remote sensing

MDPI

Article

TESR: Two-Stage Approach for Enhancement and Super-Resolution of Remote Sensing Images

Anas M. Ali [1,2,†], Bilel Benjdira [1,3,*,†], Anis Koubaa [1], Wadii Boulila [1,4] and Walid El-Shafai [2,5]

[1] Robotics and Internet-of-Things Laboratory, Prince Sultan University, Riyadh 12435, Saudi Arabia; aaboessa@psu.edu.sa (A.M.A.); akoubaa@psu.edu.sa (A.K.); wboulila@psu.edu.sa (W.B.)
[2] Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt; welshafai@psu.edu.sa
[3] SE & ICT Laboratory, LR18ES44, ENICarthage, University of Carthage, Tunis 1054, Tunisia
[4] RIADI Laboratory, University of Manouba, Manouba 2010, Tunisia
[5] Security Engineering Laboratory, Computer Science Department, Prince Sultan University, Riyadh 11586, Saudi Arabia
* Correspondence: bbenjdira@psu.edu.sa
† These authors contributed equally to this work.

Abstract: Remote Sensing (RS) images are usually captured at resolutions lower than those required. Deep Learning (DL)-based super-resolution (SR) architectures are typically used to increase the resolution artificially. In this study, we designed a new architecture called TESR (Two-stage approach for Enhancement and super-resolution), leveraging the power of Vision Transformers (ViT) and the Diffusion Model (DM) to increase the resolution of RS images artificially. The first stage is the ViT-based model, which serves to increase resolution. The second stage is an iterative DM pre-trained on a larger dataset, which serves to increase image quality. Every stage is trained separately on the given task using a separate dataset. The self-attention mechanism of the ViT helps the first stage generate global and contextual details. The iterative Diffusion Model helps the second stage enhance the image's quality and generate consistent and harmonic fine details. We found that TESR outperforms state-of-the-art architectures on super-resolution of remote sensing images on the UCMerced benchmark dataset. Considering the PSNR/SSIM metrics, TESR improves SR image quality as compared to state-of-the-art techniques from 34.03/0.9301 to 35.367/0.9449 in the scale $\times 2$. On a scale of $\times 3$, it improves from 29.92/0.8408 to 32.311/0.91143. On a scale of $\times 4$, it improves from 27.77/0.7630 to 31.951/0.90456. We also found that the Charbonnier loss outperformed other loss functions in the training of both stages of TESR. The improvement was by a margin of 21.5%/14.3%, in the PSNR/SSIM, respectively. The source code of TESR is open to the community.

Keywords: super-resolution; remote sensing images; vision transformer; self-attention; diffusion model

Citation: Ali, A.M.; Benjdira, B.; Koubaa, A.; Boulila, W.; El-Shafai, W. TESR: Two-Stage Approach for Enhancement and Super-Resolution of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 2346. https://doi.org/10.3390/rs15092346

Academic Editors: Mohib Ullah, Sultan Daud Khan and Habib Ullah

Received: 19 February 2023
Revised: 14 April 2023
Accepted: 16 April 2023
Published: 29 April 2023

1. Introduction

RS images obtained from unmanned aerial vehicles (UAVs) and satellites provide valuable insights into the Earth's surface. However, these images often have low–medium quality, which has prompted the development of algorithms to enhance their quality. The use of image enhancement algorithms on drones and satellites allows for the collection of more data over longer distances. RS images have been utilized for various applications, including updating maps [1], target detection [2–4], semantic segmentation [5], and seismic performance evaluation [6].

Obtaining high-resolution images in aerial photography is a significant challenge due to the limitations of imaging equipment [7]. Capturing high-resolution RS photographs by drones and satellites often requires expensive and high-quality imaging equipment. To address this issue, researchers have turned to SR technologies. SR techniques for RS images can be divided into two categories: one in which RS images are processed and

improved directly after being taken from the drone or satellite, but suffer from the use of additional processing devices. Another approach is to take RS images and then send them to a workstation for processing and optimization, but this suffers from communication channel issues. A major challenge in the field of super-resolution for RS images is the absence of clear and precise details in the reconstructed images.

A variety of SR algorithms exist, with traditional methods such as interpolation being commonly used. However, interpolation solely relies on pixel information [8] and often yields suboptimal results. Since the emergence of DL, it has been utilized extensively across various fields [9–11]. Several SR algorithms that surpass traditional methods have been proposed, including those in [12–15]. SR algorithms have greatly advanced with the advent of DL techniques such as Convolutional Neural Networks (CNNs) and the Generative Adversarial Networks (GAN) model [16,17]. GANs consist of a generative network and a discriminant network. GAN models have also been applied to various tasks such as image enhancement and disease diagnosis [18,19]. Despite the success of GANs, current GAN-based SR methods often struggle to produce high-quality results when applied to RS images due to issues such as pattern collapse, excessive smoothing, and artifacts. These challenges are a result of the nature of RS images, which often lack texture details.

The diffusion model [20] was first introduced a decade ago but has not yet been applied to practical use. Recently, the Denoising Diffusion Probabilistic Model (DDPM) [21] was published and has demonstrated superior performance in tasks such as SR [22], image generation [23,24], repair [25], segmentation [26,27], and deblurring [28]. Although several generative models can be employed for SR, they may not be suitable for RS images due to the requirement of accurately capturing fine details. Therefore, it is crucial to develop a system that can effectively extract and reconstruct both global and fine details in RS images.

To achieve high-quality, detailed RS images, we propose a two-stage super-resolution algorithm called TESR. Our algorithm combines the advantages of both micro- and macro-feature clustering by integrating the ViT model with the DM. The ViT is efficient at producing high-resolution images and capturing global details with its self-attention mechanism, but it needs a large dataset and cannot extract fine details from RS images. On the other hand, the DM excels at creating images that are similar to the original with fine details but struggles to capture global details and takes a long time to train. To address these issues, our proposed TESR model combines the strengths of the transformer and DM while eliminating their drawbacks. This study utilizes the UCMerced benchmark dataset [29] to evaluate the performance of the proposed model. We employed deep tuning to the pre-trained swinIR ViT model and iterative DM. To decrease diversity and randomness during training, we applied the Charbonnier loss function to both the diffusion and transformer models. Furthermore, we compared the results of our proposed model with the state-of-the-art super-resolution models in RS images. The main contributions of this research are:

1. The development of a two-stage TESR architecture for enhancement and super-resolution using a combination of the Vision Transformer and Diffusion models;
2. Proving that the Vision Transformer block was more appropriate for the super-resolution stage (generation of global details), while the Diffusion Model was more appropriate for the Enhancement stage (enhancement of fine details);
3. Outperforming other methods when tested on the super-resolution of RS images from the UCMerced dataset;
4. Demonstrating the efficiency of using the Charbonnier loss in the training of the TESR model, which emphasizes its usefulness in the super-resolution domain.

The paper is organized as follows. In Section 2, we provide an overview of previous studies in the field of super-resolution and their limitations, as well as descriptions of the vision transformer model and the diffusion model. In Section 3, we describe the TESR algorithm and its components (i.e., the SiwnIR transformer model and the iterative diffusion model), as well as the deep tuning and Charbonnier loss applied to the pre-trained models. In Section 4, we evaluate the proposed model's performance using various metrics (e.g., Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Multi-Scale

SSIM (MS-SSIM)) and compare it to the performance of previous studies. In Section 5, we discuss future work and the limitations of the proposed model. Finally, we provide our conclusions in Section 6.

2. Related Work

Based on the study of related work, the target issue of super-resolution of RS images has been found to be widely addressed using various architectures, including CNN, GAN, ViT, and DM. ViTs have been found to be particularly adept at extracting global features, which may be a challenge for CNNs. This is due to the utilization of self-attention mechanisms and the use of large-scale datasets. Self-attention mechanisms are a modern method for global feature extraction that is flexible in adapting to different data distributions. However, it is important to note that while ViTs have the potential to provide improved results, they also require a significant amount of training time and access to large datasets. Additionally, it is worth noting that ViTs may not be able to extract fine details from RS images as effectively. The DM is a widely recognized generative model that has been shown to outperform other methods. However, it has also been noted that this model may struggle to generate images that possess accurate global features. Additionally, it is important to note that the diffusion model requires a large amount of training time and can be computationally expensive.

In the field of remote sensing image super-resolution, Wan et al. [30] proposed a network that addresses this problem by utilizing an optical flow residual. This network is composed of three branches: a motion compensation unit, an optical flow residual estimation unit, and a super-resolution fusion unit. The method first extracts features from the input image using a sub-pixel convolution layer before applying the super-resolution process. González et al. [31] proposed a system for capturing low-resolution images using drones, which are then sent to a smartphone for conversion to high-resolution images using a lightweight SR architecture. Additionally, in ref. [32], Zhang et al. proposed a new mixed attention-based network called MHAN, which consists of two feature extraction and refinement networks. The extraction network uses weighted channel-wise concatenation and skip connections to gather the most information, while the refinement network is based on the High-Order Attention (HOA) mechanism to restore missing details.

Guo et al. [33] proposed a novel dense generative adversarial network called NDSR-GAN for the SR reconstruction of real aerial imagery. The generator of this network is composed of residual dense multi-level blocks that are connected by dense connections. In addition, the network uses a matrix mean discriminator to speed up training and achieve optimal results. The smooth l_1 loss is also utilized in this process. In contrast, Xiao et al. [34] proposed a two-stage image quality improvement model that first employs super-resolution using SRGAN, followed by correction and deblurring using a UNet-GAN model. Similarly, Li et al. [35] suggested a super-resolution model based on a GAN to improve UAV detection. To further enhance the edges of images and reduce weaknesses caused by convolutional layers during training, they also used an ROI extraction model. Their model incorporates edges and salient features into the convolution process through a feedback mechanism.

On the other hand, some articles have used ViT for SR on RS images. Lei et al. [36] proposed a SR framework called TransENet, which is based on a ViT. The core of this model is designed to exploit features at multiple levels. TransENet has a multi-stage architecture that can be used in conjunction with traditional SR frameworks to exploit both high- and low-frequency features across multiple bands. Its architecture includes multiple encoders to embed multi-level features and decoders to combine these features. In contrast, some authors have combined ViT and GAN models, such as Tu et al. [37], who introduced a new network called SWCGAN that combines a swin transformer [38] with convolutional layers within a GAN. This model initially uses convolution layers to extract shallow features and can adapt to different image sizes. It then uses residual swin transformer blocks to extract deep features, which are used to generate high-resolution UAV images through

enlargement. The lightweight swin transformer serves as the discriminator model for adversarial training. Similarly, there are those who have used DM for super-resolution of remote sensing images. For instance, Liu et al. [39] proposed a generative diffusion model called DMDC, which includes a complement of details. They used the structure of the SR3 model [22] to reconstruct images of generic scenes, but recognized that remote sensing images often lack detailed information. To address this issue, they proposed a detailed supplement task, which involves adding small, random black patches to the images. They also introduced a novel loss function to improve the direction of inverse diffusion in the diffusion model. There is also some research highlighting the importance of image optimization in remote sensing applications. Xueyang et al. [40] proposed a two-step approach to improving the quality of remote sensing images by using regularized histogram equalization (HE) and the discrete cosine transform (DCT). The method first improves the global contrast of the image through the use of a sigmoid function and histogram to generate a distribution function, and then enhances the local details through the adjustment of DCT coefficients. On the other hand, Ablin et al. [41] conducted a survey of existing enhancement techniques in satellite images and concluded that fusion-based enhancement techniques perform better than non-fusion-based techniques. These studies demonstrate the significance of image enhancement in remote sensing and the need for effective methods to improve the visual quality of remotely sensed images.

From the above, it can be seen that combining ViT and DM in one architecture is particularly useful for the ability to reconstruct both global and contextual details through the self-attention-based mechanism of the ViT, while DM is more suitable for reconstructing harmonic contextual and fine details. In this research, we are the first to combine ViT and DM in the task of remote sensing image super-resolution, and we apply a transfer learning approach by using models (ViT and DM) that are pre-trained on large datasets.

Many previous and current methods rely on the availability of large datasets and long training times to produce high-resolution remote-sensing images with fine details. However, these methods often fail to accurately capture the perceptual qualities of the original images. In this research, we propose an integrated algorithm that uses transfer learning with the vision transformer and diffusion model to reduce computational cost and training time while still producing high-quality remote-sensing images using a small dataset.

3. Proposed Methodology

The TESR methodology is based on different stages, as shown in Figure 1. In the first stage (the super-resolution stage), the SwinIR model is introduced for the super-resolution of RS images. The SwinIR model is one of the most popular ViT models in generic image restoration, which is applied to enlarge the remote sensing images and restore global details. In the second stage (enhancement stage), the DM is utilized, which is typically used for image generation and noise removal. The DM is applied after the SwinIR model to enhance and restore fine details of remote sensing images. TESR combines the advantages of both micro- and macro-feature clustering by integrating the SwinIR model with the DM. These stages are explained in detail below.

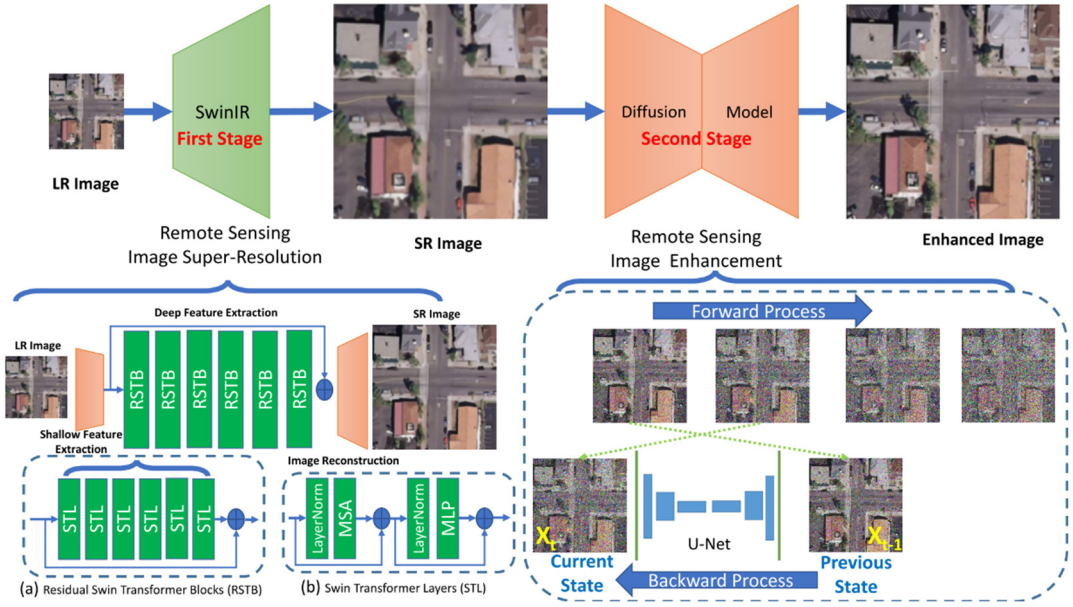

Figure 1. Block diagram for the proposed algorithm (TESR).

3.1. TESR Architecture

The frameworks for super-resolution algorithms can be divided into two categories in the literature: pre-upsampling and post-upsampling.

Pre-upsampling, as shown in Figure 2a, was initially used in deep learning-based super-resolution algorithms. In this approach, the bicubic interpolation method is first used to enlarge the low-resolution (LR) images to the same size as the original high-resolution (HR) images. The super-resolution (SR) model is then used to restore the HR images from the interpolated LR images. As a result, the SR model learns the non-linear feature mapping easily between the interpolated LR images and the original HR images. However, this approach suffers from high computational costs due to the need to perform feature extraction operations on high-dimensional image sizes.

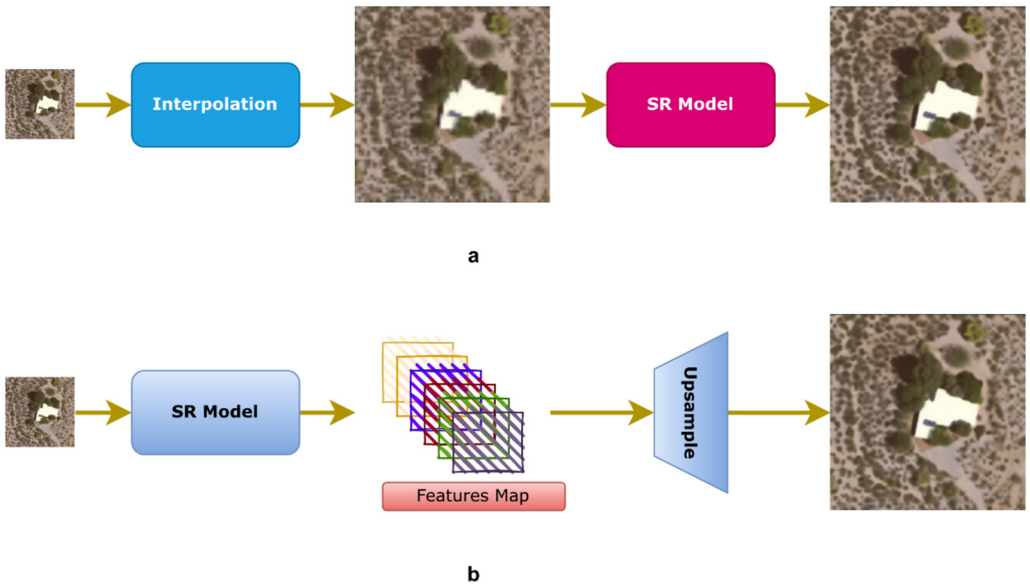

Figure 2. Illustration of different SR frameworks. (**a**) Pre-upsample framework. (**b**) Post-upsample framework.

In contrast, post-upsampling, as shown in Figure 2b, is more common because it can reduce computational costs to a minimum. In this approach, feature extraction operations are performed on low-dimensional image space, and learnable enlarged layers such as deconvolution [42] and subpixel convolution [43] are used at the end of the model. While this architecture significantly speeds up the feed-forward process, it suffers from the fact that the HR image is restored immediately after the upsampling process without any enhancements, which makes training difficult and limits the accuracy of the reconstruction. In order to achieve the best perceptual quality with the lowest computational cost, we propose a model consisting of two stages in this paper, as shown in Figure 1. The first stage enlarges the input image and extracts global details by using the ViT model, while the second stage uses DM to deeply extract fine details and generate images that are highly similar to the original images while reducing computational cost and training time.

3.1.1. Super-Resolution Stage

Vision transformers have two merits over other deep learning models. The first merit is the mechanism of self-attention, and the second is self-pre-training. The self-attention mechanism is the process of estimating the weight of an element in relation to the rest of the other elements in the same space. The images are divided into several patches, and then each patch is converted into a sequence. Then self-attention evaluates the weight of a given sequence over the rest of the sequences via dot-product. Therefore, vision transformers can extract global features from all over the image space. The SwinIR model is considered the best vision transformer in terms of its stability and learning speed. SwinIR consists of three stages, as shown in Figure 3. The first stage is called shallow feature extraction and consists of a single convolutional layer. The second stage is deep feature extraction, which consists of six Residual Swin Transformer Blocks (RSTB) and each block consists of six Swin Transformer Layers (STL). Swin Layer is a quantum leap in vision transformers due to the reduction in computational cost resulting from the use of Window Multi-head Self-Attention (W-MSA). The window technique is as simple as dividing the image into a set of patches and calculating attention only within each window. The third stage is to reconstruct the images by accumulating and merging the deep and shallow features. We can express the SwinIR model mathematically by:

$$f_{sf} = h_{sf}(i_{LR}), \tag{1}$$

where f_{sf} is the output features from the shallow features stage, h_{sf} is the transfer function of the shallow features stage, and i_{LR} is the input low-resolution image. Further, we can express the deep features stage through:

$$f_{df} = h_{df}\left(f_{sf}\right), \tag{2}$$

where f_{df} is the output features from the deep features stage and h_{df} is the transfer function of the deep features stage. Furthermore, we can express the RSTB in the deep features stage through:

$$f_{i,j} = h_{swin\ i,j}\left(f_{i,j-1}\right), \tag{3}$$

where $f_{i,j}$ is the output feature from each RSTB block, $h_{swin\ i,j}$ is the transfer function of the swin layer, i is the number of RSTB blocks, and $j = 1, 2, \ldots, L$. L is the number of swin layers. The self-attention mechanism is based on training three learnable weight matrices. The input image is converted into three similar projection matrices. The weight matrices are multiplied by the projection matrices to form the query, key, and value matrices.

$$q = x.p_q,\ k = x.p_k,\ v = x.p_v, \tag{4}$$

where q, k, and v are query, key, and value matrices, respectively. Then the query is multiplied by the key to calculate the SoftMax for the output. We can express the whole process by:

$$Attention\ map(q,k,v) = softmax\left(\frac{qk^T}{\sqrt{d_q}+B}\right)V, \tag{5}$$

where d_q is the dimension of the query vector, and k^T is the transpose key matrix. B is the learnable positional encoding. Attention in each window is calculated individually and then aggregated. SwinIR is distinguished from other ViTs by its rapid adaptation to different datasets and attention to global details. However, it suffers from negligence in estimating the fine details in RS images. Therefore, to address the drawbacks of the SwinIR transformer with RS images, we propose to use iterative DM to generate high-resolution RS images with fine details in the next sub-section.

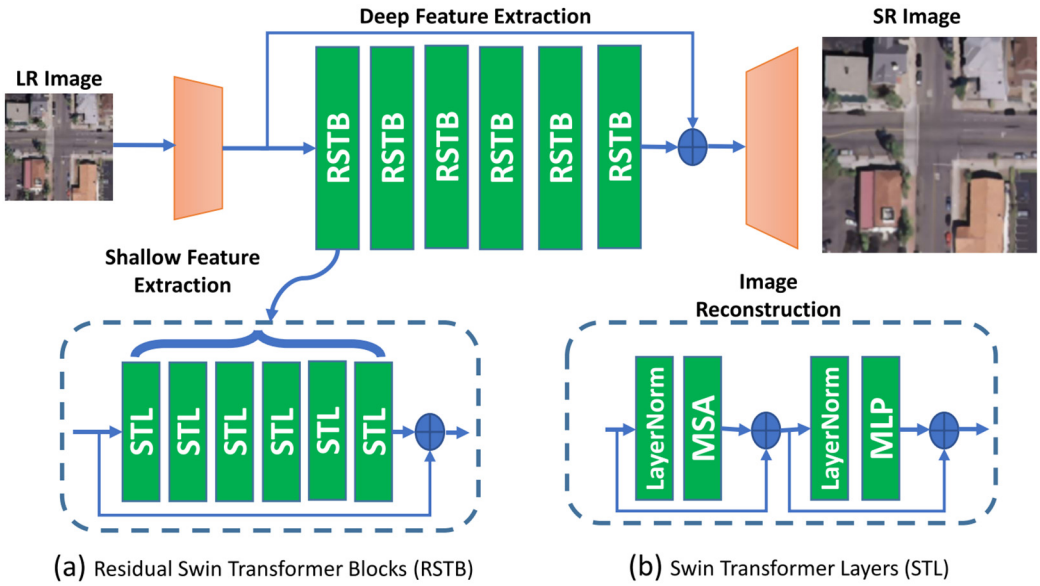

Figure 3. Block diagram for the SwinIR ViT model.

3.1.2. Enhancement Stage

The Diffusion Model consists of two processes, the first process is called the forward process and the second process is called the backward (reverse) process. Noise is added in the forward process, and the noise in the backward process is removed by the U-Net model, as shown in Figure 4. Diffusion models destroy some input, such as an image, by gradually adding Gaussian noise. The input is then restored from the noise in a reverse process also called denoising. This process is also called a Markov chain because it is a sequence of stochastic events where each time step depends on the previous time step. The forward process is fairly easy because it requires no training. All this process does is gradually add noise to the images. The forward process is referred to as a Markov chain q. The noise is added in sequential steps t to get the noisy samples. The prediction of the noise density at a given time t depends on the noise density at time $t-1$, so the noise conditional probability density can be represented as follows:

$$q(x_t/x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \tag{6}$$

where $\sqrt{1-\beta_t}x_{t-1}$ is the mean of the normal distribution, $\beta_t I$ is the variance of the normal distribution, x_t is the noisy output image, x_{t-1} is the previous less noisy image, and $\mathcal{N}$ is the normal distribution. The sequence of $\sqrt{1-\beta_t}x_{t-1}$ is a so-called variance schedule that describes how much noise we want to add in each of the time steps, which indicates that the amount of noise that is added is constant at each time step. The dilemma

in the forward-process stage is determining the noise value at a specific time step or determining the signal value at a specific time step. This dilemma is significant in the backward processing stage. So, we can define a signal at any time step according to the reparameterization technique [44] by defining $\alpha_t = 1 - \beta_t$, and the cumulative products $\overline{\alpha_t}$ of all α as $\overline{\alpha_t} = \prod_{s=0}^{t} \alpha_s$, where α_t is a signal at a certain time step. The reformulation of the above equation then becomes:

$$q(x_t/x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, (\overline{\alpha_t} - 1)I\right) \tag{7}$$

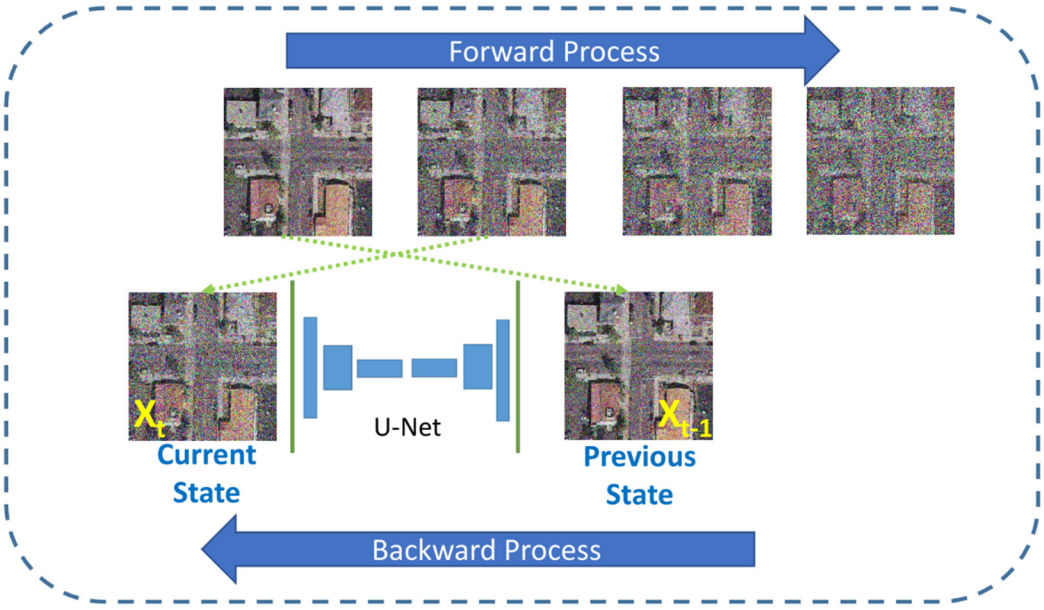

Figure 4. The main concept behind iterative DM.

In the backward process, we use U-Net with the self-attention mechanism that was referred to in the previous sub-section. U-Net is a special neural network that has a structure that is similar to the one used in an autoencoder. U-Nets are a popular model for image segmentation, and their output has the same shape as the inputs, the input passes a series of convolutional and down-sampling layers. Then input passes through a bottleneck that contains self-attention and residual connections. Finally, input passes through up-sampling layers to reach the same dimensions. The probability density of a sample at a specific time step is estimated based on its sample at the previous time step. This means reshaping for $q(x_{t-1}/x_t)$. Therefore, a noisy image $t-1$ is input into the U-Net entry, and the previous image t is used as a target. The previous state is estimated from the current state by knowing the previous gradients, which require a trainable model, which is U-Net. So, an estimate of a past state from a current state can be defined by $p_\theta(x_{t-1}/x_t)$ and is as follows:

$$p_\theta(x_{t-1}/x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)\right), \tag{8}$$

the backward process formula for all timesteps is as follows:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}/x_t), \tag{9}$$

where p_θ means neural network model or backward process. The purpose of this formula is to train the model to predict Gaussian parameters (the mean $\mu_\theta(x_t, t)$ and the covariance

matrix $\sum_\theta(x_t, t)$) for each timestep. Therefore, the mean $\mu_\theta(x_t, t)$ can be expressed in the formula from Ho et al. [21]:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}}\varepsilon_\theta(x_t, t)\right), \tag{10}$$

where $\varepsilon_\theta(x_t, t)$ is an approximate function used for predicting ε from x_t. So, the backward process uses a neural network to predict the added noise from a previous state based on the current state, as shown in Figure 4. Then the predicted noise is subtracted from the current state to produce a noise-free state.

3.2. The Training Algorithm of TESR

To train TESR, we follow the three steps, including the preprocessing phase, the pre-train phase, and the transfer learning phase, as shown in Figure 5. In the preprocessing phase, we used the UCMerced dataset [29], which contains 21 categories of remote sensing images. Moreover, the interpolation technique was applied to resize the RS images to match the size of the input images and the size of the target images to train the proposed TESR model on several different scales.

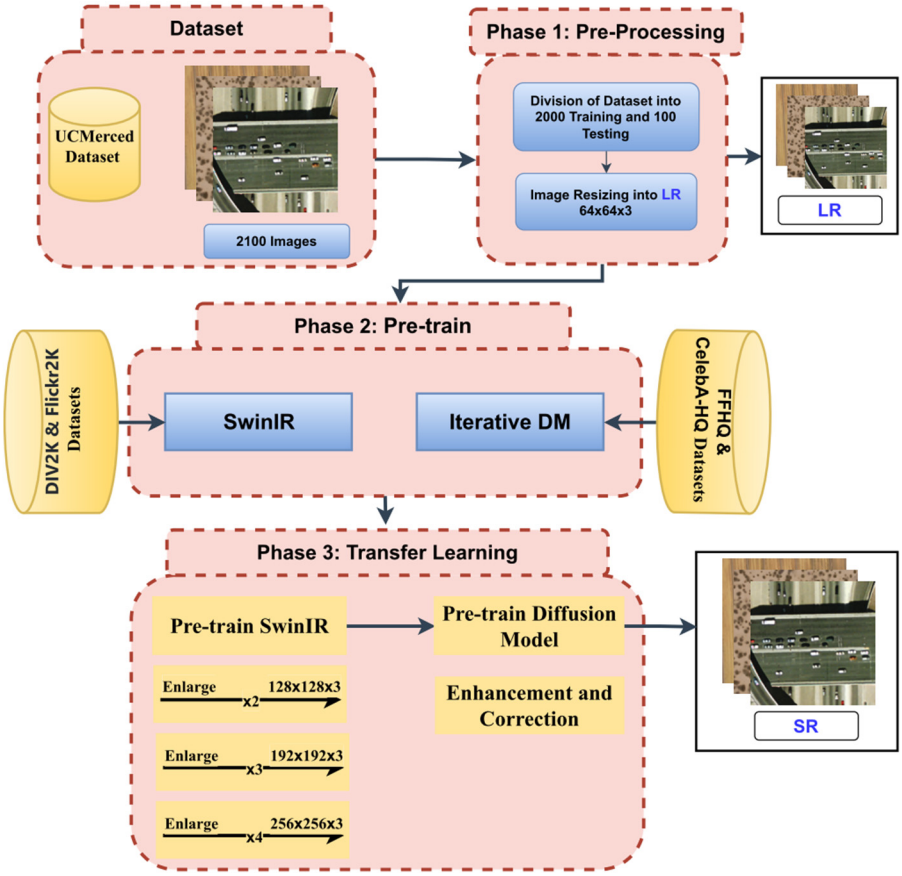

Figure 5. The three main steps to train the TESR architecture.

Transfer learning is a relatively recent strategy used for the first time on CNN models, which means transferring the knowledge and features learned by the deep learning model from a specific task and then using the trained parameters on another task or another dataset. The transfer learning strategy saves training time and is adaptable to most tasks and datasets. RS images contain fine details that are difficult to recover using conventional CNN models. Moreover, training a new model requires a huge number of RS images with high computational time and cost. In the pre-train phase, the SwinIR transformer was trained on two datasets: DIV2K and Flickr2K, which contain generic images, as well

as using pixel loss l_1. Iterative DM was also trained on two datasets, Flickr-Faces-HQ (FFHQ) [45] and CelebA-HQ [46], containing faces images and generic images, and using pixel loss l_1. In the last phase, we applied a deep tuning strategy to both SwinIR and iterative DM. Deep tuning refers to the process of applying a pre-trained deep neural network to a new task or dataset. This means using the weights of a pre-trained network as a starting point and then continuing to train the network for a new task or dataset. The main advantage of deep tuning is that it allows a network to take advantage of the knowledge and features learned during the pre-training process, which can greatly reduce the time and computational cost required to train the network from scratch. Furthermore, we have replaced the pixel loss, l_1 with Charbonnier loss, which performs better as demonstrated in the experimental section and can be expressed mathematically as follows:

$$\ell_{char}(i_p, i_{gt}) = \sqrt{\|i_r - i_o\|^2 + \varepsilon^2} \tag{11}$$

where i_p is the predicted image and i_{gt} is the ground truth image, and the constant ε in the experiments was set to 10^{-3} through empirical methods. The Charbonnier loss has been widely used in image reconstruction. The advantage of the Charbonnier loss is that it is less sensitive to outliers compared to the mean squared error (MSE) loss. The Charbonnier loss is a combination of the MSE loss and the mean absolute error (MAE) loss and therefore strikes a balance between being robust to outliers and having a good gradient for optimization. Additionally, it has a continuous and smooth gradient, which makes it easier to optimize using gradient-based optimization algorithms.

3.3. Description of the Training Algorithm

Having discussed the formulations and the training details of TESR, the steps of TESR are summarized in Algorithm 1.

Algorithm 1: TESR

1. **Procedure:**
   - **Input**: low-resolution image (*LR*), high-resolution image (*HR*), a pre-trained *SwinIR*, and Diffusion Model (*DM*);
   - **output**: Enhanced Super-Resolved Image (*ESRI*).
2. **Stage I. Image Super-Resolution:**
   - i. Load and initialize weights of the pre-trained *SwinIR* model;
   - ii. Add Charbonnier loss (*CL*) to the training loop;
   - iii. **For** each epoch **in** the range of (1, epochs), **do** the following:
     - a. Apply *LR* image to the pre-trained *SwinIR* model;
     - b. Generate SR image by *SwinIR* model;
     - c. Compare the generated image with the *HR* image using CL;
     - d. Adjust and optimize the weights of the pre-trained (deep tuning) *SwinIR* model;
   - iv. Repeat step (iii) for the different scale factors.
3. **Stage II. Image Enhancement:**
   - i. Generate FirstSR (*FSR*) images by the *SwinIR* model;
   - ii. Load and initialize weights of pre-trained *DM*;
   - iii. Add *CL* to the training loop;
   - iv. **For** each epoch **in** the range of (1, epochs), **do** the following:
     - a. Apply *FSR* image to pre-trained *DM*;
     - b. Generate *ESRI* image by *DM*;
     - c. Compare the generated *ESRI* with the *HR* image by using *CL*;
     - d. Adjust and optimize the weights of the pre-trained (deep tuning) *DM*;
   - v. Repeat step (iv) for the different scale factors.
4. **Return the *ESRI* images as the output of the algorithm.**

4. Results and Analysis

This section presents and discusses the experimental details and results of integrating the SwinIR model with the iterative diffusion model to enlarge and enhance the RS images from the UCMerced benchmark dataset. The UCMerced dataset is diverse, consisting of 21 classes and a total of 2100 images, which is considered a relatively small number. Its average resolution makes it a convenient dataset for conducting experiments and optimizing the proposed model. The source code of TESR is shared with the community at this link: https://github.com/AnasHXH/TESR.

4.1. Experimental and Analysis Details

The UCMerced dataset [29] was employed to train and evaluate the TESR model, which comprises 2100 images from 21 categories. The dataset was divided into 2000 images for training and 100 for testing. To reduce computational complexity, we applied normalization to all images so that they ranged from -1 to 1, and we also applied resizing to the dataset so that the input images had a size of $64 \times 64 \times 3$ and the target images had sizes of $[128 \times 128 \times 3, 192 \times 192 \times 3, 256 \times 256 \times 3]$. To further improve the performance, we applied a deep tuning strategy to the SwinIR transformer to enlarge the LR images. We used weights trained on two datasets: DIV2K, which contains 900 images, and Flickr2K, which contains 2,650 images, in a generic image super-resolution task. The deep tuning strategy was employed because the SwinIR model had not previously been applied to images containing many details in RS images. Furthermore, without deep tuning the model, we would lose the most important element that distinguishes SwinIR, which is the dynamic attention (window multi-head self-attention) mechanism.

The diffusion model is based on gradually adding noise in the forward process, then removing it through the use of a U-Net model with residual self-attention in the backward process. The U-Net model is trained to remove noise in small, varying amounts. Noise is added at many time steps until the image becomes pure noise. However, in the backward process, the U-Net model is trained for some, but not all, of the time steps in order to speed up the training process. The remote sensing image super-resolution problem depends on filling in and adjusting the features affected by the degradation process by adding noise and then removing it. There are several advantages to using the iterative scattering model for improving the quality of remote-sensing images:

1. Effectively recovering high-frequency (fine) details in the image, such as edges and texture;
2. Preserving structural information in the image, such as the overall shape and layout of objects;
3. Enhancing images with complex structures and noise as it adapts to the local characteristics of the image;
4. Handling images with missing or corrupted pixels by filling in missing data based on the surrounding pixels.

A deep tuning strategy was used by diffusion model weights that were trained on two Flickr-Faces-HQ (FFHQ) [43] and CelebA-HQ [44] datasets, taking 600,000 iterations to finish their training. We trained the diffusion model with 15,000 iterations on remote-sensing images in each scale factor ($\times 2$, $\times 3$, and $\times 4$). We also applied the weight-sharing strategy. The diffusion model is trained on a scale factor of $\times 2$, then the trained weights are used to train the model on a scale factor of $\times 3$, and so on. In the end, our proposed TESR model was able to extract both detailed features and global features, and TESR was able to generate remote sensing images that are very close to the original images after a relatively short training period, compared to the state-of-the-art methods.

The TESR model was trained and evaluated on an NVIDIA Quadro RTX 8000 GPU. The dimensions of the original images in the benchmark UCMerced dataset are 256×256. So, two copies of each image were created at each enlargement scale, one as an LR image and the other as an HR image. We cropped the images to 64×64 to be LR images. Images were cropped to 64×64 to serve as LR images. HR images were cropped from the original

images to 128 × 128, 192 × 192, and 256 × 256 for the ×2, ×3, and ×4 scales, respectively. In the SwinIR model, the Charbonnier loss function was used, which is particularly well-suited for reconstruction tasks. The SwinIR model was trained for 32,000 steps using a batch size of four. A fixed learning rate of 0.0002 was used with the Adam optimizer. To improve the durability and stability of the model, a series of operations were applied to the data augmentation, such as rotation, but at different angles, including 5, 10, and 15 degrees. In the iterative DM stage, we followed the model structure proposed in [22] by changing the loss function to Charbonnier loss. A 2000-time step was used. The iterative DM was trained on 20,000 iterations. The Adam optimizer was used with a fixed learning rate of 0.0001.

4.2. Performance Evaluation

Image quality assessment is the process of evaluating the visual quality of an image. It involves evaluating various aspects of the image, such as sharpness, noise, color accuracy, and overall visual appeal. There are several methods and metrics that can be used to assess the quality of an image. Some common methods include:

1. Human evaluation: this involves presenting the image to a panel of human judges, who then rate the image based on various subjective criteria;
2. Objective quality metrics: These are algorithms that analyze the image and calculate a score based on various objective criteria, such as sharpness, noise, and color accuracy. Some examples of objective quality metrics include the SSIM [47], the PSNR [48], and the MS-SSIM [49].

We used PSNR, SSIM, and MS-SSIM and the histogram to evaluate the proposed model, and the mathematical expressions are as follows:

$$SSIM(i_{gt}, i_p) = \frac{(2\mu_{i_{gt}}\mu_{i_p} + c_1)(2\sigma_{i_{gt}i_p} + c_2)}{(\mu_{i_{gt}}^2 + \mu_{i_p}^2 + c_1)(\sigma_{i_{gt}}^2 + \sigma_{i_p}^2 + c_2)}, \tag{12}$$

$$MSE(i_{gt}, i_p) = \frac{1}{t}\sum_{k=1}^{t}\left(i_{gt}(k) - i_p(k)\right)^2, \tag{13}$$

$$PSNR(i_{gt}, i_p) = 10\log_{10}\left(\frac{max^2}{MSE}\right), \tag{14}$$

$$MSSIM(i_{gt}, i_p) = \frac{1}{nm}\sum_{p=0}^{n-1}\sum_{j=0}^{m-1} SSIM(i_{gt}, i_p), \tag{15}$$

where i_p and i_{gt} are the predicted and ground truth images, μ_{i_p} and $\mu_{i_{gt}}$ are the local means of the images i_p and i_{gt}, respectively, σ_{i_p} and $\sigma_{i_{gt}}$ are the local standard deviations of the images i_p and i_{gt}, respectively, $\sigma_{i_{gt}i_p}$ is the local covariance of the images i_p and i_{gt}, c_1 and c_2 are constants used to stabilize the division with a weak denominator; max is the maximum possible pixel value of the image; t is the number of pixels of the image; and n and m are the number of rows and columns in the image, respectively. SSIM and MS-SSIM were used as evaluation indicators because they provide complementary information about image quality. SSIM provides a measure of the overall similarity between the original and reconstructed images, while MS-SSIM provides a measure of the similarity at multiple scales. A more complete understanding of the performance of our proposed method is obtained using both metrics.

4.3. Results Analysis

The goal of this paper was to improve the resolution of RS images at different scale factors by means of a TESR algorithm that combines the merits of SwinIR and iterative DM. To achieve this, the UCMerced dataset was divided into 2000 training images and 100 test images, and the images were normalized from −1 to 1 to reduce computational cost. We chose 100 images in the test set because of the prohibitively expensive computations

needed for testing the Diffusion Model in the Enhancement stage. The TESR algorithm was then applied to the images with scale factors of ×2, ×3, and ×4 to improve their resolution. The super-resolution images were compared with the ground truth images using measures of visual quality, including PSNR, SSIM, MS-SSIM, and histograms. Figure 6 shows a comparison of the visual results of each stage of the TESR model with the bicubic interpolation method.

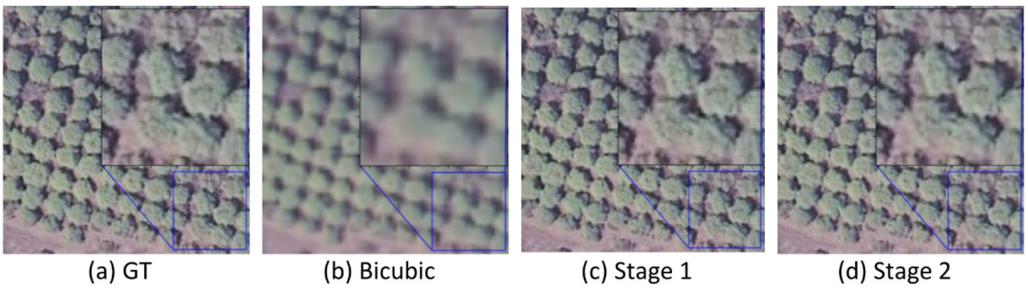

Figure 6. Result comparisons on the UCMerced dataset with different stages. (**a**) The ground-truth scene. (**b**) A bicubic interpolation scene with a ×4 factor. (**c**) SwinIR upsamples a scene with ×4 factor. (**d**) An iterative DM scene.

In this section, we will present the results obtained using our proposed model after testing it on four enlarging scales ×2, ×3, and ×4. Figure 7 shows ground-truth images, low-resolution images, and high-resolution images for each stage. Super-resolution images after the second stage (the Diffusion Model) have more details and improved visual quality compared to lower-resolution and ground-truth images. In addition, the improvement in resolution is particularly evident in fine details, such as lines of farmland and individual branches of trees. RS images are enlarged in ×2 scale from 64 × 64 to 128 × 128, in ×3 scale from 64 × 64 to 192 × 192, and in ×4 scale from 64 × 64 to 256 × 256.

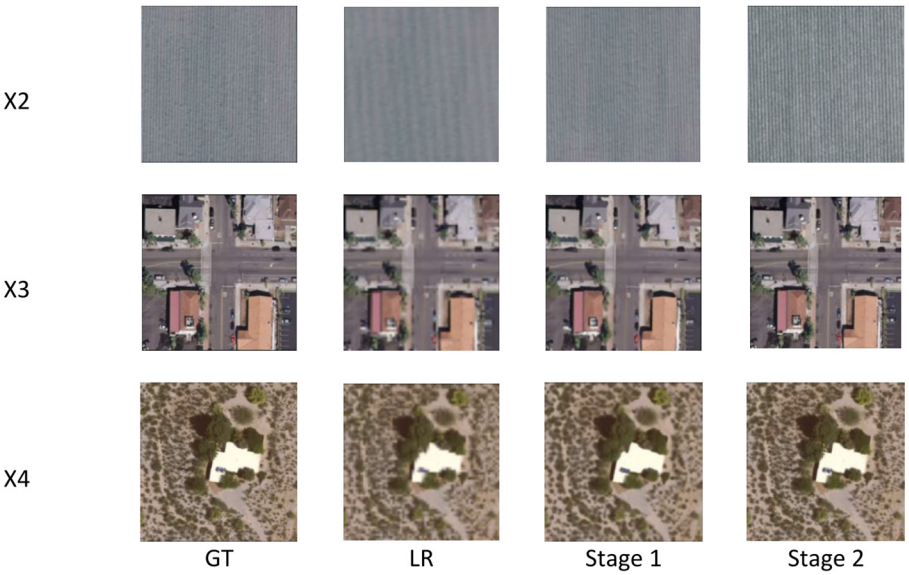

Figure 7. The sample of RS images before and after applying our proposed model includes ground-truth (GT), low-resolution (LR), and super-resolution images for each stage.

With the aim of demonstrating the efficacy of combining the SwinIR and DM models into a single algorithm, a separate evaluation of each model was performed after they were trained on the UCMerced dataset. As shown in Table 1, both models individually exhibited a high level of efficiency; however, neither of them was able to match the efficiency of the proposed TESR model at a scale of ×2.

Table 1. SwinIR and DM model evaluation separately after training on the UCMerced dataset.

Scale Factor	Deep-Tuning SwinIR		Deep-Tuning Iterative DM	
	PSNR	SSIM	PSNR	SSIM
×2	34.938	0.9232	30.256	0.90742

Table 2 shows the average values of the PSNR, SSIM, and MS-SSIM tests on 100 RS images at each up-sample scale factor. The results showed that the TESR algorithm was effective in improving the resolution of RS images, with an average PSNR improvement of 35.367 dB for ×2 scale factor, 32.311 dB for ×3 scale factor, and 31.951 dB for ×4 scale factor. The improvement in PSNR indicates that super-resolution images have much lower noise levels and higher signal-to-noise ratios than SOTA. The improvement in SSIM was similarly significant, ranging from a mean improvement of 0.9449, 0.91143, and 0.90456 for scale factors ×2, ×3, and ×4, respectively. The optimization in SSIM indicates that the super-resolution images have very close structural similarity to the high-resolution reference images. The improvement in MS-SSIM was significant in terms of its relative stability across different measurement factors.

Table 2. The evaluation of the TESR model on different scale factors.

Scale Factor	Deep-Tuning SwinIR Stage 1			Deep-Tuning Iterative DM Stage 2		
	PSNR	SSIM	MS-SSIM	PSNR	SSIM	MS-SSIM
×2	34.938	0.9232	0.9738	35.367	0.9449	0.9892
×3	30.813	0.8784	0.9385	32.311	0.91143	0.9731
×4	27.424	0.8201	0.9278	31.951	0.90456	0.9748

Figure 8 shows samples of iterative DM training images on RS images, showing forward-process and backward-process training steps. We can observe that DM is able to increase the details of the RS images from the features it learned in the first stage by adding and removing Gaussian noise at several different levels. The first image from the left (shown above) represents the output of the first stage (SwinIR). The last image (shown in the bottom right) represents the output of the second stage (DM). The images between the first and last images represent a sequence of adding noise to the images and then removing it in order to increase the fine details. Table 3 demonstrates the improved performance of the TESR algorithm compared to bicubic interpolation, SC [50], SRCNN [51], FSRCNN [52], LGCNet [53], DCM [54], DGANet-ISE [55], and TransENet [36] algorithms in terms of PSNR and SSIM metrics on the UCMerced dataset.

Table 3. The evaluation of the TESR model and SOTA, with the highest performance in bold type.

Model	×2	×3	×4
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	32.76/0.879	27.46/0.7631	25.65/0.6725
SC [50]	32.77/0.9166	28.26/0.7971	26.51/0.7152
SRCNN [51]	32.84/0.9152	28.66/0.8038	26.78/0.7219
FSRCNN [52]	33.18/0.9196	29.09/0.8167	26.93/0.7267
LGCNet [53]	33.48/0.9235	29.28/0.8238	27.02/0.7333
DCM [54]	33.65/0.9274	29.52/0.8394	27.22/0.7528
DGANet-ISE [55]	33.68/0.9344	-/-	27.31/0.7665
TransENet [36]	34.03/0.9301	29.92/0.8408	27.77/0.7630
TESR (our)	**35.367/0.9449**	**32.311/0.91143**	**31.951/0.90456**

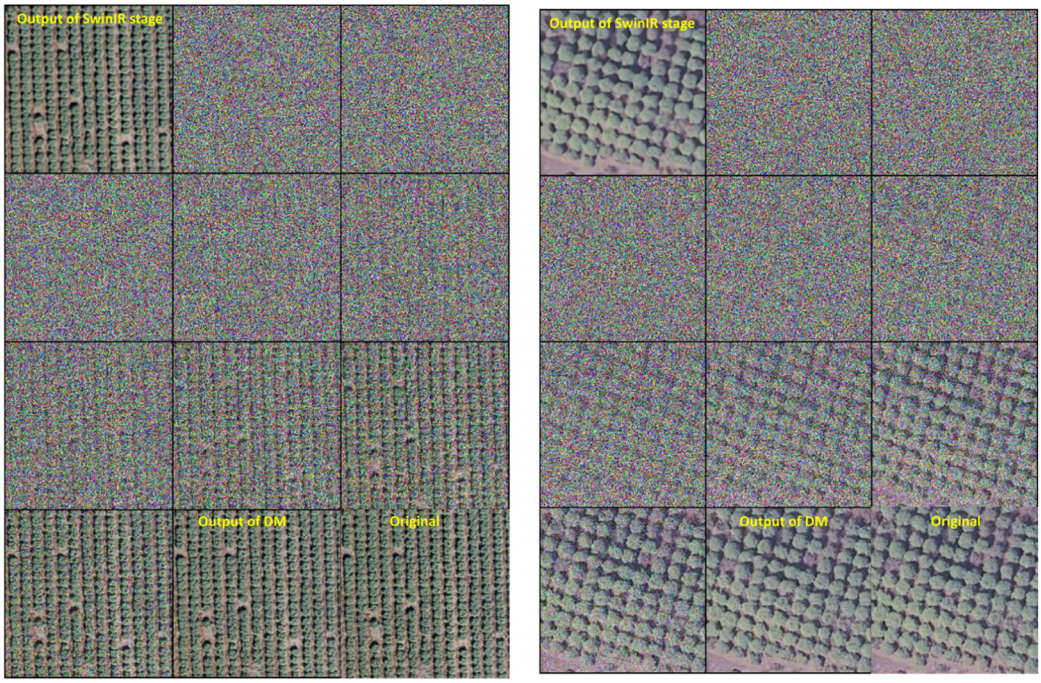

Figure 8. The sample of RS images after the second stage (iterative DM).

The histogram of the LR image showed a skewed distribution, with a majority of the pixels concentrated at low intensity values, as shown in Figure 9. This is indicative of poor image quality and a lack of detail in the shadows and highlights. After applying TESR, the histogram showed a more balanced distribution, with a wider range of intensity values represented. This indicates that TESR was able to recover more detail and improve the overall image quality. The increased contrast and dynamic range in the TESR-enhanced image are also apparent in the histogram, which shows a higher concentration of pixels at the extremes of the intensity range. Overall, the histogram results demonstrate the effectiveness of TESR in improving the quality and detail of the original image.

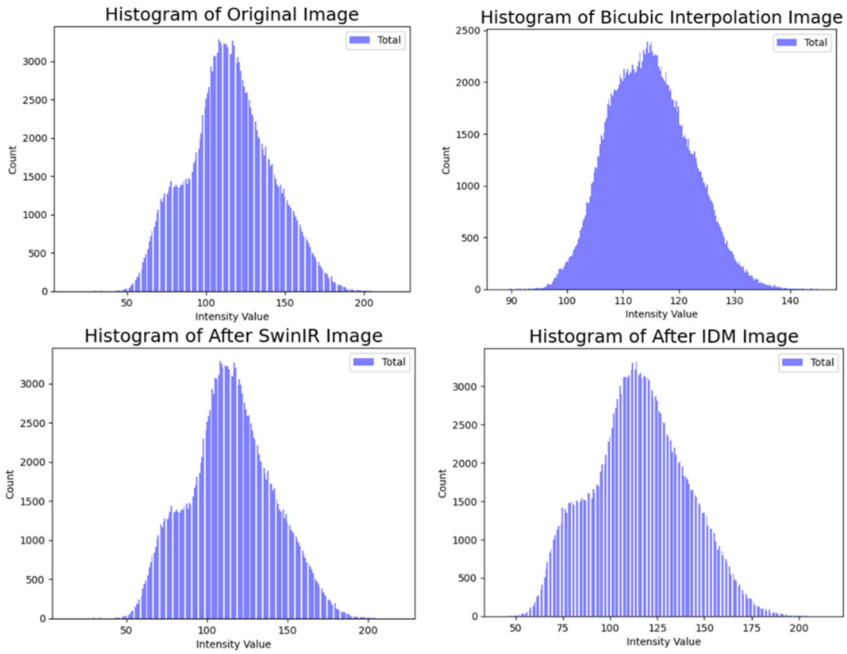

Figure 9. Illustration of the histograms of the RS original image, the interpolated LR image, an image enlarged using SwinIR, and an image enhanced using DM.

A comprehensive investigation into the performance of various loss functions, including MSE, Edge, Perceptual [56], and Charbonnier, on the TESR model was carried out at a scale factor of ×2. The objective was to determine the most suitable loss function for both the ViT and iterative DM. The results, as presented in Table 4, indicate that the Charbonnier loss function exhibited a significant advantage over the other loss indices at the scale factor of ×2.

Table 4. Investigative comparison of loss functions of the TESR model with a scale factor of ×2.

Scale Factor	MSE Loss	Edge Loss	Perceptual Loss	Charbonnier Loss
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
×2	27.662/0.78535	22.667/0.60635	27.761/0.80892	35.367/0.9449

5. Discussion

RS images are characterized by the importance of fine details in object recognition applications, and many existing algorithms struggle to recover these details accurately. In order to overcome this challenge, we propose a two-stage approach for the super-resolution of RS images. The first stage involves using the SwinIR model to enlarge the image while preserving as much global detail as possible. The second stage involves applying the iterative DM to correct and enhance the image, allowing for the restoration of missing context (fine) details and ultimately resulting in a more accurate reconstruction of the high-resolution image. Moreover, transfer learning was used to increase the efficiency of the proposed model and speed up the training process. The results of our proposed TESR model show promising performance on the UCMerced dataset in terms of both visual perception and quantitative measurements. Overall, the use of a two-stage approach for super-resolution of remote sensing images appears to be a promising approach, and our proposed model demonstrates improved performance compared to other state-of-the-art methods. The Charbonnier loss function was employed in both the SwinIR and iterative DM models. The selection of the Charbonnier loss was based on a comprehensive evaluation of various loss functions on the TESR model.

There are a few limitations to this work that should be noted. Firstly, the proposed algorithm is specifically designed for use with remote sensing images and may not perform as well on other types of images. Secondly, the iterative DM requires a significant amount of training time, so we employed transfer learning to reduce the training time. However, the DM also requires a considerable amount of time for each test image, approximately two minutes per image. As a result, we only tested the DM on a limited number of images. It should be noted that this testing period is an obstacle to the widespread use of the DM in practical applications. Finally, it should be noted that using the DM in the first stage of the TESR algorithm may introduce some noise to the reconstructed image. This could potentially be a limitation for certain applications. Therefore, we decided to use the DM in the second stage of the TESR algorithm instead.

6. Conclusions

In conclusion, the proposed two-stage TESR algorithm for remote sensing super-resolution has demonstrated promising results in terms of both visual perception and quantitative measurements. The use of SwinIR in the first stage allows for the enhancement of global details and the enlargement of the low-resolution image, while in the second stage, the iterative DM further enhances the fine-detail quality of the reconstructed image by adding and removing noise. However, there are still opportunities for further optimization and development of the TESR algorithm, particularly in terms of addressing the limitations of the DM and exploring the potential for incorporating additional techniques. Overall, the TESR algorithm represents a promising approach for improving the resolution and quality of RS images and has the potential to be a valuable tool for various applications in the field of RS.

There are several potential directions for future work on the proposed two-stage remote sensing super-resolution algorithm, TESR. One area of focus would be to apply the algorithm to a larger and more diverse dataset in order to further validate its performance. Additionally, exploring the use of the algorithm for the super-resolution of other types of images (such as medical images or microscopy images) could be an interesting direction for future research. Additionally, further research could be conducted to address the limitations of the current TESR algorithm. For example, the DM has a relatively high computational cost. Developing methods to address these issues could improve the practicality and usability of the TESR algorithm. Overall, there are many opportunities for continued development and improvement of the TESR algorithm to better meet the needs of various remote sensing applications. Furthermore, in order to assess the applicability of our proposed model, we plan to evaluate its performance using real data acquired from multiple satellites with varying spatial resolutions. This will enable us to examine the effectiveness of our method in processing images obtained from different sources with varying degrees of accuracy and to demonstrate its robustness and adaptability to various scenarios in remote sensing applications.

Author Contributions: Conceptualization, A.M.A. and B.B.; formal analysis, A.M.A. and B.B.; investigation, A.M.A. and B.B.; methodology, A.M.A. and B.B.; project administration, B.B. and A.K.; software, A.M.A.; supervision, B.B. and A.K.; validation, A.M.A.; visualization, A.M.A.; writing—original draft, A.M.A. and B.B.; writing—review and editing, B.B., A.K., W.B. and W.E.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Prince Sultan University.

Acknowledgments: The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) for this publication.

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

References

1. Lim, S.B.; Seo, C.W.; Yun, H.C. Digital Map Updates with UAV Photogrammetric Methods. *J. Korean Soc. Surv. Geod. Photogramm. Cartogr.* **2015**, *33*, 397–405. [CrossRef]
2. Sun, H.; Sun, X.; Wang, H.; Li, Y.; Li, X. Automatic Target Detection in High-Resolution Remote Sensing Images Using Spatial Sparse Coding Bag-of-Words Model. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 109–113. [CrossRef]
3. Benjdira, B.; Koubaa, A.; Boulila, W.; Ammar, A. Parking Analytics Framework Using Deep Learning. In Proceedings of the 2022 2nd International Conference of Smart Systems and Emerging Technologies, SMARTTECH, Riyadh, Saudi Arabia, 9–11 May 2022; pp. 200–205. [CrossRef]
4. Benjdira, B.; Koubaa, A.; Azar, A.T.; Khan, Z.; Ammar, A.; Boulila, W. TAU: A Framework for Video-Based Traffic Analytics Leveraging Artificial Intelligence and Unmanned Aerial Systems. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105095. [CrossRef]
5. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building Extraction Based on U-Net with an Attention Block and Multiple Losses. *Remote Sens.* **2020**, *12*, 1400. [CrossRef]
6. Tang, Y.; Zhu, M.; Chen, Z.; Wu, C.; Chen, B.; Li, C.; Li, L. Seismic Performance Evaluation of Recycled Aggregate Concrete-Filled Steel Tubular Columns with Field Strain Detected via a Novel Mark-Free Vision Method. *Structures* **2022**, *37*, 426–441. [CrossRef]
7. Wang, Z.; Jiang, K.; Yi, P.; Han, Z.; He, Z. Ultra-Dense GAN for Satellite Imagery Super-Resolution. *Neurocomputing* **2020**, *398*, 328–337. [CrossRef]
8. Xiang-Guang, Z. A New Kind of Super-Resolution Reconstruction Algorithm Based on the ICM and the Bicubic Interpolation. In Proceedings of the 2nd 2008 International Symposium on Intelligent Information Technology Application Workshop, IITA 2008 Workshop, Shanghai, China, 21–22 December 2008; pp. 817–820. [CrossRef]
9. Khan, A.R.; Saba, T.; Khan, M.Z.; Fati, S.M.; Khan, M.U.G. Classification of Human's Activities from Gesture Recognition in Live Videos Using Deep Learning. *Concurr. Comput.* **2022**, *34*, e6825. [CrossRef]
10. Ubaid, M.T.; Saba, T.; Draz, H.U.; Rehman, A.; Ghani, M.U.; Kolivand, H. Intelligent Traffic Signal Automation Based on Computer Vision Techniques Using Deep Learning. *IT Prof.* **2022**, *24*, 27–33. [CrossRef]
11. Delia-Alexandrina, M.; Nedevschi, S.; Fati, S.M.; Senan, E.M.; Azar, A.T. Hybrid and Deep Learning Approach for Early Diagnosis of Lower Gastrointestinal Diseases. *Sensors* **2022**, *22*, 4079. [CrossRef]
12. Ran, Q.; Xu, X.; Zhao, S.; Li, W.; Du, Q. Remote Sensing Images Super-Resolution with Deep Convolution Networks. *Multimed. Tools Appl.* **2020**, *79*, 8985–9001. [CrossRef]
13. Zhu, Y.; Geiß, C.; So, E. Image Super-Resolution with Dense-Sampling Residual Channel-Spatial Attention Networks for Multi-Temporal Remote Sensing Image Classification. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *104*, 102543. [CrossRef]

14. Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Lu, T.; Jiang, J. Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5799–5812. [CrossRef]
15. Xiong, Y.; Guo, S.; Chen, J.; Deng, X.; Sun, L.; Zheng, X.; Xu, W. Improved SRGAN for Remote Sensing Image Super-Resolution Across Locations and Sensors. *Remote Sens.* **2020**, *12*, 1263. [CrossRef]
16. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
17. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
18. El-Shafai, W.; Ali, A.M.; El-Rabaie, E.S.M.; Soliman, N.F.; Algarni, A.D.; Abd El-Samie, F.E. Automated COVID-19 Detection Based on Single-Image Super-Resolution and CNN Models. *Comput. Mater. Contin.* **2021**, *70*, 1141–1157. [CrossRef]
19. El-Shafai, W.; Mohamed, E.M.; Zeghid, M.; Ali, A.M.; Aly, M.H. Hybrid Single Image Super-Resolution Algorithm for Medical Images. *Comput. Mater. Contin.* **2022**, *72*, 4879–4896. [CrossRef]
20. Sohl-Dickstein, J.; Weiss, E.A.; Maheswaranathan, N.; Ganguli, S.; Edu, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2256-2265. Lille, France, 7–9 July 2015; pp. 2256–2265.
21. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Adv. Neural. Inf. Process. Syst.* **2020**, *33*, 6840–6851.
22. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image Super-Resolution Via Iterative Refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4713–4726. [CrossRef]
23. Dhariwal, P.; Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *Adv. Neural. Inf. Process. Syst.* **2021**, *34*, 8780–8794.
24. Nichol, A.Q.; Dhariwal, P. Improved Denoising Diffusion Probabilistic Models. In Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8162-8171, Virtual, 18–24 July 2021; pp. 8162–8171.
25. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations(ICLR), Virtual, 3 May 2021; Available online: https://openreview.net/forum?id=PxTIG12RRHS (accessed on 15 April 2023).
26. Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion Models for Implicit Image Segmentation Ensembles. *Proc. Mach. Learn. Res.* **2022**, *172*, 1336–1348.
27. Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; Babenko, A. Label-Efficient Semantic Segmentation with Diffusion Models. *arXiv* **2021**, arXiv:2112.03126. [CrossRef]
28. Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A.G.; Milanfar, P. Deblurring via Stochastic Refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 16293–16303.
29. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279. [CrossRef]
30. Wan, F.; Zhang, X. Super Resolution Reconstruction Algorithm of UAV Image Based on Residual Neural Network. *IEEE Access* **2021**, *9*, 140372–140382. [CrossRef]
31. González, D.; Patricio, M.A.; Berlanga, A.; Molina, J.M. A Super-Resolution Enhancement of UAV Images Based on a Convolutional Neural Network for Mobile Devices. *Pers. Ubiquitous Comput.* **2022**, *26*, 1193–1204. [CrossRef]
32. Zhang, D.; Shao, J.; Li, X.; Shen, H.T. Remote Sensing Image Super-Resolution via Mixed High-Order Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5183–5196. [CrossRef]
33. Guo, M.; Zhang, Z.; Liu, H.; Huang, Y. NDSRGAN: A Novel Dense Generative Adversarial Network for Real Aerial Imagery Super-Resolution Reconstruction. *Remote Sens.* **2022**, *14*, 1574. [CrossRef]
34. Xiao, Y.; Zhang, J.; Chen, W.; Wang, Y.; You, J.; Wang, Q. SR-DeblurUGAN: An End-to-End Super-Resolution and Deblurring Model with High Performance. *Drones* **2022**, *6*, 162. [CrossRef]
35. Li, B.; Qiu, S.; Jiang, W.; Zhang, W.; Le, M. A UAV Detection and Tracking Algorithm Based on Image Feature Super-Resolution. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 6526684. [CrossRef]
36. Lei, S.; Shi, Z.; Mo, W. Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5615611. [CrossRef]
37. Tu, J.; Mei, G.; Ma, Z.; Piccialli, F. SWCGAN: Generative Adversarial Network Combining Swin Transformer and CNN for Remote Sensing Image Super-Resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5662–5673. [CrossRef]
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 10012–10022.
39. Liu, J.; Yuan, Z.; Pan, Z.; Fu, Y.; Liu, L.; Lu, B. Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing. *Remote Sens.* **2022**, *14*, 4834. [CrossRef]
40. Fu, X.; Wang, J.; Zeng, D.; Huang, Y.; Ding, X. Remote Sensing Image Enhancement Using Regularized-Histogram Equalization and DCT. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2301–2305. [CrossRef]

41. Ablin, R.; Helen Sulochana, C.; Prabin, G. An investigation in satellite images based on image enhancement techniques. *Eur. J. Remote Sens.* **2020**, *53* (Suppl. 2), 86–94. [CrossRef]
42. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Volume 8689, pp. 818–833. [CrossRef]
43. Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
44. Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. *Found. Trends® Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
45. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
46. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2017**, arXiv:1710.10196.
47. 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]
48. Lin, W.; Jay Kuo, C.C. Perceptual Visual Quality Metrics: A Survey. *J. Vis. Commun. Image Represent.* **2011**, *22*, 297–312. [CrossRef]
49. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multi-Scale Structural Similarity for Image Quality Assessment. In Proceedings of the Conference Record of the Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402. [CrossRef]
50. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image Super-Resolution via Sparse Representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef]
51. 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]
52. Dong, C.; Loy, C.C.; Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Volume 9906 LNCS, pp. 391–407. [CrossRef]
53. Lei, S.; Shi, Z.; Zou, Z. Super-Resolution for Remote Sensing Images via Local-Global Combined Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1243–1247. [CrossRef]
54. Haut, J.M.; Paoletti, M.E.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J. Remote Sensing Single-Image Superresolution Based on a Deep Compendium Model. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1432–1436. [CrossRef]
55. Qin, M.; Mavromatis, S.; Hu, L.; Zhang, F.; Liu, R.; Sequeira, J.; Du, Z. Remote Sensing Single-Image Resolution Improvement Using A Deep Gradient-Aware Network with Image-Specific Enhancement. *Remote Sens.* **2020**, *12*, 758. [CrossRef]
56. Deng, Z.; Cai, Y.; Chen, L.; Gong, Z.; Bao, Q.; Yao, X.; Fang, D.; Yang, W.; Zhang, S.; Ma, L. RFormer: Transformer-Based Generative Adversarial Network for Real Fundus Image Restoration on a New Clinical Benchmark. *IEEE J. Biomed. Health Inf.* **2022**, *26*, 4645–4655. [CrossRef] [PubMed]