Detail-Enhancing Framework for Reference-Based Image Super-Resolution

Zihan Wang[1,3], Ziliang Xiong[2], Hongying Tang[1], Xiaobing Yuan[2]
[1]Shanghai Institution of Micro-system and Information Technology, University of Chinese Academy of Science
[2]Computer Vision Laboratory, Department of Electrical Engineering, LiU, Sweden
[3]School of Information Science and Technology, Shanghai Tech University
wangzh2@shanghaitech.edu.cn, ziliang.xiong@liu.se,
tanghy@mail.sim.ac.cn, sinowsn@mail.sim.ac.cn

Abstract

Recent years have witnessed the prosperity of reference-based image super-resolution (Ref-SR). By importing the high-resolution (HR) reference images into the single image super-resolution (SISR) approach, the ill-posed nature of this long-standing field has been alleviated with the assistance of texture transferred from reference images. Although the significant improvement in quantitative and qualitative results has verified the superiority of Ref-SR methods, the presence of misalignment before texture transfer indicates room for further performance improvement. Existing methods tend to neglect the significance of details in the context of comparison, therefore not fully leveraging the information contained within low-resolution (LR) images. In this paper, we propose a Detail-Enhancing Framework (DEF) for reference-based super-resolution, which introduces the diffusion model to generate and enhance the underlying detail in LR images. If corresponding parts are present in the reference image, our method can facilitate rigorous alignment. In cases where the reference image lacks corresponding parts, it ensures a fundamental improvement while avoiding the influence of the reference image. Extensive experiments demonstrate that our proposed method achieves superior visual results while maintaining comparable numerical outcomes.

1. Introduction

Single image super-resolution (SISR) refers to a computational imaging technique that aims to enhance the resolution and level of detail in a single low-resolution (LR) image, typically achieved by estimating and generating a corresponding high-resolution (HR) counterpart. The essence of SISR lies in the prediction of the pixel values for the required additional pixels from the information present in a single input image. Limited by the ill-posed nature of SISR, a single LR image may generate multiple different HR images, thus resulting in the artifact and hallucination in the final output against the only corresponding ground-truth (GT) image. To ensure the authenticity of the super-resolution (SR) result, it becomes imperative to incorporate supplementary information within the reference images. The semantic information present in reference images, including content and texture, is crucial in restoring input images. Besides, obtaining similar HR reference images is significantly more feasible than acquiring strictly corresponding HR ground-truth (GT) images. To summarize, transferring HR textures from related but different HR reference images to LR input images may recover a faithful result, yielding the idea of reference-based image super-resolution (Ref-SR).

The network architecture of Ref-SR typically comprises the following four components: feature extraction, feature alignment, texture transfer, and texture aggregation. Among them, correspondence is matched between the LR image and the reference image in the feature alignment procedure, which is considered to be the most crucial component. However, image pairs consisting of the LR image and reference image do not share the same resolution, which leads to misalignment. To conquer the misalignment issue, recent advancements in this procedure have shifted the research focus of spatial alignment from point-wise matching [22, 34] to patch matching [1, 8, 13, 21, 29, 31, 33] to improve the matching accuracy. Apart from that, in order to mitigate the resolution gap and obtain domain-consistent image pairs, previous methods tend to simply resize the input LR image to the same resolution of the corresponding reference image, e.g., bicubic interpolation. Lu *et al*.[13] choose to downsample reference images to fit into the matching process for the purpose of reducing computational complexity. While such an approach can somewhat alleviate the misalignment issue to some extent, it neglects the enhancement of details, potentially sabotaging the subsequent image-restoration results. In cases where there are corre-

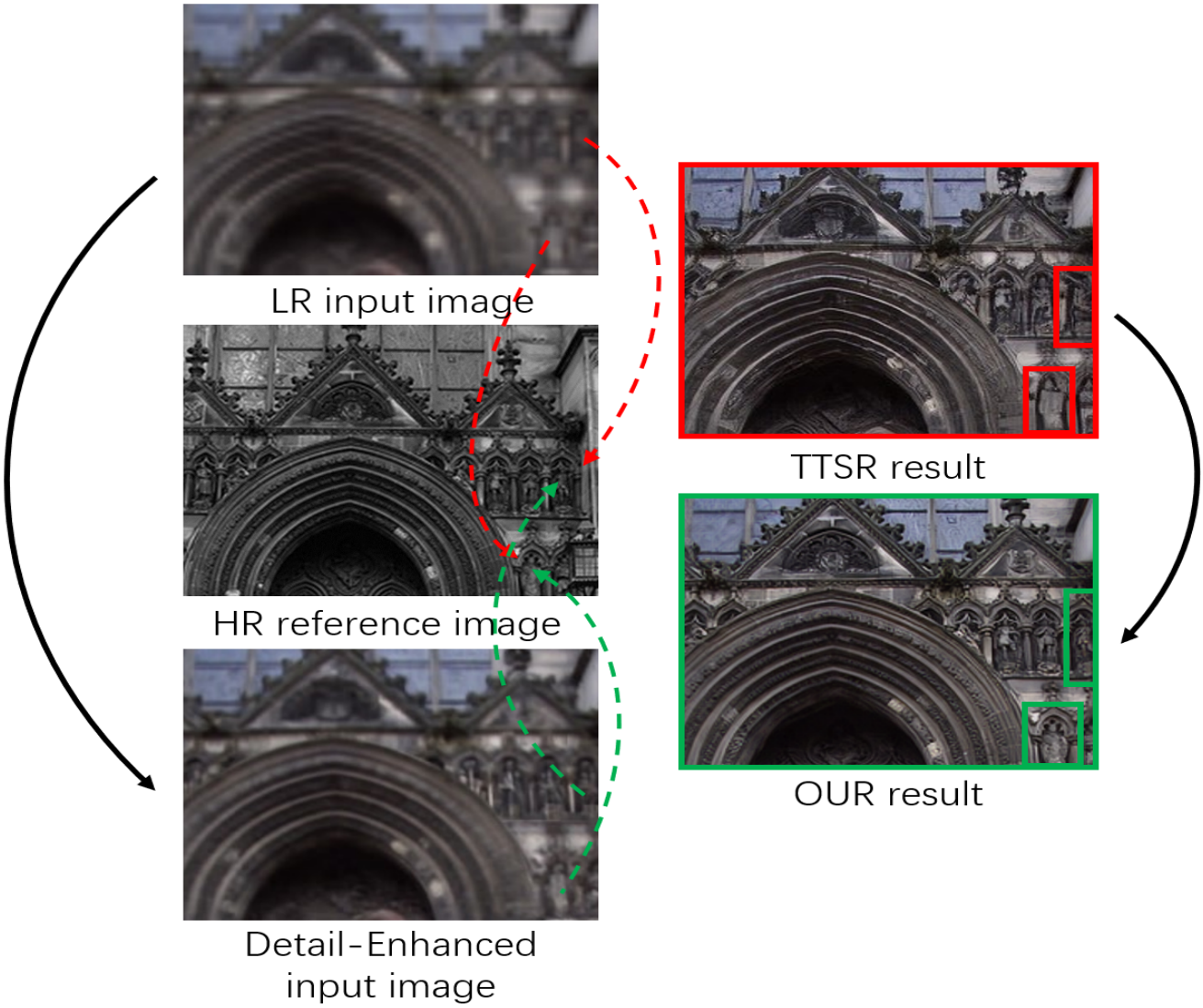

Figure 1. In the existing Ref-SR methods, such as TTSR [31], performance often deteriorates due to misalignment. To address this issue, we propose enhancing the fine-grained textures within images using a pre-trained diffusion model, thereby aiding the alignment process.

sponding features, mere resizing of LR images solely relies on the pixel values in the vicinity to predict the target pixel value, the resulting diminutive receptive field is inadequate for the comprehensive utilization of inherent information. Thus certain corresponding features between image pairs cannot be accurately aligned due to the lack of abundant details during alignment. Meanwhile, specific features within the LR image lack corresponding counterparts in the reference image, rendering them incapable of identifying aligned patches. These unaligned features remain unchanged after bicubic interpolation, hampering the visual quality of output as a result of the lack of details. Therefore, there is still room for improvement in the preprocessing of the LR image due to the absence of detail enhancement in the alignment process.

A natural idea comes up that we enhance the detail of the LR image with generative-based models ahead of alignment. Within the domain of SR, prevalent generative-based models primarily encompass two paradigms: generative adversarial networks (GANs) and diffusion models. Compared with GAN-based models, diffusion models [17] are more stable and robust to various distributions of images. Even though diffusion models exhibit ideal performance in generating details, owing to the ill-posed nature of SR, the intrinsic randomness of diffusion models, and the deficiency in generalization capability, it is prone to generate

artifacts.

To address the misalignment issue and the artifact issue altogether, we first employ theoretical analysis to elucidate the significance and positioning of details in the task of image SR. Then we propose a novel framework, which we dub, Detail-Enhancing Framework (DEF), for reference-based methods that replace the resizing of LR images with a pre-trained diffusion model. The modification of the model structure compensates for inherent limitations from both perspectives. As for Ref-based models, the introduction of diffusion models enriches the detail-wise information within the LR images, thereby benefiting the alignment between the LR image and reference image. In the meantime, to remove the artifacts, textures are transferred from reference images to guide the output of diffusion models as a procedure in the Ref-based model. Experiments have been done on five benchmark datasets, including CUFED5, Manga109, Urban100, Sun80, and WR-SR. Results demonstrate that our proposed framework attains better visual quality with comparable performance with state-of-the-art methods quantitatively.

To summarize, our primary contributions are as follows: 1) We conduct an in-depth investigation into the significant importance of detail enhancement in Ref-SR, which has been overlooked in previous approaches. 2) We proposed the Detail-Enhancing Framework (DEF) that introduces the diffusion models into the Ref-SR models, which not only facilitates a more precise alignment but also reduces artifacts for LR images after alignment. 3) Experimental results demonstrate that our proposed method achieves leading visual performance while maintaining comparable numerical fidelity.

2. Related Work

We will briefly review the historical issue of image SR in this section. Image SR can be roughly classified into two categories: 1) Single image super-resolution (SISR) and 2) Reference-based Image Super-Resolution (Ref-SR).

2.1. Single Image Super-Resolution

Restoring the information of the given LR images is the primary concern of SISR methods. The emergence and prosperity of deep learning contribute to the progress of SISR to a large extent. SRCNN [5] is the pioneer of applying deep learning to SR area. Extensive research based on CNN [9, 12, 20, 24] mostly focuses on prolonging the depth of the network. Limited by the structure of CNN and the design of the loss function, the visual quality of the result did not improve. Then some researchers resorted to GAN-based methods [10, 26, 32] and introduced perceptual loss [18] and adversarial loss [10], which greatly enhance the perceptual quality. However, GAN-based methods require a lot of time and are hard to train, so GLEAN [2] and PULSE [15] utilize latent-bank to reduce consumption and formulate specific categories of images. In the meantime, transformer [25] was also introduced to computer vision [6]. To be more accurate, the implementation in the SR field includes swinSR [11] and [3], etc. Most recently, diffusion models [16, 17] have proved to be efficient in generating details in the SR process, which include a forward process employed for training and a reverse process utilized for inference. As generative models, diffusion models [23] share difficulty in the training process with GAN-based models, but the latter ones tend to suffer the threat of mode collapse and posterior collapse due to the additional training of discriminator, making the diffusion models the more stable ones.

2.2. Reference-based Image Super-Resolution

Without additional information, SISR tends to suffer the hallucination and artifacts caused by unsubstantiated prediction of pixel value. To conquer this issue, Ref-SR transfers details from relevant reference images to input images. Crossnet [34] first proposed an end-to-end CNN network with cross-scale wrapping to achieve pixel-level alignment. However, different images may share pixels in mismatched areas which affects the construction of long-distance correlation. Thus patch-level alignment is utilized in subsequent methods. SRNTT [33] laid the foundation of patch-level transferring which has proved to be more effective. TTSR [31] inherited the cross-scale aggregation from SRNTT and introduced transformer and attention mechanism into Ref-SR, achieving more precise feature transfer. MASA [13] took the potential large disparity in distributions between the LR and reference images and computation efficiency into consideration, raising coarse-to-fine correspondence matching schemes, which is also adopted by AMSA [29]. Yet the correspondence matching still lacked robustness due to the transformation gap, so C2-matching [21] brought in contrastive learning and knowledge distillation for better performance.

To the best of our knowledge, we find the current Ref-SR methods failed to fully exploit the detail in input LR images during the matching process. Brutely resizing the LR images only narrows the resolution gap between LR images and reference images, yet ruins the correlation between them at the detail level. Inspired by recent works [1, 8, 21, 22], we apply Deep Convolutional Networks (DCN) [4, 35] in our network which requires explicit edges and contours. With the aid of our detail-enhanced input images, our approach facilitates the achievement of substantially improved alignments through the utilization of DCN. This solution effectively addresses the aforementioned challenge, offering a highly promising resolution to a significant extent.

3. Analysis of the Super-Resolution framework

3.1. Range-null space decomposition

The visual quality of the results of image SR has long been a tricky issue since it is too complex to propose a well-received metric to evaluate or improve it intentionally. Most current methods resort to narrowing the gap between pixel values, yielding an output characterized by a dearth of details, excessive smoothness, and a visual presentation that is less conducive to human perception.

Inspired by [19, 28], images can be decomposed into range-space and null-space which represent the data-consistency and realness, respectively. In a more advanced context, data consistency signifies the structural characteristics of an image, while realness tends to reflect the finer details inherent in the image. Given a noise-free image SR model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{D\times 1}$, $\mathbf{A} \in \mathbb{R}^{d\times D}$, and $\mathbf{y} \in \mathbb{R}^{d\times 1}$ denote the ground-truth (GT) image, the linear degradation operator, and the degraded image, respectively. To derive the GT $\hat{\mathbf{x}}$ from input $\mathbf{y}$, two constraints have to be set to ensure the visual quality of SR:

$$Consistency : \mathbf{A}\hat{\mathbf{x}} \equiv \mathbf{y}, Realness : \hat{\mathbf{x}} \sim q(\mathbf{x}) \tag{2}$$

Where $q(\mathbf{x})$ denotes the distribution of the GT image. By implementing Singular Value Decomposition (SVD) on A, we can solve its pseudo-inverse $\mathbf{A}^\dagger$ in matrix form, and the pseudo-inverse $\mathbf{A}^\dagger$ can be used to project the original image $\mathbf{x}$ to the range-space of $\mathbf{A}$ since

$$\mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} \tag{3}$$

conversely, $(\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})$ map $\mathbf{x}$ to null-space of $\mathbf{A}$ due to

$$\mathbf{A}(\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})\mathbf{x} = \mathbf{0} \tag{4}$$

Note that any image x can be decomposed into range-space and null-space, i.e.

$$\mathbf{x} \equiv \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + (\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})\mathbf{x} \tag{5}$$

3.2. Analysis of precise details generation

Due to the great success of PSNR-oriented models [13, 29, 31, 33], existing Ref-SR methods tend to focus on the consistency of restored images, which is highly related to the MSE between the input and the output, rather than realness oriented from details of images. The negligence of detail generation usually leads to over-smoothed results. To investigate this issue, we retrain the TTSR[31] model to evaluate whether details can be enhanced by the aggregation of the existing two kinds of model:

Ref-based models: current Ref-based algorithms can be roughly divided into three parts: feature extraction, matching, and fusion, in which fusion can be further decomposed into texture transfer and texture aggregation. By matching the most relevant patch between reference and LR images, texture can be directly transferred from HR images to LR images. This operation guarantees the data consistency of the texture that has been transferred, while there are still some drawbacks, including **texture mismatch** and **texture undermatch**.

In a reference-based dataset, textures that are similar but have slight differences may appear multiple times in the reference images. This can pose a challenge when trying to accurately match the correct texture between the reference and input images. On the other hand, reference and input images may not share the same brightness, contrast, and hue, etc. Simple execution including transferring texture without any essential adjustment can be devastating to the perceptual quality of the final output. **Texture mismatch** can occur under both circumstances.

Even under circumstances where there are multiple reference images, reference images may not cover all the textures that need to be transferred to the input images. So there could be a certain amount of unmatched textures which results in **texture undermatch**.

Generative-based models: Generative models and, more recently, Denoising Diffusion Probabilistic Models (DDPMs) are well-known for their ability to reproduce high-frequency details. Furthermore, in terms of reconstruction quality, it has been observed that DDPMs exhibit superior subjective perceived quality in comparison to regression-based methods [9, 12, 20, 24], which is ideal for refining the input images in null-space iteratively. By applying general pre-trained weight, we can drastically reduce the computational complexity and acquire stable output. however, as we have mentioned above, every image has its unique distribution, which is not completely predictable by the general pre-trained weights. In this case, generative models tend to generate fake details.

To summarize, Ref-based models utilize similar reference images to guide the restoration of LR images, but the detail deficiency in correspondence matching hinders the accuracy. In contrast, the diffusion model is fully capable of generating details, whereas specific prior information is missing, resulting in artifacts in output. By aggregating the Ref-based model and diffusion model, the details generated by the latter can be utilized to enhance the correspondence map and make up for the missing details.

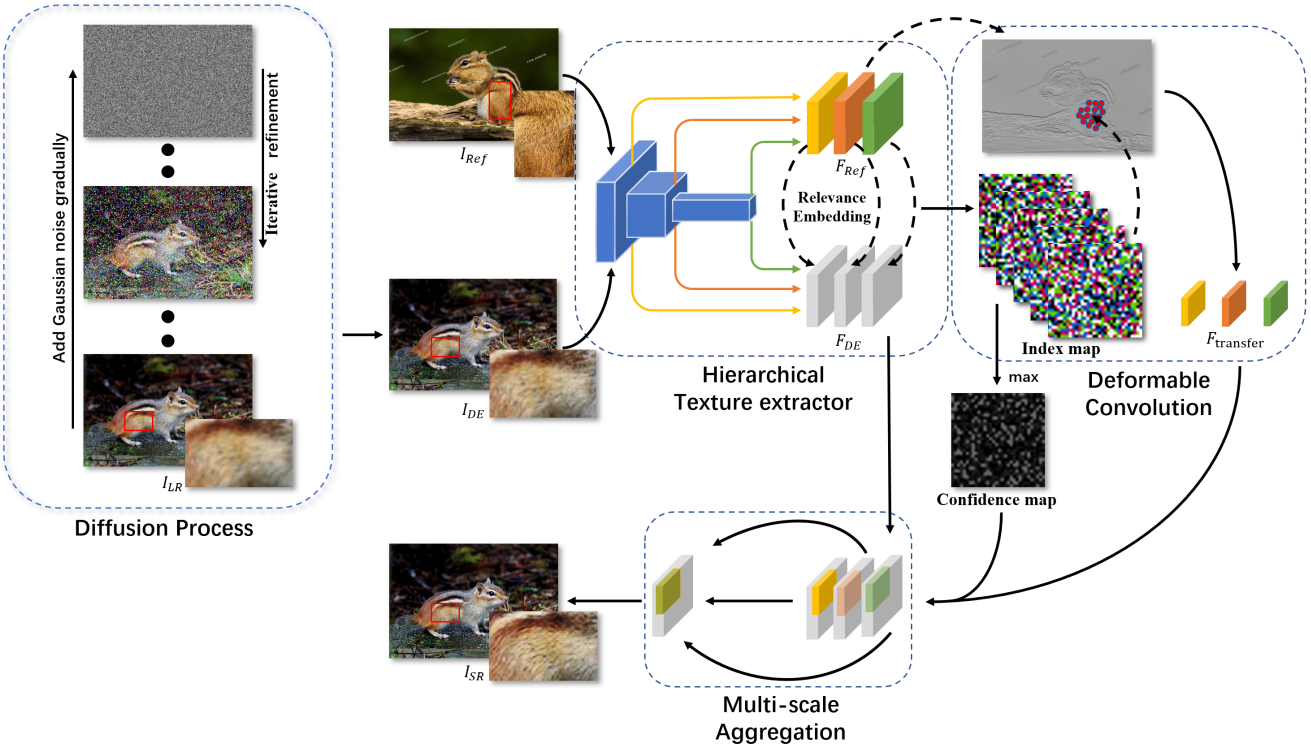

Figure 2. **Detail-Enhancing Framework overview.** For the input LR image, we commence by subjecting it to a diffusion process to enhance its fine details. Subsequently, both the detail-enhanced image and the reference image undergo feature extraction through a structurally identical network. The extracted features are aligned to obtain an index map and a confidence map, serving as the basis for the final multi-scale aggregation process.

4. Our Approach

4.1. overview

In order to resolve the detail-enhancing issue, we propose a novel framework that inherits the main structure of Ref-based models while introducing the diffusion model. The whole detail-enhancing issue can be decomposed into two subtasks: **detail-generation** and **detail-transfer**. Intuitively, the input images are deemed to go through the reverse process of the diffusion model to form the essential details. Due to its low credibility, generated details are proportionally replaced by the corresponding part of reference images, while the rest can tackle the undermatch issue mentioned above.

For **detail-generation** task, instead of applying the downsampled input images directly which compromises the prior information severely, we upsample the input images by a pre-trained diffusion model which is known for its ability to generate rich details, obtaining detail-enhanced input images.

As for **detail-transfer** task, we follow the conventional Ref-based SR procedure. Firstly, we conduct feature extraction on both detail-enhanced images and reference images:

$$\mathbf{F}_{Texture} = F_{TE}(\mathbf{I}_{\text{Ref}}, \mathbf{I}_{DE}) \tag{6}$$

where $\mathbf{F}_{Texture}$, F_{TE}, $\mathbf{I}_{\text{Ref}}$ and $\mathbf{I}_{DE}$ denote the texture feature, texture extraction module, reference images, and detail-enhanced images, respectively. When it comes to alignment, detail-enhanced input images are utilized to calculate the similarity between reference images and input images. The challenge associated with accurately computing the correspondence map is effectively alleviated by substituting detail-deficient input images with detail-enhanced input images. Finally, we use multi-scale aggregation module F_{MSA} to obtain the final result $\mathbf{I}_{SR}$ from the feature of the post-transferred image $\mathbf{F}_{DE}$ and the feature of the reference image $\mathbf{F}_{Ref}$:

$$\mathbf{I}_{SR} = F_{MSA}(\mathbf{F}_{Ref}, \mathbf{F}_{DE}) \tag{7}$$

4.2. detail enhancement

In contrast to the traditional Ref-SR framework, wherein feature extraction is initiated from both LR images and reference images from the outset, DEF introduces a novel paradigm by integrating a diffusion model as an initial step in the enhancement process. By decomposing the images

under consideration into distinct range-space and null-space components as in (5), our method prioritizes the enhancement of image details by directing primary attention toward the null-space component. We use a simple downsample operator $\mathbf{A}$ to extract the null-space information $(\mathbf{I}-\mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}$. Note that the extracted information is under the surveillance of the data consistency constraint we mentioned in (2):

$$\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\mathbf{A}^{\dagger}\mathbf{y} + \mathbf{A}(\mathbf{I}-\mathbf{A}^{\dagger}\mathbf{A})\mathbf{x_n} = \mathbf{y} + (\mathbf{A}-\mathbf{A})\mathbf{x_n} = \mathbf{y} \tag{8}$$

where $\mathbf{x_n}$ denotes the null-space information separated from the images.

The operation of extracting null-space information is performed at each step during the reverse process of the diffusion model, which entails progressively recovering the image from pure Gaussian noise. At timestep t in the reverse process, the current noisy image $\mathbf{x}_t$ undergoes a denoising operation, yielding a clean image $\mathbf{x}_{0|t}$. Then, a weaker noise is added to it to obtain the next noisy image $\mathbf{x}_{t-1}$. This iterative procedure continues until the final result $\mathbf{x}_0$ is obtained. Instead of utilizing $\mathbf{x}_{0|t}$ directly, we first decompose the $\mathbf{x}_{0|t}$ into null-space information $(\mathbf{I}-\mathbf{A}^{\dagger}\mathbf{A})\mathbf{x_{0|t}}$ and range-space information $\mathbf{A}^{\dagger}\mathbf{A}\mathbf{x_{0|t}}$. Then we replace the range-space information of $\mathbf{x}_{0|t}$ with $\mathbf{y}$ to achieve a higher data consistency. Finally, a weaker noise is added to the combination of $\mathbf{A}^{\dagger}\mathbf{y}$ and $(\mathbf{I}-\mathbf{A}^{\dagger}\mathbf{A})\mathbf{x_{0|t}}$.

Subsequent to the iterative refinement of the null-space information within the diffusion model, we derive the final output by:

$$\hat{\mathbf{x}} = \mathbf{A}^{\dagger}\mathbf{y} + (\mathbf{I}-\mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}_n \tag{9}$$

Different from other generative models, the output of diffusion models encounter rigorous constraint on image size. To evaluate our method on datasets with arbitrary image size, inspired by patch-wise methodologies, we cut the images into patches that meet the requirement of image size limitations and input them into diffusion models. A logical approach that emerges involves partitioning the images into distinct patches and subsequently concatenating them during the post-processing stage. For example, if we have an image with the size of 128*256, we can cut it into two 128*128 divisions which satisfy the input demand of the diffusion model. But this will bring significant block artifacts between each division.

To conquer the question above, instead of dividing the image into unrelated patches, we take the above example and cut the 128*256 image into four 128*64 division $[\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}]$. For the i-th turn we take $[\mathbf{y}_i^{(i)}, \mathbf{y}_i^{(i+1)}]$ as input and utilize diffusion model to get SR result $\mathbf{x}$. Then we further divide it into $[\mathbf{x}_i^{(i)}, \mathbf{x}_i^{(i+1)}]$. It is obvious that $\mathbf{x}_i^{(i+1)}$ and $\mathbf{x}_{i+1}^{(i+1)}$ which represent the i-th and $(i+1)$-th output of same region overlap in the final concatenation, so we replace the i-th output with $(i+1)$-th output to ensure the coherent between each division.

4.3. feature extraction and alignment

To achieve precise alignment between input images $\mathbf{I}_{DE} \in \mathbb{R}^{H\times W\times 3}$ and reference images $\mathbf{I}_{\text{Ref}} \in \mathbb{R}^{H^{'}\times W^{'}\times 3}$, feature of both images must be extracted. By slicing the pretrained classification model into multiple parts, we calculate the multi-scale feature of detail-enhanced images and reference images, i.e.,

$$\mathbf{F}^s_{DE} = F_{TE}(\mathbf{I}_{DE}), \mathbf{F}^s_{\text{Ref}} = F_{TE}(\mathbf{I}_{\text{Ref}}) \tag{10}$$

where $\mathbf{F}^s_{DE}$ and $\mathbf{F}^s_{\text{Ref}}$ are feature encoders at the s-th scale. Previous methods tend to preprocess the reference images by downsampling and then upsampling to match the frequency band. Since the diffusion model alleviates the resolution gap and reproduces rich details in the input LR images, upsampling is unnecessary.

The accuracy of alignment lies in the computation of similarity between corresponding patches. Cosine similarity is the most common metric for doing so. We first unfold $\mathbf{F}^s_{DE}$ and $\mathbf{F}^s_{\text{Ref}}$ into patches $\mathbf{F}^s_{DE}{}^{'} = [q_1, ..., q_{HW}]$ and $\mathbf{F}^s_{\text{Ref}}{}^{'} = [k_1, ..., k_{H'W'}]$, then we evaluate the relevance degree $r_{i,j}$ by calculating the inner product of elements in $\mathbf{F}^s_{DE}{}^{'}$ and $\mathbf{F}^s_{\text{Ref}}{}^{'}$:

$$r_{i,j} = \left\langle \frac{\mathbf{q_i}}{\|\mathbf{q_i}\|}, \frac{\mathbf{k_j}}{\|\mathbf{k_j}\|} \right\rangle \tag{11}$$

As for the i-th element in $\mathbf{F}^s_{DE}{}^{'}$, index map P_i and confidence map C_i can be obtained by:

$$P_i = \arg\max_j r_{i,j}, C_i = \max_j r_{i.j} \tag{12}$$

which represents the position in reference images to be transferred and the relevance degree it holds.

4.4. texture transfer and integration

Current Ref-SR methods encounter a pronounced decline in performance attributed to the prevalent problem of texture mismatch, as discussed in Sec.3.2. This issue encompasses not only errors arising from the alignment procedure but also inherent deficiencies in the conventional design of convolution. Unlike regular convolution kernels, the shape of textures to be transferred may not be concrete, which leads to inaccurate mapping. To address this issue, we employ the deformable convolution network (DCN) [21] with an adjustable receptive field. Given the position p_i in input images, the correspondence position p_i^k in index map P_i and the confidence c_i^k of transferred texture in confidence map C_i acquired in the alignment section can be utilized to calculate l-th scale feature $\mathbf{T}_l^i$ in this position:

$$\mathbf{T}_l^i = c_i^k \sum\nolimits_j w_j \mathbf{F}^l_{ref}(p_i^k + p_c + \Delta p_j) m_j \tag{13}$$

Where w_j denotes the convolution kernel weight, $p_c \in \{(-1,1),(-1,0),,,,,(1,1)\}$, Δp_j and m_j denote the j-th learnable offset and learnable mask, respectively. After warping $\mathbf{F}^l_{\text{Ref}}$ and l-th scale index map P_l, Δp_j and m_j can be learned by implementing convolution on the warping result w_l and l-th scale feature extracted from $\mathbf{I}_{DE}$.

Finally, the multi-scale transferred feature needs to be integrated to output the SR images. Here, we inherit the cross-scale integration module proposed by TTSR [31] which aggregates textures from lower scale to upper scale step by step. Specifically, this module exhibits ideal performance in terms of information utilization which satisfies our claims.

4.5. Implementation details

The overview network can be decomposed into two sections: 1) The diffusion model which is in charge of the SISR subtask. 2) The Ref-SR architecture which includes texture extraction and transfer.

Dataset Preprocessing. We augment the datasets by randomly rotating images within the range of 0 to 360 degrees with intervals of 90 degrees and randomly flipping images horizontally and vertically.

Implementation of Diffusion Model. We use the bicubic downsampler as the degradation operator to ensure fair comparisons. As for noise schedule and input image constraint, we choose linear noise schedule and 256*256 pretrained model. To achieve a fine-grained diffusion process during training, we set the time step to 1,000. We avoided other time-step evaluations as they would affect comparability. The linear noise schedule has the endpoints of $1-\alpha_0 = 10^{-6}$ and $1-\alpha_T = 10^{-2}$.

Training of Texture Transfer Network. For fair comparison, we train DEF on the scale of 4x, and feature extractors share the same architecture. More specifically, we train our network using Adam optimizer with parameter $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set as 1e-4, and the batch size is 9, which contains 9 LR, HR, and reference images in each batch. Note that the weight of the given extractor should be fixed, since the comparison afterward needs to be stable, and a changeable extractor can affect the performance of correspondence matching.

Loss Function. Given the focal point of our approach on enhancing the visual quality of the reconstructed image, coupled with the inherent emphasis on preserving intricate details through the utilization of spatial structure and semantic information of images, it becomes imperative to introduce reconstruction loss as an indispensable element, meticulously guiding the training process at its fundamental essence. To enhance the detail of SR images, perceptual loss, and adversarial loss are also introduced, so the overall loss function is written as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{per} + \lambda_2 \mathcal{L}_{adv} \tag{14}$$

To allocate greater emphasis on detail, we set weight coefficients of $\mathcal{L}_{rec}$, $\mathcal{L}_{per}$ and $\mathcal{L}_{adv}$ as 1, 1e-2 and 1e-4, respectively. Reconstruction loss is the only loss involved in the training process of the first two epochs for warming up the network while perceptual loss and adversarial loss are added in the following epochs to the end.

5. Experiments

5.1. Dataset and metrics

Training Dataset. The entire training process of our model is completed on the CUFED5 [27] dataset, which consists of 11871 pairs of images, including an input image and a reference image for each pair. Since the resolution of both input and reference images is 160x160, we resize the images in the input folder to 40x40 for the upcoming x4 SR.

Testing Dataset. To demonstrate the generalization ability of our network, we adopt five test sets, including CUFED5 [27], Sun80 [23], Urban100 [7], Manga109 [14] and WR-SR [21]. The test set of CUFED5 has 126 images, each image has 4 reference images with different similarity scales. Proposed by Shim *et al*. [21], WR-SR has 80 pairs of images, each containing an input image and a reference image. The source of reference images is the most relevant search result of Google. Sun80 [23] also has 80 natural images, each paired with multiple reference images. Urban100 has 100 building images and Manga109 has 109 manga images in which style is shared in most images. They are SISR datasets which have no reference image, thus we follow settings in [31]: Urban100 adopts its LR images as reference, while in Manga109 we randomly select another HR image as its reference.

Evaluation Metrics. We evaluate the outcomes achieved by both our proposed approach and alternative methods through the application of PSNR and SSIM metrics. For a more comprehensive review of these metrics, readers could refer to [30]. Specifically, these metrics are computed on the luminance (Y) channel of the YCrCb color space.

5.2. Comparison with State-of-the-art Methods

We compare our method with the previous state-of-the-art SISR methods and the single-reference Ref-SR method. SISR methods include SRCNN, EDSR, ESRGAN, and RankSRGAN, in which half of the methods we select are Gan-based due to their strong ability to generate rich details. Single-reference Ref-SR methods include SRNTT, TTSR, MASA, and C2-matching.

Quantitative Comparison. For a fair comparison, we train all the candidate methods on the CUFED5 dataset and evaluate them on the testsets of CUFED5, Manga109, Sun80, Urban100, and WR-SR. The scale factor of all mentioned methods is x4. Tab. 1 indicates that our methods

Table 1. Quantitative comparisons (PSNR and SSIM) of SR models

	Method	CUFED5	WR-SR	Urban100	Manga109	Sun80
SISR	SRCNN	25.33/0.745	27.37/0.767	24.41/0.738	27.12/0.850	28.26/0.781
	EDSR	25.93/0.777	28.07/0.793	25.51/0.783	28.93/0.891	28.52/0.792
	ESRGAN	21.90/0.633	26.07/0.726	20.91/0.620	23.53/0.797	24.18/0.651
	RankSRGAN	22.31/0.635	26.15/0.719	21.47/0.624	25.04/0.803	25.60/0.667
Ref-SR	SRNTT	25.61/0.764	26.53/0.745	25.09/0.774	27.54/0.862	27.59/0.756
	TTSR	25.53/0.765	26.83/0.762	24.62/0.747	28.70/0.886	28.59/0.774
	MASA	24.92/0.729		23.78/0.712	27.34/0.848	27.12/0.708
	C2-matcing	27.16/0.805	**27.80/0.780**	25.52/0.764	29.73/0.893	29.75/0.799
	OURS	**27.47/0.826**	27.60/0.777	**25.92/0.780**	**30.21/0.893**	**29.77/0.800**

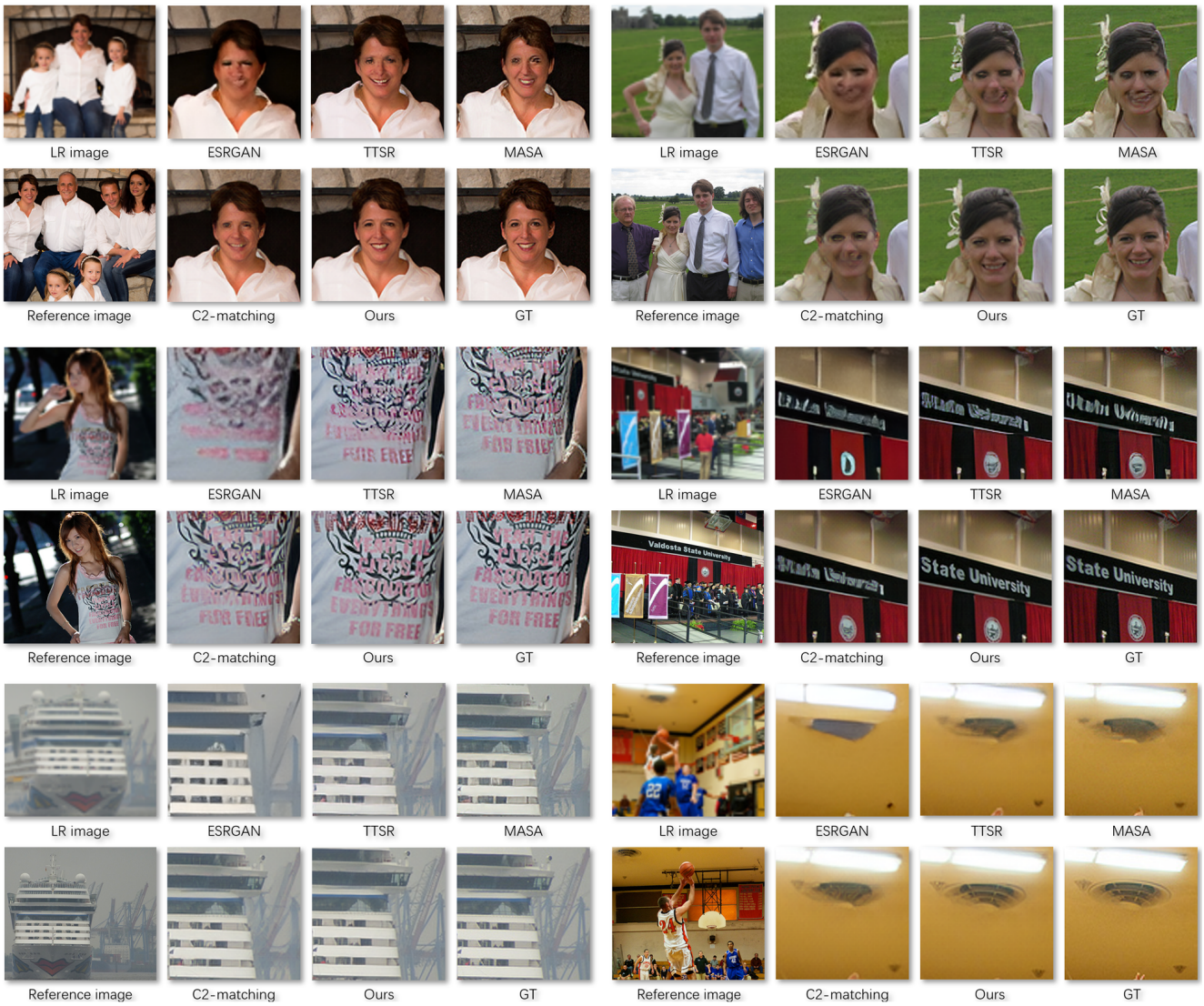

Figure 3. Visual comparison with other methods. We zoom in on the key areas for a better view.

outperform most of the previous state-of-the-art methods and achieve comparable performance against C2-matching on the WR-SR dataset, which emphasizes the superiority of the unique detail-generating structure we propose in the feature alignment and aggregation process. However, numerical inferiority does not necessarily imply a lack of detail. As previously mentioned, image restoration can be divided into two components: range-space and null-space. Numerical metrics primarily correspond to the range-space aspect, which is not the main focus of our proposed approach. The WR-SR dataset consists of 150 images selected from another dataset and website, serving as query images to retrieve 50 similar images from Google Images. Furthermore, these similar images undergo size normalization, resulting in 80 pairs of image sets. The normalization process undoubtedly compromises the details in the reference images, further affecting subsequent alignment procedures. Additionally, images sourced online exhibit differences in various aspects, such as lighting and contrast, making them more suitable for C2-matching due to their robustness under different conditions.

Qualitative Evaluation. Fig. 3 shows the visual results of our method, a SISR method, and previous state-of-the-art Ref-SR methods. We compare our method with ESRGAN, TTSR, MASA, and C2-matching. By comparing the selected part of the result from the same input LR image, it is obvious that our method can restore more accurate detail in various aspects. The first row of Fig. 3 focuses on the synthesis of natural human faces, while the focal point of the second and third rows are the recovery of letters and object textures. ESRGAN's incapacity to thoroughly exploit information from reference images results in its failure to generate reliable details. TTSR, MASA, and C2-matching can not fully utilize the information in reference images due to their detail-wise gap between input LR images and reference images, which in turn hamper the alignment and transfer procedure. For Ref-based methods, detail-enhanced input images smooth the edge of objects, which makes the alignment more accurate in the feature domain, thus reinforcing the transfer and integration procedure, exhibiting a higher visual quality image in the end.

Alleviation to the texture mismatch and texture undermatch issues. From the images in the first row, it can be observed that the reference images contain more faces than the images to be restored, and in the fourth image, the lighting conditions between the two images are dissimilar. In previous Ref-SR methods, this could easily lead to texture mismatch issues, ultimately resulting in distorted recovered information. However, it can be noted that through the pre-enhancement of fine details in the LR images, the accuracy of the final alignment is significantly improved, leading to more satisfactory restoration results. Additionally, in the first image, the background behind the individuals and the light tube in the last image does not have corresponding parts in the reference images to assist in the recovery of the original image. This presents a challenge related to texture undermatch. Similarly, following the enhancement of details in the LR images, regions that originally lacked corresponding HR information have also been partially reconstructed, ensuring an enhancement in the overall image restoration performance.

Table 2. Quantitative evaluation for ablation study of the decouple framework.

Method	DEF	DCN	PSNR	SSIM
Base			25.53	0.765
Base + DEF	✓		27.37	0.816
Base + DCN		✓	27.39	0.819
Base + DCN + DEF	✓	✓	27.47	0.826

5.3. Ablation study

In this section, we conduct an ablation study to validate the effectiveness of our improvements on the baseline. Including detail-enhancing framework and feature transfer module.

5.3.1 Detail-enhancing framework

Instead of resizing the input LR images simplistically, our detail-enhancing framework alleviates the resolution gap by applying a diffusion model before feature extraction. We re-implement TTSR as our baseline. Ablation results are shown in Tab. 2. The table reveals a substantial increase in both PSNR and SSIM values, exceeding 2dB. Previous methods usually upsample the LR image by bicubic interpolation, which exploits the surrounding 16 pixels to generate a target pixel value, matching the resolution between input images and reference images. Though the basic alignment requirement has been satisfied, the over-smooth image tends to produce artifacts in the final output. Results demonstrate that DEF outperforms the baseline by a large margin, verifying the feasibility of detail-enhancing tasks in the alignment and transfer section.

5.3.2 Feature transfer module

Due to the preprocess of reference images to obtain domain-consistent images between reference images and LR images, the baseline adopts transformer for alignment. To preserve the detail in reference images, we keep the original reference images, thus the transformer structure is unnecessary. We adopt relevance embedding in acquiring index map, and then according to the index, we upgrade the convolution networks to the deformable convolution networks, strengthening its robustness to irregular texture transfer.

The statistic in Tab. 2 exhibits significant improvement in performance in PSNR. Since detail has been sabotaged in the preprocessing of reference images, the enhancement of SSIM is limited.

6. Conclusion

In this paper, we propose a novel detail-enhancing framework to alleviate the hamper to reconstruction quality by the ill-posed nature of SR. Based on the theoretical analysis, we set two criteria in an ideal SR model to guarantee the realness and data consistency of the SR image. Specifically, we decompose the image and refine the partial content iteratively in DEF with the assistance of the diffusion model. By implementing the new framework, we are able to generate rich details in LR images and resolve the mismatch and undermatch issues in the feature alignment stage. Furthermore, the deformable convolution network is utilized to accomplish a more precise feature transfer between detail-enhanced LR images and reference images. Experiment results, especially qualitative results, demonstrate the feasibility of our proposed framework in optimizing the current Ref-SR structure.

References

[1] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, pages 325–342. Springer, 2022. 1, 3

[2] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 3

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 3

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 7

[8] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5931–5940, 2022. 1, 3

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3, 4

[10] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[11] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 3, 4

[13] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2021. 1, 3, 4

[14] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 7

[15] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 3

[16] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[17] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2, 3

[18] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 4491–4500, 2017. 3

[19] Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. Deep null space learning for inverse problems:

convergence analysis and rates. *Inverse Problems*, 35(2): 025008, 2019. 4

[20] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3, 4

[21] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8425–8434, 2020. 1, 3, 6, 7

[22] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8425–8434, 2020. 1, 3

[23] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *2012 IEEE International conference on computational photography (ICCP)*, pages 1–12. IEEE, 2012. 3, 7

[24] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, pages 4799–4807, 2017. 3, 4

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[26] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[27] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4810–4819, 2016. 7

[28] Yinhuai Wang, Yujie Hu, Jiwen Yu, and Jian Zhang. Gan prior based null-space learning for consistent super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2724–2732, 2023. 4

[29] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2768–2776, 2022. 1, 3, 4

[30] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 372–386. Springer, 2014. 7

[31] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 1, 2, 3, 4, 7

[32] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 3

[33] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7991, 2019. 1, 3, 4

[34] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 88–104, 2018. 1, 3

[35] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 3