

Article

Global-Token U-Net with Hybrid Loss for Trustworthy Medical Image Super-Resolution

Jiaqi Shang ^{1,†} , Zhiyuan Xu ^{1,†} and Dongdong Wang ^{2,*} ¹ Department of Mechanical Engineering, Hohai University, Nanjing 211100, China² School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

* Correspondence: ddong.wang.njust.cn@gmail.com

† These authors contributed equally to this work.

Abstract

Super-resolution technology significantly enhances the visual quality of low-resolution medical images, resulting in ultra-high-resolution clear images. Super-resolution technology based on artificial intelligence has achieved great success in reconstruction quality. However, like the image restoration task, super-resolution is also an ill-posed problem, and current work lacks consideration of trustworthiness. Medical image super-resolution needs to ensure clarity and, more importantly, to ensure that the output image is reliable and does not produce false details and mislead the diagnosis. To address the trustworthy issue of medical image super-resolution, we design a novel hybrid loss that combines a hinge-based adversarial term with a PSNR-based regularization. In the designed loss function, the adversarial term makes the reconstructed result close to the distribution of the true high-resolution image, thus generating more refined high-frequency textures, while the PSNR-based regularization term explicitly reduces the deviation from the ground truth. We apply this loss in the global-token U-Net backbone network and add a lightweight VGG as the discriminator for adversarial terms. We empirically verify that integrating the proposed methods can enhance the trustworthiness of medical image super-resolution technology while maintaining high reconstruction quality.

Keywords: medical image processing; super-resolution; trustworthy AI; hybrid loss function

1. Introduction

Super-resolution is an image processing technology aimed at significantly enhancing the resolution of images, thereby improving visual quality [1,2]. For medical imaging, low-resolution medical images can hinder diagnosis. Super-resolution technology can bring additional knowledge to improve the diagnostic effect [3]. As shown in Figure 1, super-resolution technology can significantly improve the visual quality of blurred medical images. Higher visual knowledge can assist human experts in diagnosis and can also be used to enhance the performance of advanced machine intelligence tasks, such as further instance segmentation and classification.

Dong et al. used convolutional neural networks to achieve universal image super-resolution [4], which was the first time that artificial intelligence technology was applied in the field of image super-resolution. Since then, super-resolution methods based on artificial intelligence have become mainstream, and many new models and methods have been proposed [5–8]. With the development of intelligent medical care, dedicated super-resolution models for medical images have also been proposed, achieving ideal results [9–11].



Academic Editor: Saed Moradi

Received: 4 December 2025

Revised: 15 February 2026

Accepted: 18 February 2026

Published: 26 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

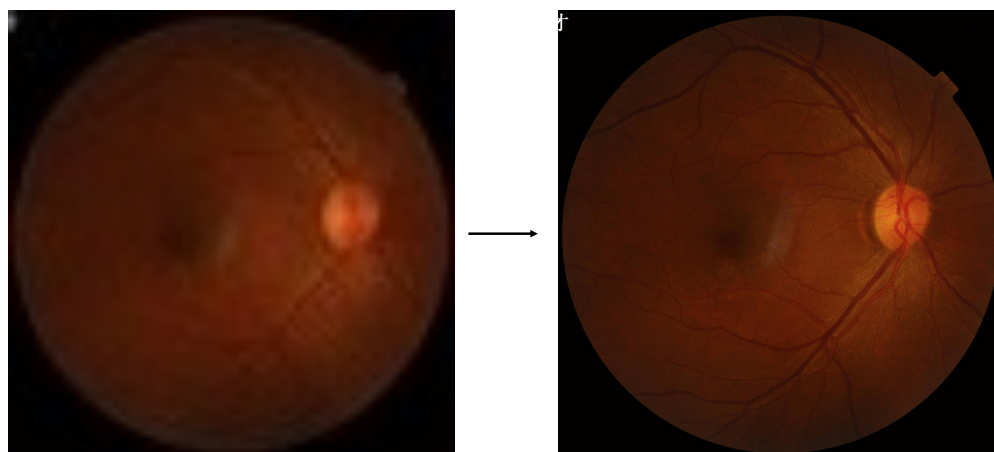


Figure 1. Medical image super-resolution example.

Deep learning has achieved great success in image restoration fields, including super-resolution [12,13]. However, all image restoration technologies based on artificial intelligence involve the process of deep scene synthesis. Scene synthesis models with high perceptual quality often generate results that are difficult to distinguish from real data; that is, details that seem clear but do not actually exist appear [14,15]. In the medical field, the consequences of this kind of image hallucination are serious [16]. Figure 2 is an extreme example of super-resolution using our other high-perception visual models. It can be seen that the super-resolution result looks very clear, but there are a lot of falsified details. This kind of medical imaging will hinder the diagnostic process and draw our attention to trustworthiness.

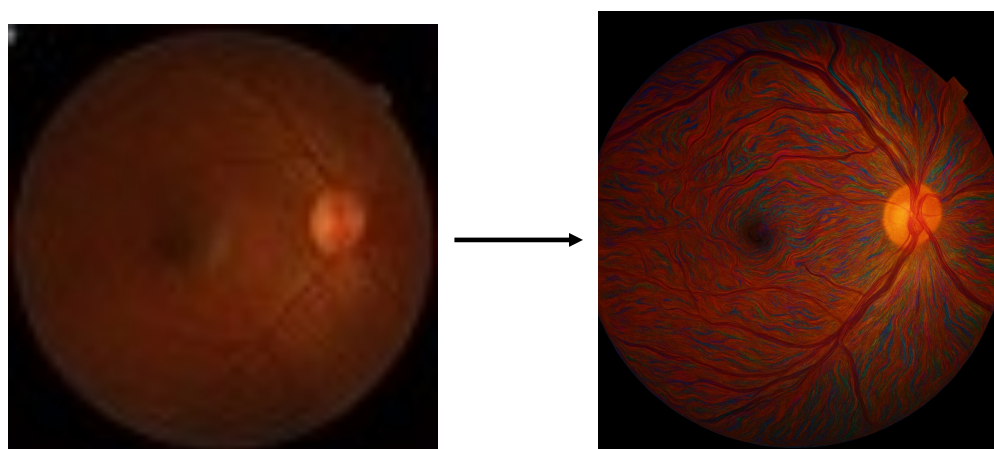


Figure 2. An extreme example of a super-resolution hallucination generated using our other visual synthesis models.

In this paper, we propose a method to enhance the trustworthiness of medical image super-resolution. Prior works have pointed out that medical images have global features [17,18]. Considering long-distance semantics, we adopt global-token U-Net as the backbone network. We design a new hybrid loss function that combines PSNR loss [19] and adversarial loss [20]. The adversarial term uses a lightweight VGG as the discriminator, mainly considering the reconstruction effect to generate fine and clear textures. The PSNR term uses explicit regularization constraints to reduce the difference between the output results and the ground truth. With this combination of hybrid loss and global modeling, we can achieve a sufficiently clear super-resolution result and improve the level of trustworthiness. The main contributions of this work are threefold:

- We revisit medical image super-resolution from the perspective of trustworthiness and identify hallucinated high-frequency details as a critical failure mode of deep learning-based restoration for ill-posed inverse problems. We argue that trustworthy medical image super-resolution should jointly pursue visual clarity and anatomical fidelity, rather than perceptual quality alone.
- We propose a global-token U-Net-based generator and a novel hybrid loss, which jointly aim to achieve trustworthy medical image super-resolution by capturing global anatomical context and suppressing hallucinated high-frequency details. The backbone augments a U-shaped convolutional network with global tokens to efficiently capture long-range anatomical context, while the hybrid loss combines a hinge-based adversarial term driven by a lightweight VGG discriminator with a PSNR-based regularization term to enhance perceptual sharpness and explicitly suppress hallucinated details.
- We conduct extensive experiments on medical image super-resolution benchmarks, showing that the proposed method maintains high reconstruction accuracy while achieving consistently perceptual quality. Qualitative and quantitative analyses further demonstrate that our approach produces fewer falsified structures and improves the trustworthiness of the reconstructed medical images.

The remainder of this paper is organized as follows: Section 2 reviews related work on universal and medical image super-resolution and discusses hallucination risks in ill-posed restoration. Section 3 presents the proposed global-token U-Net architecture and the hybrid adversarial–PSNR loss. Section 4 reports experimental settings and quantitative/qualitative results, including cross-domain evaluation and ablation studies. Finally, Section 5 concludes this paper and outlines future directions.

2. Related Work

2.1. Universal Image Super-Resolution

Early image super-resolution methods were mainly based on computer graphics and multimedia information processing technology [21]. Later, convolutional neural networks achieved outstanding performance in various computer vision tasks, and some super-resolution methods based on convolutional networks were proposed [4,5]. Modern super-resolution techniques mainly employ models with strong representational capabilities, such as transformers [8,22,23] and diffusion models [1,7].

The learning-based works include specific models for super-resolution and universal models for image restoration, and both achieve very clear reconstruction results. However, these works lack consideration of trustworthiness, which limits their application in real-world scenarios.

2.2. Medical Image Super-Resolution

Medical imaging has unique proprietary prior knowledge, so medical image super-resolution models can be designed [24–27]. Some common medical image restoration techniques also take into account super-resolution [11,18]. The acquisition of medical images involves multiple modalities and higher dimensions. Therefore, researchers have designed super-resolution models for multimodal medical imaging [10,28,29].

Hallucinations can lead to serious consequences, so the reconstruction of medical images has very strict requirements for trustworthiness [16]. This inspires us to propose novel solutions for trusted medical image super-resolution.

2.3. Ill-Posed Inverse Problems and Hallucination in Restoration

Image restoration tasks such as denoising, deblurring, inpainting, and super-resolution can all be formulated as inverse problems of an underlying degradation process. Image restoration models are often more concerned with synthetic datasets, and models that perform well in their settings may not be effective in real-world applications.

Learning-based image restoration is clearly an ill-posed problem. Restoring high visual quality from a degraded image may yield different results, and these seemingly high-quality clear images may contain details that do not exist in real scenes. Cohen et al. provide an information-theoretic analysis of hallucinations in generative restoration models, revealing that there is a certain contradiction between better perceptual quality and keeping true details [15].

In medical imaging, such hallucinated structures can have serious consequences. Bhadra et al. point out that images reconstructed by intelligent models can lead to false clear details that mislead the diagnosis [16]. As a deep synthesis reconstruction technology, super-resolution also faces similar risks. Therefore, it is important to design restoration schemes that explicitly account for the ill-posed nature of the problem and control hallucinated content. This perspective motivates us to focus on trustworthy medical image super-resolution and to develop methods that balance visual clarity with fidelity.

2.4. Discussion

Super-resolution is an inherently ill-posed inverse problem: multiple high-resolution images can correspond to the same low-resolution observation under the degradation operator, making the mapping from LR to HR non-unique and under-constrained. Based on the above analysis of hallucination, we discuss the hallucination risk for the methods in Sections 2.1 and 2.2 and summarize the results in Table 1. The synthetic paths for generating processing results using different methods are different, which leads to different risks of false details and other credibility issues. This work aims to enhance the credibility of super-resolution by imposing constraints on the loss function.

Table 1. Multi-aspect comparison of representative methods discussed in Sections 2.1 and 2.2.

Sec.	Method	Paradigm	Input/Task	Degradation Setting	Halluc. Risk [†]	Notes (Strengths vs. Trustworthiness Limitations)
2.1	GraphCutSR [21]	Optimization (graph-cut)	Multi-frame SR	Model-based; alignment-dependent	Low	Interpretable energy minimization and explicit regularization; typically faithful but limited representational power; sensitive to registration/motion errors and computationally heavier than feed-forward nets.
2.1	SRCNN [4]	CNN	Single-image SR	Usually synthetic (e.g., bicubic)	Low	Stable pixel-level fidelity as a baseline; limited capacity and tends to over-smooth; weak robustness to complex/unknown real degradations.
2.1	VDSR [5]	Deep CNN (residual)	Single-image SR	Usually synthetic (e.g., bicubic)	Low	Higher reconstruction accuracy via depth; still assumes a fixed degradation model; may sacrifice perceptual sharpness for fidelity under pixel-driven training.
2.1	EFATSR [22]	Transformer	Single-image SR	Synthetic degradations	Medium	Strong representation with long-range/frequency-aware modeling; performance and reliability can degrade under domain shift, where “plausible” high-frequency details may not be anatomically/physically grounded.
2.1	DFDAN [23]	CNN + attention	Single-image SR	Synthetic degradations	Medium	Efficient feature distillation and enlarged receptive fields; good efficiency/quality trade-off; trustworthiness largely depends on supervision realism and whether explicit consistency constraints are used.
2.1	ETCASR [8]	Transformer (efficient attention)	Stereo SR	Synthetic degradations + stereo cues	Medium	Exploits stereo correspondence for better detail recovery; requires stereo pairs and careful cross-view consistency to avoid introducing view-inconsistent “invented” structures.
2.1	DWTrans [1]	Diffusion + transformer	Single-image SR	Generative prior (learned)	High	Pretrained diffusion priors can boost perceptual realism; generative sampling increases hallucination risk unless strong data consistency and conservative objectives are enforced.
2.1	PD-CR [7]	Diffusion (constrained refinement)	Restoration/SR	Generative prior + refinement constraints	Medium–High	Constrained refinement can mitigate artifacts vs. unconstrained generative restoration; still slower and may hallucinate fine structures in ill-posed settings if constraints are insufficient.
2.2	SwinIR [18]	Transformer (Swin)	Restoration incl. SR	Mostly synthetic; can be fine-tuned	Medium	Strong universal backbone for SR/denoising/deblurring; not inherently medical-specific, so domain shift can yield anatomically implausible details without explicit fidelity constraints.
2.2	RAT [11]	Transformer (region attention)	Medical restoration incl. SR	Task-/modality-specific	Medium	Region-aware modeling can enhance salient structures; still requires explicit fidelity/anatomy constraints to suppress hallucinated content in ill-posed reconstructions.
2.2	HAT [27]	Transformer (hybrid attention)	SR backbone	Synthetic degradations	Medium	High-capacity SR transformer with strong detail synthesis; for medical usage, typically needs stronger input consistency regularization to avoid over-confident micro-textures.
2.2	SMambaUNet [25]	U-Net + Mamba	Medical SR	Modality-specific	Low–Medium	Combines U-shaped locality with long-range dependency modeling; “self-prior” can regularize outputs; trustworthiness remains data- and loss-dependent (especially under limited clinical data).
2.2	TAT [26]	Transformer (task-adaptive)	All-in-One MedIR	Multi-task/multi-degradation	Medium	Mitigates task interference and imbalance via task-adaptive weights and loss balancing; broader scope increases the need for conservative constraints to ensure anatomy-faithful SR across tasks/modalities.
2.2	InverseSR [10]	Latent diffusion	3D brain MRI SR	Generative prior (3D)	High	Diffusion priors can recover visually convincing 3D details; generative nature implies higher hallucination risk and typically demands strict data consistency and clinical validation.
2.2	CoSFNet [28]	Unified network (SR + motion)	4D MRI (motion+SR)	Motion + SR coupled model	Low–Medium	Joint motion estimation and SR improves temporal/spatial coherence; specialized pipeline and data requirements; explicit motion/consistency modeling can improve reliability.
2.2	MSFDecouple [29]	Multiscale fusion	Multimodal MRI SR	Cross-modal priors	Low–Medium	Leverages complementary modalities to stabilize SR; requires aligned multimodal inputs; potential bias if one modality is corrupted/misaligned (cross-modal “leakage”).

[†] “Hallucination risk” is a qualitative indicator mainly driven by whether the method relies on explicit generative sampling (e.g., diffusion) versus feed-forward regression and whether strong input/data consistency mechanisms are typically present.

3. Method

3.1. Model Architecture

We adopt the classic U-Net [30] architecture and add global-token-aware mechanisms at the bottleneck layer as the backbone network. The pipeline is shown in Figure 3. At the bottleneck level, given the encoder feature $\mathbf{F}_S \in \mathbb{R}^{C \times H \times W}$, we first project it into a sequence of local tokens via a 3×3 convolution and flattening:

$$\mathbf{X}^{(0)} = \Phi(\mathbf{F}_S) \in \mathbb{R}^{N \times d}, \quad N = H \cdot W, \quad (1)$$

where $\Phi(\cdot)$ denotes the projection and flattening operator, d is the token dimension, and each row of $\mathbf{X}^{(0)}$ is a local token.

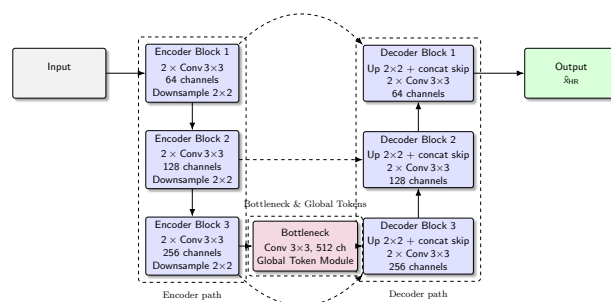


Figure 3. The architecture of global-token U-Net. At the bottleneck layer, we choose the number of attention heads to be 4 and the number of global token layers to be 4.

We introduce K learnable global tokens

$$\mathbf{G}^{(0)} \in \mathbb{R}^{K \times d}, \quad (2)$$

and form the initial token matrix by concatenation

$$\mathbf{Z}^{(0)} = \begin{bmatrix} \mathbf{G}^{(0)} \\ \mathbf{X}^{(0)} \end{bmatrix} \in \mathbb{R}^{(K+N) \times d}. \quad (3)$$

The global-token module applies T layers of multi-head self-attention to $\mathbf{Z}^{(0)}$. For $t = 1, \dots, T$, we write

$$\mathbf{Z}^{(t)} = \mathbf{Z}^{(t-1)} + \text{MSA}^{(t)}(\mathbf{Z}^{(t-1)}), \quad (4)$$

where $\text{MSA}^{(t)}(\cdot)$ denotes multi-head self-attention. Given an input token matrix $\mathbf{Z} \in \mathbb{R}^{(K+N) \times d}$, the multi-head self-attention is defined as

$$\text{MSA}^{(t)}(\mathbf{Z}) = \text{Concat}(\mathbf{H}_1^{(t)}, \dots, \mathbf{H}_M^{(t)}) \mathbf{W}_O^{(t)}, \quad (5)$$

$$\mathbf{H}_m^{(t)} = \text{Softmax}\left(\frac{\mathbf{Q}_m^{(t)}(\mathbf{K}_m^{(t)})^\top}{\sqrt{d_m}}\right) \mathbf{V}_m^{(t)}, \quad (6)$$

with query, key, and value matrices

$$\mathbf{Q}_m^{(t)} = \mathbf{Z} \mathbf{W}_{Q,m}^{(t)}, \quad \mathbf{K}_m^{(t)} = \mathbf{Z} \mathbf{W}_{K,m}^{(t)}, \quad \mathbf{V}_m^{(t)} = \mathbf{Z} \mathbf{W}_{V,m}^{(t)}, \quad (7)$$

where $m = 1, \dots, M$ indexes attention heads, $d_m = d/M$ is the head dimension, and $\mathbf{W}_{Q,m}^{(t)}, \mathbf{W}_{K,m}^{(t)}, \mathbf{W}_{V,m}^{(t)}, \mathbf{W}_O^{(t)}$ are learnable projection matrices.

After T layers, we obtain the updated token matrix $\mathbf{Z}^{(T)}$ and split it back into global and local parts:

$$\begin{bmatrix} \mathbf{G}^{(T)} \\ \mathbf{X}^{(T)} \end{bmatrix} = \mathbf{Z}^{(T)}. \quad (8)$$

Finally, the refined local tokens are reshaped back to the bottleneck feature map

$$\tilde{\mathbf{F}}_S = \Phi^{-1}(\mathbf{X}^{(T)}) \in \mathbb{R}^{C \times H \times W}, \quad (9)$$

which is then fed into the decoder path of the U-Net. The resulting global-token module is shown in Figure 4.

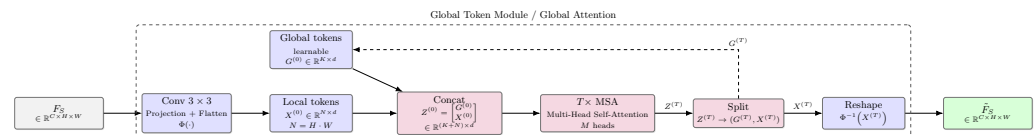


Figure 4. The detailed structure of the global token for the bottleneck module.

3.2. Hybrid Adversarial–PSNR Loss

The core of our method is to use the newly designed hybrid loss. To introduce adversarial loss, we use a lightweight VGG as the discriminator for the output of global-token U-Net, and its structure is shown in Figure 5. As such, global-token U-Net is a generator and its output also needs to be constrained by differences with the ground truth. Let G denote the generator and D the discriminator. Given a mini-batch of low-resolution inputs $\{\mathbf{y}_i\}_{i=1}^B$ and corresponding high-resolution ground-truth images $\{\mathbf{x}_i\}_{i=1}^B$, the generator produces $\hat{\mathbf{x}}_i = G(\mathbf{y}_i)$. We adopt a hinge-based adversarial loss combined with a PSNR-based regularization term.

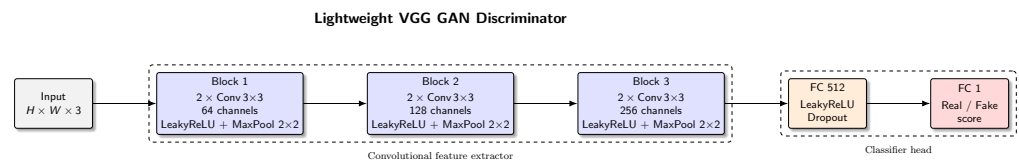


Figure 5. An additional discriminator outside the backbone network, mainly for generating adversarial losses.

Adversarial loss. This module was mainly inspired by Lucic et al. [20]. For the discriminator, the hinge adversarial loss is defined as

$$\mathcal{L}_D = \frac{1}{B} \sum_{i=1}^B \left[\max(0, 1 - D(\mathbf{x}_i)) + \max(0, 1 + D(\hat{\mathbf{x}}_i)) \right]. \quad (10)$$

The generator is trained to fool the discriminator via

$$\mathcal{L}_{\text{adv}}^G = -\frac{1}{B} \sum_{i=1}^B D(\hat{\mathbf{x}}_i). \quad (11)$$

PSNR-based regularization. This module was mainly inspired by Chen et al. [19]. For each sample, we compute the mean-squared error (MSE) between $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$ as

$$\text{MSE}_i = \frac{1}{P} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \quad P = 3 \cdot H_{\text{out}} \cdot W_{\text{out}}, \quad (12)$$

where H_{out} and W_{out} denote the height and width, respectively. Given the maximum possible pixel value $M = 255$ and a small constant $\varepsilon = 1 \times 10^{-4}$ for numerical stability, the per-sample PSNR is

$$\text{PSNR}_i = 10 \log_{10} \left(\frac{M^2}{\max(\text{MSE}_i, \varepsilon)} \right). \quad (13)$$

We define the PSNR-based loss as the negative batch-averaged PSNR:

$$\mathcal{L}_{\text{PSNR}} = -\frac{1}{B} \sum_{i=1}^B \text{PSNR}_i. \quad (14)$$

Minimizing $\mathcal{L}_{\text{PSNR}}$ is therefore equivalent to maximizing the PSNR of the reconstructed images.

Generator objective. The final generator loss is given by

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}}^G + \lambda_{\text{PSNR}} \mathcal{L}_{\text{PSNR}}, \quad (15)$$

where $\lambda_{\text{PSNR}} > 0$ is a hyperparameter that refers to a weighting coefficient.

The adversarial term encourages perceptually realistic, high-frequency details, while the PSNR-based regularization explicitly penalizes deviations from the ground truth, thereby suppressing hallucinated structures and improving the trustworthiness of the reconstructed medical images.

4. Experiments

4.1. Dataset

Our experiment is based on the FIVES dataset [31]. This is a dataset with 800 fundus images taken by the same device. It is balanced across four categories: normal eyes, diabetic retinopathy, age-related macular degeneration, and glaucoma (200 images per category). All images are scanned in a consistent style at 2048×2048 resolution. To obtain the super-resolution dataset, we use the bicubic method to downsample the images:

$$\mathbf{y}_i = \text{Down}_s^{\text{bicubic}}(\mathbf{x}_i), \quad (16)$$

where $\text{Down}_s^{\text{bicubic}}(\cdot)$ denotes bicubic downsampling by a factor of s . The resulting pairs $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^N$ constitute the supervised training set for medical image super-resolution. s is set to 8, thereby producing low-resolution images with a resolution of 256×256 . Then we divide the training set and the test set in a ratio of 8:2.

4.2. Implementation Details

We implement our method using the deep learning framework PyTorch 2.5.1 [32] and the Adam [33] optimizer. The learning rate is set to 1×10^{-4} and gradually reduced to 1×10^{-5} using the cosine annealing method. The batch size is set to 8 and the model is trained using two Nvidia VGPUs (32 G, NVIDIA Corporation, Santa Clara, CA, USA) for a fixed 800 epochs. The hyperparameter λ_{PSNR} in the designed loss function is set to 0.1.

The training curve of the hybrid loss during the training process is shown in Figure 6. As the training converges, the PSNR loss decreases more significantly, while the effect of the adversarial loss gradually decreases.

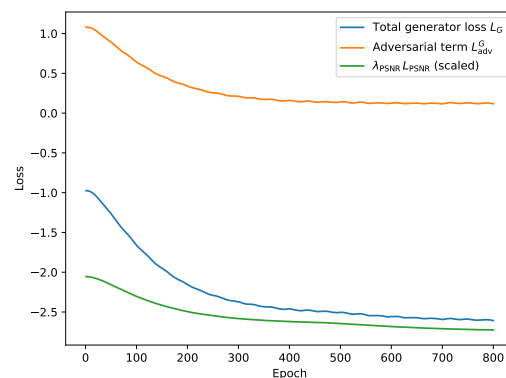


Figure 6. The training curve under the set implementation details.

4.3. Metrics

We use SSIM to assess how similar the reconstructed results are to the ground truth in terms of structure and visual quality, and we use the PSNR to measure the pixel-level fidelity [34,35].

In addition to this, to evaluate the trustworthiness, we use LR-PSNR and LR-SSIM. These two metrics are calculated by downsampling the super-resolved image to the original resolution and calculating the related metrics with the original input image. Higher LR-PSNR and LR-SSIM indicate that the super-resolution results are faithful to the original details and have higher reliability.

In addition to the metrics for measuring fidelity to the original details, we have also introduced three general indicators of trustworthiness. LR-LPIPS is computed between $\text{Down}(\hat{\mathbf{x}})$ and $\mathbf{y}$ (lower is better). GradCons is the mean L_1 error of Sobel gradients, e.g., $\|\nabla(\text{Down}(\hat{\mathbf{x}})) - \nabla(\mathbf{y})\|_{1,\text{mean}}$ (images normalized to $[0, 1]$). HF-Err is the mean L_1 error of Laplacian/high-pass responses, e.g., $\|\Delta(\text{Down}(\hat{\mathbf{x}})) - \Delta(\mathbf{y})\|_{1,\text{mean}}$ (lower is better).

4.4. Results

Table 2 compares the quantitative evaluation under the previous setting using indicators, including the general super-resolution methods and medical image super-resolution algorithms. The method we proposed demonstrates good reconstruction quality. Meanwhile, our method has reached a state-of-the-art level in LR-PSNR and LR-SSIM, demonstrating a higher degree of trustworthiness and reliability. Although PD-CR [7] and DiT4SR [36] based on diffusion models achieve higher PSNRs, their restoration of original details is still weaker.

Table 2. Quantitative comparison on the FIVES test set for $8\times$ super-resolution with additional trustworthiness metrics. PSNR/LR-PSNR are in dB. $\uparrow$ indicates higher values are better, $\downarrow$ indicates lower values are better. LR-* metrics (including LR-PSNR, LR-SSIM, LR-LPIPS) are computed between $\text{Down}(\hat{\mathbf{x}})$ and the original LR input $\mathbf{y}$. DiT4SR and LMLT models are fine-tuned from their publicly available pre-trained weights.

Method	SR Quality vs. HR		Trustworthiness/LR Consistency vs. LR Input				
	PSNR $\uparrow$	SSIM $\uparrow$	LR-PSNR $\uparrow$	LR-SSIM $\uparrow$	LR-LPIPS $\downarrow$	GradCons $\downarrow$	HF-Err $\downarrow$
SRCNN [4]	25.8	0.83	49.6	0.99	0.010	0.006	0.006
VDSR [5]	26.4	0.84	49.3	0.98	0.014	0.008	0.008
RCAN [6]	27.0	0.86	49.1	0.98	0.016	0.009	0.009
PD-CR [7]	28.1	0.87	49.0	0.97	0.025	0.015	0.016
ETCASR [8]	27.2	0.87	49.1	0.99	0.012	0.007	0.007
RAT [11]	27.3	0.86	49.0	0.99	0.013	0.007	0.008
HAT [27]	27.4	0.87	49.2	0.99	0.011	0.006	0.007
DiT4SR [36]	28.3	0.88	48.9	0.97	0.028	0.017	0.018
LMLT [37]	27.3	0.87	49.3	0.99	0.010	0.006	0.006
Ours	27.6	0.88	49.9	0.99	0.006	0.004	0.0048

The intuitive qualitative comparison is shown in Figure 7, from which it can be seen that our method maintains the original details while achieving high visual quality.

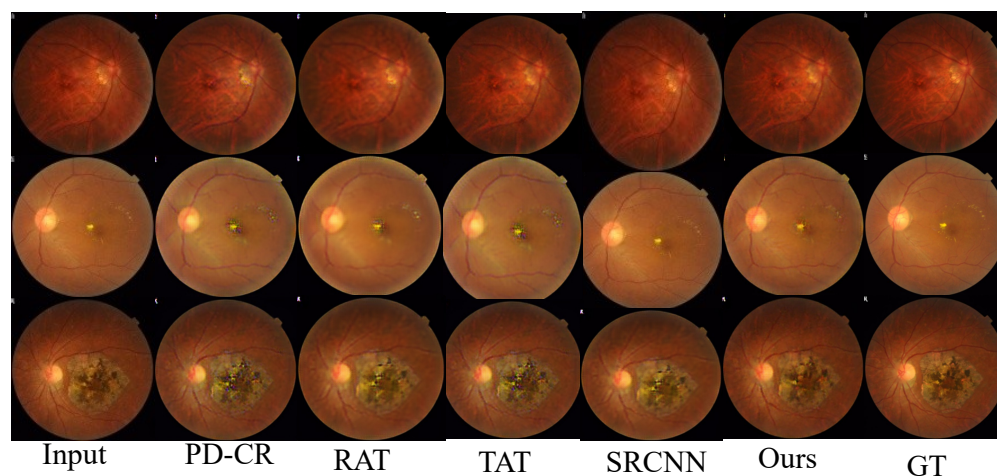


Figure 7. Visual effect comparison of super-resolution algorithms.

4.5. Cross-Domain Evaluation

To demonstrate the generalization ability of the proposed method, we conducted cross-domain tests on the trained model using low-resolution fundus images without any ground-truth references. ODIR [38] is a multi-class fundus dataset. We performed super-resolution processing on the blurred downsampled images within this dataset and obtained ideal results, indicating that the proposed method can be applied to real super-resolution tasks on different devices, as shown in Figure 8.

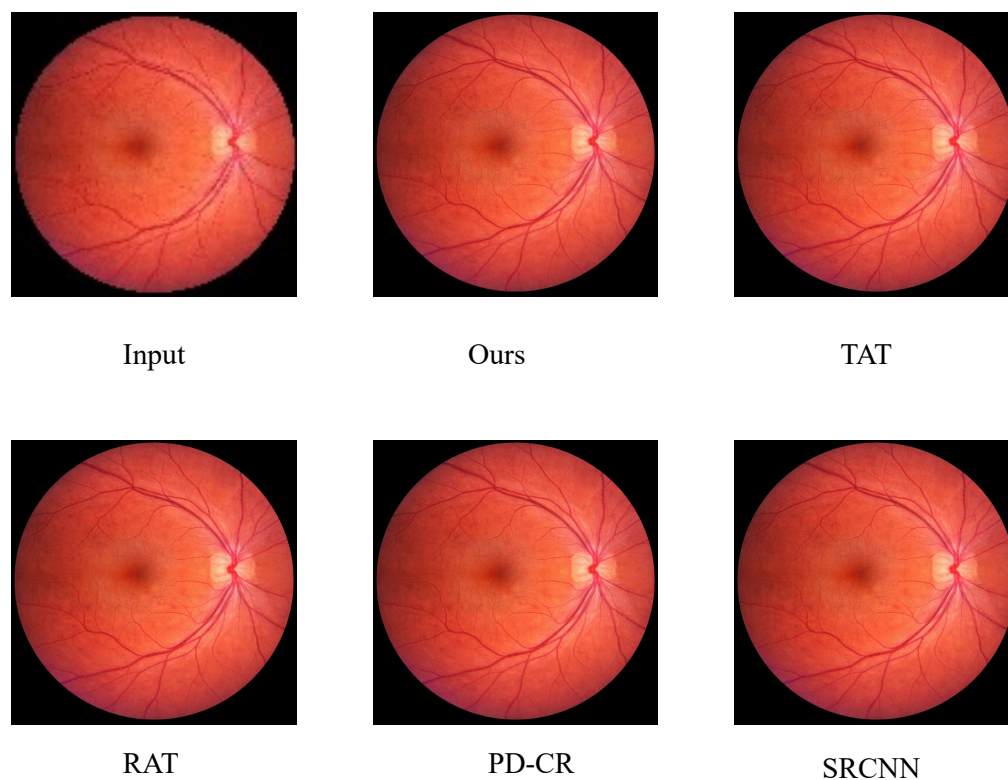


Figure 8. Cross-domain validation results without ground truth.

4.6. Ablation Study

Our method obtains competitive LR-PSNR and LR-SSIM, fully ensuring the credibility of the original semantics of medical images. The source of trustworthiness lies in the combined influence of the hybrid loss function. To verify this, we conduct ablation experiments using a single loss function (PSNR or Adv.), and the quantitative results are shown in Table 3.

Table 3. Quantitative results of the ablation study on the FIVES test set ($8\times$ super-resolution). $\uparrow$ indicates higher values are better.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LR-PSNR $\uparrow$	LR-SSIM $\uparrow$
PSNR Loss	27.4	0.87	49.02	0.99
Adv. Loss	27.3	0.88	49.12	0.99
hybrid loss	27.6	0.88	49.88	0.99

Combining the adversarial term with the PSNR-based regularization yields the best full-resolution PSNR/SSIM and restores LR-PSNR/LR-SSIM compared with the adversarial-only variant. This confirms that the proposed hybrid loss can effectively balance perceptual sharpness and input consistency, thereby enhancing the trustworthiness of medical image super-resolution. The qualitative visual quality in Figure 9 further illustrates this point.

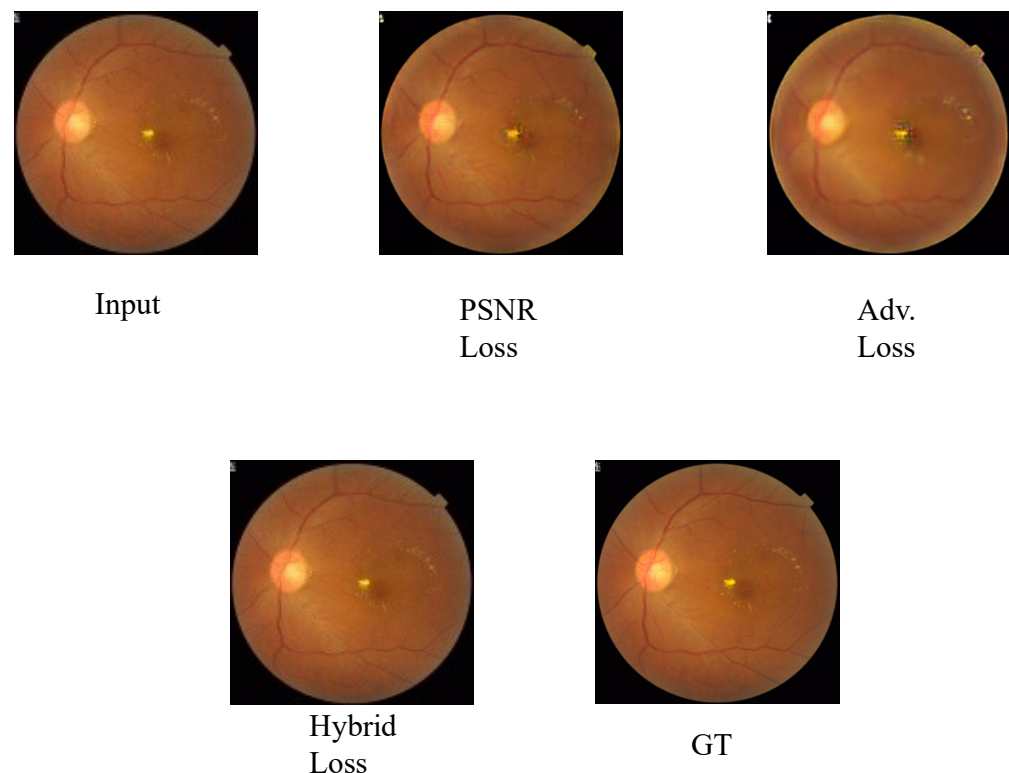


Figure 9. The result of the ablation experiment.

The evaluations of different ablations are summarized in Table 4, indicating that the proposed method can take into account both visual quality and reliability.

Table 4. Ablation study configuration: whether each component is enabled. ✓ indicates that the corresponding method achieves the listed characteristic.

Method	High Visual Quality	High Evaluation Metrics	Trustworthiness
PSNR Loss		✓	
Adv. Loss	✓		
Our Method	✓	✓	✓

4.7. Effect on High-Level Task

We give the practical auxiliary diagnosis application to demonstrate the effectiveness and credibility of the proposed method. The results show that the proposed method can effectively support the application of classification. The FIVES dataset [31] is naturally classified into four balanced categories based on different pathologies. We take all the super-resolution datasets formed after downsampling as the classification dataset. We employ a ResNet-50 network as the backbone and train it for 100 epochs for classification. Table 5 summarizes the accuracy comparison, indicating that our method can effectively assist in automatic classification diagnosis.

Table 5. Accuracy comparison. The outputs of all models are interpolated to 512×512 using bilinear interpolation as classification data. ↑ indicates higher values are better.

Input/Reconstruction Pipeline	Accuracy [%] ↑
GT	94.2
LR ($\times 8$)	86.5
Bicubic SR	88.3
SRCNN SR [4]	90.1
PD-CR SR [7]	91.3
RAT SR [11]	92.6
Ours (Global-token U-Net + hybrid)	93.0

Segmentation is another important high-level task for clinical diagnosis. We construct a segmentation dataset using the super-resolved images generated by different methods, paired with the corresponding vessel masks provided in the FIVES dataset. The dataset is split into training and testing sets with a ratio of 0.8:0.2. A classical U-Net is trained for 100 epochs under the same settings for all methods. The quantitative and qualitative segmentation results are presented in Table 6 and Figure 10, respectively. The results demonstrate that the proposed method achieves better downstream performance, indicating stronger compatibility with high-level tasks and further supporting its credibility.

Table 6. Vessel segmentation performance on FIVES using images reconstructed by different SR pipelines. All values are reported as mean $\pm$ std (sample size is omitted). Higher is better for IoU/Dice/Recall/Precision/Specificity, and lower is better for Hd/Hd95.

Method	IoU	Dice	Recall	Precision	Specificity	Hd	Hd95
InverseSR [10]	0.8437 $\pm$ 0.0910	0.9113 $\pm$ 0.0825	0.8908 $\pm$ 0.1150	0.9222 $\pm$ 0.1043	0.9951 $\pm$ 0.0017	45.4797 $\pm$ 27.8489	1.7270 $\pm$ 4.3492
TAT [26]	0.8502 $\pm$ 0.0645	0.9176 $\pm$ 0.0435	0.8911 $\pm$ 0.1150	0.9227 $\pm$ 0.1044	0.9951 $\pm$ 0.0017	45.3589 $\pm$ 26.8871	1.7248 $\pm$ 4.0370
SRCNN SR [4]	0.8515 $\pm$ 0.0645	0.9183 $\pm$ 0.0434	0.8917 $\pm$ 0.1151	0.9235 $\pm$ 0.1045	0.9952 $\pm$ 0.0017	44.4648 $\pm$ 26.4473	1.7536 $\pm$ 4.4060
PD-CR SR [7]	0.8515 $\pm$ 0.0644	0.9183 $\pm$ 0.0433	0.8918 $\pm$ 0.1151	0.9235 $\pm$ 0.1044	0.9952 $\pm$ 0.0017	43.4669 $\pm$ 26.1811	1.7403 $\pm$ 4.3984
Ours (Global-token U-Net + hybrid)	0.8524 $\pm$ 0.0644	0.9188 $\pm$ 0.0433	0.8922 $\pm$ 0.1151	0.9241 $\pm$ 0.1045	0.9952 $\pm$ 0.0017	43.4562 $\pm$ 26.0719	1.7483 $\pm$ 4.4092
GT	0.8543 $\pm$ 0.0645	0.9200 $\pm$ 0.0433	0.8931 $\pm$ 0.1152	0.9255 $\pm$ 0.1046	0.9953 $\pm$ 0.0017	42.2031 $\pm$ 23.8460	1.7845 $\pm$ 4.5530

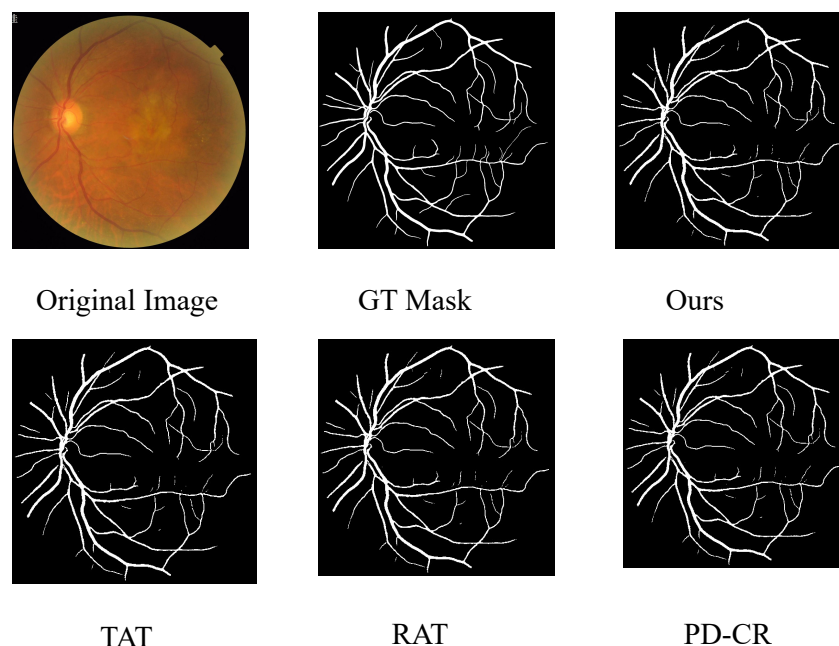


Figure 10. Vessel segmentation performance on FIVES using images reconstructed by different SR pipelines.

5. Conclusions

To address the trustworthiness issue of medical image super-resolution, we propose a comprehensive method. This method uses global-token U-Net as the backbone network, adds a lightweight VGG as the discriminator, and employs a hybrid PSNR and adversarial loss. Empirical results show that the proposed method can balance visual quality and reliability and ensure fidelity to the original details and is more suitable for auxiliary diagnosis. Future work can explore the role of other loss functions such as perceived loss. Some of the theoretical analyses of the methods presented in this paper still require exploration in future research. The validity of 3D imaging or other modalities still needs to be verified through more universal experiments.

Author Contributions: Conceptualization, J.S. and Z.X.; methodology, J.S. and Z.X.; software, J.S. and Z.X.; validation, D.W.; formal analysis, D.W.; investigation, D.W.; resources, J.S. and Z.X.; data curation, J.S. and Z.X.; writing—original draft preparation, J.S. and Z.X.; writing—review and editing, D.W.; visualization, J.S. and Z.X.; supervision, D.W.; project administration, D.W.; funding acquisition, D.W., J.S. and Z.X. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant data are included in the article. Additional data are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, G.; Xing, W.; Zhao, L.; Lan, Z.; Sun, J.; Zhang, Z.; Zhang, Q.; Lin, H.; Lin, Z. Self-Reference Image Super-Resolution via Pre-trained Diffusion Large Model and Window Adjustable Transformer. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), Ottawa, ON, Canada, 29 October–3 November 2023; pp. 7981–7992.

2. Han, L.; Zhang, X. Scalable Super-Resolution Neural Operator. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), Melbourne, Australia, 28 October–1 November 2024; pp. 10036–10045.
3. Chen, Z.; Guo, X.; Woo, P.Y.M.; Yuan, Y. Super-Resolution Enhanced Medical Image Diagnosis With Sample Affinity Interaction. *IEEE Trans. Med. Imaging* **2021**, *40*, 1377–1389. [[CrossRef](#)] [[PubMed](#)]
4. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
5. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
6. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 294–310.
7. Cho, H.; Shin, H.-K.; Jang, Y.; Ko, S.-J.; Jung, S.-W. PD-CR: Patch-Based Diffusion Using Constrained Refinement for Image Restoration. *IEEE Signal Process. Lett.* **2024**, *31*, 949–953. [[CrossRef](#)]
8. Song, J.; Sowmya, A.; Zhang, W.; Sun, C. Efficient Transformer with Compressed-Attention for Stereo Image Super-Resolution. *Knowl.-Based Syst.* **2025**, *331*, 114844. [[CrossRef](#)]
9. Umirzakova, S.; Ahmad, S.; Khan, L.U.; Whangbo, T. Medical Image Super-Resolution for Smart Healthcare Applications: A Comprehensive Survey. *Inf. Fusion* **2024**, *103*, 102075. [[CrossRef](#)]
10. Wang, J.; Levman, J.; Pinaya, W.H.L.; Tudosiu, P.-D.; Cardoso, M.J.; Marinescu, R. InverseSR: 3D Brain MRI Super-Resolution Using a Latent Diffusion Model. In *Proceedings of Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*; Springer: Cham, Switzerland, 2023; pp. 438–447.
11. Yang, Z.; Chen, H.; Qian, Z.; Zhou, Y.; Zhang, H.; Zhao, D.; Wei, B.; Xu, Y. Region Attention Transformer for Medical Image Restoration. In *Proceedings of Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*; Springer: Cham, Switzerland, 2024; pp. 603–613.
12. Yao, M.; Xu, R.; Guan, Y.; Huang, J.; Xiong, Z. Neural Degradation Representation Learning for All-in-One Image Restoration. *IEEE Trans. Image Process.* **2024**, *33*, 5408–5423. [[CrossRef](#)] [[PubMed](#)]
13. Yang, Z.; Huang, J.; Zhou, M.; Zheng, N.; Zhao, F. IRVR: A General Image Restoration Framework for Visual Recognition. *IEEE Trans. Multimed.* **2024**, *26*, 7012–7026. [[CrossRef](#)]
14. Xue, D.; Herranz, L.; Corral, J.V.; Zhang, Y. Burst Perception-Distortion Tradeoff: Analysis and Evaluation. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
15. Cohen, R.; Kligvasser, I.; Rivlin, E.; Freedman, D. Looks Too Good To Be True: An Information-Theoretic Analysis of Hallucinations in Generative Restoration Models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; pp. 22596–22623.
16. Bhadra, S.; Kelkar, V.A.; Brooks, F.J.; Anastasio, M.A. On Hallucinations in Tomographic Image Reconstruction. *IEEE Trans. Med. Imaging* **2021**, *40*, 3249–3260. [[CrossRef](#)] [[PubMed](#)]
17. Liu, Y.; Zhu, H.; Liu, M.; Yu, H.; Chen, Z.; Gao, J. Rolling-Unet: Revitalizing MLP's Ability to Efficiently Extract Long-Distance Dependencies for Medical Image Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 3819–3827.
18. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
19. Chen, L.; Lu, X.; Zhang, J.; Chu, X.; Chen, C. HINet: Half Instance Normalization Network for Image Restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2021; pp. 182–192.
20. Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; Bousquet, O. Are GANs Created Equal? A Large-Scale Study. *arXiv* **2018**, arXiv:1711.10337. [[CrossRef](#)]
21. Zhang, D.; Jodoin, P.-M.; Li, C.; Wu, Y.; Cai, G. Novel Graph Cuts Method for Multi-Frame Super-Resolution. *IEEE Signal Process. Lett.* **2015**, *22*, 2279–2283. [[CrossRef](#)]
22. Song, J.; Sowmya, A.; Sun, C. Efficient Frequency Feature Aggregation Transformer for Image Super-Resolution. *Pattern Recognit.* **2025**, *167*, 111735. [[CrossRef](#)]
23. Hu, Q.; Tang, Y.; Bu, C. Dilated Feature Distillation Attention Network for Efficient Image Super-Resolution. *Pattern Recognit.* **2025**, *168*, 111856. [[CrossRef](#)]
24. Xiao, Z.; Wen, X.; Tan, X.; Zhu, X.; Guo, K. Degradation-aware Graph Neural Network for Blind Super-Resolution. *Pattern Recognit.* **2026**, *174*, 113007. [[CrossRef](#)]
25. Ji, Z.; Zou, B.; Kui, X.; Vera, P.; Ruan, S. Self-prior Guided Mamba-UNet Networks for Medical Image Super-Resolution. In *Proceedings of Pattern Recognition*; Springer: Cham, Switzerland, 2025; pp. 160–174.

26. Yang, Z.; Zhang, J.; Yi, Y.; Liang, J.; Wei, B.; Xu, Y. TAT: Task-Adaptive Transformer for All-in-One Medical Image Restoration. In *Proceedings of Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*; Springer: Cham, Switzerland, 2026; pp. 565–575.
27. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating More Pixels in Image Super-Resolution Transformer. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 18–22 June 2023; pp. 22367–22377.
28. Zhi, S.; Wang, Y.; Xiao, H.; Bai, T.; Li, B.; Tang, Y.; Liu, C.; Li, W.; Li, T.; Ge, H.; et al. Coarse–Super-Resolution–Fine Network (CoSF-Net): A Unified End-to-End Neural Network for 4D-MRI With Simultaneous Motion Estimation and Super-Resolution. *IEEE Trans. Med. Imaging* **2024**, *43*, 162–174. [[CrossRef](#)]
29. Xu, Q.-H.; Li, B. Multiscale Fusion for Spatially Decoupled Multimodal MRI Super-Resolution Reconstruction. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–12. [[CrossRef](#)]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
31. Jin, K.; Huang, X.; Zhou, J.; Li, Y.; Yan, Y.; Sun, Y.; Zhang, Q.; Wang, Y.; Ye, J. FIVES: A Fundus Image Dataset for Artificial Intelligence Based Vessel Segmentation. *Sci. Data* **2022**, *9*, 475. [[CrossRef](#)]
32. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703. [[CrossRef](#)]
33. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**. [[CrossRef](#)]
34. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
35. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801. [[CrossRef](#)]
36. Duan, Z.-P.; Zhang, J.; Jin, X.; Zhang, Z.; Xiong, Z.; Zou, D.; Ren, J.S.; Guo, C.; Li, C. DiT4SR: Taming Diffusion Transformer for Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, HI, USA, 19–23 October 2025; pp. 18948–18958.
37. Kim, J.; Nang, J.; Choe, J. LMLT: Low-to-High Multi-Level Vision Transformer for Lightweight Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Honolulu, HI, USA, 19–23 October 2025; pp. 5509–5519.
38. Peking University. Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019). Available online: <https://odir2019.grand-challenge.org/> (accessed on 15 October 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.