electronics

MDPI

Article

Two-Stage Generator Network for High-Quality Image Inpainting in Future Internet

Peng Zhao [1,2,3,4,*], Dan Zhang [1,3,4], Shengling Geng [1,3,4] and Mingquan Zhou [1,3,4,*]

1 School of Computer Science, Qinghai Normal University, Xining 810008, China
2 School of Information Technology, Luoyang Normal University, Luoyang 471022, China
3 Academy of Plateau Science and Sustainability, People's Government of Qinghai Province & Beijing Normal University, Xining 810004, China
4 The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Normal University, Xining 810008, China
* Correspondence: zhaopeng@lynu.edu.cn (P.Z.); mqzhou@bnu.edu.cn (M.Z.)

Abstract: Sharpness is an important factor for image inpainting in future Internet, but the massive model parameters involved may produce insufficient edge consistency and reduce image quality. In this paper, we propose a two-stage transformer-based high-resolution image inpainting method to address this issue. This model consists of a coarse and a fine generator network. A self-attention mechanism is introduced to guide the transformation of higher-order semantics across the network layers, accelerate the forward propagation and reduce the computational cost. An adaptive multi-head attention mechanism is applied to the fine network to control the input of the features in order to reduce the redundant computations during training. The pyramid and perception are fused as the loss function of the generator network to improve the efficiency of the model. The comparison with Pennet, GapNet and Partial show the significance of the proposed method in reducing parameter scale and improving the resolution and texture details of the inpainted image.

Keywords: image inpainting; generative adversarial network (GAN); two-stage transformer; adaptive multi-head attention mechanism; loss function

Citation: Zhao, P.; Zhang, D.; Geng, S.; Zhou, M. Two-Stage Generator Network for High-Quality Image Inpainting in Future Internet. *Electronics* **2023**, *12*, 1490. https://doi.org/10.3390/electronics12061490

Academic Editor: Catalin Stoean

Received: 18 January 2023
Revised: 18 March 2023
Accepted: 19 March 2023
Published: 22 March 2023

1. Introduction

Image inpainting originated from manual trimming in the Renaissance, which repairs the damaged image by filling the defective areas with adequate information. It infers the unknown area from known information such as structural, statistical, semantic, etc. [1–3]. Image inpainting can be applied to image super resolution, obstruction removal and damaged image repair, and has been a popular research area in computer vision and digital image processing.

Traditional image inpainting algorithms usually attack the image blurring or incomplete information caused by various types of noise during image acquisition. They can be divided into the structure-based methods, the texture-based methods and the sparse-representation-based method [4–6]. The PDE is the core of structured image inpainting, also known as diffusion image inpainting. The PDE equation method [7] was first proposed by Bertalmio et al. to perform diffusion-based restoration of images in pixels. However, the consistency between the local and the global semantics in defective area is ignored, not to mention the consideration of the constraint of the high-level semantics. Therefore, this method is only suitable for restoring pictures with minor local defects, while texture misalignment often occurs for large areas of defects. Texture-based image inpainting algorithms [8,9] search for the optimal texture pixels in intact areas and fill them to the defective area. This method can better preserve texture and structure information in images. The image-block-based Criminisi algorithm [10] proposed by Criminisi greatly improves the speed of restoration by searching for the target pixel blocks around the defective area.

Sparse representations more effectively express known information in image inpainting algorithms. Guleryuz et al. obtained the best estimation of the defect area by incorporating an adaptive technique [11,12] into the sparse reconstruction algorithm. However, regardless of whichever traditional image inpainting method is applied, the restoration of defective images with large missing areas is unsatisfactory. Thus far, they have mainly been applied to low-resolution images. The quality of the restored images cannot meet the demanding needs due to the complexity of damaged images and the inherent ambiguity of methods. The traditional methods encounter a major challenge, especially for high-resolution images.

With the rapid development of deep-learning technology recently, the image inpainting network has much ameliorated the long-standing deficiencies of traditional approaches and significantly improved the output quality [5]. The main aspect here is to extract the relevant contextual information from various receptive fields. Since Goodfellow et al. proposed the GAN model [13,14] in 2014, it [15–17] has become one of the mainstream methods for computer image processing. It has made many achievements in the field of image inpainting and greatly promoted the development of this technology. GAN consists of a generation model and a discrimination model. The discrimination is essentially a classifier that distinguishes between real pictures and fake pictures generated by the generator network, whose function is to convert the input defective image into the output restored image. As GAN methods suffer from unstable network training and model convergence, a series of improvements have been proposed. DCGAN [18,19], proposed by Radford et al., optimizes the learning representation performance by combining a deep convolutional network and GAN. The pooling layer of DCGAN is replaced by strided convolution to optimize the learning characterization performance of GAN. In addition, partial pooling layers are also replaced by transposed convolutions so that the entire network can be differentiated. Moreover, batch normalization can improve the model performance. However, for images with large defect areas, it is not effective.

The early use of attention mechanisms [20,21] in image inpainting was inspired by the traditional idea of fast matching. Yan et al. introduced shift connection, based on the UNet structure [22,23], to move the priori image information into the decoding network layer holding the corresponding features so as to complete the missing information, which improves the image inpainting capability. Yu et al. formally introduced the attention mechanism [24] to enhance image inpainting by finding the most similar feature blocks from the background and foreground image by convolution to search for feature matching blocks at a distance. However, current computer memory and processing resources are limited, and existing deep-learning algorithms can only perform restoration of low-resolution images. Zeng et al. proposed a high-resolution image inpainting technique that includes an iterative restoration model with a feedback mechanism [25]. It divides the inpainting task into two processes: low-resolution image inpainting and upsampling. Attention networks are also incorporated, using a guided upsampling network of attention mechanisms to calculate the feature similarity of low-resolution image feature blocks and guide the reconstruction process. The above attention-mechanism-based approach improves the performance of the network and proves the effectiveness of the attention mechanism. The attention mechanism can effectively enhance the conversion of higher order image semantics across the network, so that the overall structure is more sensitive to the restoration of detailed image texture. The Transformer model [26,27] was proposed in 2017 by the Google machine translation team [28], which has an outstanding performance in the field of natural language processing. The full-attention mechanism model is used by the Transformer to replace the traditional CNN [29–31] simply through a self-attention mechanism and a forward neural network. The Transformer increases the learning capability of the model while reducing the number of computational parameters, which makes its use in high-resolution image inpainting a promising topic.

The increase in Internet transmission capacity and wide use of mobile cameras result in an increasing demand for high-resolution images and videos. However, traditional image inpainting methods for high-resolution images often yield a limited result. Although there

has been a proliferation of techniques in this field recently, many methods still suffer from a lack of model edge consistency and unclear output images. In this work, we aim to produce sufficient edge consistency and high-quality images. The main contribution of this work are:

(1) We present a high-quality image inpainting network derived from Transformer, which is a two-stage generator model based on the encoder-decoder network.
(2) We apply the adaptive multi-head attention mechanism to the fine network to control the input of the features in order to reduce the computation overhead.
(3) We fuse the pyramid and perception as the loss function of the generator network to improve the overall efficiency.

With the involvement of the adaptive multi-head attention mechanism, the computation is reduced, and the forward propagation is significantly accelerated. Pyramid loss and perceptual loss are fused to improve model learning and speed up model restoration. After a comprehensive experimental comparison, the improved algorithm has better restoration performance. The edge and semantic information is more consistent and the resolution of the restored image is sharper.

The paper is structured as follows: The architecture of GAN, self-attentive mechanisms and image inpainting are presented in Section 2. The adaptive multi-head self-attention mechanism is presented in Section 3. The network structure of the method and the loss functions of the paper are also mentioned. Section 4 is devoted to the analysis of the experimental part. Section 5 concludes the paper as a whole.

2. Related Work

2.1. GAN

The generative adversarial network is trained to reach a Nash equilibrium state. After the generator is acquired, the defective images are fed into the generator to get the restoration results. The purpose of the generator is to turn the input defective image into a restored complete image. The generated fake images are identified and distinguished from the real intact images by the discriminator. Either 0 or 1 is assigned to the fake image or the ideal output image, respectively. GAN produces more realistic images, which are widely used in the field of CV [32–34]. The generators and discriminators of the network are trained by separate loss functions.

The loss function of the discriminator network is shown below.

$$\max_{D} V(D,G) = E_{x\sim pdata(x)}[\ln(D(x)) + E_{z\sim pinput(z)}[\ln(1-D(G(z))]] \quad (1)$$

where E denotes the expectation, $pdata(x)$: the true sample, G: the generator network, D: the discriminator network and $pinput(z)$: the input to the generator network.

The adversarial losses of the generating network are shown below.

$$\min_{G}\max_{D} V(D,G) = E_{x\sim pdata(x)}[\ln(D(x)) + E_{z\sim pinput(z)}[\ln(1-D(G(z))]] \quad (2)$$

The generator is trained with a small expectation of the loss function, while the discriminator has a large one. GAN is trained in alternative iterations, where the objective functions for the discriminator and the generator are also optimized separately.

The objective function of the generator is shown below.

$$\min_{G} V(D,G) = E_{z\sim pinput(z)}[\ln(1-D(G(z))] \quad (3)$$

First, the discriminator is trained. During its training the value of $D(x)$ is as close to 1 as possible, while the value of $D(G(z))$ preferably converges to 0. After the discriminator parameters have been updated, the discriminator parameters are frozen and the generator is trained. For the training process of the generator, the $D(G(z))$ should be close to 1.

2.2. Image Inpainting

In image inpainting [35,36], the information in the filled and intact areas of the image should be texturally and semantically consistent to enable a satisfactory restoration. The input to GAN is generally a defective image and a 0–1 mask to distinguish between defective and intact regions. A U-shaped encoder-decoder network with a patchwork structure is used by the generator to pursue consistency between the input and output image sizes. As shown in Figure 1, feature output by each layer captures the multi-scale feature information. A loss function is applied to improve the final output by adjusting the output of each layer. After iterative training, the final image restored by the generator will be highly similar to the real image. Furthermore, guide the transmission of higher-order semantics across layers should be guided by incorporating the attention mechanism, which can reduce information loss and semantic bias during the computation.

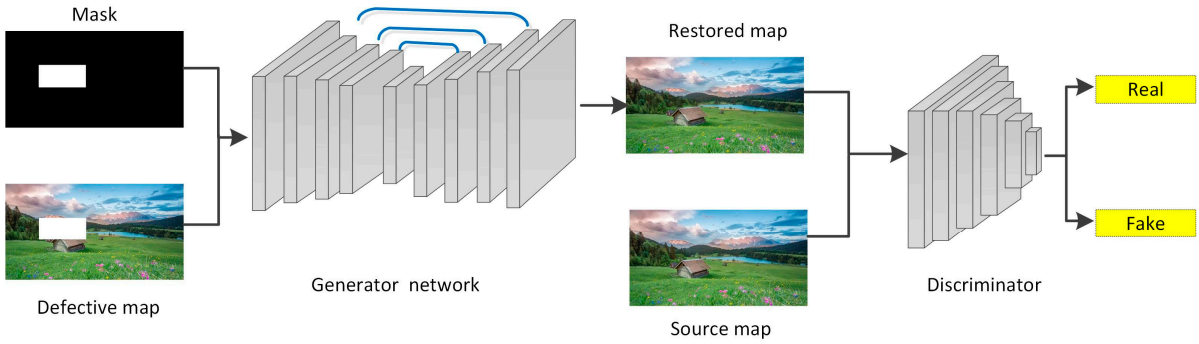

Figure 1. Overall diagram for image inpainting.

2.3. Self-Attentive Mechanism

Inpainting is obviously an ill-posed problem and hence is impossible without some assumptions about the statistics of images. It is also a precise statistical inference problem. Computer image inpainting is a process of image repair using known and a prior information in an image. The self-attention mechanism takes each pixel in the feature map as a random variable and calculates the covariance between all pixels. It can enhance or weaken the value of each predicted pixel according to its similarity with other pixels in the image, respectively. Similar pixels are used in training and prediction, and dissimilar pixels are ignored. This process requires that the semantics of the image inpainting results match the original image. The same is required for the texture consistency and edges of the image. This is a complex image-processing task. This requires the learning of deep-level features and the excellent stability of the network. Since generators consisting of general CNN sometimes do not learn sufficiently, the models can become quite complex when meeting high-information demand. Therefore, the learning stability of the neural network is strengthened by adding a self-attentive mechanism in this work.

The self-attentive mechanism [37] is a modification of the attention mechanism, which is less dependent on external information and can better capture the internal relevance of data features. The self-attentive mechanism consists of three parts: Q, K and V. Q represents the query quantity of the input model, K represents the reference quantity with high similarity to Q and V denotes the output content information corresponding to K. Depending on the processing task, the information stored in each part is different. In image-processing tasks, intrinsic information image features are used for attention interaction and the original feature map is usually mapped to Q, K and V. As shown in Figure 2, the correlation weight between Q and K are calculated and normalized using softmax. The same operation is performed for V and the output is obtained by superimposing it with the original input.

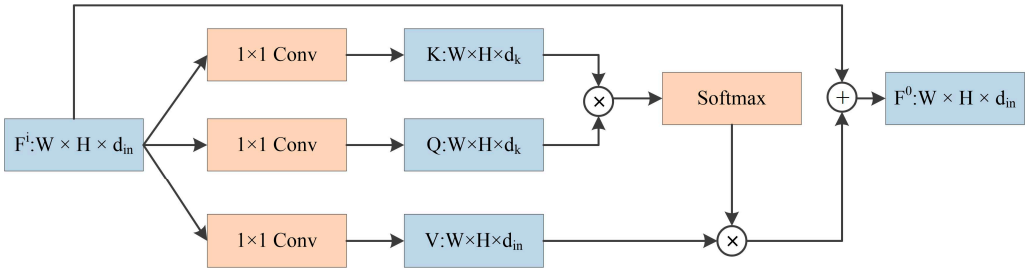

Figure 2. Architecture of the self-attention mechanism.

2.4. Pyramid Loss

Pyramid loss [38] is a loss-calculation method proposed by Zeng et al. in 2019. The final output is obtained by correcting the output of each layer. First, the feature information from each layer of the U-shaped hop network decoder is collected, and the final ground truth results are downsampled separately to calculate the L1 loss [39] and superimposed to obtain the final pyramid loss. The calculation equation is shown below.

$$L_{pd} = \sum_{l=1}^{L-1} \left|\left| x^l - h(\varphi^l) \right|\right|_1 \tag{4}$$

where l represents the output of layer l of the U-shaped network, x^l represents the downsampling of ground truth to the same dimensions as the output of layer l, h represents the 1×1 convolution of φ^l decoded into RGB images and L_{pd} is the result of an L1 loss calculation.

2.5. Perceptual Loss

Perceptual loss is a loss function proposed by Justin Johnson et al. [40] in the style-conversion task. Now, it is also widely used in tasks such as image inpainting and super-resolution. First, the loss of low-level features of pixel colo41r and edge are calculated. The potential features are extracted by the convolutional layers to obtain features similar to human perception. The features obtained by convolving the real image, typically extracted using a VGG network, are compared with the convolved features of the generated image, so as to keep the consistence of the content and high-level global structure information; then the losses are calculated.

The feature reconstruction loss function for perceptual losses is calculated as follows:

$$L_j = \|\Psi(\hat{y}) - \Psi(y)\|^2 \tag{5}$$

where Ψ denotes the pretrained network model, $\hat{y}$ indicates the generated restoration image and y denotes the original defective image. After pretraining, the network extracts the semantic information of the original image and the generated image, and the perceptual loss is obtained by calculating the two norms at the corresponding positions. Reducing the perceptual loss can effectively improve the training performance of the model.

3. Methodology

3.1. Adaptive Multi-Head Attention Mechanism

The self-attention mechanism can be treated as an interaction between different forms of the input vector in a linear projection space. The multi-head attention mechanism [41] creates projection information on the same input in several different projection spaces. The projections of the input matrix in different spaces are stitched together to enable the collection of features at multiple dimensions and directions. As shown in Figure 3, Value, Key and Query are fixed single values. There are three groups of both linear layers and scaled dot product attention units. This means that the input will be projected in 3 feature spaces. Finally, different weight coefficients are assigned to different self-attention

mechanism heads to concatenate the projection results together for result integration. However, the feature extraction method changes for different heads. Because of the decay of the attention features, the weight of the head is changed during the learning process of the model. The magnitude of change varies greatly for different heads. Therefore, the adaptive multi-head attention mechanism can adapt to the changes in weights generated by the model during the learning process. By adding a mask to the features at different locations, we can effectively lower the computation burden. Training produces a continuous mask so that the model can apply different weights for different heads, which reduces the computational burden and also allows different features to be learned.

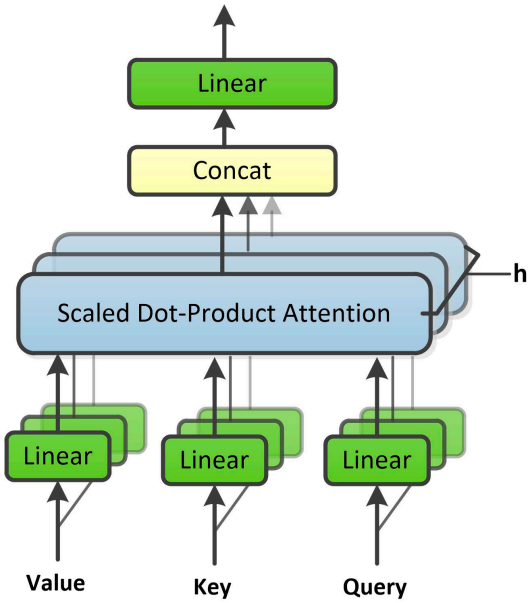

Figure 3. Architecture of the multi-head attention mechanism.

The Softmask function is shown below.

$$m_z(x) = \min[\max[\frac{1}{R}(R + z - x), 0], 1] \tag{6}$$

where $m_z(x)$ represents the function of the adaptive mask, z is the given attention weight, R is a hyperparameter that controls the adaptive mechanism and x is the input tensor.

Therefore, the formula for calculating the attention weight is as follows.

$$a_{tr} = \frac{m_z(t-r)\exp(s_{tr})}{\sum_{q=t-s}^{t-1} m_z(t-q)\exp(s_{tq})} \tag{7}$$

where a_{tr} represents the attention weight formula, S_{tr} represents the result of the self-attention mechanism before self-adaptation and t and r represent the feature information used to calculate similarity. S_{tq} represents the similarity score calculated at t and q.

The formula of s_{tr} is shown below.

$$s_{tr} = {x_t}^T {W_q}^T (W_k x_r + p_{t-r}) \tag{8}$$

where W_k and W_q are the k, q feature matrices, respectively, and p_{t-r} is the embedding result for the relevant position.

3.2. Network Structure Based on Transformer

The network structure proposed in this study is shown in Figure 4. The network includes a generator and a discriminator, where the input of the generator is a defective

image and a binary mask of the defective area. The two-stage restoration network is used, which is divided into rough network and fine network. Rough network can generate roughly restored images, and the filled content is roughly consistent with the intact area. However, there are some flaws in the texture and semantic details of the image. The input of the fine network is composed of the padding content of the rough network for the defect area and the original defect image. Fine networks can predict more detailed content in areas of image defects. The rough network consists of a convolutional layer and a self-attention mechanism layer. First, the defect image and the mask binary image are the input data to the network. The layer of self-attention mechanism in rough network can guarantee the preservation of useful information across layers. The fine network is a combination of transformer and convolution layers. First, the image and mask will pass through the position embedding layer for information embedding at the corresponding location. After the transition of four-dimensional image information to 3D features, a compression of the feature information is performed by downsampling convolution. The transformer layer computes the similarity between the features and attaches an adaptive mask to them to realize the dynamic multi-attention mechanism. The redundant calculation of the model is reduced. The last layer in the model is an embedding input. The feature map with position information is parsed into image features, and the number of layers of rough network is less than that of the fine network. In comparison, the span of up and down sampling between each two layers of the rough network is larger.

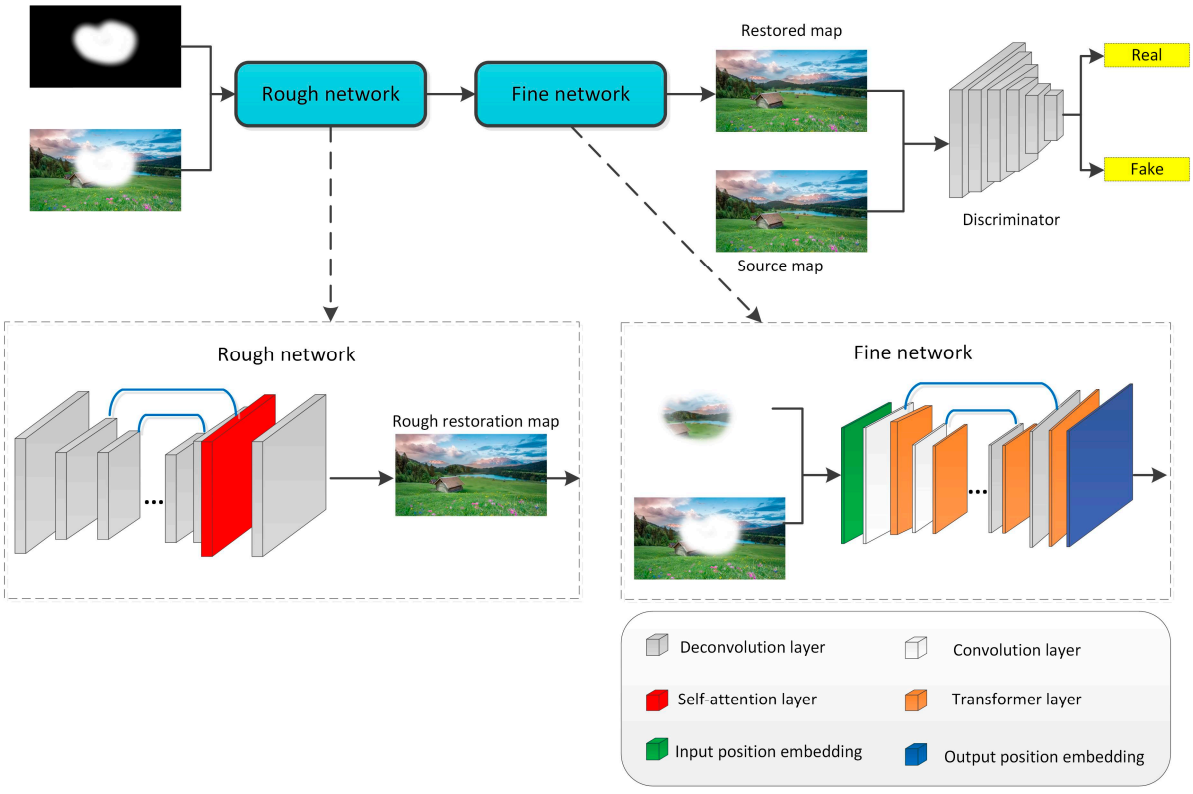

Figure 4. Network structure of the proposed method.

The discriminator is a full convolutional network consisting of five convolutional layers. Each convolutional layer is immediately followed by a normalization layer and finally by an activation layer. The role of the discriminator is to distinguish between the generated fake ones and the original true intact image, which is essentially a classification network. The resulting low-quality restored images are marked as false during the training process. The iteration process guarantees that the restored images are close to the real original image in terms of texture and content.

3.3. Loss Functions

The loss function is a key component whose choice determines the learning efficiency and direction of the model. It has three parts: the loss function of discriminator, the coarse network and the fine network. The loss function of the discriminator is the MSE loss function. In the generator loss, the L1 loss was chosen by the rough generative network because of the feature that L1 loss can increase the stability of the model learning process. The loss function of the fine generative network consists of four components, including perceptual loss, pyramidal loss, adversarial loss and L1 loss. The pre-trained VGG-19 network is responsible for extracting the higher order information and calculating the perceptual loss. The perceptual loss is calculated as shown below.

$$L_j = \frac{1}{C_j * H_j * W_j} ||V_j(\hat{Y}) - V_j(Y)||_2^2 \tag{9}$$

where C_j, H_j, W_j denote the number of feature channels, feature length and width of the jth layer feature Y, respectively. V_j denotes the pre-trained VGG19 model.

The overall loss function of the generator network is shown below.

$$L_g = \alpha L_c + \beta L_r \tag{10}$$

where L_c denotes the rough generator network losses, while L_r represents the losses of the fine network. α and β are the parameters given to adjust the weights of each loss function in the overall loss.

The loss function of the coarse network is shown below.

$$L_c = \gamma L_1 \tag{11}$$

where γ is the parameter given to adjust the corresponding weight.

The loss function of the fine network is shown below.

$$L_r = \delta L_a + \varepsilon L_p + \eta L_{pd} + \lambda L_1 \tag{12}$$

where L_a is the antagonistic loss, L_p is the perceptual loss and L_{pd} is the pyramidal loss. δ, ε, η and λ are the parameters given to adjust the weights of each loss function in the overall loss.

4. Experiments

4.1. Experimental Settings

The dataset used for the experiments in this work is the public dataset CelebA [42], and all experiments were conducted using the selected training data on CelebA. The selected training set is 1500 and the test set is 500.

The model is compared with Partial [43], Pennet [38] and GapNet [44]. Partial is the use of partial convolutions, where the convolution is masked and conditioned on only valid pixels by renormalization. Pennet is a pyramid context encoder network for image inpainting by deep generative models. A novel neural network for point cloud, dubbed GapNet, learns local geometric representations by embedding graph-attention mechanisms within stacked multi-layer-perceptron layers.

In order to objectively evaluate the experimental results, we use the peak signal to noise ratio (PSNR) and structural similarity (SSIM) [45] metrics.

4.2. Quality Assessment

In the same experimental setting, the ablation experiments of the algorithms in this paper are essential and the results are shown in Figure 5.

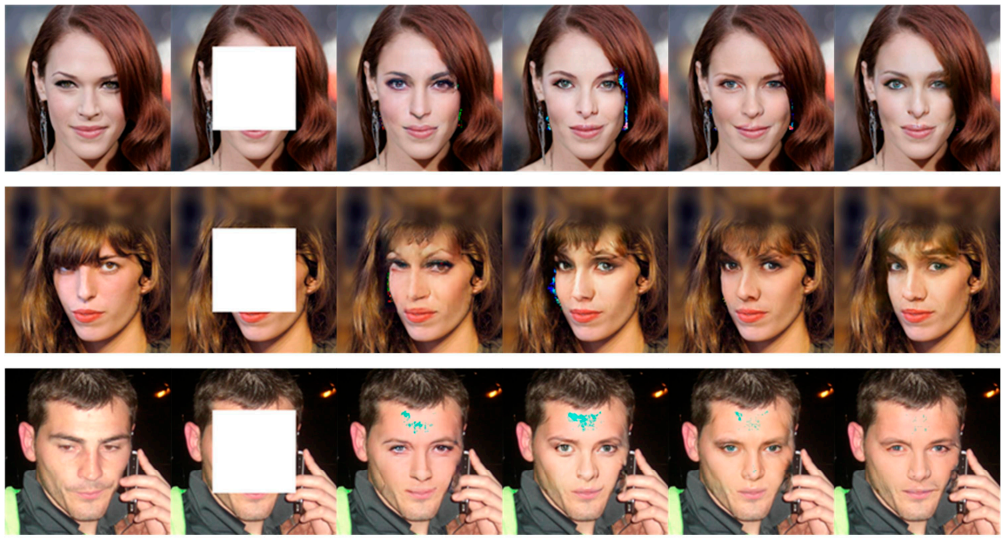

Figure 5. Ablation experiment on randomly masked square: (**a**) Ground truth; (**b**) Input image; (**c**) Non-transformer two-layer network + loss function; (**d**) Two-layer transformer + loss function; (**e**) Two-layer transformer + adaptive multi-head attention mechanism; (**f**) Two-layer transformer + adaptive multi-head attention mechanism + loss function.

From the comparison graph of the ablation experimental results in Figure 5, it can be found that the four superimposed items of two-stage transformer, adaptive multi-head attention mechanism, pyramidal loss and perceptual loss are the most effective parts and generate the highest-resolution restored images. The two-stage transformer and the adaptive multi-head attention mechanism approaches lack perceptual loss and pyramidal loss, respectively. Some colour flaws were created in the restored parts. The other combined approach has an increased computation of redundant information due to the lack of an adaptive multi-head attention mechanism. Therefore, the details of the partial restoration from the model do not match the semantic information and some colour defects are present. When all transformer layers in the restoration model are removed, the resulting non-transformer bilayer networks can obtain the perceptual loss and pyramidal loss. The lack of a full-attention mechanism model transformer and a multi-head attention mechanism results in poor repair results where pixel-level imperfections are present.

In the same experimental environment, the method of this paper was compared with Partial, Pennet and GapNet on the same set of test images, and the experimental results are shown in Figure 6.

As shown in Figure 6, the experimental results of all methods performed well in the restoration of semantic information. Among them, the overall restoration and clearness of Partial's is not as good as others. With regard to edge consistency, the gap between the processing results of Pennet compared with GapNet and the proposed method is large. The difference in colour and semantic information between the restored and intact areas of the Pennet creates a distinct edge. However, the edges of the defective regions of GapNet and our method are not obvious. The proposed method incorporates an adaptive multi-head self-attentive mechanism with the involvement of pyramidal and perceptual losses. Thus, the model reduces the computation of redundant information. The model focuses more on the generation of semantic information. In terms of edge consistency of the restoration results, our method outperforms GapNet. Regarding face details, the method proposed in this paper uses a two-stage network structure and the resolution quality of the image has strong performance compared with the other methods. Other methods do not highlight the details of the face, whereas the results from ours reveal detailed information such as wrinkles on the face.

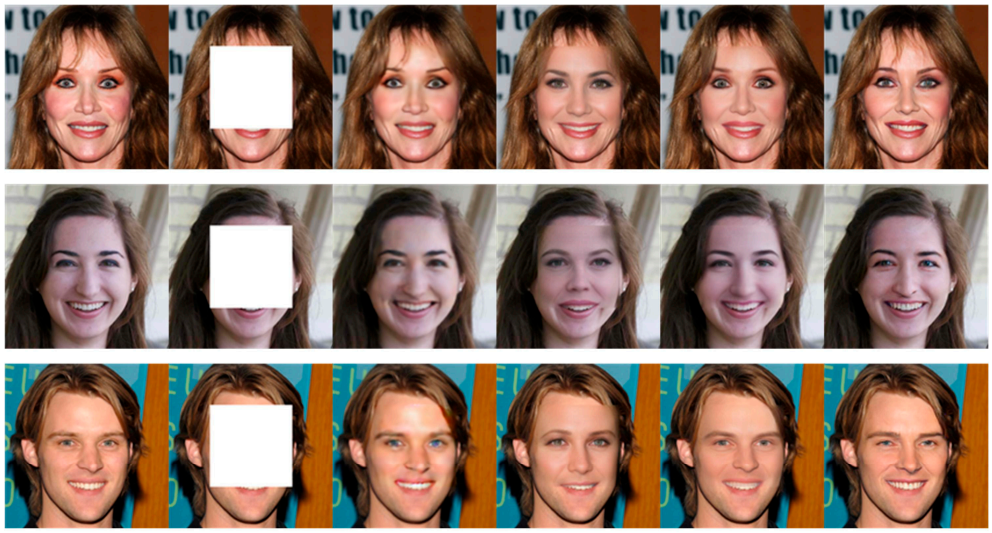

Figure 6. Results on randomly masked square: (**a**) Ground truth; (**b**) Input image; (**c**) Partial; (**d**) Pennet; (**e**) GapNet; (**f**) Ours.

4.3. Quantitative Analysis

The algorithm reduces the model's computation of redundant information through an adaptive multi-head self-attention mechanism. This boosts the calculation and alleviates the edge inconsistency arising from the model restoration results, while improving the model's restoration capability for semantic information. The transformer-based two-stage network enables the model to generate sharper images and enhances the quality of the restored images. Incorporating pyramidal loss and perceptual loss enhances the learning performance and efficiency of the model. The comparisons are implemented on the same set of test images. The experimental data are shown in Tables 1 and 2. The analysis of the experimental data shows that the algorithm in this paper is preferred in PSNR and SSIM metrics compared with other methods.

Table 1. PSNR comparison result.

Network Model	MAX PSNR/dB	MIN PSNR/dB	AVE PSNR/dB
Partial	30.75	28.86	29.12
Pennet	32.75	30.86	31.72
GapNet	34.27	32.37	33.55
Ours	37.38	34.98	36.13

Table 2. Comparison of SSIM metrics.

Network Model	MAX SSIM/%	MIN SSIM/%	AVE SSIM/%
Partial	69.61	62.42	68.23
Pennet	78.02	73.56	75.75
GapNet	90.08	85.01	88.13
Ours	94.96	89.66	92.09

Table 1 shows the results of the PSNR metric evaluation for various image inpainting contents. Our method has the highest scores in both MAX PSNR and AVE PSNR compared with other methods. Our method improves by 3.11 dB (9.0%) and 2.58 dB (7.6%), respectively, compared with the GapNet method. Compared with the Pennet method, the improvement is 4.63 dB (14.1%) and 4.41 dB (13.9%). This is an improvement of 6.63 dB (21.6%) and 7.01 dB (24.1%) compared with the Partial method.

Table 2 shows the results of the experiments on the SSIM evaluation metrics. Our method also has the highest scores on both the maximum and the mean. In particular, our method improved by 4.88 (5.4%) and 3.96 (4.5%) on the mean and maximum values compared with the GapNet method, respectively. Compared with the Pennet method it improved by 16.94 (21.7%) and 16.34 (21.6%). Compared with the Partial method, the improvement was 25.35 (36.4%) and 23.86 (35.0%).

A quantitative analysis of the experimental data shows that our method is more effective than other methods of restoration. The restoration performance is stable and there is not much difference between the maximum and minimum values. Therefore, the image inpainting model in this paper has stronger and more stable restoration performance.

Table 3 shows the results of the PSNR and SSIM metrics analysis for the ablation experiments. The analysis of the experimental data revealed that the best performance was obtained by the combination of the two-stage transformer, the adaptive multi-head attention mechanism and the loss function. The highest values were obtained on both MAX PSNR and AVE SSIM. Compared with the combination of the two-stage transformer and the adaptive multi-head attention mechanism, the improvement is 1.59 dB (4.4%) and 0.84 dB (2.4%), respectively. Compared with the two-stage transformer and loss function combination, the improvement is 0.18 dB (0.5%) and 0.86 dB (2.4%), respectively. The improvement is 4.54 dB (13.8%) and 4.18 dB (13.1%) compared with the non-transformer two-stage network and loss function combination.

Table 3. Comparison of PSNR metric for ablation experiments.

Network Model	MAX PSNR/dB	MIN PSNR/dB	AVE PSNR/dB
non-transformer two-layer network + loss function	32.84	30.62	31.95
two-layer network transformer + loss function	37.20	34.36	35.27
two-layer transformer + adaptive multi-head attention mechanism	35.79	34.67	35.29
two-layer transformer + adaptive multi-head attention mechanism + loss function	37.38	34.98	36.13

Table 4 shows the SSIM metrics for the ablation experiments. The combination of the two-stage transformer, adaptive multi-head attention mechanism and loss function yielded the highest figures for all metrics. In the maximum and mean values, compared with the combination of the two-stage transformer and adaptive multi-head attention mechanism, there is an improvement of 3.87 (4.2%) and 2.82 (3.2%), respectively. This is an increase of 1.71 (1.8%) and 2.90 (3.3%) compared with the two-stage transformer and loss function combinations. The improvement is 10.84 (12.9%) and 9.98 (12.2%) compared with the non-transformer two-stage network and loss function combination.

Table 4. Comparison of SSIM metric for ablation experiments.

Network Model	MAX SSIM/%	MIN SSIM/%	AVE SSIM/%
non-transformer two-layer network + loss function	84.12	79.68	82.11
two-layer network transformer + loss function	93.25	84.97	89.19
two-layer transformer + adaptive multi-head attention mechanism	91.09	84.18	89.27
two-layer transformer + adaptive multi-head attention mechanism + loss function	94.96	89.66	92.09

Quantitative analysis of the ablation experiments demonstrates that the restoration results of our model using combination of a two-stage transformer, an adaptive multi-head attention mechanism and a loss function have a better restoration performance compared with a method using only some of these.

5. Conclusions

For the problem of edge consistency and the blurred resolution of image inpainting, this paper proposes a transformer-based image restoration method with a two-stage generator restoration network. A self-attentive mechanism in the rough network is used to guide the transfer of higher-order semantic information across the network. It can effectively reduce the loss of image information during the transduction process. The improved transformer model and designed adaptive multi-head self-attention mechanism in the fine network reduce the number of parameters effectively in the model. The learning ability of the model and the resolution of the generated images have been improved. The fusion of pyramidal loss and perceptual loss improves the training performance of the model. Edge and semantic consistency problems in inpainted images are effectively solved. The proposed method has obtained good experimental results on CelebA compared with other methods. The applicability of the proposed method is less prominent on other datasets. Future work will involve optimizing the current inpainting framework for better performance by working on enriching the testing dataset. The developed method will be applied to other applications in the future.

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

Funding: This work was partially supported by the Natural Science Youth Foundation of Qinghai Province: 2023ZJ947Q; National Nature Science Foundation of China: 62102213; National Nature Science Foundation of China: 62262056; Independent project fund of State Key lab of Tibetan Intelligent Information Processing and Application (Coestablished by province and ministry): 2022SKL014; Key R&D and transformation plan of Qinghai Province: 2022QY203; Program for Innovative Research Team (in Science and Technology) in University of Henan Province: 22IRTSTHN016; funding scheme of Key scientific research of Henan's higher education institutions: 23A520010; Key R&D and promotion Special Project of Science and Technology Department of Henan Province: 222102210104; teaching reform research and practice project of higher education in Henan Province: 2021SJGLX502.

Data Availability Statement: All data generated or analyzed during this study are included in this article.

Acknowledgments: The authors would like to express their gratitude to the editors and anonymous reviewers for their comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Huang, J.; Kang, S.; Ahuja, N.; Kopf, J. Image completion using planar structure guidance. *ACM Trans. Graph. (TOG)* **2014**, *33*, 1–10. [CrossRef]
2. He, K.; Sun, J. Image completion approaches using the statistics of similar patches. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2423–2435. [CrossRef]
3. Li, H.; Hu, L.; Hua, Q.; Yang, M.; Li, X. Image Inpainting Based on Contextual Coherent Attention GAN. *J. Circuits Syst. Comput.* **2022**, *31*, 2250209. [CrossRef]
4. Jam, J.; Kendrick, C.; Walker, K.; Drouard, V.; Hsu, J.G.S.; Yap, M. A comprehensive review of past and present image inpainting methods. *Comput. Vis. Image Underst.* **2021**, *203*, 103147. [CrossRef]
5. Qin, Z.; Zeng, Q.; Zong, Y.; Xu, F. Image inpainting based on deep learning: A review. *Displays* **2021**, *69*, 102028. [CrossRef]
6. Zhang, X.; Zhai, D.; Li, T.; Zhou, Y.; Yang, L. Image inpainting based on deep learning: A review. *Inf. Fusion* **2023**, *90*, 74–94. [CrossRef]
7. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '00, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424. [CrossRef]
8. Yan, J.; Chen, B.; Guo, R.; Zeng, M.; Yan, H.; Xu, Z.; Wang, Y. Tongue Image Texture Classification Based on Image Inpainting and Convolutional Neural Network. *Comput. Math. Methods Med.* **2022**, *2022*, 6066640. [CrossRef]

9. Pathak, A.; Karmakar, J.; Nandi, D.; Mandal, M.K. Feature enhancing image inpainting through adaptive variation of sparse coefficients. *Signal Image Video Process.* **2022**, 1–9. [CrossRef]
10. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef] [PubMed]
11. Guleryuz, O. Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising-part I: Theory. *IEEE Trans. Image Process.* **2006**, *15*, 539–554. [CrossRef]
12. Li, Z.; Chen, A.; Miao, T. A fingerprint removal method based on fractal–criminisi technology. *Fractals* **2022**, *30*, 2250157. [CrossRef]
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Neural Information Processing Systems*; MIT Press: Montreal, QC, Canada, 2014.
14. 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]
15. Zhang, X.; Wang, X.; Shi, C.; Yan, Z.; Li, X.; Kong, B. De-gan: Domain embedded gan for high quality face image inpainting. *Pattern Recognit.* **2022**, *124*, 108415. [CrossRef]
16. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Trans. Vis. Comput. Graph.* **2022**. [CrossRef]
17. Sun, T.; Fang, W.; Chen, W.; Yao, Y.; Bi, F.; Wu, B. High-resolution image inpainting based on multi-scale neural network. *Electronics* **2019**, *8*, 1370. [CrossRef]
18. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *Comput. Ence* **2015**, *1511*, 06434.
19. Chen, X.; Zhao, J. Improved semantic image inpainting method with deep convolution generative adversarial networks. *Big Data* **2022**, *10*, 506–514. [CrossRef] [PubMed]
20. Hu, J.; Wang, H.; Wang, J.; Wang, Y.; He, F.; Zhang, J. SA-Net: A scale-attention network for medical image segmentation. *PLoS ONE* **2021**, *16*, e0247388. [CrossRef]
21. Rong, L.; Li, C. Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3675–3684.
22. Guan, S.; Hsu, K.T.; Eyassu, M.; Chitnis, P.V. Dense dilated UNet: Deep learning for 3D photoacoustic tomography image reconstruction. *arXiv* **2021**, arXiv:2104.03130.
23. Jing, J.; Wang, Z.; Rätsch, M.; Zhang, H. Mobile-Unet: An efficient convolutional neural network for fabric defect detection. *Text. Res. J.* **2020**, *92*, 30–42. [CrossRef]
24. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
25. Zeng, Y.; Lin, Z.; Yang, J.; Zhang, J.; Shechtman, E. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision 2020*; Springer: Cham, Switzerland, 2020; pp. 1–17.
26. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
27. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lui, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6836–6846.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
29. Kaur, G.; Sinha, R.; Tiwari, P.K.; Yadav, S.K.; Pandey, P.; Raj, R.; Rakhra, M. Face mask recognition system using CNN model. *Neurosci. Inform.* **2022**, *2*, 100035. [CrossRef] [PubMed]
30. Yuan, F.; Zhang, Z.; Fang, Z. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognit.* **2023**, *136*, 109228. [CrossRef]
31. Han, Q.; Liu, J.; Jung, C. Lightweight generative network for image inpainting using feature contrast enhancement. *IEEE Access* **2022**, *10*, 86458–86469. [CrossRef]
32. Maeda, H.; Kashiyama, T.; Sekimoto, Y.; Seto, T.; Omata, H. Generative adversarial network for road damage detection. *Comput.-Aided Civ. Infrastruct. Eng.* **2021**, *36*, 47–60. [CrossRef]
33. Li, H.; Zheng, Q.; Yan, W.; Tao, R.; Wen, Z. Image super-resolution reconstruction for secure data transmission in Internet of Things environment. *Math. Biosci. Eng.* **2021**, *18*, 6652–6672. [CrossRef]
34. Lu, Y.; Chen, D.; Olaniyi, E.; Huang, Y. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Comput. Electron. Agric.* **2022**, *200*, 107208. [CrossRef]
35. Xiang, H.; Zou, Q.; Nawaz, M.A.; Huang, X.; Zhang, F.; Yu, H. Deep learning for image inpainting: A survey. *Pattern Recognit.* **2023**, *134*, 109046. [CrossRef]
36. Sun, Q.; Zhai, R.; Zuo, F.; Zhong, Y.; Zhang, Y. A Review of Image Inpainting Automation Based on Deep Learning. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2022; Volume 2203, p. 012037.

37. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 7354–7363.
38. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
39. Xue, Y.; Xu, T.; Zhang, H.; Long, L.R.; Huang, X. SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation. *Neuroinformatics* **2018**, *16*, 383–392. [CrossRef]
40. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 694–711.
41. Tao, C.; Gao, S.; Shang, M.; Wu, W.; Zhao, D.; Yan, R. Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 4418–4424.
42. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
43. Liu, G.; Reda, F.; Shih, K. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.
44. Chen, C.; Fragonara, L.; Tsourdos, A. GapNet: Graph attention based point neural network for exploiting local feature of point cloud. *arXiv* **2019**, arXiv:1905.08705.
45. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369.