)} d\mathbf{x}_{1:T} \\ &\geq \mathbb{E}_{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\left[\log \frac{p_\theta\left(\mathbf{x}_0 : \mathbf{x}_T\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right]. \end{aligned} \tag{37}$$

The last step employs Jensen's inequality. For network training, the training objective is to minimize the negative VLB

$$\begin{aligned} L &= \mathbb{E}_{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\left[-\log \frac{p_\theta\left(\mathbf{x}_0 : \mathbf{x}_T\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right] \\ &= \mathbb{E}_{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\left[\log \frac{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}{p_\theta\left(\mathbf{x}_0 : \mathbf{x}_T\right)}\right]. \end{aligned} \tag{38}$$

E. Learning Strategies

Learning strategies in SR aim to leverage various information sources to enhance image resolution. Supervised and unsupervised approaches form the two primary categories of learning strategies, each addressing the SR task through different means.

1) Supervised SR: In supervised SR, the training process involves paired LR–HR image examples. These pairs serve as input–output pairs for training deep neural networks [5], [96], [97]. The network learns to capture the relationship between LR and HR images, using HR images as the target. The learning process is guided by minimizing the distance between the super-resolved image and the HR image. The loss function typically encompasses pixelwise differences, such as $\mathcal{L}_1$ [5], [96] or perceptual losses [97] that consider feature representations.

2) Unsupervised SR: Unsupervised SR [98], [99], [100], [101], [102], [103] aims to enhance image resolution without the need for explicit HR supervision during training. DASR [99] estimates degradation information using a trainable encoder in the latent feature space, with the degradation encoder trained in an unsupervised manner through contrastive learning. ZSSR [101] employs a self-supervised approach, where the network utilizes the internal recurrence of information within a single image. The unsupervised framework typically incorporates perceptual losses, adversarial losses, or cycle-consistency losses [103] to further improve the quality of the super-resolved images.

III. SINGLE-IMAGE SUPER RESOLUTION

A. Regressive Models

1) CNN-Based SISR: We categorize CNN-based methods into several more specific subcategories to organize these approaches in a structured way.

a) Image priors: Incorporating domain-specific prior knowledge is a common approach to improve the efficiency of SISR models [23], [104], [105], [106], as shown in Fig. 4. Wang et al. [107] proposed a network that combined sparse coding with expert knowledge to enhance the training process. Mei et al. [104] introduced a

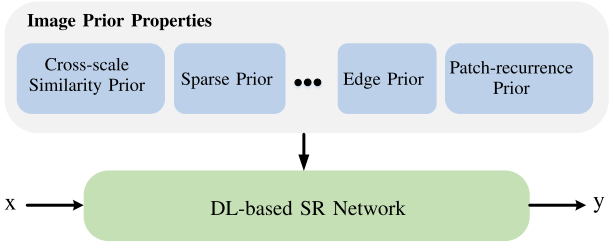

Fig. 4. *Incorporating various image priors to inform the design of the SR network.*

cross-scale nonlocal attention mechanism based on a recurrent network, considering cross-scale image correlation prior. Kong et al. [23] presented a unified framework that combined classification and SR, employing a classification module to classify subimages into different classes based on restoration difficulties computed by PSNR, and then performed SR separately for each class. Wang et al. [25] explored sparsity in image SR and developed a sparse mask network that learns sparse masks to prune redundant computation. Gu et al. [108] identified the indispensability of edge processing in SR and proposed an efficient SR approach by dividing the SR process into edge and flat parts. Nazeri et al. [105] integrated edge priors into their model and formulated a loss function that combines image content and structure terms. Jiang et al. [109] improved weight and data initialization and modified batch normalization (BN) for binary convolution networks by analyzing feature distributions. Shocher et al. [101] exploited the internal recurrence of information within a single image and trained their model with both LR and counterpart downsampling versions of the input image. Cheng et al. [106] leveraged depth information to learn the internal degradation of the image, treating distant patches as LR and short-distance patches as HR, and constructing a self-supervised model. Park et al. [110] trained their model via meta-learning and fine-tuned it at the test stage, allowing the model to adapt to the specific test image using its patch-recurrence property. Gu and Dong [70] performed attribution analysis of SR networks to understand which inputs have the greatest influence on SR performance, providing insights for designing SR networks and interpreting low-level vision deep models. Tian et al. [111] introduced a bit mapping strategy that resorts the edge information to build the lookup table (LUT) of the bit selector, resulting in an efficient network.

b) Lightweight blocks: In efforts to optimize efficiency, one prevalent strategy is the adoption of lightweight blocks that boost performance without compromising computational economy.

A common early approach, as shown in Fig. 5, is to upsample the LR image to the same size as the HR image and then pass it through SRCNN [20], which consists of three convolutional layers, to generate the super-resolved image. Building on this, FSRCNN [112] improved both model performance and computational efficiency by replacing the upsampling step in the input stage with transposed convolution [113] at the final stage. Similarly, Shi et al. [114] proposed efficient subpixel shuffling to enhance the reconstruction of HR images, as illustrated in Fig. 6.

Another line of work focuses on splitting the SR process into stages to further optimize performance. Fan et al. [115] introduced a lightweight conv-block, separating the process into LR and HR stages to extract multiscale features. Zhang et al. [116] followed with an efficient ensemble block, designed to mimic ensemble deep networks, which demonstrated improved results. In a similar vein, Luo [117] developed the lattice block, which uses two butterfly structures to integrate residual blocks.

Recent advancements have also targeted reducing the computational complexity of common SR operations. Chao et al. [118] transformed several time-consuming operations, such as clipping, repeating, and concatenating, into more efficient convolutional operations, leading to a highly efficient SR model. Lin et al. [119] proposed a scalable SR model that is memory-efficient and can handle multiple scales by dynamically switching masks based on the lottery tickets hypothesis.

Additionally, the integration of distillation techniques has emerged as an effective way to build lightweight models with enhanced feature extraction, as shown in Fig. 8. Hui et al. [120] proposed a compact network composed primarily of information distillation blocks, combining enhancement and compression units to capture both long- and short-range features. This concept was further extended by Hui et al. [121], who introduced selective fusion with multidistillation blocks to optimize the model. Meanwhile, Lee et al. [122] leveraged HR information as privileged data within a knowledge distillation network, significantly boosting model efficiency.

RLFN [123] advanced efficient SISR by utilizing a residual local feature learning method with three convolutional layers, achieving a balance between performance and inference time. This approach also incorporated a contrastive loss and a multistage warm-start training strategy, further enhancing runtime efficiency. Finally,

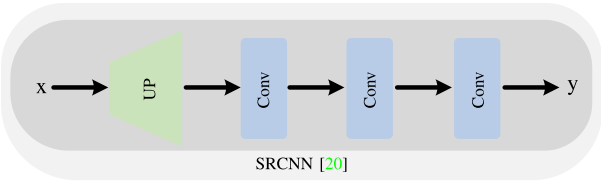

Fig. 5. ***SRCNN [20] framework with only three convolutional layers. This pioneering work was the first to successfully introduce DL into SISR.***

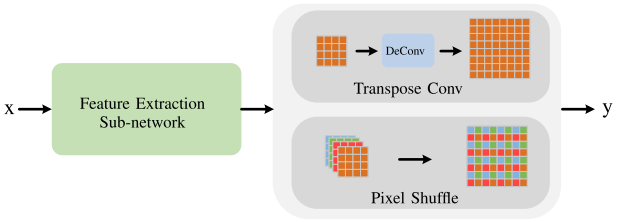

Fig. 6. *Two efficient methods for reconstructing HR images.*

Gao et al. [124] proposed diverse feature distillation blocks to improve performance while maintaining low computational complexity. They also introduced a self-calibration feature fuse block for adaptive information combination.

c) Recursive mechanisms: Recursive layers have been employed to reduce complexity and increase the receptive field, as shown in Fig. 7. Kim et al. [125] introduced a recursive layer to enlarge the receptive field and save computational resources. This recursive concept was further applied in DRRN [126] by adding residual recursive blocks to reduce the number of parameters and facilitate back-propagation. Tai et al. [127] introduced a memory block consisting of a recursive unit and a gate unit to explicitly capture persistent memory through adaptive learning. Han et al. [128] formulated the SR network as a recurrent neural network (RNN) [129] and proposed a network that allows both LR and HR features to jointly contribute to learning the mappings.

d) Pruning: Pruning techniques aim to reduce complexity by eliminating unnecessary parameters. Zhang et al. [130] proposed aligned structured sparsity learning to constrain the filter pruning process. Oh et al. [131] developed a pruning method that determines the pruning ratio for N:M structured sparsity at each layer, optimizing the tradeoff between efficiency and restoration accuracy. Wang et al. [132] perform network pruning starting from random initialization and introduce a flexible thresholding technique to preserve sparse network trainability and achieve improved performance.

e) Lookup table: Jo and Joo Kim [133] introduced a precomputed LUT that retrieves HR output values for query LR input pixels. Building upon this work, Yang et al. [134] analyzed the limitations of LUT and proposed a separable image-adaptive LUT that addresses these limitations. Ma et al. [135] proposed a series–parallel LUT to increase the receptive fields. RCLUT [136] divides the features into channel and spatial dimensions, exploring their respective relationships. This approach significantly increases the receptive field while keeping complexity low. Additionally, a convolutional block is introduced to enhance the interaction of information.

f) Neural architecture search: Neural architecture search (NAS) has been utilized to explore efficient model architectures. Song et al. [137] applied NAS to develop an efficient residual dense block. Chen et al. [138] employed NAS and decomposed upsampling operators to obtain a larger search space. Zhan et al. [26] utilized NAS to study effective pruning strategies for reducing model complexity. MoreMNAS [139] method leverages neural architecture search with a multiobjective approach, combining evolutionary computation and reinforcement learning to automatically balance restoration capacity and model simplicity, resulting in models that outperform state-of-the-art methods in terms of floating point operations (FLOPs) efficiency.

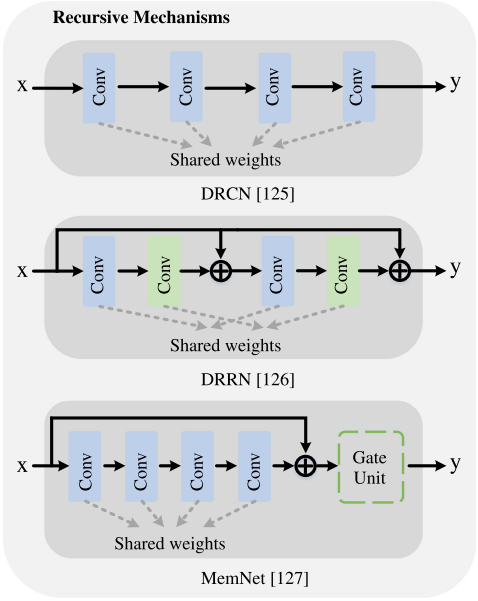

Fig. 7. *Utilizing a recursive mechanism to guide efficient structural designs. The symbol ⊕ represents the elementwise addition.*

g) Implicit neural representation: Drawing inspiration from recent advancements in 3-D reconstruction utilizing

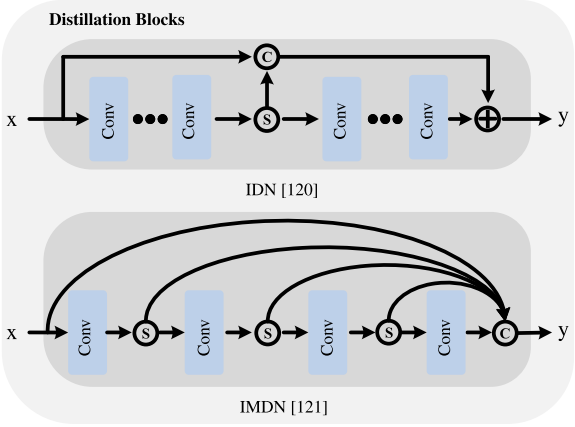

Fig. 8. *Distillation blocks. The symbols Ⓢ and Ⓒ indicate the split and concatenate operations, respectively.*

implicit neural representation [140], Chen et al. [141] introduced the concept of local implicit image function. This function predicts RGB values at specified coordinates by incorporating image coordinates and 2-D deep features in the vicinity of those coordinates. Building upon this work, Lee and Jin [142] proposed the local texture estimator, which serves as a dominant-frequency estimator for natural images. This estimator enables implicit functions to capture intricate details while ensuring the continuous reconstruction of images. Yao et al. [143] combined the implicit representation with normalizing flow and developed a visually pleasing arbitrary-scale model.

h) Diverse degradations: To address the challenges posed by multiple and spatially variant degradations, Zhang et al. [144] introduced the use of blur kernels and noise levels as inputs. Expanding on this idea, Zhang et al. [145] extended the applicability of bicubic degradation by employing a variable splitting technique to handle arbitrary blur kernels. IKC [146] addresses blind SR by iteratively correcting blur kernel estimates to reduce artifacts, and utilizes spatial feature transform layers to effectively manage multiple unknown blur kernels, leading to improved SR performance. DAN [147] addresses the blind SR problem by using an alternating optimization approach, integrating restorer and estimator modules in a single end-to-end trainable network, where the blur kernel estimation and SR image restoration mutually enhance each other for improved performance. Kim et al. [148] proposed a two-branch network that learns spatially variant information and adaptively fuses degradation kernel information. DASR [99] addresses blind SR by learning abstract degradation representations instead of relying on explicit degradation estimation, enabling flexible adaptation to various degradations and achieving state-of-the-art performance. Hui et al. [149] integrated reinforcement learning [150] into SR to guide kernel estimation using a nondifferentiable objective function. Xu et al. [151] considered multiple degrading effects and incorporated dynamic convolution to handle various types of degradations. Additionally, Luo et al. [152] transferred the blurring process from the LR space and introduced a dynamic deep linear filter to generate appropriate deblurring kernels for producing clean features.

i) Frequency processing: Frequency information has proven valuable for effective representation learning in various approaches, as shown in Fig. 9. El Helou et al. [155] introduced stochastic masking of high-frequency bands, enhancing the capability of restoring high frequencies. Xie et al. [156] explored a frequency-aware dynamic network that partitions the input based on its coefficients in the discrete cosine transform domain. Qiu et al. [157] addressed different frequencies with complexity modules and designed a block residual module to restore information from easy-to-recover to hard-to-recover frequencies. Zhong et al. [158] focused on SR in the wavelet domain, utilizing subnetworks to predict

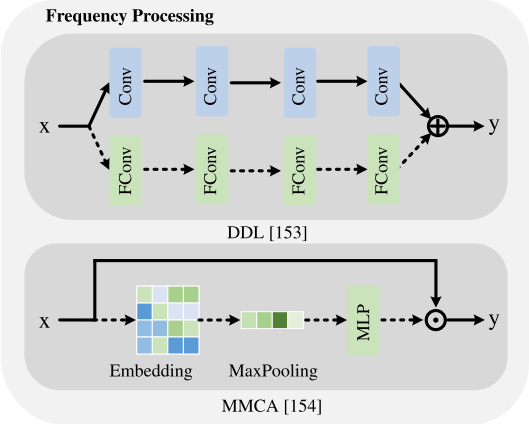

Fig. 9. ***Frequency processing. The dashed line indicates the process in the frequency domain. FConv represents the complex convolutional operation. The symbols $\oplus$ and $\odot$ indicate the elementwise addition and elementwise multiplication, respectively.***

subband coefficients and applying inverse discrete wavelet transform for HR image reconstruction. Magid et al. [154] introduced a dynamic high-pass filtering module to preserve high-frequency information and a multispectral channel attention module to recalibrate responses across different frequencies. Liu et al. [153] explored the spectral Bayesian uncertainty and introduced complex layers to perform dual-domain learning. They introduced a frequency loss defined as

$$\mathcal{L}_{\text{freq.}} = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left\|\Phi\left(X_{\text{sr}}\right)(i,j) - \Phi\left(X_{\text{hr}}\right)(i,j)\right\|_1 \tag{39}$$

where $\Phi(\cdot)$ represents the Fourier transformation, and H and W indicate the height and the width of frequency map, respectively.

j) Enlarging receptive field: Stacking more blocks and utilizing a hierarchy mechanism are introduced into SISR to increase the receptive field, as shown in Fig. 10. Kim et al. [159] proposed the use of very deep networks with 20 cascaded layers to obtain a larger receptive field. Building upon this idea, Lim et al. [161] introduced an enhanced deep SR network, which achieved significant performance improvements by removing the BN [162] layer in the residual block and adding more convolution layers. Lai et al. [160] developed the Laplacian pyramid SR network to progressively reconstruct the subband residuals of HR images. Fan et al. [163] focused on scale-invariant modeling and proposed scalewise convolution. They also built a feature pyramid representation through progressive bilinear downscaling to improve restoration performance. Bhardwaj et al. [24] introduced the over-parameterization mechanism [164], [165] to enlarge the network and achieved promising results.

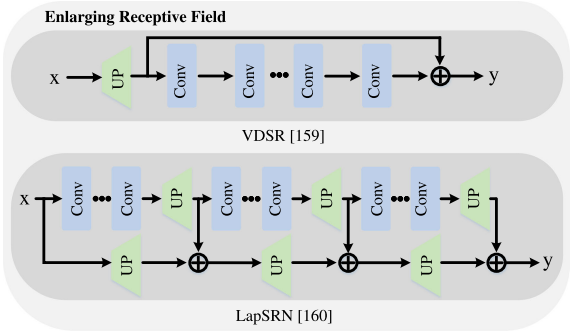

Fig. 10. *VDSR [159] stacked more than 20 layers to enlarge the receptive field, while LapSRN [160] introduced a hierarchical structure to capture diverse scale features. The symbol ⊕ indicates the elementwise addition.*

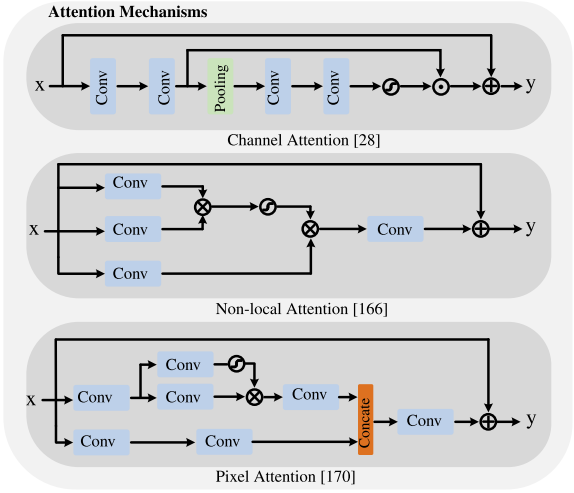

Fig. 11. *Illustration of channel attention, spatial attention, and pixel attention. The symbols ⊕, ⊗, and ⊙ indicate the elementwise addition, matrix multiplication, and elementwise multiplication, respectively.*

k) Attention mechanisms: Attention mechanisms have been widely used to improve information interaction, as shown in Fig. 11. Zhang et al. [28] proposed the residual channel attention networks to assign diverse weights to each channel and enhance performance. Zhang et al. [166] designed local and nonlocal attention blocks to capture long-range dependencies between pixels and focus more on challenging parts. Zamir et al. [167] introduced multiscale information interaction and attention mechanisms to combine contextual information and preserve HR spatial details. Dai et al. [168] introduced a second-order attention network for more powerful feature expression and feature correlation learning. Niu et al. [169] proposed a holistic attention network consisting of a layer attention module and a channel–spatial attention module to model interdependencies among layers, channels, and positions. Zhao et al. [170] proposed pixel attention to obtain 3-D attention maps, achieving promising results with low parameter cost. Mei et al. [171] combined the nonlocal attention mechanism with sparse representation and rectified nonlocal attention with spherical locally sensitive hashing to reduce computational cost.

Unlike the traditional attention mechanisms that typically focus on capturing relationships between pixels, Transformer-based self-attention mechanisms capture dependencies between patches in the input. This difference makes Transformer attention more efficient, as it operates over a higher level representation, enabling it to handle long-range dependencies with fewer computations compared to pixel-based attention. Additionally, Transformer models integrate FFNs to further enhance the feature extraction process, which leads to better performance.

l) Backprojection: Haris et al. [172] proposed deep backprojection networks, which utilized iterative up- and downsampling layers to provide an error feedback mechanism for projection errors at each stage. Michelini et al. [173] extended the iterative backprojections (IBPs) [174] to multilevel IBP and demonstrated its effectiveness comparable to classic IBP. Li et al. [21] explored the feedback mechanism present in HVS and proposed an SR feedback network to refine low-level representations using high-level information.

m) Structural improvements: Several structural improvements have been proposed to enhance the representation learning capabilities of SISR models. Building on previous work by Shocher et al. [101], which employs zero-shot learning methods, MZSR [102] integrates meta-transfer learning with zero-shot SR, allowing for rapid adaptation to specific image conditions with minimal gradient updates, and effectively utilizing both external and internal information for flexible and efficient SR. Guo et al. [175] introduced a dual regression constraint and devised a closed-loop framework to reduce the space of possible functions. Xiao et al. [176] developed the invertible rescaling net, which employed deliberately designed objectives to produce visually pleasing LR

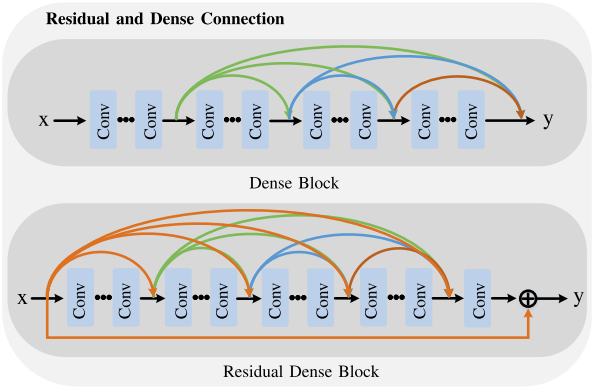

Fig. 12. *Discrepancy between dense connection and residual dense connection. The symbol ⊕ indicates the elementwise addition.*

images. Hu et al. [177] introduced the meta-upscale module, which dynamically predicts the weights of the upscale filter, enabling arbitrary scale factor SR with a single network. Wang et al. [178] incorporated scale-aware feature blocks and upsampling layers to learn scale-arbitrary SR based on fixed-scale networks. Maeda [179] proposed a deep dictionary learned from the network to handle diverse degradations not present in the training set.

n) Residual and dense connection: Several methods have explored the use of residual and dense connections to improve information interaction, as shown in Fig. 12. Lai et al. [180] introduced local skip connections, parameter sharing, and multiscale training strategies to maintain information interaction. Tong et al. [30] added dense connections to the network and demonstrated their effectiveness in achieving good performance on low-level tasks. Zhang et al. [181] combined residuals and dense connections to obtain hierarchical features and achieved favorable performance. Ahn et al. [182] designed an architecture that cascades more residual blocks and dense connections. Li et al. [183] introduced a set of predefined filters and used a CNN to learn the coefficients, enabling a linear combination of these filters to produce the final results. Liu et al. [184] proposed a residual feature aggregation framework to fully extract hierarchical features on the residual branches.

2) Transformer-Based SISR: Due to their ability to model long-range dependencies, Transformers have achieved superior SR results compared to CNNs in SR tasks. However, the high-dimensional matrix operations in Transformers make these methods significantly more complex than CNN-based approaches. Here, we categorize Transformer-based methods into three subclasses based on reducing complexity and enhancing performance, which we will introduce accordingly.

a) Efficient structural designs: Liang et al. [5] presented a strong baseline model called SwinIR for image restoration, based on the Swin Transformer [87]. The Swin Transformer architecture is depicted in Fig. 13. To reduce computational complexity, the authors designed a window-based attention mechanism and incorporated shift operations to enlarge perceptual fields. Choi et al. [185] introduced the N-Gram mechanism into the image domain to enlarge the perceptive fields. Meanwhile, SRFormer [186] evaluated that a larger window size can lead to better performance and proposed a permutation mechanism to enlarge the window size while maintaining efficient parameter scale. Gao et al. [187] introduced a lightweight bimodal network that integrates CNN and Transformer structures for SISR, achieving a better tradeoff between performance and complexity. Lu et al. [188] proposed a lightweight Transformer backbone that captures long-term dependencies between similar image patches using an efficient Transformer and an efficient multihead

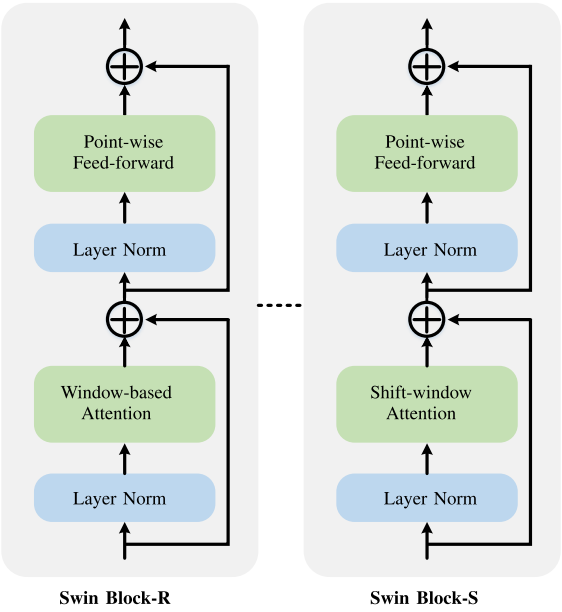

Fig. 13. *Framework of Swin Transformer. -R indicates the regular window attention, and -S indicates the shift window attention.*

attention mechanism. Zhang et al. [189] introduced shift convolution and groupwise multiscale self-attention to reduce the complexity of Transformer-based architectures. Li et al. [190] proposed a dynamic convolution and a sparse self-attention mechanism to explore the most useful similarity values to reconstruct the images. SAFMN [191] utilized multiscale strategy to explore global information and conducted a convolutional channel mixer to introduce local information.

b) Pretrained strategies: Chen et al. [192] proposed a pretrained model called the image processing Transformer, demonstrating that pretrained Transformers significantly enhance performance for low-level tasks. Li et al. [193] conducted a comprehensive analysis of pretraining effects and proposed a versatile model capable of addressing various low-level tasks.

c) Hybrid attenion blocks: Yang et al. [194] introduced a Transformer structure that utilizes soft and hard attention maps to establish feature correspondences between LR and reference images. Chen et al. [195] proposed a hybrid attention Transformer that combines channel attention and self-attention schemes. To mitigate the patch effect caused by window-based attention, the authors introduced an overlapping window partition mechanism. Wang et al. [196] considered the scale-equivariance characteristic and proposed an adaptive feature extractor module to incorporate the scale prior and a neural kriging upsampler to learn feature distance and similarity. Cao et al. [197] proposed an attention-in-attention network to implicitly explore scale nonlocal information. Omni-SR [198] leveraged the interaction

between spatial and channel information and proposed a diverse Transformer attention block to enlarge the effective receptive field. Chen et al. [199] integrated cross-scale attention and local frequency information and proposed a local implicit Transformer to adapt to arbitrary SISR. DAT [200] incorporates self-attention mechanisms across both spatial and channel dimensions and introduces a dual-feature aggregation model. Simultaneously, it leverages convolution layers and adaptive interaction modules to facilitate interactions between local and global information. CRAFT [96] investigates the discrepancy in high-frequency dependencies between CNN and Transformer structures and introduces the high-frequency enhancement module to capture high-frequency information. Simultaneously, it employs channelwise attention to blend diverse frequency characteristics for the enhancement of global representations. CFAT [201] introduced a novel attention mechanism that combines triangular window-based attention with rectangular window and channel attention. This integration enables the model to capture richer contextual information, leading to improved performance.

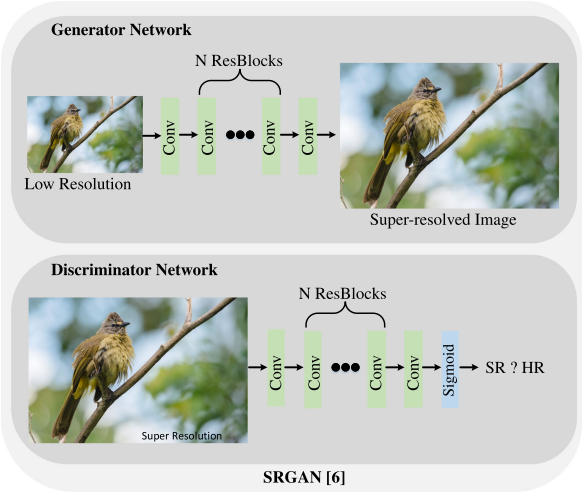

Fig. 14. *Framework of SRGAN [6]. It pioneered the utilization of an exceptionally deep ResNet [203] architecture, incorporating the concept of GANs, in order to create a perceptual loss function that faithfully emulates human perception and facilitates the attainment of photorealistic SISR.*

B. Generative Models

1) GAN-Based SISR:

a) Structural modifications: To recover finer texture details, GAN-based SISR methods have been introduced with various structural enhancements. The pioneering approach, SRGAN [6], consists of a generator and a discriminator, as depicted in Fig. 14. The generator super-resolves the LR image, while the discriminator learns to distinguish between the reconstructed image and real HR images. Building on SRGAN, Sønderby et al. [202] explored the connection between GANs and amortized variational inference, which aligns with principles of variational autoencoders. To address the problem of high-frequency noise often introduced by GAN-based methods, Park [204] proposed an additional discriminator in the feature domain, improving the quality of reconstructed textures. ESRGAN [97] further improved upon SRGAN by refining the architecture with residual-in-residual dense blocks, adopting relativistic GANs for relative realness prediction, and refining perceptual loss, leading to more realistic and visually pleasing textures.

Wang et al. [35] made significant changes to intermediate layers, focusing on recovering textures that align with semantic categories. Building on perceptual enhancements, Shang et al. [205] introduced the receptive field block [206] to produce super-resolved images with extreme perceptual sharpness. Zhang et al. [207] proposed RankSRGAN, a model that optimizes the generator using a ranker, allowing for improvements in perceptual metrics, particularly addressing the challenges posed by nondifferentiable perceptual metrics.

In contrast, Zhou and Susstrunk [208] focused on the convolution kernel mismatch between training and real-world camera blur. They introduced kernel modeling SR, which integrates blur kernel modeling directly into the training process to generate more accurate super-resolved images. To prevent structural distortions in SR, Ma et al. [209] proposed a structure-preserving GAN-based approach that maintains perceptual quality while minimizing unwanted distortions.

Another important advancement came from Maeda [210], who developed an unpaired SR method that eliminates the need for paired training data by leveraging GANs. Similarly, Umer and Micheloni [31] drew inspiration from CycleGAN [211], creating SRResCycGAN, which performs end-to-end translation from LR to HR domains without requiring paired data. Lee et al. [212] used NAS to design more efficient generator and discriminator structures, optimizing performance automatically.

Recent work has focused on further improving restoration quality under challenging conditions. Chan et al. [213] introduced a latent bank built from pretrained GANs, allowing for enhanced restoration even in extreme upscaling scenarios. Zhang et al. [214] proposed a robust model capable of handling diverse image degradations, including random shuffle blur, downsampling, and noise. Finally, Luo et al. [215] tackled the uncertainty in degradation by modeling the degradation process as a probabilistic distribution, enabling the system to learn the mapping from a prior random variable to the degradation process.

b) Introducing new losses: New loss functions have been introduced to enhance the performance of GAN-based SISR methods. Sajjadi et al. [216] proposed automated texture synthesis combined with perceptual loss to

produce more realistic images. Here,

$$\mathcal{L}_{\text{Tex.}} = \|G(\varphi(X_{\text{lr}})) - G(\varphi(X_{\text{hr}}))\|_2^2 \tag{40}$$

where the $\varphi(\cdot)$ represents the feature activation from a given VGG layer. The $G(\cdot)$ indicates the Gram matrix, and the result is calculated based on the patchwise level. Vasu et al. [217] combined distortion loss, perceptual loss, and adversarial loss to achieve less distortion and better perceptual quality. Rad et al. [218] explored the limits of perceptual loss and proposed a loss function that considered different semantic terms. It divided the object into three parts, which are background, boundary, and object, and calculated the loss, respectively. Mathematically, it can be formulated as

$$\begin{aligned}\mathcal{L}_{\text{Tar.}} = {} & \alpha \cdot \mathcal{G}_e\left(X_{\text{sr}} \circ M_{\text{OBB}}^{\text{boundary}}, X_{\text{hr}} \circ M_{\text{OBB}}^{\text{boundary}}\right) \\ & + \beta \cdot \mathcal{G}_b\left(X_{\text{sr}} \circ M_{\text{OBB}}^{\text{background}}, X_{\text{hr}} \circ M_{\text{OBB}}^{\text{background}}\right) \\ & + \gamma \cdot \mathcal{G}_o\end{aligned} \tag{41}$$

where α, β, and γ indicate the weight of corresponding loss terms, and $\mathcal{G}_e$, $\mathcal{G}_b$, and $\mathcal{G}_o$ represent the functions to calculate the space distances between two images. $M_{\text{OBB}}^{\text{background}}$ and $M_{\text{OBB}}^{\text{boundary}}$ mean the mask of each class of interest. $\circ$ denotes the pointwise multiplication. Kim et al. [219] combined SR and video frame interpolation, proposing a training strategy with multiscale temporal loss specific to the task. Fuoli et al. [220] introduced a loss function in the Fourier domain to recover missing high-frequency content and proposed a Fourier domain discriminator to better match the high-frequency distribution. Park et al. [221] designed a predict module to estimate the optimal combination of diverse losses and achieve superior perceptual quality. To better distinguish genuine high-frequency details from artifacts in image SR, WGSR [222] proposes training GAN-based models with wavelet-domain losses—where the discriminator operates solely on high-frequency wavelet subbands, and the generator is optimized via wavelet fidelity losses.

c) Introducing prior properties: Prior properties have been introduced in GAN-based SISR methods. Bulat et al. [223] enhance real-world image and face SR by employing a two-stage GAN framework that first learns to degrade HR images with unpaired data, and then trains a low-to-high GAN with paired data, achieving significant improvements over prior methods. cinCGAN [103] addresses unsupervised SISR in challenging scenarios with noise and blur by using a cycle-in-cycle network structure that combines GAN-based noise removal, upsampling, and end-to-end fine-tuning to achieve results comparable to state-of-the-art supervised models. KernelGAN [224] is an image-specific Internal-GAN that trains solely on the LR test image at test time, learning its internal patch distribution. Its generator is trained to downscale the LR image such that its discriminator cannot differentiate between the patch distributions of the downscaled and original LR images. This method determines the correct image-specific SR-kernel and enhances blind SR performance when integrated with existing SR algorithms. Soh et al. [50] focused on domain prior properties and defined a naturalness prior in the low-level domain, constraining the output image to the natural manifold to ensure high perceptual quality. Shaham et al. [36] introduced an unconditional generative model to learn the patch distribution from a single natural image. Instead of adding details from the LR image, Menon et al. [63] explored the HR image manifold and searched for images whose downscale version resembled the LR image. Pan et al. [225] leveraged the image prior captured by a GAN to restore missing semantics. Wang et al. [100] proposed a model to find indistinguishable feature maps and map them to the HR domain. Wei et al. [98] improve unsupervised SR by addressing domain gaps between synthetic and real LR images through domain-gap aware training and domain-distance weighted supervision, resulting in more realistic and natural SR outputs. SeD [226] employs a pretrained model to extract fine-grained semantic priors, enhancing the discriminator's capability through cross-attention mechanisms.

2) Diffusion-Based SISR: The inherent instability of GANs in SR tasks—particularly their tendency to produce artifacts and training difficulties—has prompted a paradigm shift toward diffusion models. These models offer distinct advantages in training stability and generation quality, leading to two principal research directions: retraining mechanisms and diffusion priors utilization.

a) Retraining mechanisms: The evolution of diffusion-based SR began with fundamental architectural innovations. SR3 [37] established the baseline approach by combining DDPMs with U-Net architectures, demonstrating superior performance over GANs through iterative multilevel refinement. Subsequent developments pursued three key improvements: 1) training efficiency; 2) computational optimization; and 3) architectural scalability. SR3+ [235] enhanced training robustness through noise-conditioning augmentation, while DiffIR [236] achieved computational efficiency via compact prior extraction. Resshift [38] further optimized the inference process through residual-based Markov chains, reducing required sampling steps to just 20. Parallel advancements in latent space operations emerged through LDM [237], which leveraged pretrained autoencoders [238] for efficient HR synthesis. The field subsequently progressed to multiscale architectures, with CDM [239] employing cascaded models for error-resistant generation and IDM [240] introducing implicit representations for continuous-scale adaptation.

b) Diffusion priors: The second major direction exploits pretrained diffusion models as powerful priors for SR, addressing three core challenges: blind

restoration, semantic consistency, and real-world degradation handling. StableSR [241] pioneered this approach through adaptive knowledge transfer from text-to-image models, establishing key components like time-aware encoders. Subsequent works expanded these foundations along two dimensions: task generalization and semantic integration. DiffBIR [242] developed a unified restoration pipeline with region-adaptive control, while PASD [243] demonstrated joint SR and stylization capabilities. The most recent innovations focus on semantic-aware processing: SUPIR [244] introduced prompt-guided restoration, and SeeSR [245] developed degradation-aware semantic prompting, effectively bridging low-level reconstruction with high-level image understanding. These approaches collectively demonstrate how diffusion priors enable unified frameworks that surpass traditional methods in both perceptual quality and task versatility.

C. Training Details

1) Datasets: There are five commonly used datasets for evaluating the performance of SISR methods: Set5 [247], Set14 [248], Urban100 [249], BSD100 [250], and Manga109 [251]. These datasets, widely adopted by the research community, serve as standard benchmarks in the field of SR. The training and evaluation protocols referenced in this work follow the common practices outlined in the previous literature, and the results discussed are based on established strategies rather than new, independent training experiments.

1) *Set5 [247]:* This dataset consists of five classic images, including baby, bird, butterfly, head, and woman.
2) *Set14 [248]:* Compared to Set5 [247], this dataset contains a wider range of categories. It includes 14 images, including both natural and manual images.
3) *Urban100 [249]:* This dataset comprises 100 images that feature various urban buildings. It includes challenging city landscapes with different frequency details.
4) *BSD100 [250]:* BSD100 consists of 100 images, including natural images as well as object-specific images such as plants, people, and food.
5) *Manga109 [251]:* Manga109 is a dataset containing 21 142 manga images drawn by 94 professional manga artists. The average size of the images is 833 $\times$ 1179 pixels. For SR, 109 images were selected from the dataset for testing purposes.

Regarding the training set, the main dataset used varies over time. In the early stages, methods used images extracted from ImageNet [252], T91 [253], and General100 [112] for training the models. After the release of DIV2K [254], almost all mainstream methods began using this dataset for training. To further improve model performance, Flicker2K [255] was also included in the training process.

1) *ImageNet [252]:* A large image dataset with over a thousand categories, commonly used for various computer vision tasks.
2) *T91 [253]:* T91 consists of 91 common natural images.
3) *General100 [112]:* This dataset contains 100 HD lossless images.
4) *DIV2K [254]:* The NTIRE 2017 competition dataset, comprising 1000 HD images with a resolution of 2k. Among them, 800 are used for training, and 100 for validation, covering diverse scenes, categories, and frequency details.
5) *Flicker2k [255]:* Another commonly used dataset, often combined with DIV2K to form DF2K, containing 2650 images with a resolution of 2k.

2) Training Settings:

a) Patch-based training: Given resource limitations and computational efficiency, training patches are extracted from the available training samples to serve as input–output pairs during the training process. The size of LR patches typically ranges from 16 $\times$ 16 to 92 $\times$ 92, depending on the model architecture and available resources.

b) Data augmentation: To enhance the diversity of training samples and improve model generalization, data augmentation techniques, such as random flipping, rotation, and color jittering, are applied to the extracted patches.

c) Optimization methods: Stochastic gradient descent (SGD) with momentum or the Adam optimizer is commonly used for training DL models. The hyperparameters, including learning rate, weight decay, and momentum, require careful tuning. Typically, the learning rate is set between 10^{-3} and 10^{-4}, depending on the network architecture and the specific method used.

d) Learning rate schedule: To facilitate efficient training, a learning rate schedule is employed to adjust the learning rate during the training process. Common schedules include step decay, learning rate warm-up [256], or cyclic learning rate [257].

e) Batch size: The batch size determines the number of samples processed together during each iteration of the training process. In light of memory constraints and computational complexity, batch sizes are commonly set between 16 and 32, striking a balance between training efficiency and memory consumption.

D. Experimental Results

We comprehensively evaluate the performance of regression-based and generative SISR methods through both objective metrics and visual comparisons.

1) Quantitative Analysis: Table 2 summarizes the PSNR and SSIM scores of various methods on synthetic test datasets (Set5 [247], Set14 [248], BSD100 [250], and Urban100 [249]), alongside model complexity (the number of parameters), training sets, architectures, venues,

Table 2 Performance Comparison of Different SISR Models on Four Benchmarks for the ×4 task. PSNR/SSIM on the Y Channel Is Reported for Each Dataset. The **Best** and Second Best Performances Are Indicated for Models Trained on DIV2K, and the **Best** and Second Best Performances Are Indicated for Models Trained on DF2K. The Results for Each Model Are Obtained From Their Respective Original Articles

Model	Params (K)	Set5 (PSNR/SSIM)	Set14 (PSNR/SSIM)	BSD100 (PSNR/SSIM)	Urban100 (PSNR/SSIM)	Training Sets	Backbone	Venue
SRCNN [161]	24	30.48/0.8628	27.50/0.7513	26.90/0.7103	24.52/0.7226	T91+ImageNet	CNN	ECCV'14
SCN [107]	-	30.86/0.8732	27.64/0.7578	27.03/0.7161	-	T91	CNN	ICCV'15
FSRCNN [112]	12	30.70/0.8657	27.59/0.7535	26.96/0.7128	24.60/0.7258	T91+General-100	CNN	ECCV'16
ESPCN [114]	25	30.52/0.8697	27.42/0.7606	26.87/0.7216	24.39/0.7241	T91+ImageNet	CNN	CVPR'17
VDSR [159]	665	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	BSD+T91	CNN	CVPR'17
DRRN [126]	207	31.68/0.8888	28.21/0.7720	27.38/0.7284	25.44/0.7638	BSD+T91	CNN	CVPR'17
LapSRN [160]	812	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	BSD+T91	CNN	CVPR'17
SRDenseNet [30]	5452	32.02/0.7819	28.50/0.7782	27.53/0.7337	26.05/0.7819	ImageNet	CNN	ICCV'17
SRGAN [6]	-	-	26.02/0.7379	25.16/0.6688	-	ImageNet	GAN	CVPR'17
IDN [120]	678	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	BSD+T91	CNN	CVPR'18
DSRN [128]	1200	31.40/0.8830	28.07/0.7700	27.25/0.7240	25.08/0.7470	T91	CNN	CVPR'18
ZSSR [101]	225	31.13/0.8796	28.01/0.7651	27.12/0.7211	-	-	CNN	CVPR'18
ESRGAN [97]	-	30.44/0.8520	-	25.29/0.6500	24.37/0.7340	DF2K+OST [35]	GAN	CVPRW'18
EDSR [161]	43000	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	DIV2K	CNN	CVPRW'17
RDN [181]	21900	32.47/0.8990	28.81/0.7871	27.72/0.7419	26.61/0.8028	DIV2K	CNN	CVPR'18
MSRN [227]	6300	32.07/0.8903	28.60/0.7751	27.52/0.7273	26.04/0.7896	DIV2K	CNN	ECCV'18
RCAN [28]	16000	32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087	DIV2K	CNN	ECCV'18
SRFeat [204]	6189	32.27/0.8938	28.71/0.7835	27.64/0.7378	-	DIV2K	GAN	ECCV'18
RDN-MetaSR [177]	-	32.38/-	28.78/-	27.71/-	26.55/-	DIV2K	CNN	CVPR'19
OISR [228]	15592	32.53/0.8992	28.86/0.7878	27.75/0.7428	26.79/0.8068	DIV2K	CNN	CVPR'19
SAN [168]	15700	32.64/0.9003	28.92/0.7888	27.78/0.7436	26.79/0.8068	DIV2K	CNN	CVPR'19
EBRN [157]	7900	**32.79**/**0.9032**	**29.01**/0.7903	**27.85**/**0.7464**	27.03/0.8114	DIV2K	CNN	ICCV'19
NatSR [50]	4800	30.98/0.8606	27.42/0.7329	26.44/0.6827	25.46/0.7602	DIV2K	GAN	CVPR'19
LatticeNet [117]	777	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873	DIV2K	CNN	ECCV'20
HAN [169]	64199	32.64/0.9002	28.90/0.7890	27.80/0.7442	26.85/0.8094	DIV2K	CNN	ECCV'20
RFANet [184]	11000	32.66/0.9004	28.88/0.7894	27.79/0.7442	26.92/0.8112	DIV2K	CNN	CVPR'20
CSNLN [104]	30000	32.68/0.9004	28.95/0.7888	27.80/0.7439	27.12/**0.8168**	DIV2K	CNN	CVPR'20
LAPAR [183]	-	32.57/0.8998	28.96/0.7908	27.84/0.7447	27.04/0.8128	DIV2K	CNN	NeurIPS'20
TPSR-D2 [212]	610	29.60/-	26.88/-	26.23/-	24.12/-	DIV2K	GAN	ECCV'20
AdderSR [229]	43000	32.13/0.8864	28.57/0.7800	27.58/0.7368	26.33/0.7874	DIV2K	CNN	CVPR'21
LUT [133]	1274	29.82/0.8478	27.01/0.7355	26.53/0.6953	24.02/0.6990	DIV2K	CNN	CVPR'21
SMSR [25]	1006	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	DIV2K	CNN	CVPR'21
RDN-LIIF [141]	-	32.50/-	28.80/-	27.74/-	26.68/-	DIV2K	CNN	CVPR'21
NLSN [171]	-	32.59/0.9000	28.87/0.7891	27.78/0.7444	26.96/0.8109	DIV2K	CNN	CVPR'21
NAPS [26]	125	31.93/0.8906	28.42/0.7763	27.44/0.7307	25.66/0.7715	DIV2K	CNN	ICCV'21
FAD-RCAN [156]	-	32.65/0.9007	28.88/0.7889	27.78/0.7437	26.86/0.8092	DIV2K	CNN	ICCV'21
ArbRCAN [178]	16600	32.55/-	28.87/-	27.76/-	26.68/-	DIV2K	CNN	ICCV'21
SPSR [230]	-	31.04/0.8772	27.07/**0.8076**	26.05/0.6818	25.23.0.9531	DIV2K	GAN	TPAMI21
EDSR-SLS [131]	363	-	28.49/-	27.51/-	25.84/-	DIV2K	CNN	CVPR'22
SPLUT [135]	18000	30.52/0.8630	27.54/0.7520	26.87/0.7090	24.46/0.7190	DIV2K	CNN	ECCV'22
RCLUT [136]	1513	30.72/0.8677	27.67/0.7577	26.95/0.7145	24.57/0.7253	DIV2K	CNN	ICCV'23
ELAN [189]	8312	32.75/0.9022	28.96/0.7914	27.83/0.7459	**27.13**/0.8167	DIV2K	Transformer	ECCV'22
ESRT [188]	751	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	DIV2K	Transformer	CVPRW'22
Omni-SR [198]	792	32.49/0.8988	28.78/0.7859	27.71/0.7415	26.64/0.8018	DIV2K	Transformer	CVPR'23
CRAFT [96]	753	32.52/0.8989	28.85/0.7872	27.72/0.7418	26.56/0.7995	DIV2K	Transformer	ICCV'23
DLGSANet [190]	761	32.54/0.8993	28.84/0.7871	27.73/0.7415	26.66/0.8033	DIV2K	Transformer	ICCV'23
SRFormer [186]	873	32.51/0.8988	28.82/0.7872	27.73/0.7422	26.67/0.8032	DIV2K	Transformer	ICCV'23
CAL-GAN [231]	-	31.18/0.8630	-	25.93/0.6760	25.29/0.7630	DIV2K	GAN	ICCV'23
WGSR [222]	-	31.51/0.8690	26.69/0.7160	26.37/0.6840	25.61/0.7770	DIV2K	GAN	CVPR'24
DBPN [172]	10000	32.47/0.8980	28.82/0.7860	27.72/0.7400	26.38/0.7946	DF2K	CNN	CVPR'18
SRCliqueNet [158]	659	32.15/0.8944	28.61/0.7818	27.61/0.7366	26.14/0.7871	DF2K	CNN	NeurIPS'19
SRFBN [21]	3500	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	DF2K	CNN	CVPR'19
DRN [175]	9800	32.74/0.9020	28.98/0.7920	27.83/0.7450	27.03/0.8130	DF2K	CNN	CVPR'20
USRGAN [232]	-	30.91/0.8660	-	25.97/0.6760	24.89/0.7500	DF2K	GAN	CVPR20
DFSA [154]	-	32.79/0.9019	29.06/0.7922	27.87/0.7458	27.17/0.8163	DF2K	CNN	ICCV'21
CRAN [233]	19940	32.72/0.9012	29.01/0.7918	27.86/0.7460	27.13/0.8167	DF2K	CNN	ICCV'21
ASSLN [130]	677	32.29/0.8964	28.69/0.7844	27.66/0.7384	26.27/0.7907	DF2K	CNN	NeurIPS'21
DCLS [152]	-	32.12/0.8890	28.54/0.7728	27.60/0.7285	26.15/0.7809	DF2K	CNN	CVPR'22
SwinIR [5]	11900	32.92/0.9044	29.09/0.7950	27.92/0.7489	27.45/0.8254	DF2K	Transformer	ICCVW'21
CAT [27]	16600	33.08/0.9052	29.18/0.7960	27.99/0.7510	27.89/0.8339	DF2K	Transformer	NeurIPS'22
SwinIR+LDL [234]	-	31.03/0.8611	27.53/0.7478	-	26.23/0.7918	DF2K	GAN	CVPR22
HAT [195]	20800	33.04/0.9056	29.23/0.7973	28.00/0.7517	27.97/0.8368	DF2K	Transformer	CVPR'23
EQSR [196]	-	32.71/-	29.12/-	27.86/-	27.30/-	DF2K	Transformer	CVPR'23
DAT [200]	11212	33.08/0.9055	29.23/0.7973	28.00/0.7515	27.87/0.8343	DF2K	Transformer	ICCV'23
SAFMN [191]	240	32.18/0.8948	28.60/0.7813	27.58/0.7359	25.97/0.7809	DF2K	Transformer	ICCV'23
CFAT [201]	22070	**33.19**/ **0.9068**	**29.30**/**0.7985**	**28.17**/**0.7524**	**28.11**/**0.8380**	DF2K	Transformer	CVPR'24

Table 3 Performance Comparison of Regression-Based and Generation-Based Models for the ×4 Task, With the Best Results for Each Category Highlighted in Red (Regression) and Blue (Generation)

Model	PSNR↑	SSIM↑	LPIPS↓	FID↓	Venue
StableSR [241]	19.11	0.559	0.171	54.77	IJCV'24
BSRGAN [214]	21.04	0.616	0.202	71.99	ICCV'21
Real-ESRGAN [246]	20.46	0.612	0.190	66.45	ICCVW'21
SeeSR [245]	20.55	0.596	0.168	51.38	CVPR'24
SUPIR [244]	18.08	0.512	0.229	70.06	CVPR'24
RRDB+LDL [234]	**23.32**	**0.717**	**0.098**	**23.05**	CVPR'22
EDSR [161]	23.90	0.737	0.204	46.60	CVPR'17
DRN [175]	24.80	0.768	0.171	40.14	CVPR'20
SwinIR [5]	24.34	0.754	0.191	44.56	ICCVW'21
CAT [27]	25.38	0.791	0.154	37.31	NeurIPS'22
SRFormer [186]	24.49	0.759	0.186	43.87	ICCV'23
CRAFT [96]	24.41	0.757	0.186	43.96	ICCV'23
HAT [195]	**26.46**	**0.817**	**0.134**	**32.83**	CVPR'23

and publication years. The results are sourced from the respective articles. We also split the comparison within the same training set to ensure fair evaluation.

Due to the synthetic nature of these datasets, generative models are underrepresented; thus, we include an additional Table 3 to explicitly compare regression and generative methods. Specifically, we randomly crop 512 × 512 HR patches from Urban100 (100 samples total) and generate corresponding LR images using bicubic ×4 downsampling. We evaluate six generation-based methods [214], [234], [241], [244], [245], [246] and seven regression-based methods [5], [27], [96], [161], [175], [186], [195] using both pixelwise (PSNR and SSIM) and perceptual metrics (LPIPS and FID).

From the tables, several key observations emerge as follows.

1) First, DL-based techniques consistently outperform the bicubic interpolation method, with a notable margin of at least 2 dB in terms of PSNR. This highlights the effectiveness of DL in enhancing image resolution and generating higher quality results.
2) The use of larger training datasets, such as DIV2K and Flickr2K, has played a significant role in improving the performance of SISR systems. The availability of more extensive and diverse datasets has allowed models to learn a wider range of image features and textures, resulting in better generalization and enhanced performance.
3) Analyzing the timeline of SISR developments, there has been a noticeable shift from CNN-based architectures to Transformer-based solutions. Transformer architectures have gradually become dominant in SISR tasks due to their ability to effectively capture long-range dependencies. Specifically, they calculate attention scores based on the entire feature map, using dimensions $(HW) \times C$, where H and W represent the height and width of the feature map, respectively, and C denotes the number of channels. However, it is important to note that achieving better performance often requires larger model parameters and increased computational resources (e.g., $\mathcal{O}((HW)^2C)$, where $(HW) \gg C$). Balancing the tradeoff between performance and complexity remains an ongoing challenge in SISR research.
4) Generation-based methods exhibit lower PSNR/SSIM but can produce visually clearer results, though with occasional fidelity loss. Interestingly, they do not always achieve better perceptual metrics (LPIPS/FID), likely due to hallucination effects visible in the super-resolved outputs.

2) Qualitative Analysis: Visual comparisons (Figs. 15 and 16) further demonstrate these trends.

From Fig. 15, the visual comparison demonstrates the progressive improvements in regression-based SR methods across different architectural paradigms. The first row (CNN-based methods: EDSR [161], HAN [169], and DRN [175]) shows competent reconstruction but with visible limitations in fine detail recovery and occasional blurring of complex textures. The second row (lightweight Transformers: SwinIR [5], CRAFT [96], and SRFormer [186]) exhibits noticeable improvements in texture preservation and edge sharpness, benefiting from more efficient attention mechanisms. The bottom row (standard Transformers: CAT [27], DAT [200], and HAT [195]) achieves the best reconstruction quality, particularly in: 1) faithfully recovering high-frequency details and 2) producing the most natural-looking results with minimal artifacts. All methods significantly outperform the bicubic baseline, with standard Transformers demonstrating superior capability in handling challenging patterns while maintaining computational efficiency. The progression from CNNs to lightweight and then standard Transformers illustrates how architectural advances have addressed traditional SR challenges, though with increasing model complexity.

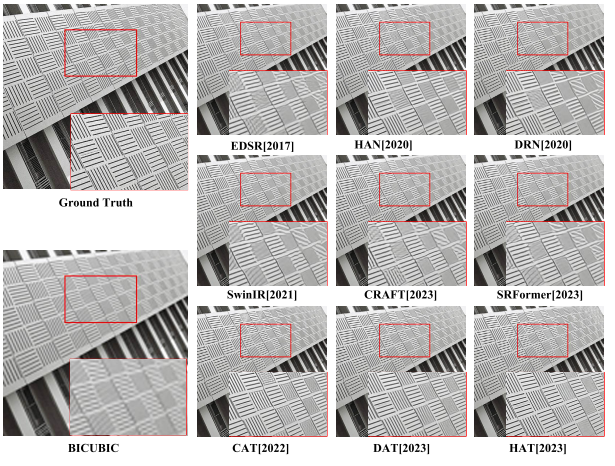

Fig. 15. *Visual comparison of regression-based SR methods: CNN-based (top), lightweight Transformers (middle), and standard Transformers (bottom). Standard Transformers achieve the best reconstruction quality.*

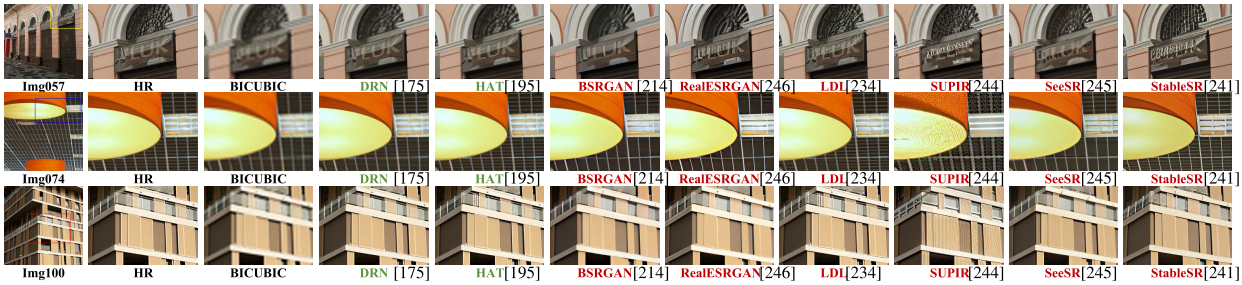

Fig. 16. *Visual comparison between regression-based and generation-based SR methods on sample images. Regression methods include CNN and Transformer approaches, while generation methods comprise GAN and diffusion-based techniques. While generation-based methods can generate clearer and more detailed context, they may easily compromise fidelity. On the contrary, regression-based methods can produce more reliable output but are often blurrier. Choosing which type of method depends on the purpose.*

From Fig. 16, the comparison demonstrates distinct characteristics between regression-based (DRN [175] and HAT [195]) and generation-based (LDL [234], SUPIR [244], SeeSR [245], and StableSR [241]) approaches. Regression methods show more conservative reconstructions that maintain better fidelity to original structures, particularly in preserving geometric patterns. Generation-based methods produce visually sharper outputs but exhibit more variations in texture synthesis and occasional artifacts in complex areas. For example, Img057 shows generative methods (especially diffusion models) generating wrong characters, while Img074 reveals color shift problems in generative outputs.

Overall, CNN/Transformer regression methods appear more consistent across different image samples, while GAN/diffusion approaches show greater variability in their enhancement styles. The results validate Transformer-based approaches' effectiveness in capturing long-range dependencies for image restoration tasks.

IV. VIDEO SUPER RESOLUTION

A. Regressive Models

1) CNN-Based VSR:

a) Efficient structural designs: Several studies have focused on lightweight network designs to reduce model complexity, as illustrated in Fig. 17. These approaches can be categorized into two primary types: recurrent modeling and convolution modeling. For recurrent modeling, BRCN [259] sets a foundation by utilizing a bidirectional recurrent convolutional network for multiframe VSR, effectively modeling temporal dependencies without the high computational cost of traditional RNNs. It replaces recurrent full connections with weight-sharing convolutional connections, enabling it to handle complex motions with reduced computational overhead. Building on this, STCN [260] further enhanced the modeling by combining deep spatial encoding with multiscale temporal components, integrating both intraframe visual patterns and interframe dynamics for higher quality reconstructions.

To improve efficiency in frame-based VSR, Sajjadi et al. [261] introduced a frame-recurrent approach, which focuses on reusing previously inferred HR frames to super-resolve the next, significantly reducing redundancy in processing consecutive frames. Similarly, RISTN [262] combined spatial and temporal modeling with a novel architecture that incorporates a lightweight residual invertible block for minimizing information loss and a recurrent convolutional model with dense connections to prevent feature degradation. Haris et al. [263] took this further by using a recurrent encoder–decoder module to integrate spatial and temporal contexts across continuous frames, demonstrating better performance for complex video sequences.

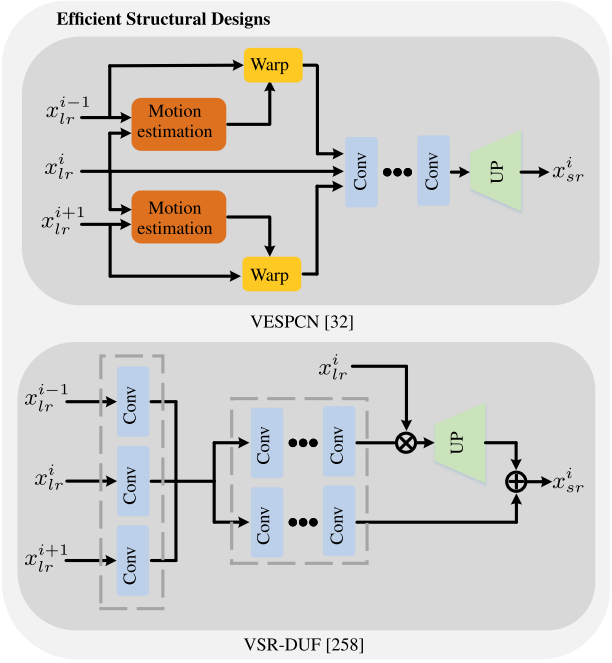

Fig. 17. *Illustration of efficient modification. Gray dashed line indicates that all convolutional layers are shared weights. ⊕ indicates the elementwise addition. ⊗ indicates the matrix product.*

Efforts to further improve computational efficiency led to RLSP [264], which uses high-dimensional latent states to propagate temporal information without the need for heavy motion compensation, achieving significant speedups over traditional methods like DUF, with a 70× improvement in runtime. Isobe et al. [265] refined this by separating input frames into structure and detail components, processing them through a two-stream recurrent unit, and using a hidden state adaptation module to mitigate error accumulation, enhancing robustness against appearance changes. More recently, Xia et al. [266] explored structured sparsity, pruning key components such as residual and recurrent blocks, resulting in even more efficient models with minimal performance loss.

For convolution modeling, early work by Caballero et al. [32] introduced spatiotemporal subpixel convolution networks, which effectively leverage temporal redundancies to improve both speed and accuracy. Building on this, RTVSR [267] proposed a real-time VSR method that uses 1-D motion convolution kernels and an enhanced gated residual unit (GEU), achieving significant gains over state-of-the-art methods while maintaining real-time performance. FFCVSR [268] further advanced this by integrating a local network for processing consecutive frames with a context network that leverages previously estimated HR frames, leading to temporally consistent, high-quality results.

In the realm of feature alignment, TDAN [269] introduced deformable convolutions to adaptively align frames at the feature level without relying on optical flow, enabling better SR results with a more compact network architecture. Jo et al. [258] built on this by using dynamic upsampling filters to compute residual images based on the local spatiotemporal neighborhood, eliminating the need for explicit motion compensation. Zhu et al. [262] added to these innovations by incorporating a residual invertible block for more efficient spatial information extraction, alongside a sparse fusion strategy for better integration of spatial–temporal features.

Recent work has also seen the exploration of more specialized architectures. Li et al. [270] introduced a fast spatiotemporal residual network using 3-D convolutions to improve VSR performance with low computational cost. Similarly, Yi et al. [271] proposed a progressive fusion network to better utilize spatiotemporal information across consecutive frames. Isobe et al. [272] took a hierarchical approach by splitting input sequences into groups based on frame rates, enhancing temporal information processing. Meanwhile, Pan et al. [273] developed a blind VSR method that jointly estimates blur kernels, motion fields, and latent images, demonstrating its effectiveness in various real-world scenarios. Finally, Li et al. [274] introduced a content-adaptive approach that applies different SR methods to different video patterns, achieving a better balance between performance and efficiency.

b) Learning better representations: VSRnet [276] introduces a CNN trained on both spatial and temporal dimensions of videos to enhance VSR, using motion-compensated frames and exploring various frame-combining architectures, demonstrating that pretraining with images allows a smaller video database to effectively train the model and achieve superior performance compared to current video and image SR algorithms. SOFVSR [277] introduces an end-to-end trainable VSR framework that enhances both image quality and optical flows, leveraging HR optical flows for accurate motion compensation and improved correspondence, thereby achieving superior accuracy and consistency in super-resolved video sequences, as validated by extensive testing on benchmark datasets. MMCNN [278] integrates an optical flow network and an image-reconstruction network with convolutional long short-term memory embedded in residual blocks to effectively utilize both spatial and temporal correlations between LR frames, demonstrating superior performance in terms of PSNR and visual quality across various scaling factors in extensive experiments. RRCN [279] leverages motion compensation, deep recurrent convolutional layers, and late fusion to effectively utilize temporal information, incorporates residual connections for enhanced accuracy, and employs a novel model ensemble strategy with a single-image SR method, demonstrating superior performance over existing state-of-the-art SR methods in quantitative visual quality assessments. MultiBoot VSR [280] employs a unique structure comprising a motion compensation-based input subnetwork, a blending backbone, and a spatial upsampling subnetwork, applied recurrently to enhance image quality progressively by reshuffling HR frames into multiple LR images, demonstrating the ability to generate temporally consistent and high-quality results without artifacts. DNLN [281] introduced deformable convolution to adaptively align features and a nonlocal structure to capture global correlations, further enhanced by residual in residual dense blocks to fully utilize hierarchical features, achieving state-of-the-art performance on benchmark datasets. The 3DSR-net [282] leverages spatiotemporal feature maps without the need for motion alignment preprocessing, maintaining temporal depth to capture the dynamic relationship between LR and HR frames, and incorporates residual learning with subpixel outputs, significantly outperforming state-of-the-art methods on the Vidset4 benchmark and addressing performance drops due to scene changes. Tao et al. [283] highlighted the importance of proper frame alignment and motion compensation and proposed a subpixel motion compensation layer in a CNN framework. Xue et al. [284] proposed a task-oriented framework to produce specific motion estimation and train with video processing simultaneously. Kim et al. [219] proposed a multiscale temporal loss to regularize spatiotemporal resolution. Chan et al. [285] introduced an offset-fidelity loss for deformable alignment in VSR to control offset learning. Isobe et al. [286] considered the temporal relation by computing the difference between frames and processing these pixels into two branches based on the

level of difference. Chen et al. [287] extended fixed-scale to arbitrary-scale VSR using implicit neural representation. Chiche et al. [288] addressed high-frequency artifacts caused by recurrent VSR models using Lipschitz stability theory to generate stable and competitive results. Chan et al. [289] designed a precleaning stage to reduce noise and artifacts caused by long-term propagation cumulative error. They also proposed a stochastic degradation scheme for a better tradeoff between speed and performance and employed longer sequences instead of large batches to allow more effective use of temporal information. MFPI [290] introduces a novel multifrequency representation enhancement (MFE) module that aggregates spatial–temporal information in the frequency domain and employs privilege training to efficiently utilize HR video data, significantly surpassing existing state-of-the-art VSR models in both performance and efficiency across multiple datasets.

c) Improving information interaction: Compared to the previous approach of extracting features in the spatial domain and performing motion compensation in the temporal domain, which can lead to a lack of joint utilization of spatiotemporal information and thus affect SR performance, D3DNet [291] leverages deformable 3-D convolutions to fully exploit the spatiotemporal dependencies of video frames, thereby enhancing VSR performance. Wang et al. [277] super-resolved optical flows and images simultaneously to obtain better performance. Yan et al. [268] utilized inter-LR frames and previous SR frames to produce temporally consistent high-quality results. Li et al. [292] proposed a temporal multicorrespondence aggregation strategy to leverage similar patches across frames and a cross-scale nonlocal correspondence aggregation scheme to explore self-similarity of images across scales. Chan et al. [275] reconsidered essential components for VSR and proposed a succinct pipeline, as shown in Fig. 18, BasicVSR, that achieved appealing improvements in speed and restoration quality compared to many state-of-the-art algorithms. Liu et al. [293] proposed a deep neural network with dual subnet and multistage communicated upsampling for VSR with large motion. Yi et al. [294] introduced a hybrid structure that utilized preceding, present, and future SR information. Li et al. [295] focused on compression VSR and proposed bidirectional recurrent warping, detail-preserving flow estimation, and Laplacian enhancement modules to handle compression properties. Yu et al. [296] proposed cross-frame nonlocal attention to alleviate performance drops in large motion scenes and introduced a memory-augmented attention module to compensate for information loss caused by large motions. Cao et al. [297] formulated space-time video super resolution (STVSR) as a joint deblurring, interpolation, and SR problem and introduced the Fourier transform and recurrent enhancement layers to recover high-frequency details. Wang et al. [298] introduced metadata to predict the compression level combined with an RNN-style decoder to produce pleasing results.

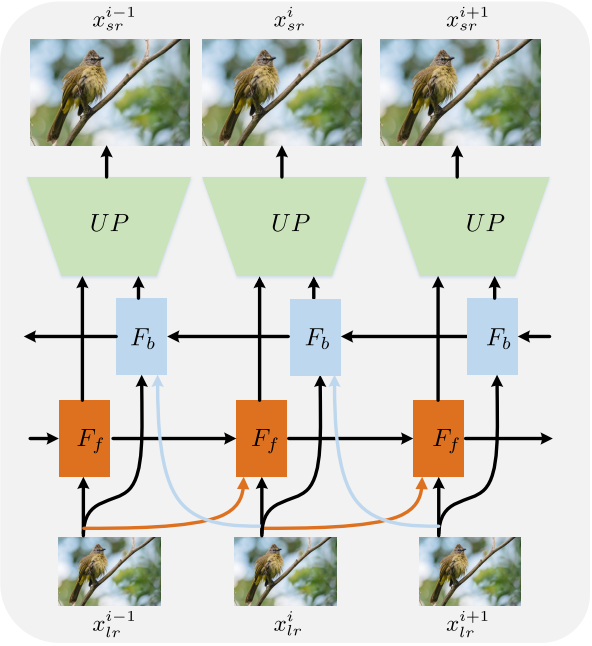

Fig. 18. ***BasicVSR [275] framework comprehensively examined the four fundamental components of VSR, namely, propagation, alignment, aggregation, and upsampling, thereby establishing a robust and comprehensive foundation for future advancements in the field of VSR.***

2) Transformer-Based VSR: Cao et al. [305] made the first attempt to adopt a Transformer for VSR. In addition, a spatial–temporal convolutional self-attention layer is proposed to exploit locality information. Liang et al. [306] proposed a video restoration Transformer with parallel frame prediction and long-range temporal dependence modeling abilities. Wang et al. [307] designed a deformable attention network and proposed a long–short term feature interpolation module, where abundant information from more neighboring frames is explored for the interpolation process of missing frame features. Geng et al. [308] addressed the slow inference speed issue by using a spatial–temporal Transformer that naturally incorporates spatial and temporal SR modules into a single model. PSRT [308] introduces an efficient patch alignment method that aligns image patches rather than pixels, enabling state-of-the-art performance across multiple benchmarks and offering new insights into the optimal use of multiframe information and alignment techniques in VSR. FTVSR++ [309] introduces a degradation-robust approach to VSR by implementing self-attention within a combined space–time–frequency domain, featuring a novel dual-frequency attention mechanism and divided attention strategy that effectively distinguish real textures from artifacts and addresses complex degradations, significantly enhancing video quality across various challenging conditions. RVRT [299] proposed clip-based feature estimation and guided deformable attention for

Fig. 19. *Visual comparison of regression-based and generation-based VSR methods on selected frames from a video sequence. The bicubic upsampling result is included as a baseline reference. Notably, regression-based methods like RVRT [299] and MIAVSR [300] tend to produce more faithful and recognizable text ("ZABITA 153"), even if slightly less sharp. In contrast, generation-based methods such as MGLDVSR [301] and UAV [302], while attempting to generate sharper details, often introduce significant artifacts or hallucinate incorrect content, as seen with the distorted or altered text in their outputs.*

clip alignment, effectively balancing model size, effectiveness, and efficiency, and achieving state-of-the-art performance in VSR, deblurring, and denoising on benchmark datasets. TTVSR [310] leverages prealigned spatiotemporal trajectories and a cross-scale feature tokenization module to efficiently model long-range dependencies and address scale variations in video sequences, outperforming state-of-the-art models in extensive benchmarks through its novel approach to reducing computational costs and enhancing the utilization of temporal information.

B. Generative Models

1) GAN-Based VSR: Lucas et al. [311] designed a new discriminator architecture optimized for the VSR problem and introduced distance losses in feature-space and pixel-space to regularize GAN training. Thawakar et al. [312] incorporated a recurrent mechanism into the GAN model to reduce the parameter count as the model depth increases. Yi et al. [313] proposed a hybrid attention structure to capture local and nonlocal spatiotemporal correlations. They also employed generative adversarial training to mitigate temporal artifacts and generate more realistic results. Chadha et al. [314] proposed a recurrent back-projection network as a generator to extract spatial and temporal information, and they combined multiple losses to capture fine details. Cao et al. [315] achieved a 7.92× performance speedup compared to state-of-the-art methods through BN layer fusion, convolutional layer acceleration, and efficient upsampling methods. Andrei et al. [316] focused on low bitrate streams and proposed a two-stage VSR strategy. The first stage removed compression artifacts, and the second stage generated detailed information using a GAN model with perceptual metrics. Wang et al. [317] introduced the Laplacian edge module to enhance intermediate results generated by the GAN model and added perceptual loss to obtain better visual quality. Chu et al. [318] introduced a temporal learning strategy to supervise both spatial and temporal relationships in VSR. They also proposed a Ping-Pong loss to alleviate temporal artifacts and obtain more natural super-resolved images. Wang et al. [319] explored spatial information and introduced a spatially adaptive loss to enhance GAN model performance. López-Tapia et al. [320] introduced the amortized maximum a posteriori (MAP) approximation to simulate multiple degradations and designed a GAN loss to eliminate artifacts. LongVideoGAN [321] enhances long-term temporal consistency and dynamic content creation by redesigning the temporal latent representation and utilizing a two-phase training strategy with varying video lengths and resolutions.

2) Diffusion-Based VSR: To address the dual challenges of spatial fidelity and temporal consistency in VSR, Zhou et al. [302] introduces a text-guided latent diffusion framework for video upscaling that ensures temporal coherence by integrating temporal layers into its architecture and employing a flow-guided recurrent latent propagation module, offering enhanced video stability and flexibility in texture creation through text prompts and adjustable noise levels, thereby achieving superior visual realism and consistency as demonstrated across various benchmarks. Yuan et al. [333] utilize an efficient diffusion-based text-to-VSR tuning approach that expands the

Table 4 Performance Comparison of Different VSR Models on the Vid [322], Vimeo-90K [284], and REDS [323] Datasets for BI Degradation, and on the Vid [322], Vimeo-90K [284], and UDM10 [271] Datasets for BD degradation, All Under the ×4 Task. PSNR/SSIM Values Are Reported for the Y Channel Across Each Dataset. "Params" Denotes the Total Number of Network parameters. The **Best** and Second Best Performances Are Highlighted for Models Trained on Vimeo-90K [284], While the **Best** and Second Best Performances Are Highlighted for Models Trained on REDS [323] + Vimeo-90K [284]. The Results for Each Model Are Obtained From Their Respective Original Articles

Methods	Params (M)	BI			BD			Training Sets	Backbone	Venue
		Vid4 [322]	Vimeo-90K [284]	REDS [323]	Vid [322]	Vimeo-90K [284]	UDM10 [271]			
Bicubic	-	23.78/0.6347	31.32/0.8684	26.14/0.7292	21.80/0.5246	31.30/0.8687	28.47/0.8253	-	-	-
BRCN [259]	-	-	-	-	24.43/0.6334	-	-	YUV25	CNN	NeurIPS'15
VSRnet [276]	0.27	24.84/0.7049	-	-	-	-	-	Myanmar	CNN	TCI'16
RRCN [279]	-	25.86/0.7591	-	-	-	-	-	Myanmar	CNN	TIP'19
VESPCN [32]	0.88	25.35/0.7557	-	-	-	-	-	CDVL	CNN	CVPR'17
SOFVSR [277]	1.71	26.01/0.7710	-	-	26.19/0.7850	-	-	CDVL	CNN	ACCV'18
RTVSR [267]	15.00	26.36/0.7900	-	-	-	-	-	harmonicinc.com	CNN	Neurocomp.'19
MultiBoot VSR [280]	60.86	-	-	31.00/0.8822	-	-	-	harmonicinc.com	CNN	CVPRW'19
VSRResFeatGAN [311]	-	25.51/0.7530	-	-	-	-	-	Myanmar	GAN	TIP'19
FFCVSR [268]	-	26.97/0.8300	-	-	-	-	-	Venice+Myanmar	CNN	AAAI'19
3DSRnet [282]	0.11	25.71/0.7588	-	-	-	-	-	largeSet	CNN	ICIP'19
DSMC [293]	11.58	27.29/0.8403	-	30.29/0.8381	-	-	-	REDS	CNN	AAAI'21
DRVSR [283]	2.17	25.52/0.7600	-	-	-	-	-	-	CNN	ICCV'17
DUF [258]	-	27.33/0.8251	-	28.63/0.8251	27.38/0.8392	36.87/0.9447	38.48/0.9605	-	CNN	CVPR'18
MMCNN [278]	10.58	26.28/0.7844	-	-	-	-	-	-	CNN	TIP'19
TecoGAN [318]	3.00	-	-	-	25.89/-	-	-	-	GAN	TOG'20
GOVSR [294]	-	-	-	-	28.41/0.8724	37.63/0.9503	30.14/0.9713	-	CNN	ICCV'21
PSRT [303]	13.4	28.07/0.8485	38.27/0.9536	32.72/0.9106	-	-	-	-	Transformer	NeurIPS22
StableVSR [39]	-	-	31.97/0.8770	27.97/0.8000	-	-	-	-	DDPM	ECCV'24
FRVSR [261]	5.1	-	-	-	26.69/0.8220	35.64/0.9319	37.09/0.9522	Vimeo-90K	CNN	CVPR'18
DNLN [281]	19.74	27.31/0.8257	-	-	-	-	-	Vimeo-90K	CNN	ACCESS'19
TOFlow [284]	1.4	25.89/0.7651	33.08/0.9054	27.98/0.7990	25.85/0.7695	34.62/0.9212	36.26/0.9438	Vimeo-90K	CNN	IJCV'19
RLSP [264]	4.21	27.55/0.8380	-	-	27.48/0.8388	36.49/0.9403	-	Vimeo-90K	CNN	ICCVW'19
PFNL [271]	3.0	26.73/0.8029	36.14/0.9363	29.63/0.8502	27.16/0.8355	-	38.74/0.9627	vimeo-90K	CNN	ICCV'19
RBPN [263]	12.2	27.12/0.8180	37.07/0.9435	30.09/0.8590	27.17/0.8205	37.20/0.9458	38.66/0.9596	Vimeo-90K	CNN	CVPR'19
MEMC-Net [324]	-	24.37/0.8380	33.47/**0.9470**	-	-	-	-	Vimeo-90K	CNN	TPAMI'19
RISTN [262]	3.67	26.13/0.7920	-	-	-	-	-	Vimeo-90K	CNN	AAAI'19
TGA [272]	5.8	27.59/0.8419	-	-	27.63/0.8423	**37.59/0.9516**	38.74/0.9627	Vimeo-90K	CNN	CVPR'20
TDAN [269]	1.97	26.24/0.7800	-	-	26.58/0.8010	-	-	Vimeo-90K	CNN	CVPR'20
STARnet [325]	111.61	-	30.83/0.9290	-	-	-	-	Vimeo-90K	CNN	CVPR'20
STVSR [326]	11.1	26.31/0.7976	-	25.27/0.8256	-	-	-	Vimeo-90K	CNN	CVPR'20
RSDN [265]	6.2	**27.79/0.8474**	37.05/0.9454	-	27.92/0.8505	37.23/0.9471	39.35/0.9653	Vimeo-90K	CNN	ECCV'20
MuCAN [292]	19.9	-	37.32/0.9465	30.88/0.8750	-	-	-	Vimeo-90K	CNN	ECCV'20
D3Dnet [291]	2.58	26.52/0.7990	-	-	-	-	-	Vimeo-90K	CNN	SPL'20
DNSTNet [327]	-	27.21/0.8220	36.86/0.9387	-	-	-	-	Vimeo-90K	CNN	Neurocomp.'20
MSFFN [328]	8.5	27.23/0.8218	**37.33**/0.9467	-	-	-	-	Vimeo-90K	CNN	TIP'21
RSTT [308]	7.67	26.43/0.7994	-	-	-	-	-	Vimeo-90K	Transformer	CVPR22
ETDM [286]	8.4	-	-	**32.15/0.9024**	**28.81/0.8725**	-/-	**40.11/0.9707**	Vimeo-90K	CNN	CVPR'22
EDVR [329]	20.60	27.35/0.8264	37.61/0.9489	31.09/0.8800	-	-	-	REDS+Vimeo-90K	CNN	CVPRW'19
RRN [330]	3.4	-	-	-	27.69/0.8488	-	38.96/0.9644	REDS+Vimeo-90K	CNN	BMVC'20
BasicVSR [275]	6.3	27.24/0.8251	37.18/0.9450	31.42/0.8909	27.96/0.8553	37.53/0.9498	39.96/0.9694	REDS+Vimeo-90K	CNN	CVPR'21
IconVSR [275]	8.7	27.39/0.8279	37.47/0.9476	31.67/0.8948	28.04/0.8570	37.84/0.9524	40.03/0.9694	REDS+Vimeo-90K	CNN	CVPR'21
BasicVSR++ [304]	7.3	27.79/0.8400	37.79/0.9500	32.39/0.9069	29.04/0.8753	38.21/0.9550	40.72/0.9722	REDS+Vimeo-90K	CNN	CVPR'22
RVRT [299]	10.8	27.99/0.8462	38.15/0.9527	32.75/0.9113	**29.54/0.8810**	38.59/0.9576	40.90/0.9729	REDS+Vimeo-90K	Transformer	NeurIPS'22
TTVSR [310]	6.8	-	-	32.12/0.9021	28.40/0.8643	37.92/0.9526	40.41/0.9712	REDS+Vimeo-90K	Transformer	CVPR'22
SSL [266]	1.0	27.15/0.8208	36.82/0.9419	31.06/0.8933	27.56/0.8431	37.06/0.9458	39.35/0.9665	REDS+Vimeo-90K	CNN	CVPR'23
MFPI [290]	7.3	28.11/0.8481	**38.28/0.9534**	**32.81**/0.9106	29.34/0.8781	**38.70/0.9579**	**41.08/0.9741**	REDS+Vimeo-90K	CNN	ICCV'23
FTVSR++ [309]	10.8	**28.80/0.8680**	-	32.42/0.9070	-	-	-	REDS+Vimeo-90K	Transformer	TPAMI'23
VideoGigaGAN [331]	369	26.78/-	35.97/-	30.46/-	27.04/-	35.30/-	36.57/-	REDS+Vimeo-90K	GAN	ArXiv'24
MIAVSR [300]	6.35	28.20/0.8507	38.22/0.9532	32.78/**0.9220**	-	-	-	REDS+Vimeo-90K	Transformer	CVPR'24
SATeCo [332]	-	27.44/0.8420	38.22/0.9532	31.62/0.8932	-	-	-	REDS+Vimeo-90K	DDPM	CVPR'24

capabilities of text-to-image SR models into video generation by inflating model weights and integrating a temporal adapter, enabling the creation of high-quality, temporally coherent videos with a balance between computational efficiency and SR performance, as validated by testing on the Shutterstock video dataset. StableVSR [39] employs diffusion models enhanced with a temporal conditioning module and framewise bidirectional sampling to synthesize realistic and temporally consistent details, achieving superior perceptual quality and temporal consistency in upscaled videos. MGLDVSR [301] integrates pretrained latent diffusion models with a motion-guided loss and a temporal module in the decoder, optimizing the latent sampling path to ensure content consistency and temporal coherence in HR video generation from LR inputs, significantly surpassing existing methods in perceptual quality on real-world VSR benchmarks. Imagen Video [334] uses a cascade of video diffusion models, including interleaved spatial and temporal SRs, to produce high-definition videos with high fidelity, controllability, and diverse stylistic and 3-D capabilities. SATeCo [332] proposes lightweight spatial feature adaptation (SFA) and temporal feature alignment (TFA) modules. These leverage pixelwise affine modulation and tubelet-based cross-attention, respectively, to guide denoising and reconstruction—while keeping the pretrained diffusion model frozen for efficiency.

C. Training Details

1) Datasets: The evaluation of VSR methods is conducted using the following datasets.

1) *REDS [323]:* The NTIRE19 challenge dataset, consisting of sequences with 100 frames each and a resolution of 720 × 1280.
2) *Vimeo-90K [284]:* A large-scale, high-quality video dataset comprising 89 800 video clips with a resolution of 448 × 256.
3) *Vid4 [322]:* This dataset includes four categories: walk, city, calendar, and foliage. The resolutions are 720×480 for foliage and walk, 720×576 for calendar, and 704 × 576 for city.
4) *UDM10 [271]:* A video dataset containing 522 video sequences for training, collected from ten videos.
5) *UDF [258]:* This dataset comprises 351 Internet videos with diverse content, including wildlife, activity, and landscape.
6) *SPMC-11 [283]:* It consists of 975 sequences from high-quality 1080-pixel HD video clips, mostly shot with high-end cameras and containing rich details in natural-world and urban scenes.

2) Training Settings:

a) Patch extraction: LR video frames are partitioned into fixed-size patches, such as 64 × 64, either nonoverlapping or with overlap. Overlapping patches are favored to encompass information near boundaries, enhancing the SR outcome.

b) Data augmentation: VSR employs spatial and temporal transformations on training data to boost diversity and generalization. Spatial transformations encompass random cropping, flipping, and rotation, while temporal transformations involve jittering, frame shuffling, and skipping.

c) Batch size and temporal span: The batch size typically ranges from 4 to 16 video sequences based on available GPU memory and computational resources. The temporal span, ranging from 5 to 15 frames, captures temporal coherence, ensuring sharper and temporally consistent HR video frames.

d) Learning schedule: A stepwise learning rate decay is often used, reducing the learning rate by a factor of 10 after a set number of epochs. For instance, a typical training schedule runs for 200 epochs, starting with an initial learning rate of 10^{-3} that is decayed to 10^{-4} at the 100th epoch.

D. Experimental Results

1) Quantitative Analysis: The performance of various VSR methods is systematically evaluated in Table 4 using PSNR and SSIM metrics under two degradation scenarios: bicubic (BI) and blur downsampling (BD). The results reveal several important trends. First, VSR models consistently demonstrate higher parameter counts than their SISR counterparts, reflecting the inherent complexity of temporal modeling. Second, model performance shows strong dependence on training data scale, with methods trained on larger datasets like REDS and Vimeo-90K achieving superior metrics. Most notably, Transformer-based architectures have established dominance in recent years, with their long-range attention mechanisms proving particularly effective for temporal modeling. However, this advantage comes with substantial computational costs that create deployment challenges for resource-constrained applications.

2) Qualitative Analysis: Visual comparisons in Fig. 19 provide complementary insights beyond numerical metrics. Regression-based approaches such as BasicVSR++ [304] and PSRT [303] exhibit excellent temporal consistency and faithfully reconstruct fine details like the "ZABITA" text sequences. In contrast, generation-based methods like MGLDVSR [301] and UAV [302], while producing perceptually sharper results, frequently fail to accurately reconstruct critical information (e.g., the frame numbers "153"/"155") and exhibit visible temporal instability. Transformer-based solutions, particularly RVRT [299], stand out by maintaining structural integrity across frames while achieving superior detail preservation compared to both CNN-based and generative approaches. All modern methods significantly outperform the bicubic baseline, though with distinct characteristic artifacts.

3) Future Prospects: The experimental results collectively suggest that VSR technology remains far from saturation, with multiple promising research directions. Improving computational efficiency represents a critical challenge, particularly for Transformer architectures where memory and processing demands limit practical deployment. Alternative training paradigms that reduce dependence on massive labeled datasets could help democratize VSR development. Furthermore, hybrid approaches that combine the fidelity of regression methods with the perceptual strengths of generative techniques may help bridge the current quality gap. These advancements could potentially address the fundamental tradeoffs between accuracy, efficiency, and visual quality that currently define the VSR landscape.

V. STEREO SUPER RESOLUTION

A. Regressive Models

1) CNN-Based SSR:

a) Learning better representations: Jeon et al. [33] learned a parallax prior from stereo image datasets by jointly training two-stage networks for luminance and chrominance, respectively, and enhanced the spatial resolution of stereo images significantly more than SISR methods. Song et al. [335] imposed stereo-consistency constraints on integrating stereo information.

b) Improving information interaction: Wang et al. [73] introduced a generic parallax attention mechanism with a global receptive field along the epipolar line to handle different stereo images with large disparity variations, as shown in Fig. 20. Ying et al. [341] proposed a generic stereo attention module to extend arbitrary SISR networks

Table 5 Performance Comparison of Different SSR Models on Three Benchmarks for the ×4 task. PSNR/SSIM Values on the Y Channel Are Reported for Each Dataset. "Params" Represents the Total Number of Network Parameters. The **Best** Performance Is Highlighted for Models Trained on Flickr1024 [336], While the **Best** Performance Is Highlighted for Models Trained on Middlebury [337] + Flickr1024 [336]. The Results for Each Model Are Obtained From Their Respective Original Articles

Model	Params (M)	KITTI2015 [338] (PSNR/SSIM)	KITTI2012 [339] (PSNR/SSIM)	Middlebury [337] (PSNR/SSIM)	Training Sets	Backbone	Venue
Bicubic	-	23.90/0.7100	24.64/0.7334	26.39/0.7564	-	-	-
StereoSR [33]	1.06	25.12/0.7679	25.94/0.7839	28.24/0.8133	Middlebury+KITTI+Tsukuba	CNN	CVPR'18
DASSR [340]	1.1	25.35/0.8740	**26.96/0.8820**	**29.83/0.9090**	Flickr1024	CNN	CVPR'20
SRRes+SAM [341]	1.73	25.55/0.7825	26.35/0.7957	28.76/0.8287	Flickr1024	CNN	SPL'20
CVCNet [342]	0.99	25.55/0.7801	26.35/0.7935	28.65/0.8231	Flickr1024	CNN	TMM'21
SSRDE-FNet [343]	2.24	25.74/0.7884	26.61/0.8028	29.29/0.8407	Flickr1024	CNN	ACMMM'21
NAFSSR [344]	1.53	**25.78/0.7927**	26.62/0.8051	29.27/0.8447	Flickr1024	CNN	CVPRW'22
Steformer [345]	1.34	25.74/0.7906	26.61/0.8037	29.29/0.8424	Flickr1024	Transformer	TMM'23
EDSR [6]	38.9	25.38/0.7811	26.26/0.7954	29.15/0.8383	Middlebury+Flickr1024	CNN	CVPR'17
RCAN [28]	15.4	25.53/0.7836	26.36/0.7968	29.20/0.8381	Middlebury+Flickr1024	CNN	ECCV'18
PASSRnet [73]	1.35	25.34/0.7722	26.18/0.7874	28.36/0.8153	Middlebury+Flickr1024	CNN	CVPR'19
IMSSRnet [346]	6.89	25.59/-	26.44/-	29.02/-	Middlebury+Flickr1024	CNN	TCSVT'20
iPASSR [347]	1.42	25.61/0.7850	26.47/0.7993	29.07/0.8363	Middlebury+Flickr1024	CNN	CVPRW'21
SIRFormer [348]	1.48	25.75/0.7882	26.53/0.7998	29.23/0.8396	Middlebury+Flickr1024	Transformer	ACMMM'22
SwinFIRSSR [349]	24.09	26.15/0.8062	27.06/0.8175	30.33/0.8676	Middlebury+Flickr1024	Transformer	ArXiv'22
MSSFNet [350]	1.82	26.07/0.7990	26.88/0.8098	29.67/0.8498	Middlebury+Flickr1024	Transformer	ArXiv'24
LSSR [351]	1.11	26.12/0.7997	26.93/0.8097	29.86/0.8489	Middlebury+Flickr1024	CNN	ACMMM'24
SCGLANet [352]	25.29	**26.94/0.8268**	**27.10/0.8204**	**30.23/0.8628**	Middlebury+Flickr1024	GAN	ESWA'24

for SSR. Zhang et al. [353] incorporated a disparity-based constraint mechanism into the generation of SR images with an additional atrous parallax-attention module. Yan et al. [340] proposed a unified stereo image restoration framework that explicitly learns the inherent pixel correspondence between stereo views and restores stereo images by leveraging cross-view information at the image and feature level. Lei et al. [346] proposed an interaction module-based stereoscopic network to effectively utilize the correlation information in stereoscopic images. Wang et al. [347] proposed a symmetric bidirectional parallax attention module and an inline occlusion handling scheme to effectively interact with cross-view information. Chen et al. [354] proposed a cross parallax attention stereo network that can perform SSR for multiple scale factors and both views. Zhu et al. [342] proposed a cross-view capture network for stereo image SR by using both global contextual and local features extracted from both views. Xu et al. [355] incorporated bilateral grid processing in a CNN framework and proposed a bilateral stereo network. Dan et al. [356] proposed a disparity feature alignment module to exploit the disparity information for feature alignment and fusion. Dai et al. [343] proposed a disparity estimation feedback network, which simultaneously handles stereo image SR and disparity estimation in a unified framework and leverages their interaction to further improve performance. Chu et al. [344] inherited a strong and simple image restoration model, NAFNet, for single-view feature extraction and extended it by adding cross-attention modules to fuse features between views to adapt to binocular scenarios.

2) Transformer-Based SSR: Imani et al. [357] designed a self-attention and optical flow-based feed-forward layer to make the Transformer suitable for stereo VSR. Jin et al. [358] adopted the Swin Transformer as the backbone and incorporated it with the bidirectional parallax attention module to maximize auxiliary information given the binocular mechanism. Lin et al. [345] resorted to the Transformer to capture reliable stereo information and proposed a cross-to-intraattention mechanism to further extract information. Yang et al. [348] leveraged long-range context dependencies and proposed the stereo alignment Transformer (SAT) to align the views, and the stereo fusion transformer (SFT) to aggregate cross-view information. PFT-SSR [359] proposed a fusion strategy based on the Transformer structures and designed two branches to explore the interview and the intraview information. SIR-Former [348] leverages an SAT for global correspondence alignment and an SFT for cross-view information

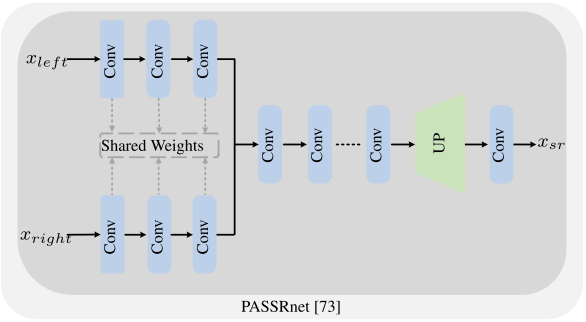

Fig. 20. ***Framework of PASSRnet [73]. It introduced a compelling and efficient parallax-attention mechanism, along with the creation of the extensive dataset Flickr1024, both of which have had a profound impact on subsequent SSR research. The results for each model are obtained from their respective original articles.***

aggregation, enhancing performance across various image restoration tasks.

B. Generative Models

Ma et al. [360] proposed a perception-oriented StereoSR framework to restore stereo images with better subjective quality by employing feedback from the quality assessment of the StereoSR results. Dinh et al. [361] proposed a model to simultaneously solve denoising and SR problems and introduced matchability attention to improve the interaction between generated stereo images. RealSCGLAGAN [352] is a novel stereo image SR method that integrates an implicit stereo information discriminator and a hybrid degradation model to enhance visual quality while preserving disparity consistency. DiffSteISR [362] is a pioneering stereo image reconstruction framework that integrates techniques like time-aware stereo cross attention, omni attention control, and stereo semantic extraction to enhance texture and semantic consistency in super-resolved stereo images.

C. Training Details

1) Datasets: There are five commonly used datasets for evaluating the performance of SSR methods: Middlebury [337], Tsukuba [363], KITTI 2012 [339], KITTI 2015 [338], and Flickr1024 [336].

1) *Middlebury [337]:* This dataset consists of input images captured under multiple exposures and ambient illuminations, with and without a mirror sphere present to capture the lighting conditions.
2) *Tsukuba [363]:* The dataset contains 1800 full-color stereo pairs per illumination condition, along with ground-truth disparity maps. It was created using an animated stereo camera and includes 1-min videos at 30 frames per second (FPS).
3) *KITTI 2012 [339]:* This dataset selects a subset of sequences where the environment is static. It includes 194 training and 195 test image pairs.
4) *KITTI 2015 [338]:* It consists of stereo videos of road scenes captured by a calibrated pair of cameras mounted on a car. Ground truth for optical flow and disparity is obtained from a 3-D laser scanner combined with the egomotion data of the car.
5) *Flickr1024 [336]:* This is the largest dataset for SSR, containing 1024 high-quality image pairs that cover diverse scenarios. The dataset provides realistic cases that align with daily photography scenarios.

2) Training Settings:

a) Patch extraction: In SSR, patch extraction involves dividing LR stereo image pairs into smaller fixed-size patches, usually ranging from 32×32 to 64×64 pixels. For instance, the StereoLR dataset divides each LR stereo image pair into 30k patches of size 41×41 pixels, creating training samples for the SR model.

b) Data augmentation: To increase the diversity of training samples, data augmentation is applied to both the left and right images in stereo pairs. Common techniques, such as random rotations, flips, color transformations, and additional strategies like random horizontal shifting, Mixup, CutMix, and CutMixup, are employed to further enhance the model's performance.

c) Optimization: Standard optimization algorithms like SGD, Adam, and RMSprop are commonly used to iteratively update the model weights and minimize the disparity error. For instance, a recent SSR method used the Adam optimizer with a learning rate of 10^{-4} for 100 epochs, leading to an average disparity error reduction of 25%.

d) Learning schedule: A typical learning rate schedule involves starting with an initial learning rate of 10^{-3} and then reducing it by a factor of 10^{-1} after every 10k training iterations. This facilitates effective model convergence and improves accuracy and efficiency.

e) Batch size: Batch sizes in SSR typically range from 4 to 32, with smaller batch sizes employed in memory-constrained scenarios or when dealing with large-resolution stereo images. A common choice is a batch size of 8, which strikes a balance between computational efficiency and memory requirements. These training settings are crucial for achieving high-quality results in this field.

D. Experimental Results

We conduct comprehensive experiments to evaluate the performance of SSR methods in comparison with traditional SISR approaches. The quantitative results are summarized in Table 5, which includes standard evaluation metrics (PSNR and SSIM), model complexity (the number of parameters), training datasets, network architectures, and publication venues.

The experimental results demonstrate clear advantages of dedicated SSR methods over traditional SISR approaches when processing stereo imagery. As shown in Table 5, while both types of methods can achieve comparable performance in terms of standard metrics like PSNR and SSIM, the stereo-specific designs demonstrate superior capability in maintaining view consistency and handling disparity variations. Traditional SISR methods such as EDSR [6] and RCAN [28], despite their good performance on single-image tasks, exhibit significant limitations when directly applied to stereo pairs. These approaches process left and right views independently, failing to exploit the inherent correlation between stereo images. In contrast, modern SSR methods like NAFSSR [344] and SwinFIRSSR [349] explicitly address these challenges through specialized architectures. They incorporate stereo matching components and disparity-aware processing modules that maintain view consistency during SR. Visualization results, as shown in Fig. 21, particularly highlight this advantage. Specifically, we visualize the outputs of CNN and Transformer-based SISR methods. From these

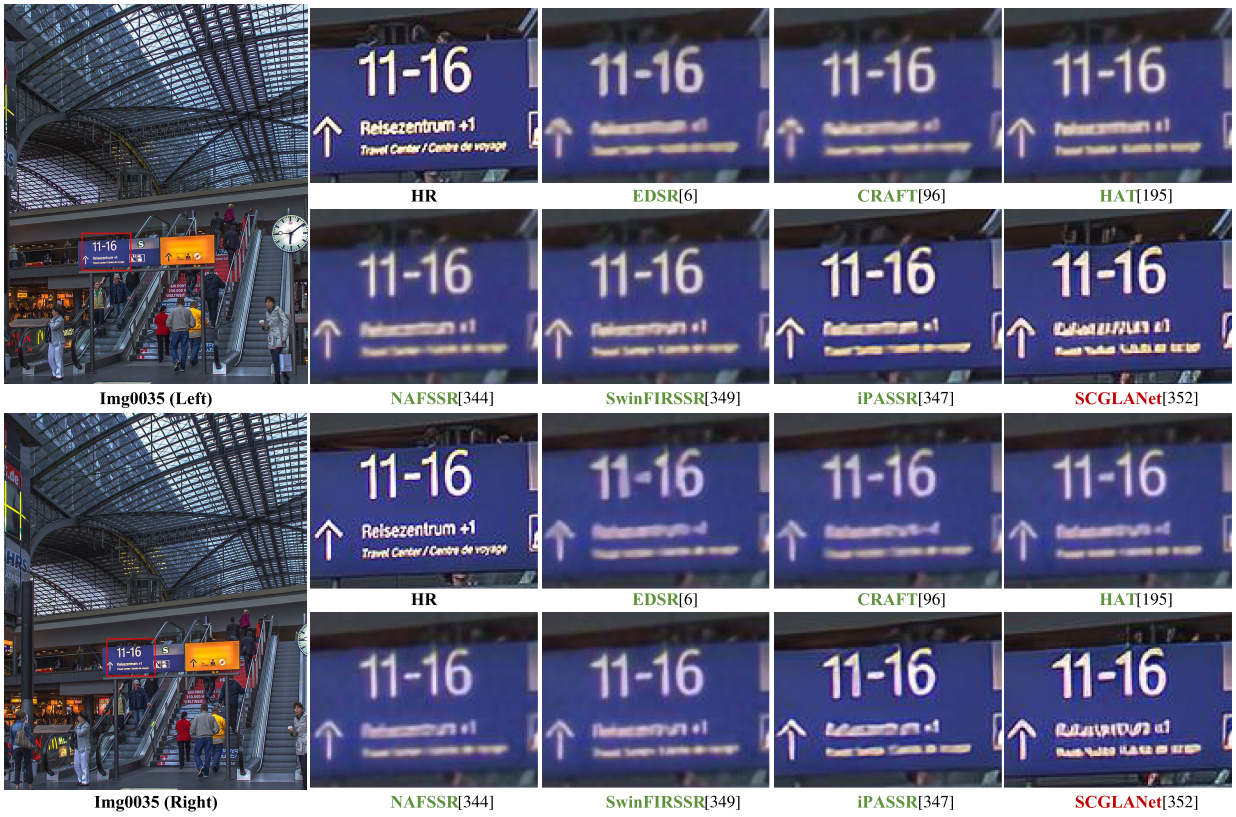

Fig. 21. *Visual comparison of regression-based and generation-based SSR methods. regression-based methods such as EDSR [6], CRAFT [96], and NAFSSR [344] generally succeed in making the primary text "11-16" legible and attempt to reconstruct the finer"Reisezentrum +1" text, albeit with noticeable blur. Conversely, the generation-based method SCGLANet [352], while also clearly rendering "11-16", tends to introduce significant artifacts and unfaithful, sharp textures in place of the smaller textual details, often rendering them illegible and unnatural compared to the HR or even the regression-based outputs.*

results, SISR methods (EDSR [6] and HAT [195]) produce more artifacts and blurriness. Although Transformer-based methods like HAT possess strong long-term dependence capture capabilities, their outputs still exhibit some artifacts. In contrast, SSR methods produce more coherent and geometrically consistent HR outputs compared to their SISR counterparts. In addition, it can be observed that generative models (e.g., SCGLANet [352]) can produce more perceptually pleasant results than regression-based models.

VI. LIGHT FIELD IMAGE SUPER RESOLUTION

Light-field images capture both the spatial and angular information of light rays, providing a richer representation of the scene compared to traditional 2-D images. This additional angular information enables post-capture image manipulation such as refocusing, depth estimation, and perspective changes. However, the resolution of light-field images is typically lower due to hardware limitations, necessitating the use of SR techniques to enhance both spatial and angular resolutions simultaneously.

A. Regressive Models

1) CNN-Based LFSR:

a) Efficient structural designs: LFCNN [365] enhances both spatial and angular resolutions of light-field images using a deep CNN, by first applying spatial SR to subaperture images and then generating novel views with angular SR, culminating in end-to-end fine-tuning. Yoon et al. [34] introduced a method that uses CNNs to simultaneously upsample the spatial and angular resolutions of light field images. Yuan et al. [366] presented a combined deep CNN framework that incorporates an SISR module and an EPI enhancement module to achieve geometrically consistent super-resolved light field images. Yeung et al. [367] proposed an hourglass-shaped model that performs feature extraction at the LR level to reduce computation and memory costs. They utilized spatial–angular separable convolutions for efficient extraction of joint spatial–angular features. Zhu et al. [368] developed an EPI-based CNN-LSTM network specifically designed for simultaneous spatial and angular SR of light fields. Wang et al. [369] introduced an approach to extract and decouple spatial and angular features. Meng et al. [370] formulated LFSR as tensor restoration and developed a two-stage restoration

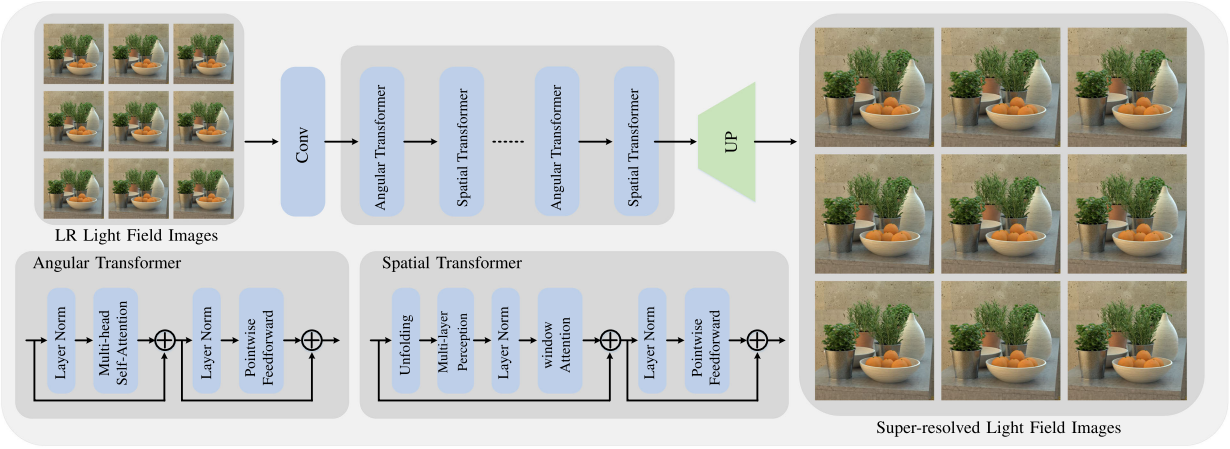

Fig. 22. ***Framework of LFT [364]. It introduced a straightforward yet highly efficient Transformer-based model that incorporated an angular Transformer block and a spatial Transformer block. This design enabled the extraction of abundant nonlocal information, ultimately resulting in exceptional performance and establishing a new state of the art.***

framework based on 4-Dconvolution. Cheng et al. [371] introduced a zero-shot learning framework for LFSR, where the reference view is super-resolved solely based on examples extracted from the input LR light field. Wang et al. [372] developed a mechanism to disentangle coupled information in light field image processing, leading to the design of three networks (DistgSSR, DistgASR, and DistgDisp) for spatial SR, angular SR, and disparity estimation, respectively. These networks leverage the disentangling mechanism for improved performance in LFSR tasks. Jin et al. [373] individually super-resolved each view of a light field image by leveraging complementary information among views through combinatorial geometry embedding. Jin et al. [374] utilized depth estimation and light field blending modules to leverage the intrinsic geometry information. They also introduced a loss based on the gradient of EPIs to preserve the light field parallax structure. Zhang et al. [375] proposed an end-to-end LFSR framework called MEG-Net, which utilizes multiple epipolar geometries to simultaneously super-resolve all views in a light field by exploring subpixel information in different spatial and angular directions.

b) Improving information interaction: Wang et al. [376] proposed an implicitly multiscale fusion scheme for LFSR by accumulating contextual information from multiple scales. Zhang et al. [377] introduced residual convolutional networks for reconstructing HR light fields. Cheng et al. [378] proposed a framework for LFSR that exploits both internal and external similarities, demonstrating their complementary nature. Liu et al. [379] proposed LF-IINet, an intra–interview interaction network consisting of two parallel branches that model global interview information and correlations among intraview features, respectively. Wang et al. [380] designed an angular deformable alignment module for feature-level alignment and proposed a collect-and-distribute approach for bidirectional alignment between the center-view feature and each side-view feature. Meng et al. [381] explored the geometric structural properties in light field data and proposed a second-order residual network to learn geometric features for reconstruction. LFSR-AFR [382] improves light-field image SR by utilizing spatial and angular SR networks with a trainable disparity estimator and an adaptive feature remixing (AFR) module to enhance multiview features, achieving superior performance on various datasets. DDAN [383] enhances light-field image SR by employing a dense dual-attention network that adaptively captures and fuses discriminative features across different views and channels, leveraging a chain structure and dense connections to improve SR performance. Xiao et al. [384] conducted a series of preprocessing steps involving blending crop patches from each view of LR/HR to produce a blended patch and pasting it to the corresponding position of HR/LR. This approach achieves improved SR results while keeping the model unchanged.

2) Transformer-Based LFSR: Liang et al. [364] designed an angular Transformer and a spatial Transformer to model the relationship among different views and capture local and nonlocal context information, s, as shown in Fig. 22. Wang et al. [385] treated LFSR as a sequence learning problem and proposed a detail-preserving Transformer that uses gradient maps of light fields to guide the sequence learning. Shabbir et al. [386] introduced a Transformer module to extract texture features from the reference image in a four-step reconstruction process. Guo et al. [387] focused on data generation and proposed a new raw data generation method, followed by the use of Transformer models to aggregate angular and nonlocal cross-view information. Wang and Lu [388] proposed intra- and inter-Transformers to dynamically exploit information between subaperture images. EPIT [389] takes into

account the spatial and angular correlation within the EPI and introduces a Transformer-based structure to capture this relationship.

B. Generative Models

Ko et al. [391] employed a GAN-based structure to generate finer texture details in skin for high up-scaling factors. They introduced content and adversarial loss to enhance image quality. Ruan et al. [392] proposed a Wasserstein GAN with gradient penalty to further improve generative qualities. Additionally, they introduced perceptual loss on the epipolar plane to enhance depth information generation. Zhu et al. [393] proposed an autoencoder structure to extract neighboring view information and employed a discriminator model with WGAN loss to enhance sharpness in reconstructed light fields. Chen et al. [394] generated a 4-D LF from a single image using GAN and devised a multistage training strategy combined with multiple losses to constrain the training process. Meng et al. [395] introduced a generative model called LFGAN, which incorporated high-dimensional convolution layers into a GAN framework to learn high correlations among neighboring LF views. Guo et al. [396] designed a strategy for selecting foreground elemental images (EIs) to improve rendering performance and utilized a GAN to individually super-resolve the EIs for better image quality. Wafa and Nasiopoulos [397] proposed a residual dense structure as the generator and introduced a patch discriminator to enhance visual quality. LFSRDiff [398] leverages a disentangled UNet to effectively capture and fuse spatial and angular information, producing diverse and realistic SR outcomes.

C. Training Details

1) Datasets: Five commonly used datasets are employed for training and evaluating the performance of LFSR methods: EPFL [399], HCInew [400], HCIold [401], INRIA [402], and STFgantry [403].

1) *EPFL [399]:* This dataset contains images in the light field raw (LFR) file format captured by the Lytro Illum camera. The LFR files store raw, uncompressed lenslet images before the demosaicing process, with 10 bits per pixel precision in little-endian format.
2) *HCInew [400]:* It consists of 24 carefully designed synthetic densely sampled 4-D light fields with highly accurate disparity ground truth.
3) *HCIold [401]:* This dataset is characterized by a dense sampling of light fields, which is suitable for current plenoptic cameras.
4) *INRIA [402]:* The synthetic light fields "butterfly" and "stillife" are taken from the HCI database. The others were captured with a Lytro camera, and the subaperture images were extracted using the toolbox of Dansereau et al. [404].
5) *STFgantry [403]:* The Computer Graphics Laboratory at Stanford University has acquired several light fields for research in computer graphics and vision.

2) Training Settings:

a) Patch extraction: In LFSR, patch extraction involves dividing the 4-D light field data into smaller patches to create training samples. Spatial patches of size, e.g., 32×32, are commonly extracted from each view, and angular patches of size, e.g., 9×9 views, are sampled to account for the multiview nature. This process results in a large number of training samples, contributing to the effectiveness of training LFSR models.

b) Data augmentation: Commonly used transformations include random shifts, rotations, and flips applied to the 4-D light field data. For instance, on the EPFL dataset, data augmentation can increase the number of training samples from 500 to over 5000 by applying random shifts and flips in both spatial and angular dimensions.

c) Optimization: In LFSR, popular optimizers, such as Adam and SGD, are commonly used for parameter updates during training. An average learning rate of Adam is set to 10^{-3} to 10^{-4}, while SGD is used with an average learning rate of 10^{-2} to 10^{-3}.

d) Learning schedule: A common learning rate schedule for LFSR models starts with an initial learning rate of 10^{-3} and uses a step decay strategy, reducing the learning rate by a factor of 10^{-1} after every 50 000 iterations. This approach helps achieve convergence and avoid overfitting while training LFSR models on datasets like EPFL with an average training time of 150 000 iterations.

e) Batch size: Due to the high dimensionality of light field data, memory limitations, and computational complexity, the batch size is typically smaller compared to SISR models. Common batch sizes for LFSR range from 2 to 8 light field samples per iteration.

D. Experimental Results and Analysis

We evaluate the competing methods on five widely used benchmark datasets, assessing their performance using PSNR and SSIM metrics while analyzing model complexity through parameter counts. The implementation details, including backbone architectures and source publications, are systematically documented for reference.

As quantitatively demonstrated in Table 6, LFSR models consistently outperform direct SISR approaches across all metrics. Notably, current LFSR methods exhibit suboptimal performance on the EPFL dataset (PSNR $<$30 dB), revealing significant room for improvement in this domain.

A detailed comparison reveals that Transformer-based architectures achieve state-of-the-art performance by effectively exploiting global attention mechanisms. However, this superior performance comes at the cost of substantially higher computational complexity, potentially limiting their deployment in resource-constrained scenarios. Qualitative results in Fig. 23 demonstrate that DL-based approaches,

Table 6 Performance Comparison of Different LFSR Models on Five Benchmarks for the ×4 Task. PSNR/SSIM Values on the Y Channel Are Reported for Each Dataset. "Params" Represents the Total Number of Network Parameters. The Best and Second Best Performances Are Indicated in Bold and Underlined

Model	Params (M)	EPFL (PSNR/SSIM)	HCInew (PSNR/SSIM)	HCIold (PSNR/SSIM)	INRIA (PSNR/SSIM)	STFgantry (PSNR/SSIM)	Backbone	Venue
Bicubic	-	25.14/0.8324	27.61/0.8517	32.42/0.9344	26.82/0.8867	25.93/0.8452	-	-
VDSR [159]	0.66	27.25/0.8777	29.31/0.8823	34.81/0.9515	29.19/0.9204	28.51/0.9009	CNN	CVPR'16
EDSR [161]	38.9	27.84/0.8854	29.60/0.8869	35.18/0.9536	29.66/0.9257	28.70/0.9072	CNN	CVPR'17
RCAN [28]	15.4	27.88/0.8863	29.63/0.8886	35.20/0.9548	29.76/0.9276	28.90/0.9131	CNN	ECCV'18
DDAN [383]	0.51	29.19/0.9178	31.60/0.9211	37.36/0.9722	31.44/0.9513	30.72/0.9419	CNN	TCSVT'18
LFSSR [367]	1.77	28.27/0.9118	30.72/0.9145	36.70/0.9696	30.31/0.9467	30.15/0.9426	CNN	TIP'18
resLF [377]	8.64	28.27/0.9035	30.73/0.9107	36.71/0.9682	30.34/0.9412	30.19/0.9372	CNN	CVPR'19
LF-ATO [373]	1.36	28.25/0.9115	30.88/0.9135	37.00/0.9699	30.71/0.9484	30.61/0.9430	CNN	CVPR'20
LF-InterNet [369]	5.48	28.67/0.9162	30.98/0.9161	37.11/0.9716	30.64/0.9491	30.53/0.9409	CNN	ECCV'20
LF-DFnet [380]	3.99	28.77/0.9165	31.23/0.9196	37.32/0.9718	30.83/0.9503	31.15/0.9494	CNN	TIP'20
LF-IINet [379]	4.89	29.11/0.9188	31.36/0.9208	37.62/0.9734	31.08/0.9515	31.21/0.9502	CNN	TMM'21
MEGNet [375]	1.77	28.74/0.9160	31.10/0.9177	37.28/0.9716	30.66/0.9490	30.77/0.9453	CNN	TIP'21
DistgSSR [372]	3.58	28.99/0.9195	31.38/0.9217	37.56/0.9732	30.99/0.9519	31.65/0.9535	CNN	TPAMI'22
DPT [385]	3.78	28.93/0.9170	31.19/0.9188	37.39/0.9721	30.96/0.9503	31.14/0.9488	Transformer	AAAI'22
LFT [364]	1.16	29.25/0.9210	31.46/0.9218	37.63/0.9735	31.20/0.9524	31.86/0.9548	Transformer	SPL'22
LFSAV [390]	1.54	**29.37/0.9223**	31.45/0.9217	37.50/0.9721	31.27/**0.9531**	31.36/0.9505	Transformer	TCI'22
EPIT [389]	1.47	29.34/0.9197	**31.51/0.9231**	**37.68/0.9737**	**31.37**/0.9526	**32.18/0.9571**	Transformer	ICCV'23

particularly LFSR methods, generate significantly sharper textures compared to traditional interpolation techniques. While conventional SISR methods like RCAN show competitive results, all specialized LFSR implementations deliver visually satisfactory reconstruction quality.

Compared to traditional SISR, LFSR remains a relatively understudied domain, with only a limited body of dedicated research to date. This presents compelling opportunities for future exploration, particularly in two key directions: 1) the development of computationally efficient architectures optimized for edge-device deployment, where resource constraints demand lightweight yet high-performance solutions and 2) the investigation of advanced angular–spatial feature fusion strategies to better harness the unique 4-D structure of light field data, which remains underexploited in current methods.

VII. REMAINING ISSUES AND FUTURE DIRECTIONS

A. Unlocking the Power of SR-Specific Priors

While many existing methods focus on network architecture—such as model depth and spatial or channel correlations—to improve performance, they often overlook the importance of domain-specific priors. As a result, these models tend to be more generic and less tailored for SR tasks. There is a clear need to explore SR-specific priors to enhance both quantitative and qualitative performances [23], [104], [105], [106], as shown in Fig. 4.

Incorporating edge priors, for example, can significantly improve the restoration of sharp, high-quality images [105], [108]. Specifically, using edge loss helps the network focus on edges and textures, yielding more visually pleasing results. Additionally, modules designed

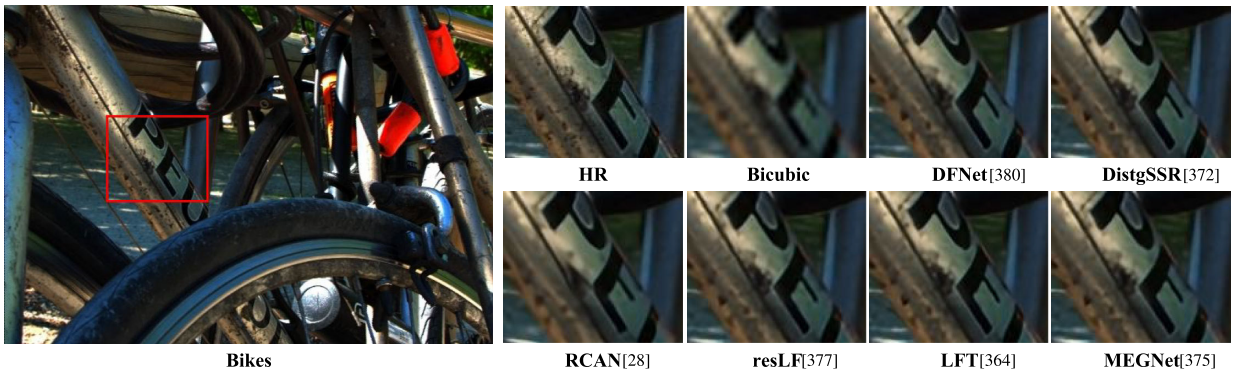

Fig. 23. *Visual comparison of LFSR methods. Dedicated LFSR methods (like LFT [364] or MEGNet [375]) generally achieve sharper results and more accurate texture reconstruction by leveraging the multiview information inherent in the light field data. In contrast, traditional single-image methods like RCAN [28], while an improvement over bicubic, cannot fully exploit these angular cues, resulting in comparatively less detailed and often blurrier outputs in this context.*

to extract and integrate edge information into later stages allow the network to learn more useful features [405].

Similarly, exploiting the inherent sparsity in SR, a sparse mask can be predicted from feature maps, enabling sparse convolution to skip unnecessary positions and reduce model complexity [25]. Integrating recurrence priors [101] can further enhance SR by considering near/far distances through patch cropping and applying self-supervised methods to improve performance.

B. Advancements in Architectural Structures for SR

The field of SR has witnessed the emergence of numerous architectural structures, each contributing to the progress of the field. Initially, basic convolutional layers were employed, and they were later enhanced by incorporating residual connections and attention mechanisms. More recently, Transformer-based architectures have emerged as the leading models, exhibiting remarkable performance in SR tasks. Alongside, diffusion models [37], [235], [240] have introduced an iterative diffusion process to enhance LR images, gradually refining image details by propagating information between adjacent pixels. This iterative approach results in higher levels of image clarity. Furthermore, the exploration of state-space models (SSMs), particularly the Mamba architecture, has demonstrated significant potential in SR tasks by efficiently modeling long-range dependencies with linear computational complexity [406], [407], [408], [409]. Additionally, graphic neural networks (GNNs) have yielded promising results in SR tasks, while NAS techniques have been employed to discover more efficient and effective structures, despite the time-consuming nature of the process. Although the design of new structures poses challenges, it represents a crucial direction for advancing the SR community, pushing the boundaries of performance and fostering innovation.

C. Toward Lightweight SR: Balancing Model Complexity and Performance

Despite the significant performance advancements achieved in SR, the practical implementation of these models is hindered by their large parameter sizes and high computational costs. It is crucial to develop lightweight SR models that can be deployed on resource-constrained devices. Several methods have been explored to reduce model complexity, including knowledge distillation, reparameterization, NAS, and quantization [410], [411], [412]. However, these methods often suffer from performance degradation due to information loss during the simplification process. Designing lightweight models that can maintain high performance is still a challenging task that necessitates further research and development.

D. Beyond Singular Degradation: Modeling Diverse Processes for Enhanced SR

Current SR methods commonly focus on modeling a single degradation process, such as Gaussian blur or bicubic blur. However, real-world data present a broad spectrum of factors that contribute to image degradation, including camera aging, diverse optical settings, and combinations of multiple degradation processes. Some approaches have attempted to tackle this issue by enumerating various degradation processes and integrating them into the training process. Others have redefined the degradation process and employed generative models to estimate the degradation kernel. Nevertheless, comprehensive research in this area is still lacking, and finding suitable solutions for modeling diverse degradation processes remains a challenge that needs to be addressed.

E. Bridging the Gap: Advancing Arbitrary SR for Diverse Scaling Factors

While numerous SR methods primarily concentrate on fixed-scale SR tasks, there is an increasing demand for arbitrary SR, where the scaling factor is not predefined. Recent works have made strides in exploring arbitrary SR and have achieved promising results [141], [142], [177]. These approaches involve modifying the weight kernel to adaptively handle diverse scales or utilizing implicit neural expressions to regress pixel values for arbitrary scaling factors. However, there still exists a performance gap between fixed-scale and arbitrary SR models. Further exploration and research are necessary to enhance the performance of arbitrary SR and bridge this gap effectively.

F. Exploring the Synergy Between Low-Level and High-Level Tasks in SR

Low-level tasks in SR have a dual objective: to restore images and to provide robust inputs for high-level tasks. Regrettably, there is a paucity of research that explores the relationship and interaction between low- and high-level tasks. Establishing connections between these two tasks can enable seamless information exchange across different levels and potentially foster the fusion of different models. This direction holds promise and bears significant value for future implementations in SR.

G. Addressing Data Challenges in SR: Toward Dedicated Large-Scale Low-Level Datasets

In contrast to high-level tasks that often benefit from access to large-scale datasets, the field of SR lacks such abundant resources, which presents a challenge to its development. Although some approaches have suggested leveraging datasets like ImageNet for pretraining models, the image quality in such datasets may not be adequate for low-level tasks, where image quality plays a crucial role. Constructing dedicated large-scale low-level datasets would serve as a major boost to SR performance.

VIII. CONCLUSION

We present a comprehensive and in-depth review of SR methods, covering a wide range of areas including SISR,

VSR, SSR, and LFSR. This article encompasses over 150 SISR methods, nearly 70 VSR approaches, and approximately 30 techniques for SSR and LFSR. We categorize SR methods into two primary groups: regression-based and generative models, and further subdivide them according to their backbone architectures. For regression-based methods, we discuss CNN-based and Transformer-based approaches, while for generative models, we explore those using GANs and diffusion models. Each category is further refined based on specific architectural innovations and techniques. This article also delves into critical topics such as datasets, evaluation protocols, empirical performance, and implementation challenges. Moreover, we highlight key open issues that require further exploration, aiming to inspire continued research and innovation in the SR domain. We hope this article will serve as a valuable resource for researchers, providing insights and guidance for future advancements in SR technologies. ■

REFERENCES

[1] H. Liu et al., "Video super-resolution based on deep learning: A comprehensive survey," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 5981–6035, Dec. 2022.

[2] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1034–1040, May 2016.

[3] D. Qiu, Y. Cheng, and X. Wang, "Medical image super-resolution reconstruction algorithms based on deep learning: A survey," *Comput. Methods Programs Biomed.*, vol. 238, Aug. 2023, Art. no. 107590.

[4] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, Jan. 2009.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[6] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[7] Z. Zhang and V. Sze, "FAST: A framework to accelerate super-resolution processing on compressed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1015–1024.

[8] Y. Ben Sahel and Y. C. Eldar, "Self-STORM: Deep unrolled self-supervised learning for super-resolution microscopy," 2024, *arXiv:2403.16974*.

[9] G. Dardikman-Yoffe and Y. C. Eldar, "Learned SPARCOM: Unfolded deep super-resolution microscopy," *Opt. Exp.*, vol. 28, no. 19, pp. 27736–27763, 2020.

[10] O. Solomon, Y. C. Eldar, M. Mutzafi, and M. Segev, "SPARCOM: Sparsity based super-resolution correlation microscopy," *SIAM J. Imag. Sci.*, vol. 12, no. 1, pp. 392–419, Jan. 2019.

[11] R. J. G. van Sloun et al., "Super-resolution ultrasound localization microscopy through deep learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 829–839, Mar. 2021.

[12] T. Tirer, R. Giryes, S. Y. Chun, and Y. C. Eldar, "Deep internal learning: Deep learning from a single input," *IEEE Signal Process. Mag.*, vol. 41, no. 4, pp. 40–57, Jul. 2024.

[13] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[14] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[15] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Classification (CVPR)*, vol. 1, Jun. 2004, pp. 275–282.

[16] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[20] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 57–65.

[21] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3862–3871.

[22] X. Zhang, H. Zeng, and L. Zhang, "Edge-oriented convolution block for real-time super resolution on mobile devices," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4034–4043.

[23] X. Kong, H. Zhao, Y. Qiao, and C. Dong, "ClassSR: A general framework to accelerate super-resolution networks by data characteristic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12016–12025.

[24] K. Bhardwaj et al., "Collapsible linear blocks for super-efficient super resolution," in *Proc. Mach. Learn. Syst. (MLSys)*, 2021, pp. 529–547.

[25] L. Wang et al., "Exploring sparsity in image super-resolution for efficient inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4915–4924.

[26] Z. Zhan et al., "Achieving on-mobile real-time super-resolution with neural architecture and pruning search," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4801–4811.

[27] Z. Chen, Y. Zhang, J. Gu, Y. Zhang, L. Kong, and X. Yuan, "Cross aggregation transformer for image restoration," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 25478–25490.

[28] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[29] B. Xia, Y. Hang, Y. Tian, W. Yang, Q. Liao, and J. Zhou, "Efficient non-local contrastive attention for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2759–2767.

[30] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4799–4807.

[31] R. M. Umer and C. Micheloni, "Deep cyclic generative adversarial residual convolutional networks for real image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 484–498.

[32] J. Caballero et al., "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2848–2857.

[33] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1721–1730.

[34] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 848–852, Jun. 2017.

[35] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.

[36] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4569–4579.

[37] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[38] Z. Yue, J. Wang, and C. C. Loy, "ResShift: Efficient diffusion model for image super-resolution by residual shifting," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 13294–13307.

[39] C. Rota, M. Buzzelli, and J. van de Weijer, "Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2023, pp. 36–53.

[40] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.

[41] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.

[42] B. B. Moser, A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio, and A. Dengel, "Diffusion models, image super-resolution, and everything: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 11793–11813, Jul. 2025.

[43] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Science Rev.*, vol. 232, Sep. 2022, Art. no. 104110.

[44] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[45] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33, 2020, pp. 6840–6851.

[46] Y. Nikankin, N. Haim, and M. Irani, "SinFusion: Training diffusion models on a single image or video," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 1–11.

[47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[48] A. Vaswani et al., "Attention is all you need," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[49] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–22.

[50] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho, "Natural and realistic single image super-resolution with explicit natural manifold discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8114–8123.

[51] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.

[52] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–21, 2020.

[53] K. Li, S. Yang, R. Dong, X. Wang, and J. Huang, "Survey of single image super-resolution reconstruction," *IET Image Process.*, vol. 14, no. 11, pp. 2273–2290, 2020.

[54] Y. Li, B. Sixou, and F. Peyrin, "A review of the deep learning methods for medical images super resolution problems," *IRBM*, vol. 42, no. 2, pp. 120–133, Apr. 2021.

[55] H. Chen et al., "Real-world single image super-resolution: A brief review," *Inf. Fusion*, vol. 79, pp. 124–145, Mar. 2022.

[56] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learning-based single image super-resolution," *PeerJ Comput. Sci.*, vol. 7, p. e621, Jul. 2021.

[57] X. Wang et al., "A review of image super-resolution approaches based on deep learning and applications in remote sensing," *Remote Sens.*, vol. 14, no. 21, p. 5423, Oct. 2022.

[58] S. Liu, P. Hoess, and J. Ries, "Super-resolution microscopy for structural cell biology," *Annu. Rev. Biophys.*, vol. 51, no. 1, pp. 301–326, May 2022.

[59] D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal, "Image super-resolution: A comprehensive review, recent trends, challenges and applications," *Inf. Fusion*, vol. 91, pp. 230–260, Mar. 2023.

[60] K. Prakash, B. Diederich, R. Heintzmann, and L. Schermelleh, "Super-resolution microscopy: A brief history and new avenues," *Philos. Trans. Roy. Soc. A*, vol. 380, no. 2220, 2022, Art. no. 20210110.

[61] A. A. Baniya, T.-K. Lee, P. W. Eklund, and S. Aryal, "A survey of deep learning video super-resolution," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 4, pp. 2655–2676, Aug. 2024.

[62] J. Li et al., "A systematic survey of deep learning-based single-image super-resolution," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–40, Oct. 2024.

[63] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2434–2442.

[64] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 537–546.

[65] N. M. Gottschling, V. Antun, A. C. Hansen, and B. Adcock, "The troublesome kernel: On hallucinations, no free lunches, and the accuracy-stability tradeoff in inverse problems," *SIAM Rev.*, vol. 67, no. 1, pp. 73–104, Mar. 2025.

[66] M. L. de Leeuw den Bouter, G. Ippolito, T. P. A. O'Reilly, R. F. Remis, M. B. van Gijzen, and A. G. Webb, "Deep learning-based single image super-resolution for low-field MR brain images," *Sci. Rep.*, vol. 12, no. 1, p. 6362, Apr. 2022.

[67] Q. Lyu, C. You, H. Shan, and G. Wang, "Super-resolution MRI through deep learning," 2018, *arXiv:1810.06776*.

[68] J. Andrew et al., "Super-resolution reconstruction of brain magnetic resonance images via lightweight autoencoder," *Informat. Med. Unlocked*, vol. 26, Feb. 2021, Art. no. 100713.

[69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6629–6640.

[70] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9195–9204.

[71] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. 1st Int. Conf. Image Process.*, vol. 2, 1994, pp. 168–172.

[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[73] L. Wang et al., "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12250–12259.

[74] R. Chen, H. Sheng, D. Yang, S. Wang, Z. Cui, and R. Cong, "Take your model further: A general post-refinement network for light field disparity estimation via BadPix correction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 2023, pp. 331–339.

[75] H. Sheng et al., "LFNAT 2023 challenge on light field depth estimation: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3473–3485.

[76] H. Sheng, R. Cong, D. Yang, R. Chen, S. Wang, and Z. Cui, "UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7880–7893, Nov. 2022.

[77] R. Cong, D. Yang, R. Chen, S. Wang, Z. Cui, and H. Sheng, "Combining implicit-explicit view correlation for light field semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9172–9181.

[78] H. Sheng, S. Zhang, X. Liu, and Z. Xiong, "Relative location for light field saliency detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1631–1635.

[79] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[80] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[81] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 21–37.

[82] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[83] J. Shin Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. So Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2167–2176.

[84] K. Fu, Q. Zhao, and I. Y.-H. Gu, "RefiNet: A deep segmentation assisted refinement network for salient object detection," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 457–469, Feb. 2019.

[85] W. Liu, C. Zhang, G. Lin, and F. Liu, "CRNet: Cross-reference networks for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4165–4173.

[86] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[87] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[88] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[89] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–16.

[90] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "MaX-DeepLab: End-to-end panoptic segmentation with mask transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5463–5474.

[91] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, Feb. 2024, Art. no. 127063.

[92] X. Liu, H.-F. Yu, I. Dhillon, and C.-J. Hsieh, "Learning to encode position for transformer with continuous dynamical model," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6327–6335.

[93] S. Luo, S. Li, S. Zheng, T. Liu, L. Wang, and D. He, "Your transformer may not be as powerful as you expect," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 4301–4315.

[94] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, Apr. 2024.

[95] C. Luo, "Understanding diffusion models: A unified perspective," 2022, *arXiv:2208.11970*.

[96] A. Li, L. Zhang, Y. Liu, and C. Zhu, "Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12480–12490.

[97] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2019, pp. 63–79.

[98] Y. Wei, S. Gu, Y. Li, R. Timofte, L. Jin, and H. Song, "Unsupervised real-world image super resolution via domain-distance aware training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13385–13394.

[99] L. Wang et al., "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10576–10585.

[100] W. Wang, H. Zhang, Z. Yuan, and C. Wang, "Unsupervised real-world super-resolution: A domain adaptation perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4298–4307.

[101] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.

[102] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3516–3525.

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

[104] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5690–5699.

[105] K. Nazeri, H. Thasarathan, and M. Ebrahimi, "Edge-informed single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3275–3284.

[106] X. Cheng, Z. Fu, and J. Yang, "Zero-shot image super-resolution with depth guided internal degradation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 265–280.

[107] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 370–378.

[108] J. Gu et al., "Super-resolution by predicting offsets: An ultra-efficient super-resolution network for rasterized images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 583–598.

[109] X. Jiang, N. Wang, J. Xin, K. Li, X. Yang, and X. Gao, "Training binary neural network without batch normalization for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1700–1707.

[110] S. Park, J. Yoo, D. Cho, J. Kim, and T. H. Kim, "Fast adaptation to super-resolution networks via meta-learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 754–769.

[111] S. Tian, M. Lu, J. Liu, Y. Guo, Y. Chen, and S. Zhang, "CABM: Content-aware bit mapping for single image super-resolution network with large input," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1756–1765.

[112] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.

[113] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, *arXiv:1603.07285*.

[114] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[115] Y. Fan et al., "Balanced two-stage residual networks for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1157–1164.

[116] L. Zhang et al., "Nonlinear regression via deep negative correlation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 982–998, Mar. 2021.

[117] X. Luo, "LatticeNet: Towards lightweight image super-resolution with lattice block," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 272–289.

[118] J. Chao et al., "Equivalent transformation and dual stream network construction for mobile image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14102–14111.

[119] J. Lin, X. Luo, M. Hong, Y. Qu, Y. Xie, and Z. Wu, "Memory-friendly scalable super-resolution via rewinding lottery ticket hypothesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14398–14407.

[120] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 723–731.

[121] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.

[122] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 465–482.

[123] F. Kong et al., "Residual local feature network for efficient super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 766–776.

[124] G. Gao, W. Li, J. Li, F. Wu, H. Lu, and Y. Yu, "Feature distillation interaction weighting network for lightweight image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 661–669.

[125] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[126] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3147–3155.

[127] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[128] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1654–1663.

[129] Q. Liao and T. Poggio, "Bridging the gaps between residual learning, recurrent neural networks and visual cortex," 2016, *arXiv:1604.03640*.

[130] Y. Zhang, H. Wang, C. Qin, and Y. Fu, "Aligned structured sparsity learning for efficient image super-resolution," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 2695–2706.

[131] J. Oh, H. Kim, S. Nah, C. Hong, J. Choi, and K. M. Lee, "Attentive fine-grained structured sparsity for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17652–17661.

[132] J. Wang, H. Wang, Y. Zhang, Y. Fu, and Z. Tao, "Iterative soft shrinkage learning for efficient image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12556–12565.

[133] Y. Jo and S. Joo Kim, "Practical single-image super-resolution using look-up table," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 691–700.

[134] C. Yang, M. Jin, Y. Xu, R. Zhang, Y. Chen, and H. Liu, "SepLUT: Separable image-adaptive lookup tables for real-time image enhancement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2022, pp. 201–217.

[135] C. Ma, J. Zhang, J. Zhou, and J. Lu, "Learning series-parallel lookup tables for efficient image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 305–321.

[136] G. Liu, Y. Ding, M. Li, M. Sun, X. Wen, and B. Wang, "Reconstructed convolution module based look-up tables for efficient image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12183–12192.

[137] D. Song, C. Xu, X. Jia, Y. Chen, C. Xu, and Y. Wang, "Efficient residual dense block search for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12007–12014.

[138] Y. Chen, C. Gao, E. Robb, and J. Huang, "NAS-DIP: Learning deep image prior with neural architecture search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 442–459.

[139] X. Chu, B. Zhang, H. Ma, R. Xu, and Q. Li, "Fast, accurate and lightweight super-resolution with neural architecture search," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 59–64.

[140] C. Jiang et al., "Local implicit grid representations for 3D scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2020, pp. 6001–6010.

[141] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jul. 2021, pp. 8628–8638.

[142] J. Lee and K. H. Jin, "Local texture estimator for implicit representation function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1919–1928.

[143] J.-E. Yao, L.-Y. Tsao, Y.-C. Lo, R. Tseng, C.-C. Chang, and C.-Y. Lee, "Local implicit normalizing flow for arbitrary-scale image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1776–1785.

[144] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.

[145] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.

[146] 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. (CVPR)*, Jun. 2019, pp. 1604–1613.

[147] Z. Luo, Y. Huang, L. Shang, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 5632–5643.

[148] S. Y. Kim, H. Sim, and M. Kim, "KOALAnet: Blind super-resolution using kernel-oriented adaptive local adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10606–10615.

[149] Z. Hui, J. Li, X. Wang, and X. Gao, "Learning the non-differentiable optimization for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2093–2102.

[150] T. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.

[151] Y.-S. Xu, S.-Y.-R. Tseng, Y. Tseng, H.-K. Kuo, and Y.-M. Tsai, "Unified dynamic convolutional network for super-resolution with variational degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12493–12502.

[152] Z. Luo, H. Huang, L. Yu, Y. Li, H. Fan, and S. Liu, "Deep constrained least squares for blind image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17621–17631.

[153] T. Liu, J. Cheng, and S. Tan, "Spectral Bayesian uncertainty for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18166–18175.

[154] S. A. Magid et al., "Dynamic high-pass filtering and multi-spectral attention for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4288–4297.

[155] M. E. Helou, R. Zhou, and S. Süsstrunk, "Stochastic frequency masking to improve super-resolution and denoising networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 749–766.

[156] W. Xie, D. Song, C. Xu, C. Xu, H. Zhang, and Y. Wang, "Learning frequency-aware dynamic network for efficient super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4308–4317.

[157] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4180–4189.

[158] Z. Zhong, T. Shen, Y. Yang, Z. Lin, and C. Zhang, "Joint sub-bands learning with clique structures for wavelet domain super-resolution," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 165–175.

[159] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[160] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.

[161] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[162] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing

internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[163] Y. Fan, J. Yu, D. Liu, and T. S. Huang, "Scale-wise convolution for image restoration," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10770–10777.
[164] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," in *Proc. 35th Int. Conf. Mach. Learn.*, Jul. 2018, pp. 244–253.
[165] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.
[166] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–18.
[167] S. W. Zamir et al., "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 492–511.
[168] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.
[169] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 191–207.
[170] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 56–72.
[171] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3517–3526.
[172] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.
[173] P. N. Michelini, H. Liu, and D. Zhu, "Multigrid backprojection super-resolution and deep filter visualization," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, 2018, pp. 4642–4650.
[174] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP, Graph. Models Image Process.*, vol. 53, no. 3, pp. 231–239, May 1991.
[175] Y. Guo et al., "Closed-loop matters: Dual regression networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5407–5416.
[176] M. Xiao et al., "Invertible image rescaling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 126–144.
[177] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-SR: A magnification-arbitrary network for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1575–1584.
[178] L. Wang, Y. Wang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning a single network for scale-arbitrary super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4801–4810.
[179] S. Maeda, "Image super-resolution with deep dictionary," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 464–480.
[180] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.
[181] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[182] N. Ahn, B. Kang, and K. A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 256–272.
[183] W. Li, K. Zhou, L. Qi, N. Jiang, J. Lu, and J. Jia, "LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 20343–20355.
[184] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2359–2368.
[185] H. Choi, J. Lee, and J. Yang, "N-gram in Swin transformers for efficient lightweight image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2071–2081.
[186] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, "SRFormer: Permuted self-attention for single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12734–12745.
[187] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 913–919.
[188] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 457–466.
[189] X. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 649–667.
[190] X. Li, J. Dong, J. Tang, and J. Pan, "DLGSANet: Lightweight dynamic local and global self-attention network for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12746–12755.
[191] L. Sun, J. Dong, J. Tang, and J. Pan, "Spatially-adaptive feature modulation for efficient image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13144–13153.
[192] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.
[193] W. Li, X. Lu, S. Qian, and J. Lu, "On efficient transformer-based image pre-training for low-level vision," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 1089–1097.
[194] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.
[195] X. Chen, X. Wang, J. Zhou, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22367–22377.
[196] X. Wang, X. Chen, B. Ni, H. Wang, Z. Tong, and Y. Liu, "Deep arbitrary-scale image super-resolution via scale-equivariance pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1786–1795.
[197] J. Cao et al., "CiaoSR: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1796–1807.
[198] H. Wang, X. Chen, B. Ni, Y. Liu, and J. Liu, "Omni aggregation networks for lightweight image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22378–22387.
[199] H.-W. Chen, Y.-S. Xu, M.-F. Hong, Y.-M. Tsai, H.-K. Kuo, and C.-Y. Lee, "Cascaded local implicit transformer for arbitrary-scale super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18257–18267.
[200] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12278–12287.
[201] A. Ray, G. Kumar, and M. H. Kolekar, "CFAT: Unleashing triangular windows for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26120–26129.
[202] C. K. Sønderby, J. A. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised MAP inference for image super-resolution," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–17.
[203] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[204] S.-J. Park, "SRFeat: Single image super-resolution with feature discrimination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 439–455.
[205] T. Shang, Q. Dai, S. Zhu, T. Yang, and Y. Guo, "Perceptual extreme super resolution network with receptive field block," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1778–1787.
[206] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.
[207] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "RankSRGAN: Generative adversarial networks with ranker for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3096–3105.
[208] R. Zhou and S. Susstrunk, "Kernel modeling super-resolution on real low-resolution images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2433–2443.
[209] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7769–7778.
[210] S. Maeda, "Unpaired image super-resolution using pseudo-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 288–297.
[211] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[212] R. Lee et al., "Journey towards tiny perceptual super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 85–102.
[213] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14245–14254.
[214] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.
[215] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Learning the degradation distribution for blind image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 6063–6072.
[216] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4491–4500.
[217] S. Vasu, N. T. Madam, and A. N. Rajagopalan, "Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 114–131.
[218] M. S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran, "SROBB: Targeted perceptual loss for single image super-resolution,"

in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2710–2719.
[219] S. Y. Kim, J. Oh, and M. Kim, "FISR: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11278–11286.
[220] D. Fuoli, L. Van Gool, and R. Timofte, "Fourier space losses for efficient perceptual image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2340–2349.
[221] S. H. Park, Y. S. Moon, and N. I. Cho, "Perception-oriented single image super-resolution using optimal objective estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1725–1735.
[222] C. Korkmaz, A. M. Tekalp, and Z. Dogan, "Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5926–5936.
[223] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 185–200.
[224] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. NIPS*, 2019, pp. 284–293.
[225] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7474–7489, Nov. 2022.
[226] B. Li et al., "SeD: Semantic-aware discriminator for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25784–25795.
[227] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–532.
[228] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng, "ODE-inspired network design for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1732–1741.
[229] D. Song, Y. Wang, H. Chen, C. Xu, C. Xu, and D. Tao, "AdderSR: Towards energy efficient image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15648–15657.
[230] C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7898–7911, Nov. 2022.
[231] J. Park, S. Son, and K. M. Lee, "Content-aware local GAN for photo-realistic super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10585–10594.
[232] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3217–3226.
[233] Y. Zhang, D. Wei, C. Qin, H. Wang, H. Pfister, and Y. Fu, "Context reasoning attention network for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4278–4287.
[234] J. Liang, H. Zeng, and L. Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5657–5666.
[235] H. Sahak, D. Watson, C. Saharia, and D. Fleet, "Denoising diffusion probabilistic models for robust image super-resolution in the wild," 2023, *arXiv:2302.07864*.
[236] B. Xia et al., "DiffIR: Efficient diffusion model for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13095–13105.
[237] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.
[238] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.
[239] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *J. Mach. Learn. Res.*, vol. 23, no. 47, pp. 1–33, 2022.
[240] S. Gao et al., "Implicit diffusion models for continuous super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10021–10030.
[241] J. Wang, Z. Yue, S. Zhou, K. C. K. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *Int. J. Comput. Vis.*, vol. 132, no. 12, pp. 5929–5949, Dec. 2024.
[242] X. Lin et al., "DiffBIR: Toward blind image restoration with generative diffusion prior," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 430–448.
[243] T. Yang, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2023, pp. 74–91.
[244] F. Yu et al., "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25669–25680.
[245] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "SeeSR: Towards semantics-aware real-world image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25456–25467.
[246] 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. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.
[247] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2012, pp. 1–10.
[248] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.
[249] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 5197–5206.
[250] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 416–423.
[251] Y. Matsui et al., "Sketch-based Manga retrieval using Manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.
[252] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[253] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 111–126.
[254] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 126–135.
[255] R. Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 114–125.
[256] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–12.
[257] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.
[258] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3224–3232.
[259] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 235–243.
[260] J. Guo and H. Chao, "Building an end-to-end spatial–temporal convolutional network for video super-resolution," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 31, 2017, pp. 4053–4060.
[261] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6626–6634.
[262] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, and Z. Xue, "Residual invertible spatio-temporal network for video super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 5981–5988.
[263] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3897–3906.
[264] D. Fuoli, S. Gu, and R. Timofte, "Efficient video super-resolution through recurrent latent space propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3476–3485.
[265] T. Isobe, X. Jia, S. Gu, S. Li, S. Wang, and Q. Tian, "Video super-resolution with recurrent structure-detail network," in *Proc. Eur. Conf. Comput.*, Aug. 2020, pp. 645–660.
[266] B. Xia et al., "Structured sparsity learning for efficient video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22638–22647.
[267] B. Bare, B. Yan, C. Ma, and K. Li, "Real-time video super-resolution via motion convolution kernel estimation," *Neurocomputing*, vol. 367, pp. 236–245, Nov. 2019.
[268] B. Yan, C. Lin, and W. Tan, "Frame and feature-context video super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5597–5604.
[269] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3360–3369.
[270] S. Li, F. He, B. Du, L. Zhang, Y. Xu, and D. Tao, "Fast spatio-temporal residual network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10514–10523.
[271] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3106–3115.
[272] T. Isobe et al., "Video super-resolution with temporal group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8008–8017.
[273] J. Pan, H. Bai, J. Dong, J. Zhang, and J. Tang, "Deep blind video super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.
[274] G. Li et al., "Towards high-quality and efficient video super-resolution via spatial–temporal data overfitting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10259–10269.
[275] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential

components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4947–4956.

[276] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[277] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for video super-resolution through HR optical flow estimation," in *Proc. 14th Asian Conf. Comput. Vis.*, Perth, WA, Australia, 2019, pp. 514–529.

[278] Z. Wang et al., "Multi-memory convolutional neural network for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2530–2544, May 2019.

[279] D. Li, Y. Liu, and Z. Wang, "Video super-resolution using non-simultaneous fully recurrent convolutional network," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1342–1355, Mar. 2019.

[280] R. Kalarot and F. Porikli, "MultiBoot VSR: Multi-stage multi-reference bootstrapping for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2060–2069.

[281] H. Wang, D. Su, C. Liu, L. Jin, X. Sun, and X. Peng, "Deformable non-local network for video super-resolution," *IEEE Access*, vol. 7, pp. 177734–177744, 2019.

[282] S. Y. Kim, J. Lim, T. Na, and M. Kim, "Video super-resolution based on 3D-CNNS with consideration of scene change," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2831–2835.

[283] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4472–4480.

[284] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[285] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding deformable alignment in video super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 973–981.

[286] T. Isobe et al., "Look back and forth: Video super-resolution with explicit temporal difference modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17411–17420.

[287] Z. Chen et al., "VideoINR: Learning video implicit neural representation for continuous space-time super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2037–2047.

[288] B. N. Chiche, A. Woiselle, J. Frontera-Pons, and J.-L. Starck, "Stable long-term recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 827–836.

[289] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "Investigating tradeoffs in real-world video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5952–5961.

[290] F. Li, L. Zhang, Z. Liu, J. Lei, and Z. Li, "Multi-frequency representation enhancement with privilege information for video super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12768–12779.

[291] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.

[292] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "MuCAN: Multi-correspondence aggregation network for video super-resolution," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 335–351.

[293] H. Liu, P. Zhao, Z. Ruan, F. Shang, and Y. Liu, "Large motion video super-resolution with dual subnet and multi-stage communicated upsampling," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2021, pp. 2127–2135.

[294] P. Yi et al., "Omniscient video super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4429–4438.

[295] Y. Li, P. Jin, F. Yang, C. Liu, M.-H. Yang, and P. Milanfar, "COMISR: Compression-informed video super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2523–2532.

[296] J. Yu, J. Liu, L. Bo, and T. Mei, "Memory-augmented non-local attention for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17834–17843.

[297] J. Cao et al., "Towards interpretable video super-resolution via alternating optimization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 393–411.

[298] Y. Wang, T. Isobe, X. Jia, X. Tao, H. Lu, and Y.-W. Tai, "Compression-aware video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2012–2021.

[299] J. Liang et al., "Recurrent video restoration transformer with guided deformable attention," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 378–393.

[300] X. Zhou, L. Zhang, X. Zhao, K. Wang, L. Li, and S. Gu, "Video super-resolution transformer with masked inter&intra-frame attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25399–25408.

[301] X. Yang, C. He, J. Ma, and L. Zhang, "Motion-guided latent diffusion for temporally consistent real-world video super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2023, pp. 224–242.

[302] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. Loy, "Upscale-A-video: Temporal-consistent diffusion model for real-world video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2535–2545.

[303] S. Shi, J. Gu, L. Xie, X. Wang, Y. Yang, and C. Dong, "Rethinking alignment in video super-resolution transformers," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 36081–36093.

[304] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5972–5981.

[305] J. Cao, Y. Li, K. Zhang, and L. Van Gool, "Video super-resolution transformer," 2021, *arXiv:2106.06847*.

[306] J. Liang et al., "VRT: A video restoration transformer," *IEEE Trans. Image Process.*, vol. 33, pp. 2171–2182, 2024.

[307] H. Wang, X. Xiang, Y. Tian, W. Yang, and Q. Liao, "STDAN: Deformable attention network for space-time video super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 10606–10616, Aug. 2024.

[308] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: Real-time spatial temporal transformer for space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17441–17451.

[309] Z. Qiu, H. Yang, J. Fu, D. Liu, C. Xu, and D. Fu, "Learning degradation-robust spatiotemporal frequency-transformer for video super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14888–14904, Dec. 2023.

[310] C. Liu, H. Yang, J. Fu, and X. Qian, "Learning trajectory-aware transformer for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5687–5696.

[311] A. Lucas, S. Lopez-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3312–3327, Jul. 2019.

[312] O. Thawakar, P. W. Patil, A. Dudhane, S. Murala, and U. Kulkarni, "Image and video super resolution using recurrent generative adversarial network," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[313] P. Yi, Z. Wang, K. Jiang, J. Jiang, T. Lu, and J. Ma, "A progressive fusion generative adversarial network for realistic and consistent video super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2264–2280, May 2022.

[314] A. Chadha, J. Britto, and M. M. Roja, "ISeeBetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks," *Comput. Vis. Media*, vol. 6, no. 3, pp. 307–317, Sep. 2020.

[315] Y. Cao, C. Wang, C. Song, Y. Tang, and H. Li, "Real-time super-resolution system of 4K-video based on deep learning," in *Proc. IEEE 32nd Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2021, pp. 69–76.

[316] S. S. Andrei, N. Shapovalova, and W. Mayol-Cuevas, "SUPERVEGAN: Super resolution video enhancement GAN for perceptually improving low bitrate streams," *IEEE Access*, vol. 9, pp. 91160–91174, 2021.

[317] J. Wang, G. Teng, and P. An, "Video super-resolution based on generative adversarial network and edge enhancement," *Electronics*, vol. 10, no. 4, p. 459, Feb. 2021.

[318] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning temporal coherence via self-supervision for GAN-based video generation," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 1–75, Aug. 2020.

[319] X. Wang, A. Lucas, S. Lopez-Tapia, X. Wu, R. Molina, and A. K. Katsaggelos, "Spatially adaptive losses for video super-resolution with GANs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1697–1701.

[320] S. López-Tapia, A. Lucas, R. Molina, and A. K. Katsaggelos, "A single video super-resolution GAN for multiple downsampling operators based on pseudo-inverse image formation models," *Digit. Signal Process.*, vol. 104, Sep. 2020, Art. no. 102801.

[321] T. Brooks et al., "Generating long videos of dynamic scenes," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 31769–31781.

[322] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[323] S. Nah et al., "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1996–2005.

[324] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 933–948, Mar. 2021.

[325] M. Haris, G. Shakhnarovich, and N. Ukita, "Space-time-aware multi-resolution video enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2856–2865.

[326] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3370–3379.

[327] W. Sun, J. Sun, Y. Zhu, and Y. Zhang, "Video super-resolution via dense non-local spatial–temporal convolutional network," *Neurocomputing*, vol. 403, pp. 1–12, Aug. 2020.

[328] H. Song, W. Xu, D. Liu, B. Liu, Q. Liu, and D. N. Metaxas, "Multi-stage feature fusion network for video super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 2923–2934, 2021.

[329] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced

deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1954–1963.

[330] T. Isobe, F. Zhu, and S. Wang, "Revisiting temporal modeling for video super-resolution," in *Proc. Brit. Mach. Vis. Conf.*, 2020, pp. 3476–3485.

[331] Y. Xu et al., "VideoGigaGAN: Towards detail-rich video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 2139–2149.

[332] Z. Chen et al., "Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 9232–9241.

[333] X. Yuan, J. Baek, K. Xu, O. Tov, and H. Fei, "Inflation with diffusion: Efficient temporal adaptation for text-to-video super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 489–496.

[334] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.

[335] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12031–12038.

[336] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "*Flickr1024*: A large-scale dataset for stereo image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3852–3857.

[337] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.

[338] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[339] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[340] B. Yan, C. Ma, B. Bare, W. Tan, and S. Hoi, "Disparity-aware domain adaptation in stereo image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13176–13184.

[341] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 496–500, 2020.

[342] X. Zhu, K. Guo, H. Fang, L. Chen, S. Ren, and B. Hu, "Cross view capture for stereo image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 3074–3086, 2022.

[343] Q. Dai, J. Li, Q. Yi, F. Fang, and G. Zhang, "Feedback network for mutually boosted stereo image super-resolution and disparity estimation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1985–1993.

[344] X. Chu, L. Chen, and W. Yu, "NAFSSR: Stereo image super-resolution using NAFNet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2022, pp. 1239–1248.

[345] J. Lin, L. Yin, and Y. Wang, "Steformer: Efficient stereo image super-resolution with transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 8396–8407, 2023.

[346] J. Lei et al., "Deep stereoscopic image super-resolution via interaction module," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3051–3061, Aug. 2021.

[347] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo, "Symmetric parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 766–775.

[348] Z. Yang, M. Yao, J. Huang, M. Zhou, and F. Zhao, "SIR-former: Stereo image restoration using transformer," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6377–6385.

[349] D. Zhang, F. Huang, S. Liu, X. Wang, and Z. Jin, "SwinFIR: Revisiting the SwinIR with fast Fourier convolution and improved training for image super-resolution," 2022, *arXiv:2208.11247*.

[350] H. Gao and D. Dang, "Learning accurate and enriched features for stereo image super-resolution," *Pattern Recognit.*, vol. 159, Mar. 2025, Art. no. 111170.

[351] H. Gao, J. Yang, Y. Zhang, J. Yang, B. Ma, and D. Dang, "Learning optimal combination patterns for lightweight stereo image super-resolution," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 5566–5574.

[352] Y. Zhou et al., "Towards real world stereo image super-resolution via hybrid degradation model and discriminator for implied stereo image information," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124457.

[353] T. Zhang, Y. Gu, X. Huang, E. Tu, and J. Yang, "Stereo endoscopic image super-resolution using disparity-constrained parallel attention," 2020, *arXiv:2003.08539*.

[354] C. Chen, C. Qing, X. Xu, and P. Dickinson, "Cross parallax attention network for stereo image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 202–216, 2022.

[355] Q. Xu, L. Wang, Y. Wang, W. Sheng, and X. Deng, "Deep bilateral learning for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 28, pp. 613–617, 2021.

[356] J. Dan, Z. Qu, X. Wang, and J. Gu, "A disparity feature alignment module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 28, pp. 1285–1289, 2021.

[357] H. Imani, M. B. Islam, and L.-K. Wong, "A new dataset and transformer for stereoscopic video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 706–715.

[358] K. Jin et al., "SwiniPASSR: Swin transformer based parallax attention network for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 919–928.

[359] H. Guo, J. Li, G. Gao, Z. Li, and T. Zeng, "PFT-SSR: Parallax fusion transformer for stereo image super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[360] C. Ma, B. Yan, W. Tan, and X. Jiang, "Perception-oriented stereo image super-resolution," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2420–2428.

[361] V. Q. Dinh, T.-D. Nguyen, and P. H. Nguyen, "Stereo domain translation for denoising and super-resolution using correlation loss," in *Proc. 7th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, Nov. 2020, pp. 261–266.

[362] Y. Zhou et al., "DiffSteISR: Harnessing diffusion prior for superior real-world stereo image super-resolution," *Neurocomputing*, vol. 623, Mar. 2025, Art. no. 129437.

[363] M. Peris, S. Martull, A. Maki, Y. Ohkawa, and K. Fukui, "Towards a simulation driven stereo vision system," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 1038–1042.

[364] Z. Liang, Y. Wang, L. Wang, J. Yang, and S. Zhou, "Light field image super-resolution with transformers," *IEEE Signal Process. Lett.*, vol. 29, pp. 563–567, 2022.

[365] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 24–32.

[366] Y. Yuan, Z. Cao, and L. Su, "Light-field image superresolution using a combined deep CNN based on EPI," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1359–1363, Sep. 2018.

[367] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2319–2330, May 2019.

[368] H. Zhu, M. Guo, H. Li, Q. Wang, and A. Robles-Kelly, "Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 3019–3033, Jun. 2021.

[369] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 290–308.

[370] N. Meng, H. K.-H. So, X. Sun, and E. Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 873–886, Mar. 2021.

[371] Z. Cheng, Z. Xiong, C. Chen, D. Liu, and Z.-J. Zha, "Light field super-resolution with zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2021, pp. 10010–10019.

[372] Y. Wang et al., "Disentangling light fields for super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425–443, Jan. 2023.

[373] J. Jin, J. Hou, J. Chen, and S. Kwong, "Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2260–2269.

[374] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11141–11148.

[375] S. Zhang, S. Chang, and Y. Lin, "End-to-end light field spatial super-resolution network using multiple epipolar geometry," *IEEE Trans. Image Process.*, vol. 30, pp. 5956–5968, 2021.

[376] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4274–4286, Sep. 2018.

[377] S. Zhang, Y. Lin, and H. Sheng, "Residual networks for light field image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11046–11055.

[378] Z. Cheng, Z. Xiong, and D. Liu, "Light field super-resolution by jointly exploiting internal and external similarities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2604–2616, Aug. 2020.

[379] G. Liu, H. Yue, J. Wu, and J. Yang, "Intra-inter view interaction network for light field image super-resolution," *IEEE Trans. Multimedia*, vol. 25, pp. 256–266, 2023.

[380] Y. Wang et al., "Light field image super-resolution using deformable convolution," *IEEE Trans. Image Process.*, vol. 30, pp. 1057–1071, 2021.

[381] N. Meng, X. Wu, J. Liu, and E. Y. Lam, "High-order residual network for light field super-resolution," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 11757–11764.

[382] K. Ko, Y. J. Koh, S. Chang, and C.-S. Kim, "Light field super-resolution via adaptive feature remixing," *IEEE Trans. Image Process.*, vol. 30, pp. 4114–4128, 2021.

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

[384] Z. Xiao, Y. Liu, R. Gao, and Z. Xiong, "CutMIB: Boosting light field super-resolution via multi-view image blending," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2023, pp. 1672–1682.

[385] S. Wang, T. Zhou, Y. Lu, and H. Di, "Detail-preserving transformer for light field image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2522–2530.

[386] J. Shabbir, M. Z. Alam, and M. U. Mukati, "Learning texture transformer network for light field super-resolution," in *Proc. Eur. Conf. Vis. Media Prod. (CVMP)*, 2022, p. 1.
[387] X. Guo, X. Sang, B. Yan, D. Chen, and P. Wang, "Light field image super-resolution based on raw data with transformers," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 39, no. 12, pp. 2131–2141, 2022.
[388] Z. Wang and Y. Lu, "Multi-granularity aggregation transformer for light field image super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 261–265.
[389] Z. Liang, Y. Wang, L. Wang, J. Yang, S. Zhou, and Y. Guo, "Learning non-local spatial-angular correlation for light field image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12342–12352.
[390] Z. Cheng, Y. Liu, and Z. Xiong, "Spatial-angular versatile convolution for light field reconstruction," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 1131–1144, 2022.
[391] M. Ko, D. Kim, and K. Kim, "GAN-based super resolution for accurate 3D surface reconstruction from light field skin images towards haptic palpation," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 392–397.
[392] L. Ruan, B. Chen, and M. L. Lam, "Light field synthesis from a single image using improved Wasserstein generative adversarial network," in *Proc. Annu. Eur. Assoc. Comput. Graph. Conf. (EG)*, 2018, pp. 19–20.
[393] M. Zhu, A. Alperovich, O. Johannsen, A. Sulc, and B. Goldluecke, "An epipolar volume autoencoder with adversarial loss for deep light field super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1853–1861.
[394] B. Chen, L. Ruan, and M.-L. Lam, "LFGAN: 4D light field synthesis from a single RGB image," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–20, Feb. 2020.
[395] N. Meng, T. Zeng, and E. Y. Lam, "Spatial and angular reconstruction of light field based on deep generative networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4659–4663.
[396] X. Guo et al., "Real-time optical reconstruction for a three-dimensional light-field display based on path-tracing and CNN super-resolution," *Opt. Exp.*, vol. 29, no. 23, pp. 37862–37876, 2021.
[397] A. Wafa and P. Nasiopoulos, "Light field GAN-based view synthesis using full 4D information," in *Proc. Eur. Conf. Vis. Media Prod.*, Dec. 2022, pp. 1–7.
[398] W. Chao, J. Zhao, F. Duan, and G. Wang, "LFSRDiff: Light field image super-resolution via diffusion models," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
[399] M. Rerábek and T. Ebrahimi, "New light field image dataset," in *Proc. Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2016, pp. 1–10.
[400] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.
[401] S. Wanner, S. Meister, and B. Goldlucke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. 18th Int. Workshop Vis., Modeling Vis.*, 2013, pp. 225–226.
[402] M. Le Pendu, X. Jiang, and C. Guillemot, "Light field inpainting propagation via low rank matrix completion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1981–1993, Apr. 2018.
[403] V. Vaish and A. Adams, "The (new) Stanford light field archive," Dept. Comput. Graph. Lab., Stanford Univ., Stanford, CA, USA, Tech. Rep. 7, 2008, vol. 6, no. 7.
[404] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.
[405] F. Fang, J. Li, and T. Zeng, "Soft-edge assisted network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 4656–4668, 2020.
[406] X. Di et al., "QMambaBSR: Burst image super-resolution with query state space model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 23080–23090.
[407] X. Lei, W. Zhang, and W. Cao, "DVMSR: Distillated vision mamba for efficient super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 6536–6546.
[408] Y. Huang, T. Miyazaki, X. Liu, and S. Omachi, "IRSRMamba: Infrared image super-resolution via mamba-based wavelet transform feature modulation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5005416.
[409] Y. Ren, X. Li, M. Guo, B. Li, S. Zhao, and Z. Chen, "MambaCSR: Dual-interleaved scanning for compressed image super-resolution with SSMs," 2024, *arXiv:2408.11758*.
[410] H. Li, "PAMS: Quantized super-resolution via parameterized max scale," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 564–580.
[411] C. Hong, H. Kim, S. Baik, J. Oh, and K. M. Lee, "DAQ: Channel-wise distribution-aware quantization for deep image super-resolution networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 2675–2684.
[412] C. Hong et al., "CaDyQ: Content-aware dynamic quantization for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 367–383.

ABOUT THE AUTHORS

Le Zhang (Member, IEEE) received the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2016.

He subsequently worked as a Postdoctoral Fellow with the Advanced Digital Sciences Center (ADSC), Singapore, from 2016 to 2018. He was a Research Scientist with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore, from 2018 to 2021. He is a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests lie in computer vision and machine learning.

Dr. Zhang has received multiple honors and awards, including the 2022 Norbert Wiener Review Award from IEEE/CAA JOURNAL OF AUTOMATICA SINICA and best paper awards at IEEE HPCC 2022 and IEEE ICIEA 2022. He currently serves as a Senior Action Editor for NEURAL NETWORKS and an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He has also served as a Guest Editor for several special issues in leading journals, including IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON BIG DATA, and *Pattern Recognition*.

Ao Li received the B.S. degree from Chongqing University, Chongqing, China, in 2017, and the M.S. degree from Chongqing University of Posts and Telecommunications, Chongqing, in 2020. He is currently working toward the Ph.D. degree at the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

His research interests primarily focus on image super-resolution.

Qibin Hou (Member, IEEE) received the Ph.D. degree from the School of Computer Science, Nankai University, Tianjin, China, in 2019.

Then, he worked as a Research Fellow with the National University of Singapore, Singapore. He is currently an Associate Professor with the School of Computer Science, Nankai University. He has published more than 30 articles in top conferences/journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (T-PAMI), CVPR, ICCV, and NeurIPS. His research interests include deep learning and computer vision.

Ce Zhu (Fellow, IEEE) received the B.S. degree in electronic and information engineering from Sichuan University, Chengdu, China, in 1989, and the M.Eng. and Ph.D. degrees from Southeast University, Nanjing, China, in 1992 and 1994, respectively.

He was a Postdoctoral Researcher with the Chinese University of Hong Kong, Hong Kong, in 1995, and the University of Melbourne, Melbourne, VIC, Australia, from 1996 to 1998. He was with Nanyang Technological University, Singapore, for 14 years from 1998 to 2012, where he was a Research Fellow, a Program Manager, an Assistant Professor, and an Associate Professor in 2005. He has been with the University of Electronic Science and Technology of China (UESTC), Chengdu, as a Professor since 2012. He is also the Dean of Glasgow College, a joint school between the University of Glasgow, Glasgow, U.K., and UESTC, China. His research interests include video coding and communications, video analysis and processing, 3D video, visual perception and applications.

Dr. Zhu has served on the editorial boards of multiple international journals, including as an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BROADCASTING, IEEE SIGNAL PROCESSING LETTERS, an Editor of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, and an Area Editor of *Signal Processing: Image Communication*. He has also served as a Guest Editor of several special issues in international journals, including as a Guest Editor (twice) of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He was an APSIPA Distinguished Lecturer from 2021 to 2022 and also an IEEE Distinguished Lecturer of Circuits and Systems Society from 2019 to 2020. He is currently the Chair of IEEE ICME Steering Committee (2024–2025) and the Chair of IEEE Chengdu Section (2024–2028). He is a co-recipient of 10 paper/demo awards at international conferences, including the most recent Best Demo Award in IEEE ICME 2025 and in IEEE MMSP 2022, Best Paper Award in IEEE BMSB 2025, and Best Paper Runner-Up Award in IEEE ICME 2020.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, USA, in 2002.

She is a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel, where she heads the center for biomedical engineering and signal processing and holds the Dorothy and Patrick Gorman Professorial Chair. She is also a Visiting Professor with MIT, a Visiting Scientist with the Broad Institute, Cambridge, USA, and an Adjunct Professor with Duke University, Durham, USA, and was a Visiting Professor with Stanford.

Dr. Eldar is a member of Israel Academy of Sciences and Humanities and a EURASIP Fellow. She has received many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), and the IEEE Kiyo Tomiyasu Award (2016). She was a Horev Fellow of the Leaders in Science and Technology Program, the Technion and an Alon Fellow. She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), and the Award for Women with Distinguished Contributions. She received several best paper awards and best demo awards together with her research students and colleagues, was selected as one of the 50 most influential women in Israel, and was a member of the Israel Committee for Higher Education. She is an Editor-in-Chief of *Foundations and Trends in Signal Processing*, a member of several IEEE Technical Committees and Award Committees, and heads the Committee for Promoting Gender Fairness in Higher Education Institutions in Israel.