Iterative Token Evaluation and Refinement for Real-World Super-resolution

Chaofeng Chen[1], **Shangchen Zhou**[1], **Liang Liao**[1], **Haoning Wu**[1],
Wenxiu Sun[2], **Qiong Yan**[2], **Weisi Lin**[1]

[1] S-Lab, Nanyang Technological University
[2] SenseTime Research
{chaofeng.chen, liang.liao, wslin}@ntu.edu.sg, shangchenzhou@gmail.com,
haoning001@e.ntu.edu.sg, {sunwx, yanqiong}@tetras.ai

Abstract

Real-world image super-resolution (RWSR) is a long-standing problem as low-quality (LQ) images often have complex and unidentified degradations. Existing methods such as Generative Adversarial Networks (GANs) or continuous diffusion models present their own issues including GANs being difficult to train while continuous diffusion models requiring numerous inference steps. In this paper, we propose an Iterative Token Evaluation and Refinement (ITER) framework for RWSR, which utilizes a discrete diffusion model operating in the discrete token representation space, *i.e.*, indexes of features extracted from a VQGAN codebook pre-trained with high-quality (HQ) images. We show that ITER is easier to train than GANs and more efficient than continuous diffusion models. Specifically, we divide RWSR into two sub-tasks, *i.e.*, distortion removal and texture generation. Distortion removal involves simple HQ token prediction with LQ images, while texture generation uses a discrete diffusion model to iteratively refine the distortion removal output with a token refinement network. In particular, we propose to include a token evaluation network in the discrete diffusion process. It learns to evaluate which tokens are good restorations and helps to improve the iterative refinement results. Moreover, the evaluation network can first check status of the distortion removal output and then adaptively select total refinement steps needed, thereby maintaining a good balance between distortion removal and texture generation. Extensive experimental results show that ITER is easy to train and performs well within just 8 iterative steps.

Introduction

Single-image super-resolution (SISR) aims to restore high-quality (HQ) outputs from low-quality (LQ) inputs that have been degraded through processes such as downsampling, blurring, noise, and compression. Previous studies (Liang et al. 2021; Zamir et al. 2022; Chen et al. 2023) have achieved remarkable progress in enhancing LQ images degraded by a single predefined type of degradation, thanks to the emergence of increasingly powerful deep networks. However, in real-world LQ images, multiple unknown degradations are typically present, making previous methods unsuitable for such complex scenarios.

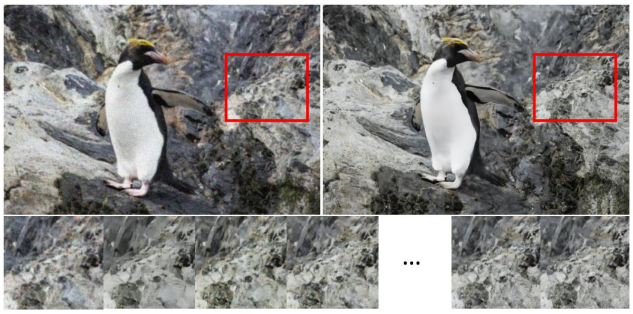


Figure 1: Example result with the proposed ITER. Left top: input LQ image; Right top: SR result with ITER; Bottom: results from $t = T$ to $t = 0$, and t is the iterative step index of the reverse discrete diffusion process. We can observe that the textures are gradually enriched with iterative refinement. To obtain satisfactory results, our ITER requires only a total iteration step of $T \leq 8$. (Zoom in for best view)

Real-world super-resolution (RWSR) is particularly ill-posed because details are usually corrupted or completely lost due to complex degradations. In general, the RWSR can be divided into two subtasks: *distortion removal* and *conditioned texture generation*. Many existing approaches, such as (Wang et al. 2018b; Zhang et al. 2019a), follow the seminal SRGAN (Ledig et al. 2017) and rely on Generative Adversarial Networks (GANs). Typically, these methods require the joint optimization of various constraints for the two subtasks: 1) reconstruction loss for distortion removal, which is usually composed of pixel-wise L1/L2 loss and feature space perceptual loss; 2) adversarial loss for texture generation. Effective training of these models often involves tedious fine-tuning of hyper-parameters between restoration and generation abilities. Moreover, most models have a fixed preference for restoration and generation and cannot be flexibly adapted to LQ inputs with different degradation levels. Recently, approaches such as SR3 (Saharia et al. 2022) and LDM (Rombach et al. 2022) have turned to the popular diffusion model (DM) for realistic generative ability. Although DMs are easier to train and more powerful than GANs, they require hundreds or even thousands of iterative steps to generate outputs. Additionally, current DM-based methods have only been shown to be effective on images with moderate

distortions. Their performance on severely distorted real-world LQ images remains to be validated.

In this paper, we introduce a new framework for RWSR based on a conditioned discrete diffusion model, called Iterative Token Evaluation and Refinement (ITER). ITER incorporates several critical designs to address the challenges of RWSR. Firstly, we formulate the RWSR task as a discrete token space problem, utilizing a pretrained codebook of VQGAN (Esser, Rombach, and Ommer 2021), instead of pixel space regression. This approach offers two advantages: 1) A small discrete proxy space reduces the ambiguity of image restoration, as demonstrated in (Zhou et al. 2022); 2) Generative sampling in a limited discrete space requires fewer iteration steps than denoising diffusion sampling in an infinite continuous space, as shown in (Bond-Taylor et al. 2022; Gu et al. 2022; Chang et al. 2022). Secondly, in contrast to previous GAN and DM methods, we explicitly separate the two sub-tasks of RWSR and address them with token restoration and token refinement modules, respectively. For the first task, we use a simple token restoration network to predict HQ tokens from LQ images. For the second task, we use a conditioned discrete diffusion model to iteratively refine outputs from the token restoration network. This approach facilitates optimizing each module and enables flexible trade-offs between restoration and generation. Finally, we propose to include a token evaluation block in the condition diffusion process. Unlike previous discrete diffusion models (Bond-Taylor et al. 2022; Chang et al. 2022) which directly rely on token prediction probability to select tokens to keep in each de-masking step, we introduce a evaluation block to check whether each tokens are correctly refined or not. This allows our model to better select good tokens in each step during iterative refinement process, and therefore improve the final results. Additionally, the token evaluation block enables us to adaptively select the total refinement steps to balance restoration and texture generation by evaluating the initially restored tokens. We can use fewer refinement steps for good initial restoration results to avoid over-textured outputs. The experiments demonstrate that our proposed ITER framework can effectively remove distortions and generate realistic textures without tedious GAN training in an efficient manner, requiring *less than 8 iterative refinement steps.* Please refer to Fig. 1 for an example. In summary, our contributions are as follows:

- We propose a novel framework, ITER, that addresses the two sub-tasks of RWSR in discrete token space. Compared to GAN, ITER is much easier to train and more flexible at inference time. Compared to DM-based methods, it requires fewer iteration steps and has demonstrated effectiveness on real-world LQ inputs with complex degradations.
- We propose an iterative evaluation and refinement approach for texture generation. The newly introduced token evaluation block allows the model to make better decisions on which tokens to refine during the iterative refinement process. Furthermore, by evaluating the quality of initially restored tokens, ITER is able to adaptively balance distortion removal and the texture generation in the final results by using different refinement steps. Besides, the user can also manually control the visual effects of outputs through a threshold value without the need for retraining the model.

Related Works

In this section, we provide a brief overview of SISR and generative models utilized in SR. We also recommend recent literature reviews (Anwar, Khan, and Barnes 2020; Liu et al. 2022, 2023) for more comprehensive summaries.

Single Image Super-Resolution. Recent SISR for bicubic downsampled LQ images has made remarkable progress with the improvement of network architectures. Methods such as (Kim, Lee, and Lee 2016a,b; Lim et al. 2017; Ledig et al. 2017; Zhang et al. 2018c) introduced deeper and wider networks with more skip connections, showing the power of residual learning (He et al. 2016). Attention mechanisms, including channel attention (Zhang et al. 2018b), spatial attention (Niu et al. 2020; Chen et al. 2020), and non-local attention (Zhang et al. 2019b; Mei, Fan, and Zhou 2021; Zhou et al. 2020), have also been found to be beneficial. Recent works employing vision transformers (Chen et al. 2021; Liang et al. 2021; Zhang et al. 2022; Chen et al. 2023) have surpassed CNN-based networks by a large margin, thanks to the ability to model relationships in a large receptive field.

Latest works have focused on the challenging task of RWSR. Some methods (Fritsche, Gu, and Timofte 2019; Wei et al. 2021; Wan et al. 2020; Maeda 2020; Ji et al. 2020; Wang et al. 2021a; Zhang et al. 2021a; Mou et al. 2022; Liang, Zeng, and Zhang 2022) implicitly learn degradation representations from LQ inputs and perform well in distortion removal. However, their generalization ability is limited due to the complexity of the real-world degradation space. BSRGAN (Zhang et al. 2021b) and Real-ESRGAN (Wang et al. 2021c) adopt manually designed large degradation space to synthesize LQ inputs and have proven to be effective. Li *et al*. (Li et al. 2022) proposed learning degradations from real LQ-HQ faces and then synthesizing training datasets. Although these methods improve distortion removal, they rely on unstable adversarial training to generate missing details, which may result in unrealistic textures.

Generative Models for Super-Resolution. Many works employ GAN networks to generate missing textures for real LQ images. StyleGAN (Karras et al. 2020) works well for real face SR (Yang et al. 2021; Wang et al. 2021b; Chan et al. 2021). Pan *et al*. (Pan et al. 2020) used a BigGAN generator (Brock, Donahue, and Simonyan 2019) for natural image restoration. The recent VQGAN(Esser, Rombach, and Ommer 2021) demonstrates superior performance in image synthesis and is shown to be effective in real SR of both face (Zhou et al. 2022) and natural images (Chen et al. 2022).

The latest works with diffusion models (Saharia et al. 2022; Rombach et al. 2022; Gao et al. 2023; Wang et al. 2023) are more powerful than GAN, but they are based on continuous feature space and require many iterative sampling steps. In this work, we take advantage of the discrete diffusion models (Gu et al. 2022; Bond-Taylor et al. 2022;

Chang et al. 2022), which is powerful in texture generation and efficient at inference time. To the best of our knowledge, we are the first work to show the potential of discrete diffusion models on image restoration.

Methodology

In this work, we propose a new iterative token sampling approach for texture generation in RWSR. Our pipeline operates in the discrete representation space pre-trained by VQGAN, which has been shown to be effective in image restoration (Chen et al. 2022; Zhou et al. 2022). Our framework consists of three stages:

- **Stage I: HQ images to discrete tokens.** Different from previous works based on continuous latent diffusion models, our method is based on discrete latent space. Therefore, we need to pretrain a vector-quantized autoencoder (VQVAE) (Esser, Rombach, and Ommer 2021) with discrete codebook to encode input HQ images I_h, such that I_h can be transformed to discrete tokens, denoted as S_h.
- **Stage II: LQ images to tokens with distortion removal.** Instead of directly encoding LQ images I_l with pretrained VQVAE, we propose to train a separate distortion removal encoder for I_l. It helps to remove obvious distortions in LQ input I_l and encode it to a relatively clean discrete token space S_l.
- **Stage III: Texture generation with discrete diffusion.** After obtaining the discrete representations S_l and S_h, we formulate the texture generation as a discrete diffusion model between S_l and S_h. The key difference with our method is that we include an additional token evaluation block to improve the decision-making process for which tokens to refine during the reverse diffusion process. In such manner, the proposed ITER not only generates realistic textures but also permits adaptable control over the texture strength in the final output.

Details are given in the following sections.

HQ Images to Discrete Tokens

Following VQGAN (Esser, Rombach, and Ommer 2021), the encoder E_H takes the input high-quality (HQ) image $I_h \in \mathbb{R}^{H\times W\times 3}$ in RGB space and encodes it to latent features $Z_h \in \mathbb{R}^{m\times n\times d}$. Subsequently, Z_h is quantized into discrete features $Z_c \in \mathbb{R}^{m\times n\times d}$ by identifying its nearest neighbors in the learnable codebook $\mathcal{C} = \{c_k \in \mathbb{R}^d\}_{k=0}^{N-1}$:

$$Z_c^{(i,j)} = \underset{c_k\in\mathcal{C}}{\arg\min} \|Z_h^{(i,j)} - c_k\|_2. \tag{1}$$

The corresponding indices $k \in \{0,\dots,N-1\}$ determine the token representation of the inputs $S_h \in \mathbb{Z}_0^{m\times n}$. Finally, the decoder reconstructs the image from the latent $I_{rec} = D_H(Z_c) = D_H(E_H(I_h))$. Instead of using the original VQGAN (Esser, Rombach, and Ommer 2021), we replace the non-local attention with Swin Transformer blocks (Liu et al. 2021) to reduce memory cost for large resolution inputs.

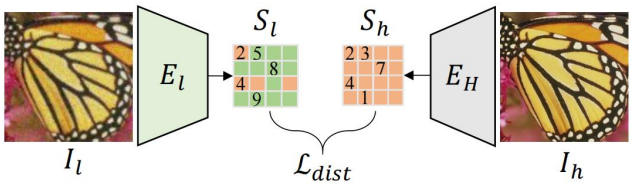


Figure 2: Training of E_l to encode I_l to token space S_l.

LQ Images to Tokens with Distortion Removal

It is straightforward to also encode I_l with pretrained E_H in the first stage. However, since I_l contains complex distortions, the encoded tokens are also noisy, increasing the difficulties of restoration in the following stage. Inspired by recent works (Chen et al. 2022; Zhou et al. 2022), we realize that a straightforward token prediction can eliminate evident distortions. Hence, we introduce a preprocess subtask to remove distortions when encoding I_l into token space. Specifically, we employ an LQ encoder E_l to directly predict the HQ code indexes S_h as illustrated in Fig. 2:

$$S_l = E_l(I_l), \quad \mathcal{L}_{dist} = -S_h^i \log(S_l^i), \tag{2}$$

Through this approach, I_l can be encoded into a comparatively clean token space with the learned E_l.

Texture Generation with Discrete Diffusion

Although the distortions in S_l are effectively removed, generating missing details through Eq. (2) is a challenging task because the generation of diverse natural textures is highly ill-posed and essentially a one-to-many endeavor. To address this issue, we propose an iterative token evaluation and refinement approach, named as ITER, for RWSR, following the generative sampling pipeline outlined in (Chang et al. 2022; Lezama et al. 2022). As ITER is based on the discrete diffusion model (Bond-Taylor et al. 2022; Gu et al. 2022), we will first provide a brief overview of it.

Discrete Diffusion Model. Given an initial image token $\mathbf{s}_0 \in \mathbb{Z}_0$, the forward diffusion process establishes a Markov chain $q(\mathbf{s}_{1:T}|\mathbf{s}_0) = \prod_{t=1}^{T} q(\mathbf{s}_t|\mathbf{s}_{t-1})$, which progressively corrupts $\mathbf{s}_0$ by randomly masking $\mathbf{s}_0$ over T steps until $\mathbf{s}_T$ is entirely obscured. Conversely, the reverse process is a generative model that incrementally "unmasks" $\mathbf{s}_T$ to the data distribution $p(\mathbf{s}_{0:T}) = p(\mathbf{s}_T)\prod_{t=1}^{T} p_\theta(\mathbf{s}_{t-1}|\mathbf{s}_t)$. According to (Bond-Taylor et al. 2022; Chang et al. 2022; Lezama et al. 2022), the "unmasking" transit distribution p_θ can be approximated by learning to predict the authentic $\mathbf{s}_0$, given any arbitrarily masked version $\mathbf{s}_t$:

$$\underset{\theta}{\arg\min} - \log p_\theta(\mathbf{s}_0|\mathbf{s}_t). \tag{3}$$

Following (Chang et al. 2022), during the forward process, $\mathbf{s}_t$ is obtained by randomly masking $\mathbf{s}_0$ at a ratio of $\gamma(r)$, where $r \in \text{Uniform}(0,1]$, and $\gamma(\cdot)$ represents the mask scheduling function. In the reverse process, $\mathbf{s}_t$ is sampled according to the prediction probability $p_\theta(\mathbf{s}_t|\mathbf{s}_{t+1},\mathbf{s}_T)$. The masking ratio is computed using the predefined total sampling step T, *i.e.*, $\gamma(\frac{t}{T})$ where $t \in \{T,\dots,1\}$.

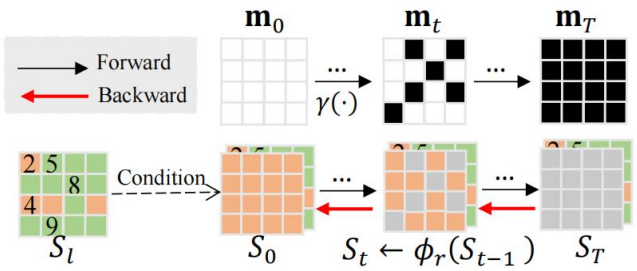


Figure 3: Illustration of forward and backward diffusion process with the conditioned discrete diffusion model. The condition inputs of ϕ_r are omitted here for simplicity.

Algorithm 1: Training of ITER

Input: S_l, S_h, schedule function $\gamma(\cdot)$, learning rate η, networks ϕ_r and ϕ_e

repeat
  $r \sim \text{Uniform}(0, 1]$
  $N \leftarrow$ token numbers in S_h
  $\mathbf{m}_t \leftarrow \text{RandomMask}(\lceil \gamma(r) \cdot N \rceil)$
  $S_t \leftarrow S_h \odot \mathbf{m}_t + (1 - \mathbf{m}_t) \odot S_T$
  $\theta_r \leftarrow \theta_r - \eta \nabla_{\theta_r} \mathcal{L}_r$ ▷ Update ϕ_r
  $\theta_e \leftarrow \theta_e - \eta \nabla_{\theta_e} \mathcal{L}_e$ ▷ Update ϕ_e
until converge

Network Training. As depicted in Fig. 3, the proposed ITER model is a conditioned version of the discrete diffusion model. It is a Markov chain that goes from ground truth tokens S_h (*i.e.*, S_0) to fully masked tokens S_T while being conditioned on S_l. The reverse diffusion step $p_\theta(\mathbf{s}_{t-1}|\mathbf{s}_t)$ is learned with the refinement network ϕ_r using the following objective function:

$$\mathcal{L}_r = -S_h \log\big(\phi_r(S_t, S_l, \mathbf{m}_t)\big), \tag{4}$$

where $\mathbf{m}_t$ is the random mask in corresponding forward diffusion step, and tells ϕ_e which tokens need to be refined.

The difference is that we introduce an extra token evaluation network ϕ_e to learn which tokens are good tokens for both S_t and S_l with the objective function below:

$$\mathcal{L}_e = -\mathbf{m}_t \log\big(\phi_e(S_t)\big) - \mathbf{m}_l \log\big(\phi_e(S_l)\big), \tag{5}$$

where $\mathbf{m}_l$ are the ground truth sampling masks for S_l.

Adaptive Inference of ITER

As illustrated in Algorithm 2, the inference process of ITER can be a standard reverse diffusion from S_T to S_0 with the condition S_l. However, in our framework, the initially restored tokens S_l already contain good tokens and may not require the entire reverse process. With the aid of the token evaluation network ϕ_e, it is possible to select the appropriate starting time step T_s for the reverse diffusion process by assessing the number of good tokens in S_l using $\mathbf{m}_l = \phi_e(S_l)$, as shown below:

$$\mathbf{m}_s^i = \begin{cases} 1 & \text{if } p_{\phi_e}(\mathbf{m}_l^i = 1) \geq \alpha; \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Algorithm 2: Adaptive Inference of ITER

Input: $I_l, T = 8, \gamma(\cdot)$, networks E_l, D_H, ϕ_r and ϕ_e

$S_l \leftarrow E_l(I_l)$ ▷ Initial restoration
$N \leftarrow$ token numbers in S_l
$T_s \leftarrow T$
if use adaptive inference **then**
  $\mathbf{m}_s \leftarrow \phi_e(S_l)$ with α, Eq. (6)
  while $\left\lceil \left(1 - \gamma\left(\frac{T_s - 1}{T}\right)\right) \cdot N \right\rceil < \sum \mathbf{m}_s$ **do**
    $T_s \leftarrow T_s - 1$ ▷ Find start time step
  end while
  Initialize with Eq. (7)
end if
for $t = T_s \cdots 1$ **do**
  $k \leftarrow \left\lceil \left(1 - \gamma\left(\frac{t-1}{T}\right)\right) \cdot N \right\rceil$ ▷ Number to sample
  $S_{t-1} \leftarrow$ sample $p_{\phi_r}(S_{t-1}|S_t, S_l, \mathbf{m}_t)$ ▷ Refine
  $\mathbf{m}_{t-1} \leftarrow$ sample k from $p_{\phi_e}(\mathbf{m}_{t-1} = 1|S_{t-1})$ ▷ Evaluate
  $S_{t-1} \leftarrow S_{t-1} \odot \mathbf{m}_{t-1} + S_T \odot (1 - \mathbf{m}_{t-1})$
end for
return $I_{sr} \leftarrow D_H(S_0)$ ▷ Get SR result.

where α is the threshold value, and $\mathbf{m}_s$ is the binary mask for the starting time step T_s. We can quickly determine the appropriate T_s by comparing the mask ratio indicated by $\gamma(\cdot)$, see Algorithm 2 for further details. We can then initialize S_t and $\mathbf{m}_t$ using the following equations:

$$S_t = \mathbf{m}_s \odot S_l + (1 - \mathbf{m}_s) \odot S_T, \quad \mathbf{m}_t = \mathbf{m}_s. \tag{7}$$

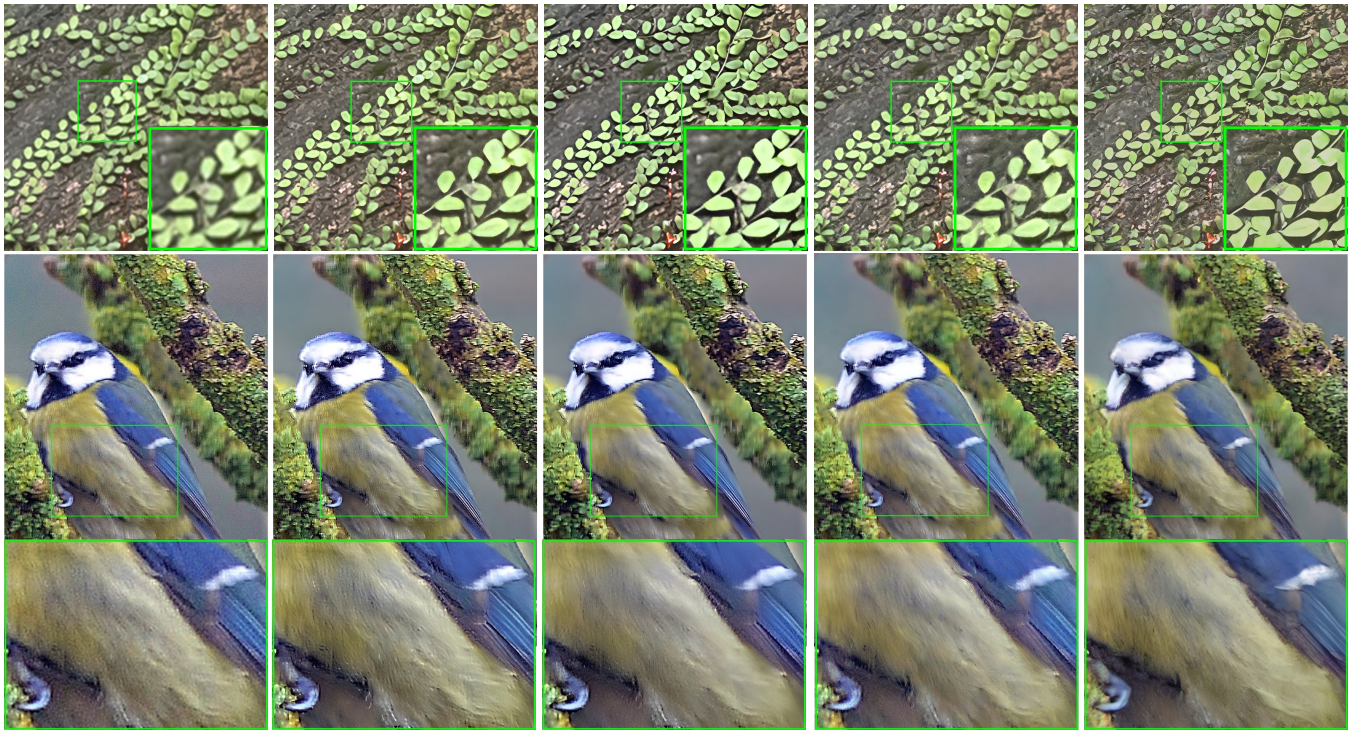
Finally, we follow the typical reverse diffusion process to compute the "unmasking" distribution p_{ϕ_r}, where $t \in \{T_s, \ldots, 1\}$. The final outcome is obtained by $Isr = D_H(S_0)$. The proposed adaptive inference strategy not only makes ITER more efficient but also avoids disrupting the initial good tokens in S_l.

Implementation Details

Training Dataset. Our training dataset generation process follows that of Real-ESRGAN (Wang et al. 2021c), in which we obtain HQ images sourced from DIV2K (Agustsson and Timofte 2017), Flickr2K (Lim et al. 2017), and OutdoorSceneTraining (Wang et al. 2018a). These images are cropped into non-overlapping patches of size 256×256 to serve as HQ images. Meanwhile, the corresponding LQ images are produced using the second-order degradation model proposed in (Wang et al. 2021c).

Testing Datasets. We evaluate the performance of our model on multiple benchmarks that include real-world LQ images such as RealSR (Wang et al. 2021b), DRealSR (Wei et al. 2020), DPED-iphone (Ignatov et al. 2017), and RealSRSet (Zhang et al. 2021b). Additionally, we create a synthetic dataset using the DIV2K validation set to validate the effectiveness of different model configurations.

Training and Inference Details. ITER is composed of three networks, namely E_l, ϕ_r, and ϕ_e, trained with cross-



(a) LQ (×4) (b) FeMaSR (c) MM-RealSR (d) LDM-BSR (e) **ITER (Ours)**

Figure 4: Visual comparison between recent approaches and the proposed ITER on real LQ images.

Datasets	Bicubic	BSRGAN	Real-ESRGAN	SwinIR-GAN	FeMaSR	MM-RealSR	LDM-BSR	**Ours**
RealSR	6.24 / 8.16	5.74 / 4.51	4.83 / 4.54	4.76 / 4.65	4.74 / 4.51	4.69 / 4.50	5.56 / 4.75	**4.67 / 4.47**
DRealSR	6.58 / 8.58	6.14 / 4.78	4.98 / 4.77	4.71 / 4.74	4.20 / 4.30	4.82 / 4.76	5.14 / 4.46	**4.15 / 4.27**
DPED-iphone	6.01 / 7.48	5.99 / 4.55	5.44 / 5.02	4.95 / 4.78	5.11 / 4.36	5.56 / 5.36	5.89 / 4.61	**4.84 / 4.23**
RealSRSet	7.98 / 7.35	5.49 / 4.79	5.65 / 4.92	5.30 / 4.68	**5.18 / 4.31**	5.25 / 4.59	6.03 / 4.60	5.29 / 4.62

Table 1: Quantitative comparison (NIQE ↓ / PI ↓) on real-world benchmarks. Results of BSRGAN and Real-ESRGAN are taken from (Wang et al. 2021c), and others are tested with official codes.

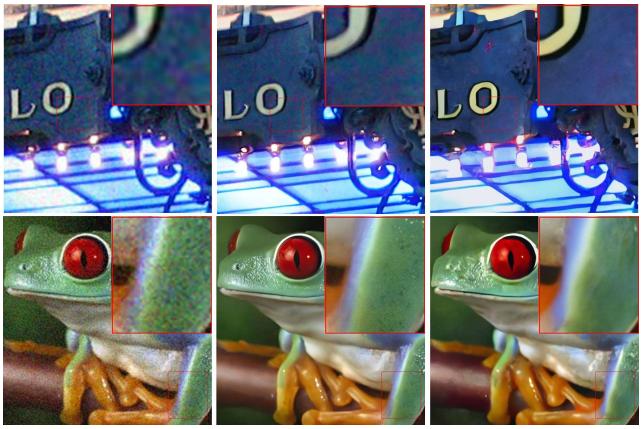
entropy losses in Eqs. (2), (4) and (5). In theory, the optimal strategy comprises training E_l foremost, succeeded by ϕ_e and ϕ_r sequentially. Nevertheless, we discovered that training them concurrently works well in practice, thereby leading to a significant reduction in overall training time. The prominent Adam optimizer (Kingma and Ba 2014) is employed to optimize all three networks, with specific parameters of $lr = 0.0001$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. Each batch contains 16 HQ images of dimensions 256×256, paired with their corresponding LQ images. All networks are implemented by PyTorch (Paszke et al. 2019) and trained for 400k iterations with 4 Tesla V100 GPUs.

Experiments

Comparison with Other Methods

We perform a comprehensive comparison of ITER against several state-of-the-art GAN-based approaches, including BSRGAN (Zhang et al. 2021b), Real-ESRGAN (Wang et al. 2021b), SwinIR-GAN (Liang et al. 2021), FeMaSR (Chen et al. 2022), and MM-RealSR (Mou et al. 2022). Specifically, BSRGAN, Real-ESRGAN, and MM-RealSR employ the RRDBNet backbone proposed by (Wang et al. 2018b), whereas SwinIR-GAN utilizes the Swin transformer architecture, and FeMaSR utilizes the VQGAN prior. Regarding diffusion-based models, we compare with the most popular work, LDM-BSR (Rombach et al. 2022), which operates in the latent feature space using the denoising diffusion models. The model is finetuned with the same dataset for fair comparison. SR3 (Saharia et al. 2022) is not included in comparison due to the unavailability of public models.

We use two different no-reference metrics, namely NIQE (Mittal, Soundararajan, and Bovik 2012) and PI (perceptual index) (Blau et al. 2018), to evaluate the performance of different approaches. NIQE is widely used in previous works involving RWSR, such as (Wang et al. 2021b; Zhang et al. 2021a; Mou et al. 2022), while PI has been extensively used in recent low-level computer vision workshops, including the renowned NTIRE (Cai et al. 2019; Zhang et al. 2020; Gu et al. 2021) and AIM (Ignatov et al. 2019, 2020).



(a) LQ input (b) LDM-BSR (c) **ITER (Ours)**

Figure 5: Problem of LDM-BSR without explicit distortion removal. (Zoom in for best view)

Comparison with GAN Methods. As demonstrated in Tab. 1, our ITER yields the best performance in 3 out of 4 benchmarks as demonstrated, and the results in the last RealSRSet are also competitive. These results demonstrate the clear superiority of ITER over existing GAN-based methods. The visual examples depicted in Fig. 4 illustrate why ITER performs better. We can observe that the textures in the images generated by ITER look more natural and realistic. On the other hand, the results from other GAN-based approaches are either over-smoothed (first row in Fig. 4) or over-sharpened (second row). GAN-based methods often encounter difficulties in generating realistic textures for different distortion levels. Moreover, they are generally harder to train and more likely to produce artifacts when not well-tuned. In conclusion, compared to GAN-based methods, our proposed ITER exhibits better performance and is more straightforward to train.

Comparison with LDM-BSR. As can be seen from Tab. 1, it is evident that although LDM-BSR utilizes a diffusion-based model, its performance is worse than that of ITER. In Fig. 5, it is apparent why quantitative results of LDM-BSR are suboptimal for the RWSR task. Although LDM-BSR is capable of generating sharper edges for the blurry LQ inputs, it struggles with eliminating complex noise degradations in both examples. On the other hand, our proposed ITER does not face such challenges and can produce outputs with greater clarity while maintaining reasonably natural textures. This can be attributed to two main reasons. Firstly, LDM-BSR incorporates continuous diffusion models, while ITER relies on discrete representations. Prior studies (Zhou et al. 2022; Chen et al. 2022) have shown that a pre-trained discrete proxy space offers benefits for intricate distortions. Secondly, ITER explicitly filters out the distortions during the encoding of LQ images into token space before diffusion processing. As a result, ITER avoids generating additional textures similar to what can occur in LDM-BSR, as demonstrated in the second example.

(a) LQ inputs (b) w/o Refinement (c) w/ Refinement

Figure 6: Comparison of results with and without iterative refinement. We can observe that the results only with distortion removal present overly smoothed textures and inconsistent color. After iterative refinement, the textures are enriched and the color is also corrected.

Ablation Study and Model Analysis

We performed a thorough analysis of various configurations of our model using a synthetic DIV2K validation test set. Firstly, we evaluated the effectiveness of refinement network in adding textures to the initial results S_l. Secondly, we assessed the necessity of the token evaluation block. Finally, we demonstrated how the token evaluation block can be exploited to manage the model preference toward removing distortions or generating textures. We utilized the PSNR metric to evaluate the quality of distortion removal and used the widely recognized perceptual metric LPIPS (Zhang et al. 2018a) to measure the performance of texture generation. The incorporation of these two metrics allowed us to assess the extent to which the proposed ITER adjusts the visual effects of its outputs in accordance with the threshold value α, as stated in Eq. (6).

Effectiveness of Iterative Refinement. We first evaluate the effectiveness of the iterative refinement network for texture generation. As illustrated in Fig. 6, the results obtained without the iterative refinement stage exhibit an over-smoothed texture and inconsistency in color. This could be attributed to the inherent limitations of token classification when confronted with complex distortions present in diverse natural images. In contrast, the results with iterative refinement are more realistic. Noticeable enhancements in texture richness and color correction are observed. These observations provide compelling evidence that the iterative refinement network plays an crucial role in our framework.

Necessity of Token Evaluation. An alternative method to decide which tokens to retain or refine involves directly selecting the top-k tokens in S_t with higher confidence, as implemented in MaskGIT (Chang et al. 2022). However, our experimental findings indicate that the top-k mask selection is trapped with local propagation. This is due to the fact that under the greedy selection strategy, the refinement network ϕ_r tends to assign higher confidence to neighboring tokens of previous selections. As illustrated in Fig. 8, the masks consistently expand around the previous step, resulting in some regions (indicated by black mask) being refined until

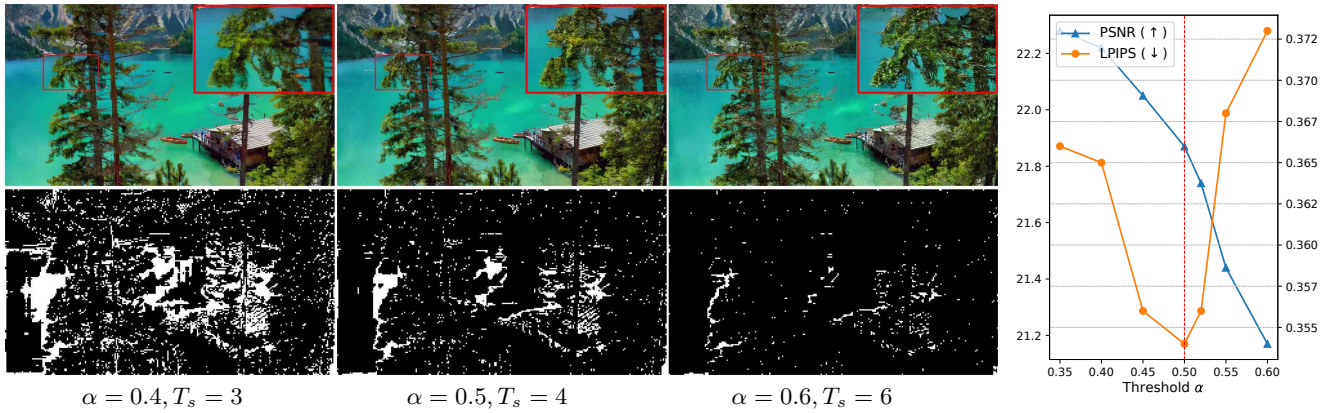


Figure 7: Results with different threshold. Left: visual examples of final results (top) and masks at start time step (bottom). Bigger α leads to stronger texture effect because more refinement steps are conducted. Right: LPIPS/PSNR with different α.

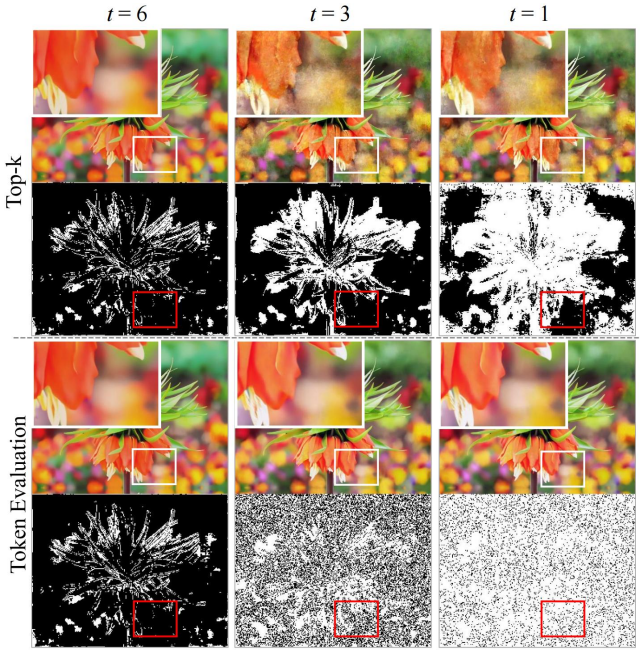


Figure 8: The top-k masking technique suffers from the local propagation problem, which is effectively avoided by the proposed token evaluation block.

the last step. This approach is unfavorable in the iterative texture generation process because it corrupts some good-looking regions with unnecessary refinement. Our hypothesis is that low-level vision tasks exhibit the locality property where neighboring features are naturally more correlated. Although the networks have large receptive fields with Swin transformer blocks, it still prefers to propagate information to neighbor features, resulting in higher confidence scores surrounding previous selections.

The use of the proposed token evaluation network ϕ_e allows the iterative refinement process to avoid the local propagation trap. As demonstrated in Fig. 8, the masks are distributed more evenly, leading to more consistent results.

Balance Restoration and Generation. In Fig. 7, we have presented an example of the results with different threshold α. It is evident from the results that a larger α will lead to the identification of fewer valid tokens, thereby necessitating more refinement steps, or in other words, a larger start time step T_s. Consequently, larger α create images with stronger textures. In Fig. 7, we have provided quantitative results for the different α thresholds, where the effectiveness of each threshold can be seen in the score curves of LPIPS and PSNR. We have observed that smaller α produce enhanced PSNR scores, which is a clear indication of a better ability to eliminate distortion. As for texture generation performance, the optimal LPIPS score of $\alpha = 0.5$ was achieved since both excessively strong and overly weak textures can negatively impact the perceptual quality. In practice, we can adjust α to obtain the desired results without having to modify the network, resulting in a more adaptable framework during inference than GAN-based techniques, which are unmodifiable once the training process is completed.

Conclusion

We presents a novel framework named ITER that utilizes iterative evaluation and refinement techniques for texture generation in real-world image super-resolution. Unlike GANs, which require painstake training, we incorporate discrete diffusion generative pipelines with token evaluation and refinement blocks for RWSR. This new approach simplifies training with just cross-entropy losses and allows for greater flexibility in balancing distortion removal and texture generation during inference. Furthermore, our ITER has demonstrated superior performance with ≤ 8 iterations, highlighting the vast potential of discrete diffusion models in RWSR.

Acknowledgements

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *CVPRW*.

Anwar, S.; Khan, S.; and Barnes, N. 2020. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3): 1–34.

Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; and Zelnik-Manor, L. 2018. The 2018 PIRM challenge on perceptual image super-resolution. In *ECCVW*, 0–0.

Bond-Taylor, S.; Hessey, P.; Sasaki, H.; Breckon, T. P.; and Willcocks, C. G. 2022. Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes. In *ECCV*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Cai, J.; et al. 2019. NTIRE 2019 Challenge on Real Image Super-Resolution: Methods and Results. *CVPRW*.

Chan, K. C.; Wang, X.; Xu, X.; Gu, J.; and Loy, C. C. 2021. GLEAN: Generative latent bank for large-factor image super-resolution. In *CVPR*, 14245–14254.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. MaskGIT: Masked Generative Image Transformer. In *CVPR*.

Chen, C.; Gong, D.; Wang, H.; Li, Z.; and Wong, K.-Y. K. 2020. Learning Spatial Attention for Face Super-Resolution. In *IEEE TIP*.

Chen, C.; Shi, X.; Qin, Y.; Li, X.; Han, X.; Yang, T.; and Guo, S. 2022. Real-World Blind Super-Resolution via Feature Matching with Implicit High-Resolution Priors. In *ACM MM*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-Trained Image Processing Transformer. In *CVPR*.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating More Pixels in Image Super-Resolution Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22367–22377.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*, 12873–12883.

Fritsche, M.; Gu, S.; and Timofte, R. 2019. Frequency separation for real-world super-resolution. In *ICCVW*, 3599–3608.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10021–10030.

Gu, J.; et al. 2021. NTIRE 2021 Challenge on Perceptual Image Quality Assessment. *CVPRW*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Ignatov, A.; Kobyshev, N.; Timofte, R.; Vanhoey, K.; and Van Gool, L. 2017. DSLR-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 3277–3285.

Ignatov, A.; et al. 2019. AIM 2019 Challenge on RAW to RGB Mapping: Methods and Results. *ICCVW*.

Ignatov, A.; et al. 2020. AIM 2020 Challenge on Learned Image Signal Processing Pipeline. *ECCVW*, 152–170.

Ji, X.; Cao, Y.; Tai, Y.; Wang, C.; Li, J.; and Huang, F. 2020. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, 466–467.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*, 8110–8119.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016a. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 1646–1654.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016b. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 1637–1645.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 4681–4690.

Lezama, J.; Chang, H.; Jiang, L.; and Essa, I. 2022. Improved masked image generation with token-critic. *ECCV*.

Li, X.; Chen, C.; Lin, X.; Zuo, W.; and Zhang, L. 2022. From Face to Natural Image: Learning Real Degradation for Blind Image Super-Resolution. In *ECCV*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *ICCVW*.

Liang, J.; Zeng, H.; and Zhang, L. 2022. Efficient and Degradation-Adaptive Network for Real-World Image Super-Resolution. In *ECCV*.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 136–144.

Liu, A.; Liu, Y.; Gu, J.; Qiao, Y.; and Dong, C. 2022. Blind image super-resolution: A survey and beyond. *IEEE TPAMI*.

Liu, M.; Wei, Y.; Wu, X.; Zuo, W.; and Zhang, L. 2023. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *Science China Information Sciences*, 66(5): 1–28.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*.

Maeda, S. 2020. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, 291–300.

Mei, Y.; Fan, Y.; and Zhou, Y. 2021. Image Super-Resolution With Non-Local Sparse Attention. In *CVPR*, 3517–3526.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3): 209–212.

Mou, C.; Wu, Y.; Wang, X.; Dong, C.; Zhang, J.; and Shan, Y. 2022. MM-RealSR: Metric Learning based Interactive Modulation for Real-World Super-Resolution. *ECCV*.

Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; and Shen, H. 2020. Single image super-resolution via a holistic attention network. In *ECCV*, 191–207. Springer.

Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C. C.; and Luo, P. 2020. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 262–277. Springer.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, volume 32, 8026–8037.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE TPAMI*.

Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; and Wen, F. 2020. Bringing old photos back to life. In *CVPR*, 2747–2757.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. In *arXiv preprint arXiv:2305.07015*.

Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; and Guo, Y. 2021a. Unsupervised Degradation Representation Learning for Blind Super-Resolution. In *CVPR*, 10581–10590.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021b. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*, 9168–9178.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021c. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. *ICCVW*.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018a. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018b. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 0–0.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 101–117. Springer.

Wei, Y.; Gu, S.; Li, Y.; Timofte, R.; Jin, L.; and Song, H. 2021. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, 13385–13394.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *CVPR*, 672–681.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*.

Zhang, J.; Lu, S.; Zhan, F.; and Yu, Y. 2021a. Blind Image Super-Resolution via Contrastive Representation Learning. *arXiv preprint arXiv:2107.00708*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021b. Designing a practical degradation model for deep blind image super-resolution. *ICCV*.

Zhang, K.; et al. 2020. NTIRE 2020 Challenge on Perceptual Extreme Super-Resolution: Methods and Results. *CVPRW*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhang, W.; Liu, Y.; Dong, C.; and Qiao, Y. 2019a. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *CVPR*, 3096–3105.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022. Efficient Long-Range Attention Network for Image Super-resolution. In *ECCV*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 286–301.

Zhang, Y.; Li, K.; Li, K.; Zhong, B.; and Fu, Y. 2019b. Residual Non-local Attention Networks for Image Restoration. In *ICLR*.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018c. Residual dense network for image super-resolution. In *CVPR*, 2472–2481.

Zhou, S.; Chan, K. C.; Li, C.; and Loy, C. C. 2022. Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. In *NeurIPS*.

Zhou, S.; Zhang, J.; Zuo, W.; and Loy, C. C. 2020. Cross-Scale Internal Graph Neural Network for Image Super-Resolution. In *NeurIPS*.