symmetry

MDPI

Article

Enhanced Example Diffusion Model via Style Perturbation

Haiyan Zhang and Guorui Feng *

School of Communication & Information Engineering, Shanghai University, Shanghai 200444, China; haiyan_zhang@shu.edu.cn
* Correspondence: grfeng@shu.edu.cn

Abstract: With the extensive applications of neural networks in several fields, research on their security has become a hot topic. The digitization of paintings attracts our interest in the security of artistic style classification tasks. The concept of symmetry is commonly adopted in the construction of deep learning models. However, we find that low-quality artistic examples can fool high-performance deep neural networks. Therefore, we propose the enhanced example diffusion model (EDM) for low-quality paintings to symmetrically generate high-quality enhanced examples with positive style perturbations, which improves the performance of the deep learning-based style classification model. Our proposed framework consists of two parts: a style perturbation network that transforms the inputs into the latent space and extracts style features to form a positive style perturbation, and a conditional latent diffusion model that generates high-quality artistic features. High-quality artistic images are combined with positive style perturbations to generate artistic style-enhanced examples. We conduct extensive experiments on synthetic and real datasets, and find the effectiveness of our approach in improving the performance of deep learning models.

Keywords: enhanced examples; diffusion model; artistic images; adversarial examples; image restoration

1. Introduction

Artistic painting is a form of expression that uses highly summarized and refined figurative graphics to design objective objects, which carries the connotation of human social development and has a very important cultural and historical value. The difference between artistic images and natural images is that artistic images have unique artistic styles. There are many styles of artistic images, such as Van Gogh's style, Monet's style, abstract expressionism, and figurative expressionism. With the digitization of paintings, it is worthwhile to study artistic paintings as objects in the computer vision domain.

Convolutional neural networks (CNNs), as a crucial technique in computer vision, are widely adopted in the processing of digital images, such as image classification [1,2], image translation [3,4], and object detection [5,6]. In the field of artistic images, neural networks can be applied to artistic image generation (e.g., style transfer) [7–9], artistic style classification [10,11], etc. Although neural networks are successful in multiple domains, extensive research finds that they are vulnerable to data errors, natural noise, and carefully designed examples, which can fool high-performance deep learning models. An adversarial attack on an image classification model is usually to add small perturbations to images so that the target classifier makes incorrect judgments. The perturbed images are called the adversarial examples [12]. The success of adversarial attacks brings security threats to the real-life application of neural networks. Artistic paintings often suffer from serious quality degradation due to inappropriate preservation conditions and the environment, such as blurring, noise, and fading. These degradations are equivalent to adding perturbations to high-quality artworks. Low-quality artistic images affect the visual appearance and reduce the classification accuracy of the deep learning-based classifier which affects our

check for updates

Citation: Zhang, H.; Feng, G. Enhanced Example Diffusion Model via Style Perturbation. *Symmetry* **2023**, *15*, 1074. https://doi.org/10.3390/sym15051074

Academic Editor: Alice Miller

Received: 13 April 2023
Revised: 5 May 2023
Accepted: 8 May 2023
Published: 12 May 2023

judgments of the artistic styles at the same time. Therefore, enhancing the style of low-quality paintings is a valuable issue that can be applied to the fields of artistic image restoration and robustness improvement of deep learning-based style classifiers.

The enhancement of low-quality paintings is the process of generating high-quality paintings that are visually consistent with the original low-quality images. The process can be considered an image-to-image translation task. A typical approach is to use a deep generative model to learn the distribution of the output image given the input. Deep generative models are excellent at learning complex distributions [13]. Generative adversarial networks (GANs) [14,15], variational autoencoders (VAEs) [16,17], and normalizing flows (NFs) [18,19] have produced excellent results in image generation. Moreover, diffusion models [20,21] have attracted attention recently with their superior performance in computer vision. Perona et al. [22] first used the diffusion process to smooth images and detect edges. Since Ho et al. [20] proposed denoising diffusion probabilistic models (DDPM), which we will call the "diffusion model" for brevity, many extended methods have been proposed to be used in the image domain, such as image-to-image translation [23,24], super-resolution [25,26], and image in-painting [27,28].

Motivated by these observations, we propose an enhanced example diffusion model EDM for the low-quality artistic images to restore them, while we inject positive style perturbations to generate enhanced examples. The style-enhanced examples can ultimately improve the performance of the style classifier and the robustness of the model. We compress the input image into a latent space and use DDPM to restore the low-quality artwork features. Then, we inject positive style perturbations into the restored image to achieve style enhancement. We separate the training into two phases: first, we train a style perturbation network (SP-Net), which contains an encoder that can extract the input painting features, and a decoder that can reconstruct the processed image features in the latent space; second, we train a conditional latent diffusion model (CLDM) in the latent space to remove the noise from the input image feature space. The latent diffusion model (LDM) [25] achieves effective feature denoising and also reduces the complexity of the network. In summary, this paper makes the following contributions:

- We propose an artistic style enhancement method to improve the adversarial robustness of deep learning models for low-quality artistic images. To the best of our knowledge, this is a new approach for improving the accuracy of artistic style classifiers.
- We use the latent diffusion model to denoise the features of the input paintings and add positive style perturbations in the pixel space to produce artistic style-enhanced examples.
- We show that our method can generate high-quality artistic examples and improve the accuracy of the deep learning-based style classifier.

The remainder of the paper is organized as follows. In Section 2, we introduce the related works of this paper. Section 3 describes the principles of DDPM to provide the theoretical basis for the proposed method. The details of our proposed method are shown in Section 4. Experimental results and analyses are demonstrated in Section 5. Finally, we conclude this paper and present prospects for further work in Section 6.

2. Related Works

In this section, we introduce some related works, which summarize the current state of research and representative approaches on adversarial attack and image restoration.

2.1. Adversarial Attack

Szegedy et al. [12] demonstrate that deep neural networks are sensitive to adversarial examples. For example, in an image classification task, adding some slight perturbations to the input image can result in different outputs from a well-performing neural classifier. Many algorithms have been proposed to investigate the vulnerability of neural networks. Goodfellow et al. [29] propose the fast gradient sign method (FGSM), which utilizes the gradient information of the target classifier to generate adversarial examples. This method is extended by iterative FGSM (I-FGSM) [30] to iteratively perform the FGSM attack. The projected gradient descent (PGD) [31] randomly selects an initial noise near a positive example as the beginning of the iterative attack. DeepFool [32] successfully attacks the target neural classifier with minimal perturbations from an optimization perspective. Carlini et al. [33] propose three adversarial attack algorithms based on different norms and aim to solve the minimization perturbation problem.

The above methods are proposed on the premise that all information about the target model is known (i.e., white-box attack). In the black-box case, Chen et al. [34] generate adversarial examples by estimating the gradient of the target model with limited information. However, this method has a low success rate of attack. Therefore, the prior-guided random gradient-free method (P-RGF) [35] is proposed to solve this problem, which makes better use of the transfer-base prior to estimating the gradient. Cheng et al. [36] transform the decision-based black-box attack problem (i.e., only the prediction class of a given input image from the target model is available) into a continuous real-valued optimization problem. Croce et al. [37] propose a versatile framework based on the random search that reduces the problem in a black-box setting to a discrete problem instead of continuous optimization one.

2.2. Image Restoration

There are a lot of degradations that can affect the photographs, including some degradations during the shooting process and other degradations that occur over time and with environmental influences. The existing image degradation can be broadly categorized into unstructured degradation and structured degradation. The former contains blurriness, noise, fading, JPEG compression, etc., while the latter contains spots, scratches, holes, etc. [38].

For unstructured degradation, traditional restoration methods mostly employ different image prior constraints, such as non-local self-similarity [39], local smoothness [40], and sparsity [41]. Deep learning-based restoration methods transform image restoration into an image translation and learn the mapping between the low-quality and high-quality image pairs. Many state-of-the-art methods have been proposed. Zhang et al. [42] propose the fast and flexible denoising CNN (FFDNet), which uses a single network for image denoising with different noise levels. Fang et al. [43] perform the non-blind image deblurring task by using a deep neural network (DNN) as an implicit prior of the image. Saharia et al. [26] complete the image super-resolution via the diffusion model. The structured degradation image restoration is often regarded as an image inpainting task. Liu et al. [44] mask hole regions during DNN learning and focus only on features in non-hole regions. Ren et al. [45] make the texture in the hole regions available for synthesis from features of patches with similar structures by estimating the appearance flow. Shao et al. [46] and Xu et al. [47] restore images with the conditional GAN and new perceptual losses. RePaint [27] utilizes the diffusion model to generate high-quality results for any inpainting form.

However, in the real world, images usually suffer from complex degradations. The reinforcement learning restoration method (RL-Restore) [48] dynamically selects different networks to handle different degradations. Suganuma et al. [49] execute different convolution operations for different degradations using the attention mechanism. To improve the performance of the restoration on real photos, Wan et al. [50] propose a latent space restoration method that learns the domain translation between real old photos, synthetic degraded images, and the corresponding ground truth images, which generalizes the restoration to real old photographs well.

3. Preliminaries: Denoising Diffusion Probabilistic Models

Diffusion models [20] have emerged as a very powerful family of deep generative models that break the long-standing dominance of GANs [14] in image synthesis tasks [51]. In this paper, we use the diffusion model as a method to generate high-quality artistic features. DDPM is a parametric Markov chain of length T trained using variational inference, which consists of two processes: the forward process of gradually adding Gaussian noise to the input data x_0 and the reverse process of denoising the noisy data x_T by training a neural network. Next, we briefly review the definitions of DDPM from Ho et al. [20].

The forward process defined in DDPM is to gradually add Gaussian noise to the input image $x_0 \sim q(x_0)$ to generate Gaussian white noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ in T timesteps. Each step of the forward process is as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where β_t is the variance of the Gaussian noise added at timestep t.

The reverse process gradually removes the noise by a learnable Markov chain to restore the original image x_0, which can be modeled by a neural network that predicts the parameters of a Gaussian distribution:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

where θ denotes the neural network model parameter, and $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ denote the mean and variance of the Gaussian distribution, respectively. Then, the posterior probability $q(x_{t-1}|x_t, x_0)$ can be calculated using the Bayesian formula:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \tag{3}$$

Assuming that $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$, the expressions for $\tilde{\mu}_t(x_t, x_0)$ and $\tilde{\beta}_t$ in Equation (3) are as follows:

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{4}$$

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{5}$$

The training objective of the reverse process can be optimized with a variable lower bound:

$$\begin{aligned} \mathbb{E}[-\log p_\theta(x_0)] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \\ &= \mathbb{E}_q\left[-\log p(x_T) - \sum_{t\geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] =: L \end{aligned} \tag{6}$$

As extended by Ho et al. [20], Equation (6) can be rewritten as:

$$\mathbb{E}_q\left[\underbrace{D_{KL}(q(x_T|x_0)||p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \underbrace{- \log p_\theta(x_0|x_1)}_{L_0}\right] \tag{7}$$

where $D_{KL}(\cdot)$ denotes the function to calculate the KL divergence. For the sum of L_T, $L_{t-1}(t = 2, \ldots, T)$, L_0 forms the variational lower bound L. There is an exception for L_0, L_{t-1}, and L_T as they are KL divergences between the two Gaussian distributions since $q(x_{t-1}|x_t, x_0)$ is also Gaussian [20], so they can be evaluated in closed forms.

The above derivations of DDPM are important theoretical bases for the restoration of artistic images in Section 4.

4. Methodology

There are several challenges that need to be addressed when enhancing low-quality paintings. Some paintings are damaged due to an improper preservation environment, which affects the appreciation and the viewers' or classifiers' judgments of the artistic style. Restoring damaged paintings by hand is very time-consuming and requires a high level of expertise. In addition, since the degradations of paintings in real environments are very complex, it is difficult to realistically model and collect a large amount of representative training data.

To address these challenges, we propose an enhancement network EDM to enhance the degraded paintings and improve the robustness of the classification network. For the problem of fewer training data, we resolve it by manually synthesizing low-quality artistic images. Firstly, we use the encoder in SP-Net to transform the input paintings into the latent space. Because the domain gap is closed in the latent space [50], the above transformation can be better applied to the real degraded paintings. The perturbation of the external environment on the artistic image is represented as a very large "noise" in the feature mapping of the network [52]. We use CLDM to denoise the low-quality artistic features. Secondly, we add positive artistic style perturbations to the restored paintings so that these samples with incorrectly classified styles can be successfully classified by the style classifier, and the correctly classified ones retain the original correct classification.

Figure 1 illustrates the overall architecture of EDM, which contains two main parts: the style perturbation network SP-Net and the conditional latent diffusion model CLDM. Here, X_{sr} denotes the source image, which is the original high-quality painting, X_{old} is the low-quality painting, X_{rec} is the restored painting, X_{sp} denotes the positive perturbation, and X_{se} is the corresponding output, i.e., the enhanced example. The encoder and decoders of the auto-encoder are denoted by $E(\cdot)$, $D_0(\cdot)$, and $D_1(\cdot)$, respectively. Further, the content and style parts are denoted by $(\cdot)_C$ and $(\cdot)_S$, respectively.

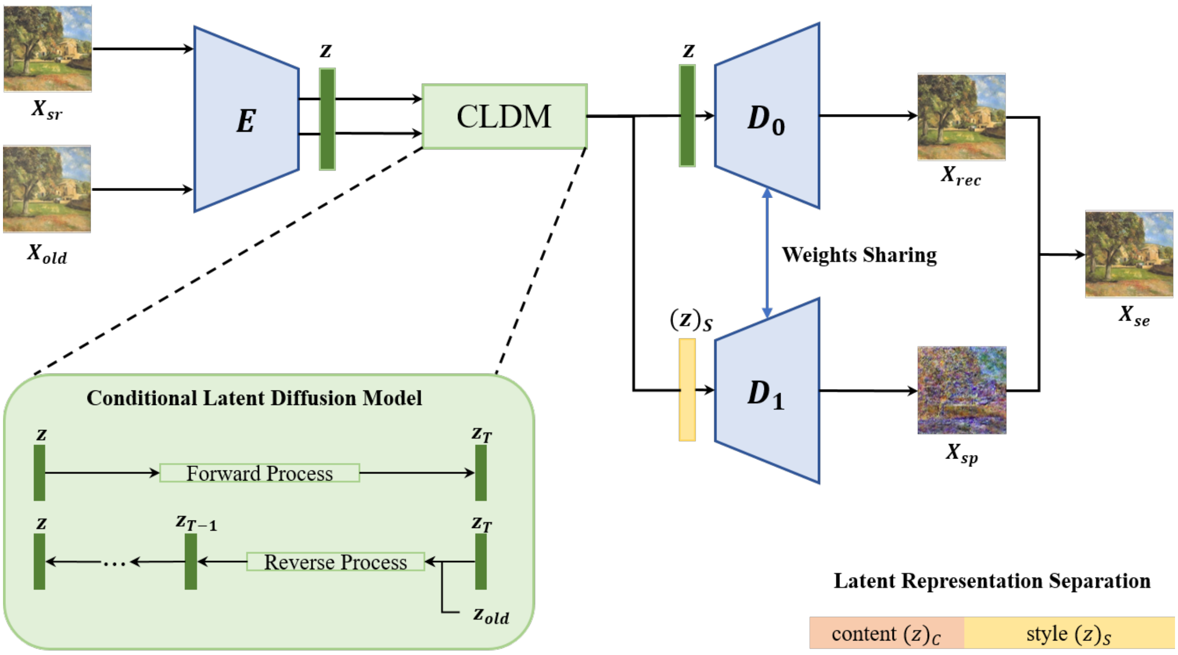

Figure 1. An overview of our framework. EDM contains the style perturbation network (SP-Net) and the conditional latent diffusion model (CLDM). SP-Net is composed of an encoder and two decoders, one branch for decoding images, and the other branch for extracting style features and generating style perturbations. The forward process of CLDM generates pure noise, and the reverse process performs denoising.

4.1. Style Perturbation Network

Our SP-Net is based on the work proposed by Esser et al. [53], which is composed of an encoder and two decoders. The auto-encoder is the neural backbone of SP-Net, which is a typical encoder-decoder structure. The encoder compresses the input data and extracts the most representative information from the input, which aims to decrease the dimensionality of the input information and thus reduce the processing burden of CLDM. The decoder decompresses the important features into raw information. Two decoders divide the network into two branches: one branch is the reconstruction branch that is used to reconstruct the paintings from the latent representations and another branch is the style perturbation branch that is used to generate style perturbations. It is worth mentioning that the two decoders share weights. Moreover, inspired by Rombach et al. [25], we introduce a vector quantization layer [54] in the decoder to prevent arbitrarily high-variance latent spaces. This model can be interpreted as a vector quantized generative adversarial network (VQGAN) [53], with the difference that the quantization layer of this model is absorbed by the decoder.

More precisely, the encoder E encodes the source painting X_{sr} and the old painting X_{old} into latent representations $z_{sr} = E(X_{sr})$ and $z_{old} = E(X_{old})$. CLDM restores the input z_{old} to generate the high-quality feature z_0, which is similar to z_{sr}:

$$z_0 = CLDM(z_{sr}, z_{old}) \tag{8}$$

The details of CLDM are described later. Svoboda et al. [55] separate the latent code z into the content part $(z)_C$ which contains the content information of the painting (e.g., objects, scale, etc.), and the style part $(z)_S$ which contains the style information presented in the painting's content (e.g., shapes of objects, textures, etc.). In the reconstruction branch, the decoder D_0 reconstructs the painting from z_0, giving $X_{rec} = D_0(z_0)$. In the style perturbation branch, we separate the restored latent code z_0 into the content part and style part by the method proposed by Svoboda et al. [55]:

$$z_0 = [(z_0)_C, (z_0)_S] \tag{9}$$

Then, we feed the style features $(z_0)_S$ into the decoder D_1 to generate a positive style perturbation X_{sp}, giving $X_{sp} = D_1((z_0)_S)$. Finally, the style perturbation X_{sp} and the reconstruction image X_{rec} are combined to generate an artistic style-enhanced example X_{se}:

$$X_{se} = X_{rec} + \lambda_{sp} X_{sp} = D_0(z_0) + \lambda_{sp} D_1((z_0)_S) \tag{10}$$

where $\lambda_{sp} \in [0, 1]$ is the coefficient that affects the intensity of the style perturbation X_{sp}.

To ensure that SP-Net can successfully reconstruct artistic images, we use the adversarial loss L_{adv} [55] and the perceptual loss L_{per} [56] to optimize the SP-Net:

$$\begin{aligned} L_{adv} = {} & \mathbb{E}_{X_{sr}}\left[(C(X_{sr}) - \mathbb{E}_{X_{se}} C(X_{se}) + 1)^2\right] \\ & + \mathbb{E}_{X_{se}}\left[(\mathbb{E}_{X_{sr}} C(X_{sr}) - C(X_{se}) + 1)^2\right] \end{aligned} \tag{11}$$

$$L_{per} = \sum_j \|\phi_j(X_{sr}) - \phi_j(X_{se})\|_2^2 \tag{12}$$

where $C(\cdot)$ denotes the discriminator, which is used to determine whether X_{sr} and X_{se} belong to the same artistic style. $\phi_j(\cdot)$ is the activations of the jth layer of the pretrained 16-layer VGG network ϕ [57]. In addition, we introduce a style loss to guarantee the style consistency of X_{sr} and X_{se}:

$$L_{style} = \mathbb{E}_{(z_{sr})_s}\left[\|(E(X_{se}))_S - (E(X_{sr}))_S\|_2^2\right] \tag{13}$$

The above three losses compose the full objective of SP-Net, which is defined as follows:

$$L = L_{per} + L_{style} + \lambda L_{adv} \tag{14}$$

where λ is the weight of the adversarial loss L_{adv}, which is calculated adaptively based on the gradients of L_{per} and L_{adv} as follows [53]:

$$\lambda = \frac{\nabla_{G_L}[L_{per}]}{\nabla_{G_L}[L_{adv}] + \delta} \tag{15}$$

where $\nabla_{G_L}[\cdot]$ denotes the gradient of its input with respect to the last layer L of the decoder, and δ is set to 10^{-6} for the stability of the values [53].

4.2. Conditional Denoising Diffusion Model

Diffusion Model. The neural network used in the reverse process can be interpreted as a sequence of denoising autoencoders $\epsilon_\theta(x_t, t); t = 1 \dots T$, which are trained to predict the noise added to the current noisy input x_t. According to the derivation of Ho et al. [20], the training objective of the model can be simplified to the following equation:

$$L_{DDPM} = \mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2] \tag{16}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $t \sim Uniform(\{1, \dots, T\})$.

Conditional Latent Diffusion Model. We use a trained auto-encoder to transform the input data into a valid low-dimensional latent space. The latent space is more suitable for a likelihood-based diffusion model than the pixel space [25]. The latent representations of X_{sr} and X_{old} are denoted by z_{sr} and z_{old}, respectively. z_0 is the corresponding denoising output of CLDM.

Similar to the diffusion model, the role of CLDM is to learn the parameter approximation of the conditional distribution $p(z|z_{old})$. CLDM generates a target representation z_0 in T refinement steps. In the forward process, the pure noise representation z_T is generated by gradually adding Gaussian noise to the input z_{sr}. Moreover, one can characterize the distribution of z_t at the arbitrary timestep t by the rewritten Equation (1) [20]:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t)\mathbf{I}) \tag{17}$$

where $\bar{\alpha}_t$ is the same as the definition in Section 3. The model of the reverse process starts from pure noise z_T and iteratively optimizes the output to obtain a sequence $(z_{T-1}, z_{T-2}, \cdots, z_1, z_0)$ according to learned conditional distributions $p_\theta(z_{t-1}|z_t, z_{old})$, such that $z_0 \sim p(z|z_{old})$ [26]:

$$p_\theta(z_{t-1}|z_t, z_{old}) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, z_{old}, t), \Sigma_\theta(z_t, z_{old}, t)) \tag{18}$$

where $\mu_\theta(z_t, z_{old}, t)$ and $\Sigma_\theta(z_t, z_{old}, t)$ denote the mean and variance of $p_\theta(z_{t-1}|z_t, z_{old})$, respectively. The reverse process is implemented with a conditional denoising auto-encoder $\epsilon_\theta(z_t, t, z_{old})$ [25], which takes a noisy representation z_t and the conditional input z_{old} as inputs. Following [20], the expression of z_t is given by:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \tag{19}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The loss function for training $\epsilon_\theta(z_t, t, z_{old})$ is:

$$L_{CLDM} = \mathbb{E}_{t,E(X_{sr}),\epsilon,E(X_{old})}[||\epsilon - \epsilon_\theta(z_t, t, z_{old})||_2^2] \tag{20}$$

where $E(\cdot)$ denotes the encoder of SP-Net. The denoising model ϵ_θ is trained to estimate the noise added to the current noisy representation z_t. As reported by Saharia et al. [26], we can parameterize the mean $\mu_\theta(z_t, z_{old}, t)$ of $p_\theta(z_{t-1}|z_t, z_{old})$ as:

$$\mu_\theta(z_t, z_{old}, t) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(z_t, t, z_{old})\right) \tag{21}$$

where α_t and β_t are the same as the definition in Section 3. We set the variance of $p_\theta(z_{t-1}|z_t, z_{old})$ to $\beta_t\mathbf{I}$. Thus, each iteration of CLDM can be implemented by the following equation [26]:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(z_t, t, z_{old})\right) + \sqrt{\beta_t}\epsilon_t \tag{22}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. By iterating over T timesteps, we can transform the low-quality features z_{old} into high-quality features z_0 and achieve feature restoration in the latent space.

In this section, we provide detailed descriptions of our proposed approach. Subsequently, we will construct a series of experiments to demonstrate the effectiveness of our method.

5. Experiments and Discussions

In this section, we experimentally evaluate the proposed EDM. As artistic style enhancement is a relatively unexplored area, the proposed method is compared with other state-of-the-art methods for old photo restoration. We illustrate the effectiveness of our approach based on extensive quality analyses and quantitative evaluations. Furthermore, to support the choice of architecture and style perturbations, ablation experiments are performed to demonstrate the effectiveness of various components and how they impact the results.

5.1. Implementation

5.1.1. Datasets

The source paintings are sampled from the Wikiart dataset. Since it is difficult to collect a large number of representative degraded artistic paintings, we synthesize low-quality paintings using paintings sampled from the Wikiart dataset. To generate realistic defects, we follow the data generation method used by Wan et al. [50] and added Gaussian blur, Gaussian white noise, JPEG compression, and paper texture to the original high-quality paintings. In addition, we collected a small number of real degraded artistic drawings which are used to test the generalization of the model.

5.1.2. Training Details

In our work, we separated the training process into two distinct phases: first, SP-Net is trained to compress the data into the latent representation and generate style perturbations. Then, CDLM is trained to restore low-quality artistic features in the learned latent space. Both training processes are implemented on a GeForce RTX 3090 GPU.

Training of SP-Net. The training dataset of SP-Net is the original high-quality paintings sampled from the Wikiart dataset. These paintings are cropped and resized to 256×256. The size of the latent representation is 64×64. SP-Net uses the Adam optimizer [58] with an initial learning rate of 4.5×10^{-6} and a batch size of 2. We chose the batch size of 2 only due to limited computing resources.

Training of CLDM. The latent representations of the source paintings and the old paintings are the training datasets of CLDM, where the latent representations of the old paintings participate in the training of the reverse process as conditions. In the latent space, the size of the inputs and conditions are 64×64. The neural backbone of the diffusion model is implemented by a time-conditional UNet [59], which uses the Adam optimization scheme [58] with a learning rate of 1×10^{-6} and a batch size of 16. Moreover, we set $T = 1000$ timesteps in our work.

5.2. Experimental Results

We propose an enhanced example diffusion model EDM for low-quality paintings, and this work consists of two core tasks: artistic image restoration and artistic style enhancement. Therefore, the visual effect of the results and the style accuracy are two important metrics to evaluate the performance of our proposed method. However, to the best of our knowledge, there are fewer studies similar to ours. To demonstrate the effectiveness of our method, we compare EDM with image-to-image translation (denoted as Pix2pix [60], Palette [24]), old photo restoration (denoted as Wan et al. [50]), and the commercial tool (denoted as Meitu [61], Photoshop). For fair comparisons, we train all methods (except the commercial tool) with the same dataset (Wikiart), and test them on our synthesized old artistic images.

5.2.1. Quality Analysis

Theoretically, our method will produce high-quality artistic images, whose artistic styles are visually similar to the source paintings. To prove the effectiveness of our method to enhance the appearance of low-quality paintings, we design experiments on the Wikiart test images with the size of 512 × 512. As the old paintings for the test here are synthetic, we use the corresponding source paintings as references. Figure 2 shows the old paintings, source paintings, and the restoration results of all methods. The Pix2pix method can restore the degradation to some extent, however, it is not very effective in removing the obvious noise. Moreover, this method generates some undesirable artifacts, which leads to the destruction of the original texture. Palette also uses the diffusion model, but is visually inferior to our method. Some degradations remain in the results. This is due to the fact that their method is performed in the pixel domain. The method proposed by Wan et al. and Meitu can realize image restoration better. However, the results produced by both methods are too smooth, which leads to the loss of the fine structure. Moreover, the method proposed by Wan et al. may change the color of the artistic image. The results produced by Photoshop are similar to those of Meitu. Photoshop can repair images better. However, for artistic images, the results are too smooth and the style characteristics of the original painting are lost. In comparison, our method generates clean, clear artistic images with rich texture details. In addition to successfully addressing synthetic degradations, our method also enhances the style of the artistic image appropriately. In general, our method produces visually pleasant results.

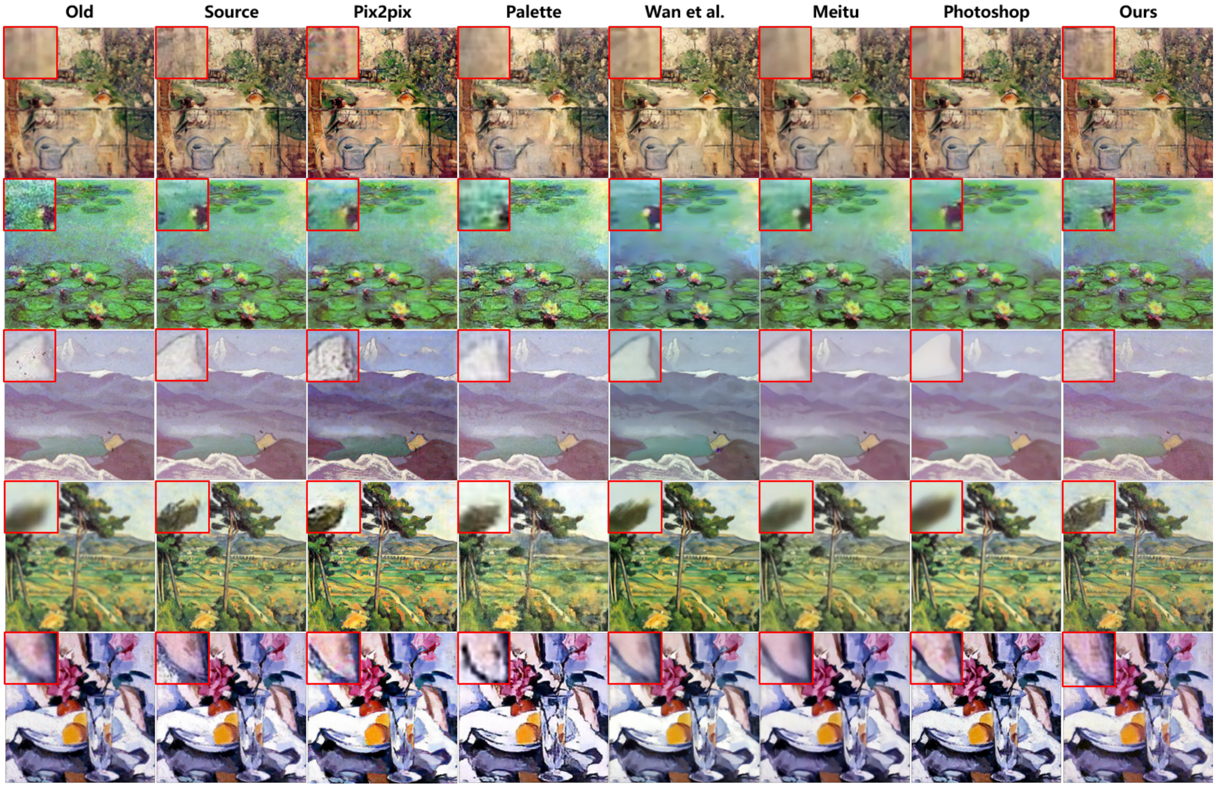

Figure 2. Style-enhanced results of our method and all comparison methods. From the first to the last column, in order, are old paintings, source paintings, the results generated by Pix2pix [60], Palette [24], Wan et al. [50], Meitu [61], Photoshop, and our method. It shows that the results generated by our method are more similar to the style of the source images and have clearer texture details than other methods.

5.2.2. Quantitative Analysis

Artistic images are characterized by unique styles, such as Van Gogh's style and Monet's style. The proposed EDM is designed to improve the accuracy of deep learning-based style classifiers for degraded artistic images. Therefore, we test different methods on the synthetic paintings from the Wikiart dataset and adopt style accuracy and the peak signal-to-noise ratio (PSNR) as evaluation metrics for comparison. The style accuracy is measured by an artistic style classifier, which indicates the proportion of paintings that are correctly classified with respect to their styles. We collect paintings from nine artists with about 5000 high-quality paintings as the training dataset and train the style classifier with two, four, and seven classes of paintings, respectively. For each artist's style, we synthesize about 1000 low-quality artistic images as the test set for the classifier. Table 1 shows the quantitative comparison results. PSNR is used to compare low-level differences between the enhanced output and ground truth. Meitu unsurprisingly achieves the best PSNR score since this method generates smooth and noiseless paintings. Our method ranks second-best. The method proposed by Wan et al. also obtains a good score. The style accuracy is used to evaluate whether the results can improve the performance of the classifier. Our method achieves the best scores. Conversely, the method proposed by Wan et al. and Meitu perform poorly in terms of style accuracy since their results are overly smooth, which destroys the artistic style of the paintings to some extent. In all, the results generated by our method have a better quality and can significantly improve the performance of the classification model.

Table 1. Quantitative comparison on the different number of style categories. For brevity, the style accuracy is abbreviated to "acc". Upward arrows indicate that a higher score denotes a better enhancement performance or image quality.

Class Metric	**Two**			**Four**			**Seven**		
	Old-acc (%)↑	**Enhanced-acc (%)↑**	**PSNR↑**	**Old-acc (%)↑**	**Enhanced-acc (%)↑**	**PSNR↑**	**Old-acc (%)↑**	**Enhanced-acc (%)↑**	**PSNR↑**
Pix2pix [60]	75.31	77.50	21.02	57.95	64.77	21.32	49.86	53.14	21.88
Palette [24]		77.30	21.86		62.53	22.62		54.76	22.76
Wan et al. [50]		72.43	23.26		65.91	23.56		45.24	23.53
Meitu [61]		61.57	**24.89**		42.05	**24.18**		44.64	**25.03**
Ours		**88.65**	23.58		**76.14**	23.85		**57.14**	23.89

5.3. Ablation Studies

There are two key components of our method. The first one is that we use the diffusion model in the latent space rather than pixel space to achieve the restoration of low-quality paintings. Another one is that we introduce a style perturbation branch in our framework for generating positive style perturbations to enhance the robustness of the style classifier. We verify their effectiveness in the overall model architecture by removing or replacing these components.

5.3.1. Space Conversion

In this work, we first convert the input into the latent space with an auto-encoder and then enhance the low-quality paintings into high-quality ones with the diffusion model. According to the theory of Wan et al. [50], converting low-quality paintings into latent space can better generalize the method to real old artistic images. We collect some old and real paintings and apply them in the pixel space-based and latent space-based diffusion models. As shown in Figure 3, the diffusion model in the latent space can better enhance the low-quality artwork, and the model in the pixel space can only remove some obvious noise. Furthermore, the application of the diffusion model in the latent space can reduce the complexity. We test the training and sampling times of both models on a single GeForce RTX 3090 GPU, and Table 2 reports the training time in hours per epoch and test time in minutes per sample at resolution 256 × 256. We observe a speedup of at least 1.4× for training and at least 12.7× for testing between pixel space-based and latent space-based models.

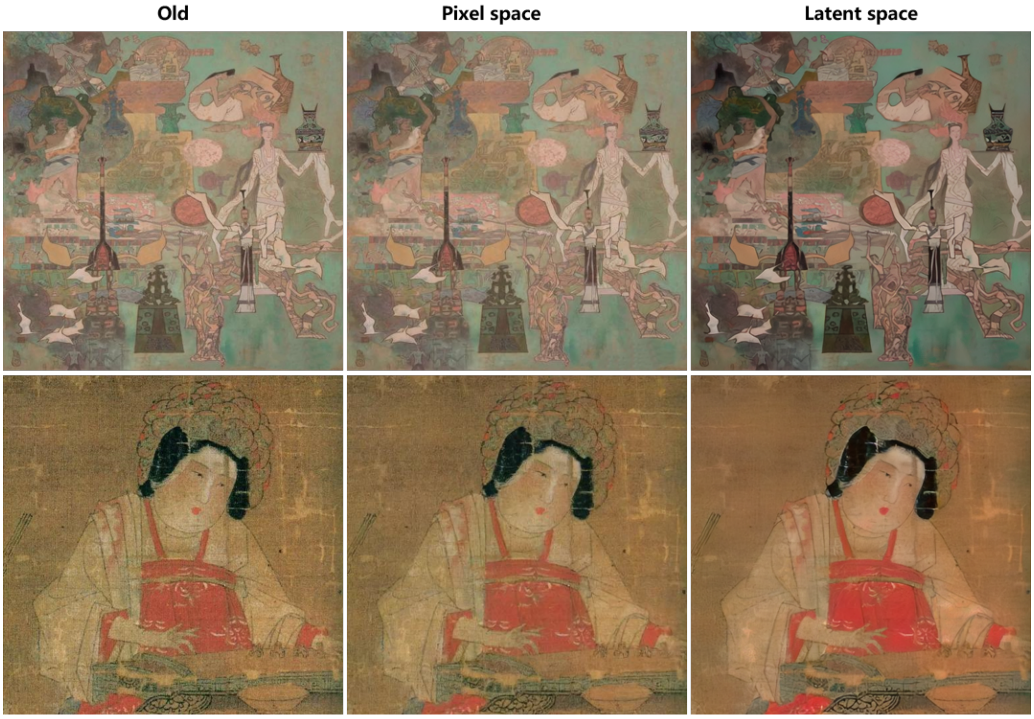

Figure 3. Visual comparisons of latent space and pixel space on real old data. It shows that enhancing the low-quality paintings in the latent space can better generalize to the real old paintings.

Table 2. Efficiency comparison of pixel space-based and latent space-based diffusion models. Downward arrows indicate that less time corresponds to higher efficiency.

Model	**Training (hours/epoch)↓**	**Test (minutes/sample)↓**
Pixel space-based	0.706	1.53
Latent space-based	**0.504**	**0.12**

5.3.2. Style Perturbation

Our method in this paper aims to enhance the artistic style of low-quality paintings to produce good visual effects while improving the robustness of the style classifier. We introduce CLDM to generate high-quality artistic images and add positive style perturbations to the paintings, which form artistic style-enhanced examples to improve the accuracy of the style classifiers. Figure 4 demonstrates that the positive style perturbation can enhance the texture details of the artistic image. Table 3 shows that the enhanced examples can improve the classification accuracy of the style classifier. Overall, it is effective to add positive style perturbations to the artistic images in improving texture details and style accuracy.

Table 3. Style accuracy of results with/without style perturbation (SP) on a different number of style categories.

Model	**Acc (%)↑**		
	Two	**Four**	**Seven**
EDM without SP	85.37	74.28	55.69
EDM with SP	**88.65**	**76.14**	**57.14**

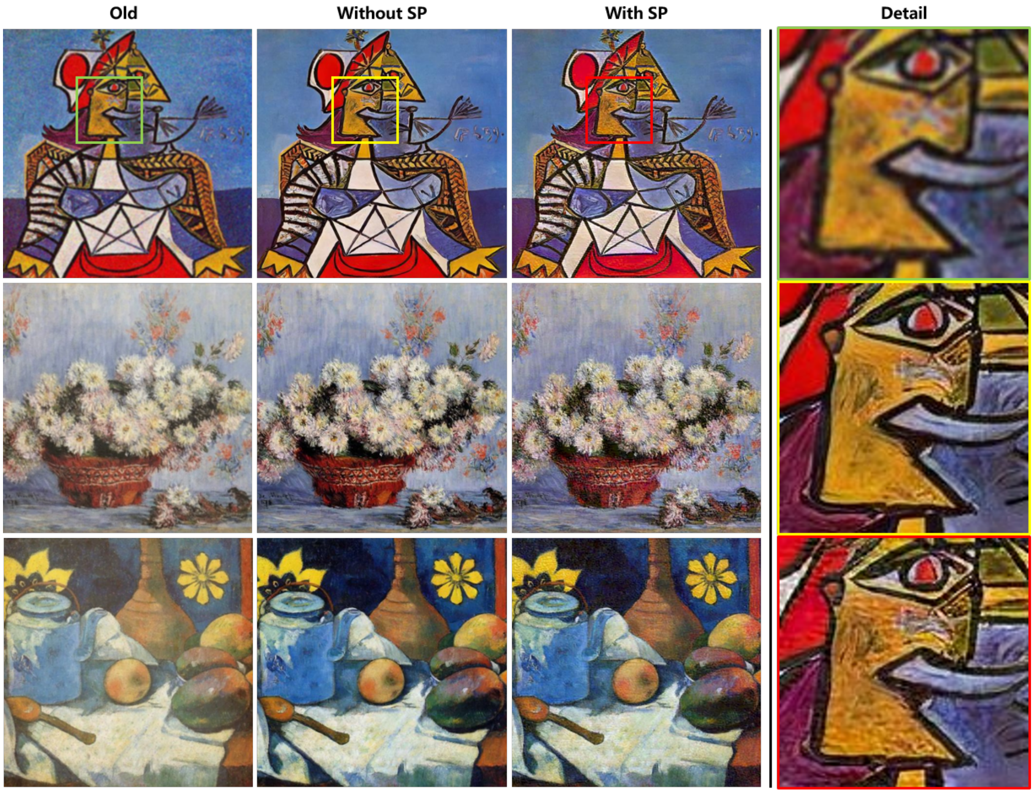

Figure 4. Visual comparison of enhanced results with/without style perturbation. The details of the paintings in row 1 are shown in the rightmost column. It shows that the results with style perturbations have richer texture details.

6. Conclusions

We propose an enhanced example diffusion model called EDM, an effective way to significantly enhance the artistic style of low-quality paintings and improve the robustness of the style classification model. The overall system contains a style perturbation network SP-Net and a conditional latent diffusion model CLDM. SP-Net transforms the inputs into the latent space and extracts the artistic style to produce positive style perturbations. CLDM denoises low-quality inputs in the latent space to generate high-quality artistic features. The combination of high-quality paintings and positive style perturbations forms artistic style-enhanced examples, which can further enhance the artistic style and improve the robustness of the deep learning-based models without degrading the quality of restored paintings. Furthermore, our approach can be generalized to real degraded paintings. Experimental results show that the accuracy of style-enhanced examples generated by our approach improves by about 13% on average on the style classifier compared to degraded paintings. Moreover, EDM performs exceptionally well in terms of visual quality. In the future, we will explore other ways to enhance artistic images, and study enhancement problems in more scenes to further improve the robustness of the classification model.

Author Contributions: Supervision, G.F.; Writing—original draft, H.Z.; Writing—review & editing, H.Z. and G.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9908, pp. 630–645.
3. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
4. Yang, S.; Jiang, L.; Liu, Z.; Loy, C.C. Unsupervised image-to-image translation with generative prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 18332–18341.
5. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
6. Afouras, T.; Asano, Y.M.; Fagan, F.; Vedaldi, A.; Metze, F. Self-supervised object detection from audio-visual correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 10575–10586.
7. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
8. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
9. Zhang, Y.; Li, M.; Li, R.; Jia, K.; Zhang, L. Exact feature distribution matching for arbitrary style transfer and domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 8035–8045.
10. Chen, X.; Yan, X.; Liu, N.; Qiu, T.; Ni, B. Anisotropic stroke control for multiple artists style transfer. In Proceedings of the 28th ACM International Conference on Multimedia, MM 2020, Virtual, 12–16 October 2020; pp. 3246–3255.
11. Brunner, G.; Konrad, A.; Wang, Y.; Wattenhofer, R. MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer. In Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 23–27 September 2018; pp. 747–754.
12. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
13. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
14. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
15. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2016, Barcelona, Spain, 5–10 December 2016; pp. 2172–2180.
16. Kingma, D.P.; Welling, M. An introduction to variational autoencoders. In *Foundations and Trends® in Machine Learning*; Now Publishers: Delft, The Netherlands, 2019; Volume 12, pp. 307–392.
17. Vahdat, A.; Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2020, Virtual, 6–12 December 2020; pp. 19667–19679.
18. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using Real NVP. In Proceedings of the International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
19. Kingma, D.P.; Dhariwal, P. Glow: Generative Flow with Invertible 1x1 Convolutions. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2018, Montreal, QC, Canada, 3–8 December 2018; pp. 10236–10245.
20. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2020, Virtual, 6–12 December 2020; Volume 33, pp. 6840–6851.
21. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, PMLR, ICML 2015, Lille, France, 6–11 July 2015; pp. 2256–2265.
22. Perona, P.; Malik, J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 629–639.
23. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 14347–14356.
24. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, SIGGRAPH 2022, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.

25. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
26. 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**, 4713–4726. [CrossRef] [PubMed]
27. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 11461–11471.
28. Gao, R.; Song, Y.; Poole, B.; Wu, Y.N.; Kingma, D.P. Learning Energy-Based Models by Diffusion Recovery Likelihood. In Proceedings of the International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021.
29. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
30. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial Machine Learning at Scale. In Proceedings of the International Conference on Learning Representations. OpenReview.net, ICLR 2017, Toulon, France, 24–26 April 2017.
31. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
32. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
33. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, IEEE, SP 2017, San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
34. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec 2017, Dallas, TX, USA, 3 November 2017; pp. 15–26.
35. Cheng, S.; Dong, Y.; Pang, T.; Su, H.; Zhu, J. Improving black-box adversarial attacks with a transfer-based prior. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
36. Cheng, M.; Le, T.; Chen, P.; Zhang, H.; Yi, J.; Hsieh, C. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. In Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
37. Croce, F.; Andriushchenko, M.; Singh, N.D.; Flammarion, N.; Hein, M. Sparse-rs: A versatile framework for query-efficient sparse black-box adversarial attacks. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual, 22 February–1 March 2022; Volume 36, pp. 6437–6445.
38. Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; Wen, F. Bringing old photos back to life. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 2747–2757.
39. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Non-local sparse models for image restoration. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, 27 September–4 October 2009; pp. 2272–2279.
40. Weiss, Y.; Freeman, W.T. What makes a good model of natural images? In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, Minneapolis, Minnesota, USA, 18–23 June 2007; pp. 1–8.
41. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef] [PubMed]
42. Zhang, K.; Zuo, W.; Zhang, L. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Trans. Image Process.* **2018**, *27*, 4608–4622. [CrossRef] [PubMed]
43. Fang, Y.; Zhang, H.; Wong, H.S.; Zeng, T. A robust non-blind deblurring method using deep denoiser prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2022, New Orleans, LA, USA, 19–20 June 2022; pp. 735–744.
44. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision, ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 85–100.
45. Ren, Y.; Yu, X.; Zhang, R.; Li, T.H.; Liu, S.; Li, G. Structureflow: Image inpainting via structure-aware appearance flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 181–190.
46. Shao, C.; Li, X.; Li, F.; Zhou, Y. Large Mask Image Completion with Conditional GAN. *Symmetry* **2022**, *14*, 2148. [CrossRef]
47. Xu, J.; Li, F.; Shao, C.; Li, X. Face Completion Based on Symmetry Awareness with Conditional GAN. *Symmetry* **2023**, *15*, 663. [CrossRef]
48. Yu, K.; Dong, C.; Lin, L.; Loy, C.C. Crafting a toolchain for image restoration by deep reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2443–2452.

49. Suganuma, M.; Liu, X.; Okatani, T. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2018; pp. 9039–9048.
50. Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Wen, F.; Liao, J. Old photo restoration via deep latent space translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2071–2087. [CrossRef] [PubMed]
51. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; Yang, M.H. Diffusion models: A comprehensive survey of methods and applications. *arXiv* **2022**, arXiv:2209.00796.
52. Xie, C.; Wu, Y.; Maaten, L.v.d.; Yuille, A.L.; He, K. Feature denoising for improving adversarial robustness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2018; pp. 501–509.
53. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 12873–12883.
54. Van Den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. In Proceedings of the Advances in Neural Information Processing Systems NeurIPS 2017, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
55. Svoboda, J.; Anoosheh, A.; Osendorfer, C.; Masci, J. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 13816–13825.
56. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2016, ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9906, pp. 694–711.
57. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
58. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
59. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings, Part III 18, Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
60. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
61. Meitu. Available online: https://www.meitu.com/en (accessed on 18 March 2023).

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.