Hindawi
Mobile Information Systems
Volume 2022, Article ID 5165962, 11 pages
https://doi.org/10.1155/2022/5165962

Research Article

Multichannel Cross-Scale Semantic Coherent Attention Network for Image Inpainting

Changjun Zou and Lintao Ye

East China Jiaotong University, Nanchang 330013, China

Correspondence should be addressed to Changjun Zou; zoucj2006@163.com

Received 15 October 2022; Revised 13 December 2022; Accepted 16 December 2022; Published 31 December 2022

Academic Editor: Zahid Mehmood

This paper investigates a cross-scale space semantic feature coherent image inpainting approach since it is challenging for the existing image inpainting methods to fuse the semantic feature information effectively. Firstly, we learn the feature semantic relevance step-by-step from the high-level semantic feature map's attention mechanism and then we apply what we have learned to the preceding low-level feature map. In order to preserve the visual and semantic coherence of image repair, the missing content can be filled by changing attention from deep to shallow in a multiscale manner. A broader receptive field is generated by partial convolution, and semantic feature relevance is achieved using a multiscale cross feature space feature attention mechanism based on semantic attention. This technique improves the extensibility and continuity of the restored images by reconstructing the semantic information of different feature spaces, not only taking into account the reuse of existing semantic space features but also including across feature spaces. The experimental results demonstrated an improvement in PSNR, SSIM, and L1 performance by 10.50%, 0.13%, and 47.09%, respectively, with clear benefits.

1. Introduction

In order to make the restored image look very natural and be difficult to tell apart from the undamaged image, image inpainting requires the algorithm to fill up the missing areas of the image in accordance with the image itself or the training set information. According to existing researches, it will be quite obvious as long as there is even a tiny discrepancy between the filled content and the undamaged area. As a result, in order to achieve high-quality image inpainting, it is necessary for both the content semantics and the generated image texture to be sufficiently real and clear.

Currently, there are two primary categories in which image inpainting techniques fall: the first is the traditional texture generation method. The fundamental concept is to fill in the missing area by selecting identical pixel patches from the area of the image that is undamaged. The alternative approach uses deep learning to encode the image as a feature of highly dimensional hidden space, which is subsequently decoded to provide a fully recovered image. The missing areas of the damaged image must be filled in with appropriate information in order to achieve high-quality image inpainting. The present approaches either generate semantically consistent patches from the context of the region or fill the region by replicating image patches, oblivious to the importance of both visual and semantic credibility. As a result, these two techniques have some drawbacks when it comes to maintaining adequate semantics and distinct texture.

The motivation of this research is to further enhance the semantic consistency of image restoration, gradually understand the regional semantic relevance from the attention in the high-level semantic feature map, and apply the understood attention to the prior low-level feature map. It can guarantee the visual and semantic coherence of image repair since the missing content can be filled by moving attention from deep to shallow in a multiscale manner. Besides, attention mechanism in neural network is a resource optimization allocation scheme that assigns computing resources to more important tasks first and solves the

problem of information overload when computing resources are limited, especially for the automobile systems [1, 2] in self-driving application.

We have developed a robust strategy for learning semantic feature maps across feature spaces. For missing areas, the generation model can produce results with semantic consistency. We proposed a framework for multiscale image inpainting based on a deep learning model; it emphasizes a cross-scale semantic correlation image inpainting technique that takes into account both the current feature scale space and the cross-scale feature space. By utilizing a cross feature space feature attention mechanism and semantic attention mechanism, we achieved semantically-coherent image restoration.

This approach achieves high-quality images by realizing image restoration from a semantic standpoint and combining multiscale feature space information. Additionally, the results of the experiment demonstrate that our technique performs better in terms of PSNR, SSIM, and L1 performance metrics. Our primary contributions are as follows:

(1) In this paper, a cross-scale method for semantically-coherent image restoration with four scales is proposed. Cross-scale semantic feature extraction is realized with our novel method. High-quality image restoration with semantics coherence is achieved through our search and generation strategy.

(2) A reconstruction module called cross-scale coherent semantic attention (CCSA) is proposed. Attention score to reconstruct the sibling features of the lower-level semantic network module is calculated. Reasoning operations is utilized to depict the useful regions. With this technique, the semantic features of several feature spaces can be combined, and the feature information is then transferred to the subsequent layer for feature fusion. The experimental results demonstrate that the cross-scale reconstruction technique improves PSNR, SSIM, and L1 performance by 5.34%, −0.14%, and 33.86%, respectively.

(3) A semantic residual attention (SRA) module is proposed, which could further enhance the network's performance and increase the semantic coherence of image restoration through the semantic residual structure, as well as reducing network residual error. This approach enhances PSNR and L1 performance by 3.43% and 27.51%, respectively.

2. Related Research

The approaches for image restoration could mainly be divided into two types. The first one is the classical texture synthesis method, while the second is the deep learning method [3].

2.1. Classical Matching Approach. Training set is not required for such method, for example, the DIP approach [4]. Only one damaged image is needed for the entire procedure, which may then be utilized for image restoration. The TV (total variation) model [5] was enhanced with the CDD model [6], which addresses the issue with the TV model's inability to restore the visual connectedness of images. When attempting to find the best match using Criminsi's traditional violent block matching method [7, 8], the outcome is not always pleasing. Because we only consider how closely the portion outside the hole matches the other images when looking for the best match. Barnes' PatchMatch [9] is a very clever patch matching technique that accelerates patch matching by taking use of the local correlation of images. Although this technique can attain the overall approximate optimal, it cannot guarantee that every patch will find the best match. Because they require a lot of processing to achieve pixel level filling and patching, these traditional approaches are typically slow. The absence of semantic knowledge and in-depth understanding of visuals are another significant flaw in such methods. The restoration of complicated semantic scenes cannot be handled by this strategy, and it is difficult to produce semantically plausible solutions.

2.2. Deep Learning-Based Regular Filling. An unsupervised visual feature learning system driven by context-based pixel prediction is Pathak's context-encoder [10–12]. It can generate acceptable results for semantic filling and it is used to generate content for any image area based on its surroundings. Global and local discriminators are introduced by the GL technique [13, 14]. Local and global consistent images can be produced using this technique. Any shape of a missing region can be filled using the entire convolution neural network. This strategy has greater benefits than patch-based approaches such as PatchMatch [9]. The color difference, blurring, and other flaws are improved by Liu's partial convolution technique [15, 16]. There are certain benefits to this paradigm for irregular holes.

2.3. Deep Learning-Based Progressive Filling. For instance, edge guided repair methods [17, 18] needs the determination of the edge in advance, and various parameters will result in varied edge features, which will influence the repair results. The prior one shot fill model is not the same as the RFR model [19]. The RNN framework and this network are comparable. The first input to the RFR module is the feature map, and the second input to the RFR module is the output results. After a number of cycles, the subsequent stage of feature fusion will be initiated in this manner. In Zhang's PGN [20], progressive filling at the image level was accomplished by connecting GANs together using LSTM. With partial convolution and expansion methods, Guo's FRRN [21] stacks 8 full resolution residual modules to achieve progressive filling. These processes frequently require a lot of computational resources and are time-consuming.

2.4. Attention-Based Deep Learning. The deep learning model could produce semantically consistent results for missing areas by utilizing advanced semantic feature

learning. Nevertheless, it remains difficult to get aesthetically realistic outcomes from small potential features. Using the similar texture of the feature map source area to fill in the target area, Yu's Deepfillv1 [22] method, for instance, proposes an improved GCA structure based on contextual attention [23]. The content learned from the contextual attention layer is the key feature information which is could be used to repair the missing area for a damaged image. Gated convolution is used in the enhanced Deepfillv2 [24]. When the damaged area is in free form, the gated convolutions are optimized to produce gaps near the filling edge. It is suggested to divide the image into patches and then identify each local region using a spectral normalized discriminator. The attention transfer network (ATN), which is designed to transfer the features of the known area to the missing area to achieve a better filling impact, is used in the pyramid type layer-by-layer repair [25], the generator adopts the structure of encoding and decoding, and the encoder adopts the pyramid type encoder. Diversified repair [26–28] developed a novel framework based on the probability principle that combines prior conditions and potential variables and has several parallel paths in order to produce multivariate results with appropriate confidence. The image is changed into a hidden space by a variational automatic coder [29, 30], and an image restoration operation is then carried out in the hidden space. According to the realistic and diversity dynamic balancing repair approach [30], pixels near the hole center should have more degrees of freedom while those close to the hole edge should be more predictable. It can dynamically balance the authenticity and diversity within the missing area [31], making the generated content more diversified towards the hole center and the hole boundary more similar to the adjacent image content. By learning this patch match behavior to a generator without attention through joint training to assist context reconstruction tasks, Zeng et al. [32] proposed the context reconstruction assisted repair and encouraged the generated output to be reasonable even when it is reconstructed from the surrounding areas. Wide-ranging focus [33], a novel attention perception layer (AAL), is introduced to better use the high-frequency properties of long-distance correlation in order to enhance the appearance consistency between the visible region and the generated area.

Few studies have been conducted on multiscale semantic feature fusion, and the majority of approaches now in use only take into account of image restoration with one scale. Therefore, it is important to investigate semantic consistency image inpainting techniques from a cross-scale space perspective.

3. Our Method

3.1. Overall Structure of Our Method. Figure 1 depicts the overall structure of our network, which is primarily composed of several basic blocks as shown in Figure 2 (BBs) connected by cross-scale coherent semantic attention (CCSA) and semantic residual attention (SRA) blocks. The present scale's feature information is learned by each BB individually, and the semantic coherent attention module

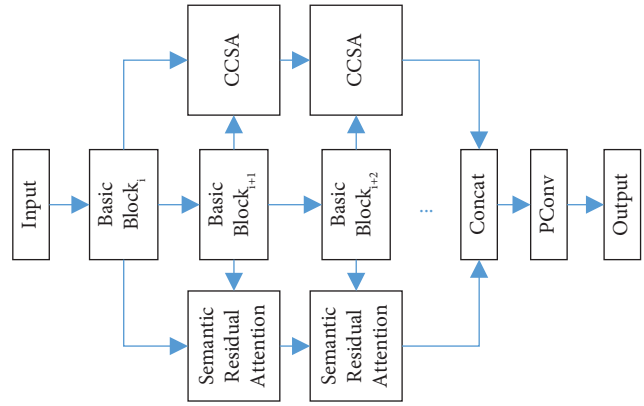


Figure 1: Overall structure of our cross-scale coherent semantic attention network.

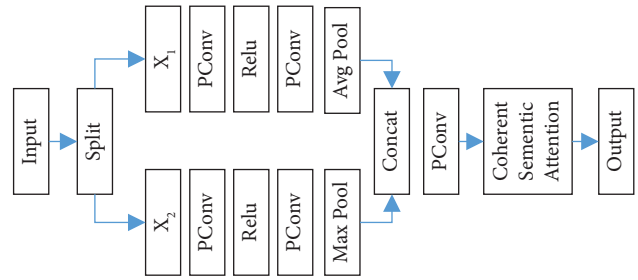


Figure 2: Structure of our basic block.

and semantic residual attention module connect several scale space. We split each input into two paths in each BB structure. Pixel-wise concat is used to combine the output from last two BB blocks. To restore more information while maintaining visual performance, the two channels are pooled maximum and on average.

The semantic correlation attention method in the backbone network realizes the cross-scale semantic correlation learning. This cross-scale semantic correlation can make use of feature at several scales. The main purpose of this structure is to achieve the cross-scale propagation of feature information between two adjacent BBs. In our network, four BBs are included, each BB represents a distinct scale space and essentially satisfies the requirements, and three cross-scale attention structures are consequently needed to enable semantic feature transmission.

In order to further reduce the semantic residuals between modules and enhance network performance, the semantic residual attention module mainly realizes the transmission of semantic residuals across adjacent BB modules. The experimental results demonstrate that the introduction of the semantic residual module improves the network's overall performance, demonstrating the semantic residual module's value in raising the semantic residual of the network.

Search and generation are the two key steps in the realization of semantic attention learning. Image restoration with semantic cross scale and associated functionality is realized. Our network does not directly employ the

convolutional layer for feature learning. Instead, we employ partial convolution to achieve a bigger receptive field and boost learning effectiveness even more.

3.2. Feature Reconstruction Based on Semantic Coherent Attention. We believe that it is insufficient to reconstruct M solely by taking into account the relationship between M and M' (which represent the known area and the missing area in the feature map, respectively) in the feature map, as this ignores the correlation between the generated image patches, particularly the semantic correlation, which may result in lacking ductility and continuity in the restoration results.

We investigate the semantic residual and semantic correlation between the generated restoration image blocks in order to resolve this weakness and propose a SCA layer. As for illustration, the SCA layer implementation includes search and generation steps. Figure 3 illustrates how the SCA layer works, with M and M' representing, respectively, the known area and the missing area in the feature map.

In order to initialize m_i during the search, the RSA layer searches for the closest matching context patch m_i in the known region M for the i^{th} patch m_i in M.

Then, in order to recover the m_i during generation, we set m_i as the primary component and all previously generated patches as the secondary part. The two sections' weights are determined using the following cross-correlation measures:

$$\begin{aligned} D\max_i &= \frac{\langle m_i, m_i' \rangle}{\|m_i\| \cdot \|m_i'\|}, \\ D\text{ad}_i &= \frac{\langle m_i, m_{i-1} \rangle}{\|m_i\| \cdot \|m_{i-1}\|}, \end{aligned} \tag{1}$$

where $D\text{ad}_i$ denotes the similarity between two created adjacent patches and $D\max_i$ represents the similarity between m_i and the context area's most similar patch m_i'. The weights of the context patch part and all previously created patch parts are normalized as $D\max_i$ and $D\text{ad}_i$, respectively. The following are the two steps.

3.2.1. Search. In order to apply the convolution filter to M, we first extract the patch from M and transform it to a convolution filter. We can find the correlation between each patch in M and every patch in M by using this procedure. Based on this, we initialize each generated patch m_i with the context patch $m_{i,}$ which is the most comparable to it and, for the subsequent operation, we give it the maximum cross-correlation value $D\max_i$.

3.2.2. Generation. We start the generation process from the upper left patch of M (marked with m_1 in Figure 3). Dad_1 is 0 and m_1 has never had a patch, so we simply replace m_1 with m_1', making $m_1 = m_1'$. Although the preceding patch, m_1, serves as an additional reference for the subsequent patch, m_2, we treat m_1 as a convolution filter in order to get the cross-correlation measure, $D\text{ad}_2$, between m_1 and m_2. Then, to update the m_2 value, $D\text{ad}_2$ and $D\max_2$ are merged and adjusted to weights of m_1 and m_2, respectively. The steps of the generation process, from m_1 to m_n, can be summed up as follows:

$$\begin{cases} m_1 = m_1', & D\text{ad}_1 = 0, \\ m_i = \dfrac{D\text{ad}_i}{D\text{ad}_i + D\max_i} \times m_{i-1} + \dfrac{D\max_i}{D\text{ad}_i + D\max_i} \times m_i', & i \geq 2. \end{cases} \tag{2}$$

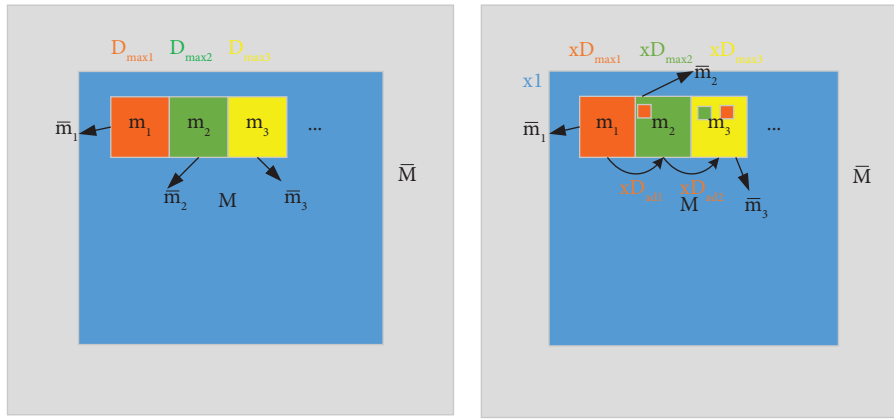
This process is a recursive process. The method described above can be used to determine the repair area.

3.3. Image Reconstruction Based on Cross-Scale Semantic Coherent. We propose employing the semantic correlation feature between high-level and low-level semantic modules to reconstruct feature maps in order to preserve as much low-level semantic information as possible. We utilize some reasoning operations to depict the useful regions since we are confident that the high-level semantic network module must deal with smaller missing regions (relative to low-level regions). In more detail, the low-level semantic module's feature map's patches are deconvoluted using the similarity score, which is then used to reconstruct the filled feature map based on the features from the high-level semantic module.

Formally, we assume that the cross-scale semantic feature reconstruction network's i-layer feature of the j^{th} network module is f_{ij}. The following definition enumerates the sibling features shared by nearby modules:

$$\text{sim}^{i,j}_{x,y,x',y'} = \left\langle \frac{f^{i,j}_{x,y}}{\left\| f^{i,j}_{x,y} \right\|}, \frac{f^{i,j}_{x',y'}}{\left\| f^{i,j}_{x',y''} \right\|} \right\rangle, \tag{3}$$

where $\text{sim}^{r,i,j}_{x,y,x',y'}$ is the measure of similarity between (x, y) and (x', y') that is unknown. The adjacent pixels are smoothed to further enhance the continuity and smoothness between them:



(a) (b)

Figure 3: Schematic diagram of semantic correlation calculation procedure. (a) Search. (b) Generation.

$$\mathrm{sim}_{x,y,x',y'}^{r,i,j} = \frac{\sum_{p,q\in\{-k,\ldots,k\}} \mathrm{sim}_{x+p,y+q,x',y'}^{i,j}}{k\times k}. \tag{4}$$

The output mapping of hierarchical modules represents various semantic levels for various semantic properties. In order to keep the semantic information from the preceding module, we additionally include a trainable parameter λ.

$$\mathrm{score}_{x,y,x',y'}^{i,j} = \lambda\cdot \mathrm{score}_{x,y,x',y'}^{i,j} + (1-\lambda)\cdot \mathrm{score}_{x,y,x',y'}^{i,j-1}. \tag{5}$$

Finally, the final attention score was used to reconstruct the sibling features of the lower-level semantic network module as follows:

$$\widehat{f}_{x,y}^{i,j-1} = \sum_{x'\in\_1\ldots W\},y'\in\{1\ldots H\}} \mathrm{score}_{x,y,x',y'}^{i,j} f_{x',y'}^{i,j-1}. \tag{6}$$

3.4. Multiscale Feature Selection and Fusion. A deeper module is then employed to extract features from the feature map. Cross-scale procedures can keep the deep network's low-level semantic information flowing. It might, however, include some deceptive background details. With this technique, we intended to use multiscale feature extraction to extract information from a wide receptive field. Four distinct scales are employed to extract features. To preserve the balance between performance and efficiency, we specifically use distinct expansion rates for different scale extractions to obtain a 3×3 convolution kernel. We consider the convolution operation g_r^k, which has a kernel size of k and an expansion rate of r. Thus, the following is a definition of the feature selection operation:

$$ff_{x,y}^{i,j} = \sum_{k,r\in\{1,2,4,8\}} g_r^k\left(f_{x,y}^{i,j}\right), \tag{7}$$

where $ff_{\max}^{i,j}$ and $ff_{\mathrm{mean}}^{i,j}$ are the maximum and average values for each channel that must also be determined, respectively. The computation of each scale's attention score $s_{r,x,y}^{k,i,j}$ may then be performed, where the scale and the value are [1, 2, 4, 8]. Finally, the following formula can be used to get the cumulative output:

$$\widehat{F}_{x,y}^{i,j} = g_r^k\left(f_{x,y}^{i,j}\right)\times s_{r,x,y}^{k,i,j}. \tag{8}$$

Low-level semantic information may be lost and low-level semantics may be destroyed when feature mapping travels through low-level semantic modules. In order to ensure that low-level semantic information can be transmitted throughout the network, the high-level semantic module must be paired with the low-level semantic feature module. In order to achieve this purpose, we reconstruct the feature as well as the feature from the high-level semantic module to link through the channel, and the core size is 1×1 as the output feature. The output characteristic can be expressed as follows, assuming that the original input characteristic is given as F:

$$F_{x,y}^{i,j} = \Phi\left(\left|F_{x,y}^{i,j}, \widehat{F}_{x,y}^{i,j}\right|\right). \tag{9}$$

4. Experiment Results

4.1. Training Platform, Data, and Evaluation Metrics. This research compares regularly used test datasets in order to validate our image inpainting strategy. Urban100 [34], DTD [35], and CelebA [36], are the test datasets. The main training performance evaluation metrics are PSNR [37], SSIM [38], and L1 error. PSNR is a peak signal-to-noise ratio that serves as an objective measure for image evaluation. PSNR is the most popular and widely used approach for evaluating image quality objectively. The structural similarity index (SSIM) is an image quality evaluation metric that compares image brightness, contrast, and structure. The training platform and related parameters employed in this technique are shown in Table 1.

Training settings: Adam, learning_rate = 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and Keras 2.7 training platform.

In the experiments comparing the performance of other methods, the comparison methods were replicated

Table 1: Training platform and related parameters.

CPU	Intel i9 12900K	GPU	RTX3080
CPU memory	64 G, DDR5, 4800 MHz	GPU memory	GDDR6 12 G
Operation system	Windows 10	Training platform	Keras 2.7

according to the literature, and the training process was conducted on the same platform and data sets. The number of iterations of each training procedure is 500, and the total training epoch is 1000.

4.2. Comparison with Existing Methods. Our proposed method shows good results on various datasets, as shown in Table 2. Almost all of the metrics are optimal on the testing datasets. For PSNR metric, our method achieves the best in all data sets for different mask ratios. For SSIM metric, our method achieves the best in all data sets for different mask ratios, except on Urban100 dataset with mask ratio 0.4, our method ranks second with 0.9883, while PSR method ranks first with 0.9884. For L1 error, our method has the best performance, except on CelebA dataset, ranking second with mask ratios 0.3 and 0.4.

From this, it can be stated that, while the new technique does not produce the best results in all datasets, it does so in the majority of them, and its performance is significantly enhanced when compared to the original methods, demonstrating the new method's clear benefits.

The table compares the PSNR, SSIM, and L1 performance metrics of several approaches to more clearly illustrate the impact of comparison between them. CelebA is one of them, and it serves as the training data set. 30000 pairs of training data are obtained after preprocessing, of which 10% are used as the test data set.

Model size and inference time for the image with a size of 128 ∗ 128: Inference time for our method is 4.69 ms, which is slightly higher than that of the Pconv method and PSR method. The estimated inference time for the image with a resolution of 720P (1080 ∗ 720) should be less than 263.8 ms (56.25 patches with a resolution of 128 ∗ 128). Inference time details of different methods are shown in Table 3.

5. Ablation Experiment

The single variable control principle serves as the theoretical foundation. We control the modification of the single variable and leave the other variables unaltered in each group of trials so that the impact of a single variable or single structure on system performance can be examined.

Four groups of comparison experiments were designed in order to verify the operation of each module of the cross-scale semantic feature restoration approach. RES0ATT0 is set as the baseline, followed by RES1ATT1, RES0ATT1, RES1ATT0, and RES0ATT0. RES1ATT1 stands for using cross-scale feature attention and semantic residuals. Similar to semantic residuals, cross-scale semantic reconstruction is not included in RES0ATT0. The impact of utilizing cross-scale feature reconstruction and semantic residuals is demonstrated in RES1ATT1. The results demonstrated that the network's PSNR, SSIM, and L1 performance has improved as a result of the addition of the aforementioned two components. The PSNR, SSIM, and L1 metrics have improved by 10.5%, 0.13%, and 47.09%, respectively, over the benchmark technique RES0ATT0, with the PSNR and L1 indicators showing the most improvement. The detailed results are shown in Table 4.

The result that RES1ATT1 has a beneficial performance when compared to RES0ATT1 and RES1ATT0 indicates that the two new structures play an important role in promoting the network performance. However, with a single structure, the PSNR, SSIM, and L1 indicators improved by 5.34%, −0.14%, and 33.86% and 3.43%, −0.10%, and 27.51%, respectively, in comparison to the benchmark network. Among them, both PSNR and L1 performance metrics have improved significantly, especially the L1 performance, while SSIM indicators have decreased slightly, but the decline is almost negligible.

Figure 4 illustrates the PSNR results for various experiment settings. Semantic residual structure is added in the RES1ATT1 and RES0ATT1 strategies. When compared to the other two, PSNR of RES0ATT1 increased more quickly in the beginning, but after around 400 epochs, the rate of growth slowed down and RES1ATT1 overtook it. Similar to the RES1ATT0 approach, the RES1ATT1 method overtook the RES1ATT0 method after around 500 epochs. The RES0ATT0 approach performs the worst out of all the strategies, showing that the cross-scale semantic feature learning structure and the semantic residual structure both work well in promoting the network performance.

In Figure 5, similar findings are also illustrated. The similarity of two figures can be explained by L1 error since it indicates the overall level of inaccuracy between images. This study demonstrates the benefit of incorporating semantic residual structure and cross-scale feature attention structure by demonstrating that the L1 error of RES1ATT1 is caused by other settings.

Naturally, the SSIM indications show comparable results. The SSIM indicators are less distinguishable between the outcomes than the previous two indicators because they have been approaching saturation for a long period. As a result, no in-depth comparison of SSIM is provided here, but Table 3 shows the average value of the last 20 outcomes.

5.1. Visual Performance Comparison. This experiment investigates the visual experimental results of RDN, Deepfill, PCONV, RFR, PSR, and other approaches in order to further validate the comparison of the visual performance of various image restoration techniques. With 500 iterations of each epoch, the epoch is set as 1000 and all models run on the same training and validation datasets. PSNR, SSIM, and L1 are the primary performance evaluation metrics. The results

Table 2: Comparison result of different methods.

Index	Dataset	Urban100			DTD			CelebA		
	Mask ratio	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
PSNR↑	RDN	25.8047	26.0044	27.5418	24.8502	26.1248	28.5307	25.3175	27.4980	29.2314
	DeepFill	24.9095	24.1774	24.8977	22.9332	25.4656	25.6100	28.4158	29.3279	28.8496
	Pconv	20.5415	20.1884	20.2059	21.1439	21.1842	20.5190	24.2375	24.0591	24.8409
	RFR	34.0451	37.3325	33.4948	29.9784	28.3995	30.6775	35.5869	32.5108	29.2596
	PSR	34.2673	36.6664	36.5071	35.8867	35.2267	34.9433	37.7040	33.6886	34.2690
	Our	34.6323	39.3942	37.8651	37.6460	36.2432	35.9936	38.6528	37.4934	36.6306
SSIM↑	RDN	0.8207	0.8203	0.8689	0.7442	0.8024	0.8571	0.8251	0.8752	0.9086
	DeepFill	0.7659	0.7274	0.7610	0.6847	0.7490	0.7344	0.8815	0.8982	0.8639
	Pconv	0.6405	0.6081	0.6229	0.6126	0.6365	0.6845	0.7659	0.7585	0.8067
	RFR	0.9747	0.9788	0.9512	0.9259	0.9092	0.9513	0.9288	0.9570	0.7723
	PSR	0.9798	0.9884	0.9834	0.9767	0.9746	0.9756	0.9845	0.9713	0.9779
	Our	0.9856	0.9883	0.9846	0.9794	0.9777	0.9773	0.9851	0.9813	0.9797
L1↓	RDN	0.0261	0.0316	0.0233	0.0379	0.0308	0.0234	0.0259	0.0195	0.0141
	DeepFill	0.0412	0.0494	0.0480	0.0520	0.0505	0.0431	0.0274	0.0250	0.0231
	Pconv	0.0743	0.0847	0.0733	0.0769	0.0722	0.0812	0.0480	0.0470	0.0453
	RFR	0.0178	0.0177	0.0192	0.0173	0.0148	0.0159	0.0182	0.0145	0.0131
	PSR	0.0163	0.0148	0.0133	0.0125	0.0116	0.0161	0.0107	0.0119	0.0134
	Our	0.0135	0.0096	0.0109	0.0100	0.0114	0.0107	0.0127	0.0132	0.0116

Table 3: Model parameters for image inpainting with size of 128 ∗ 128. Inference time is taken from the average of 20 inference tests.

Method	Parameter size (M)	Inference time (ms)
Pconv	196	3.90
RFR	119	10.15
PSR	110	3.12
OUR	100	4.69

Table 4: Ablation experiment result.

Method	Res	Attention	PSNR↑	SSIM↑	L1↓
RES1ATT1	√	√	37.6460	0.9794	0.0100
	Improvement compared with baseline		10.50%	0.13%	47.09%
RES0ATT1	×	√	35.8867	0.9767	0.0125
	Improvement compared with baseline		5.34%	−0.14%	33.86%
RES1ATT0	√	×	35.2383	0.9771	0.0137
	Improvement compared with baseline		3.43%	−0.10%	27.51%
RES0ATT0	×	×	34.0685	0.9781	0.0189
			—	—	—

↑ means the bigger, the better; ↓ means the smaller, the better.

of the experiments demonstrate that the strategy based on cross-scale semantic feature attention produces the best performance.

As an illustration, Figure 6 shows that our novel technique generates results that are 33.518, 0.982, and 0.014, while the RFR method and PSR method yield results that are 29.602, 0.930, and 0.023 and 32.372, 0.966, and 0.016, respectively. The outcomes demonstrate that our strategy outperforms other methods in terms of performance metrics. Figures 6~9 display the similar outcomes.

6. Conclusions

In view of the difficulty in semantic level image inpainting in previous image repair methods, this paper proposes a cross-scale semantic feature image repair method to

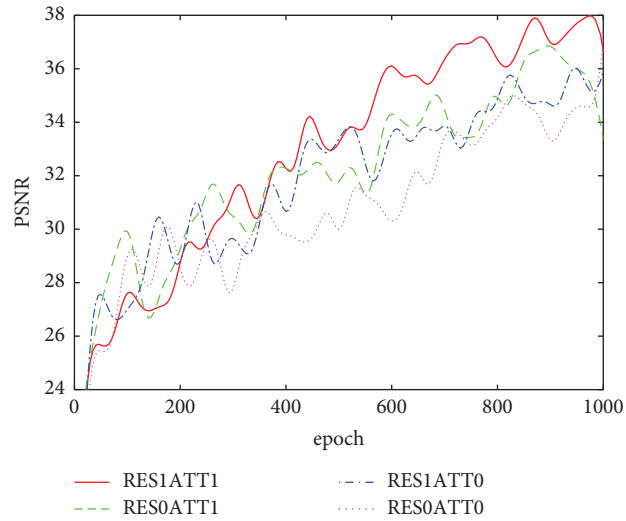


Figure 4: PSNR result for different setting.

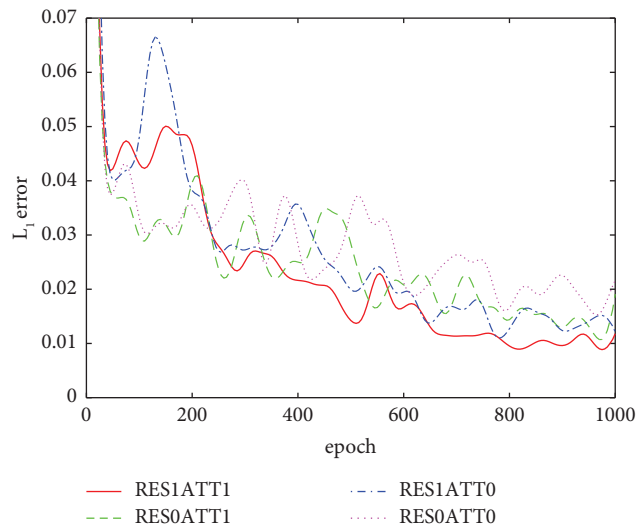


Figure 5: L1 error result for different setting.

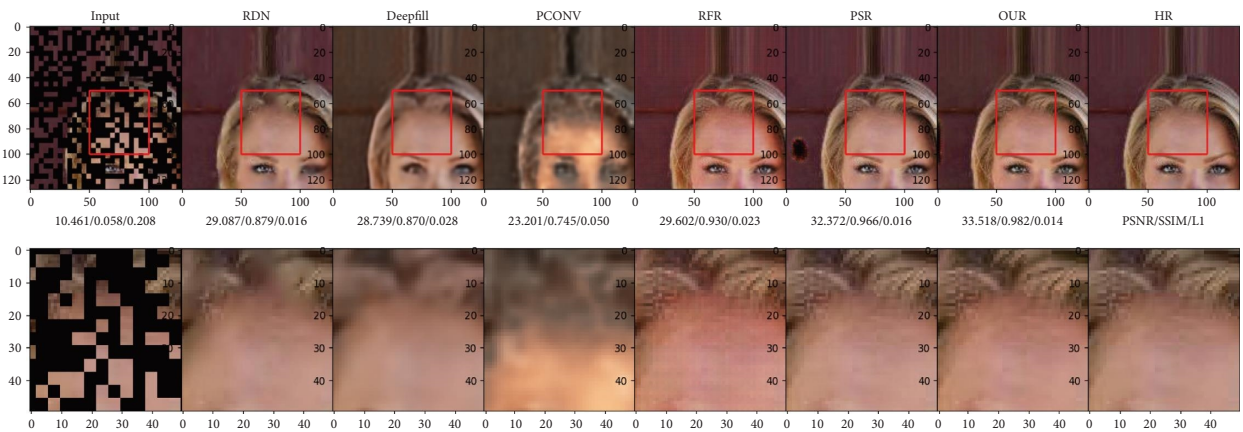


Figure 6: Inpainting result of Image 3 from dataset CelebaA.

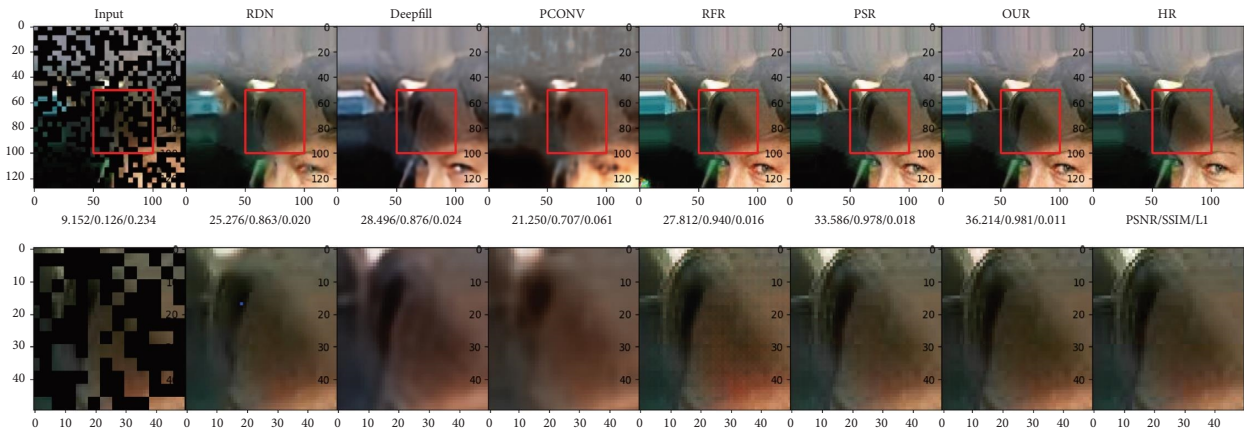


Figure 7: Inpainting result of Image 14 from dataset CelebaA.

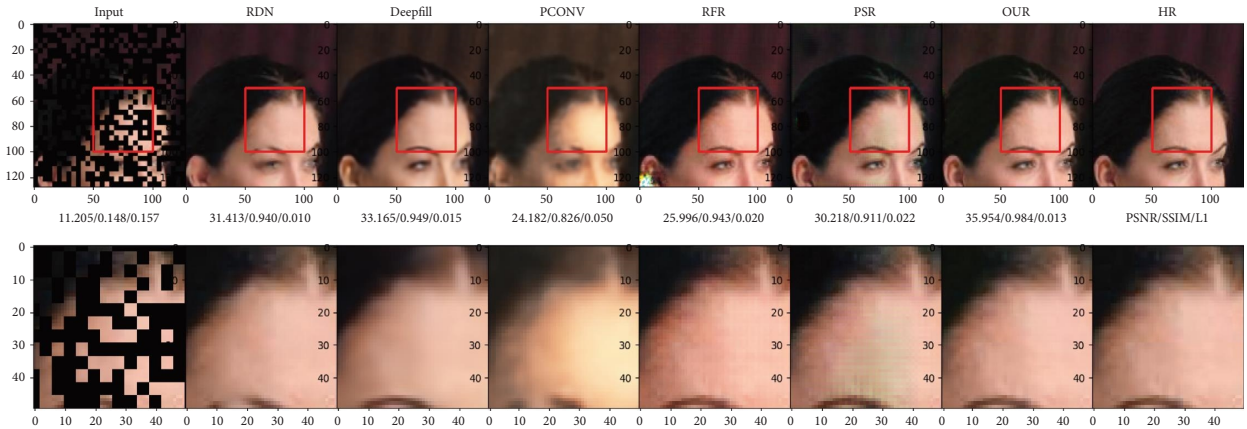


Figure 8: Inpainting result of Image 195 from dataset CelebaA.

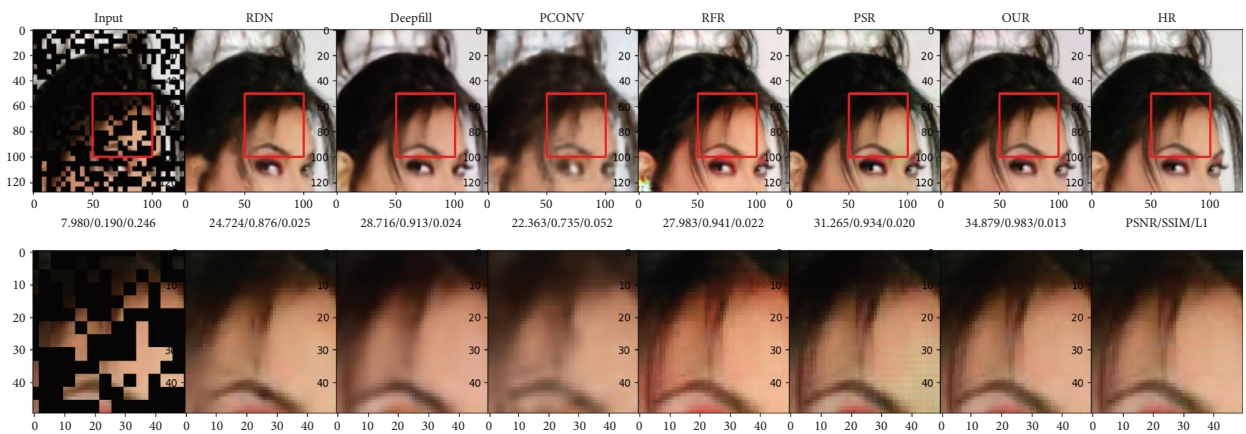


Figure 9: Inpainting result of Image 219 from dataset CelebaA.

improve the lack of ductility and continuity of the existing methods.

This approach can capture semantic feature information from various scale space in addition to the semantic feature information of the current scale space, which can help the image inpainting process. Higher quality image restoration is possible using the semantic feature information. The results of the experiment indicated that integrating the cross-scale semantic feature restoration method can accelerate the spread of semantic features, which is advantageous for the application of semantic level image restoration.

Data Availability

Data are available on request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge the support from the National Natural Science Foundation of China (No. 62162027); Science and Technology Project of Jiangxi Provincial Department of Education (No. GJJ210646); and Key R&D Projects of Jiujiang City (No. 2020069).

References

[1] J. Zhao, X. Sun, Q. Li, and X. Ma, "Edge caching and computation management for real-time internet of vehicles: an online and distributed approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 4, pp. 2183–2197, April 2021.

[2] J. Zhao, Q. Li, Y. Gong, and K. Zhang, "Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7944–7956, 2019.

[3] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: a review," *Neural Processing Letters*, vol. 51, no. 2, pp. 2007–2028, 2020.

[4] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, Salt Lake City, UT, USA, June 2018.

[5] M. Fuchs and J. Müller, "A higher order TV-type variational problem related to the denoising and inpainting of images," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 154, pp. 122–147, 2017.

[6] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *Journal of Visual Communication and Image Representation*, vol. 12, no. 4, pp. 436–449, 2001.

[7] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based inpainting," *IEEE Transactions on Image Processing*, vol. 13, pp. 1200–1212, 2004.

[8] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, WI, USA, June 2003, https://www.microsoft.com/en-us/research/publication/object-removal-by-exemplar-based-inpainting.

[9] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: a randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–11, 2009.

[10] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "context encoders: feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2016*, Vegas, NV, USA, June 2016.

[11] R. Gao and K. Grauman, "On-demand learning for deep image restoration," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1095–1104, Venice, Italy, October 2017.

[12] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in *Proceedings of the 16th European Conference*, pp. 725–741, Glasgow, UK, August 2020.

[13] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.

[14] W. Quan, R. Zhang, Y. Zhang, Z. Li, J. Wang, and D. M. Yan, "Image inpainting with local and global refinement," *IEEE Transactions on Image Processing*, vol. 31, pp. 2405–2420, 2022.

[15] G. Liu, F. A. Reda, and K. J. Shih, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, September 2018.

[16] J. Zhang, L. Niu, D. Yang et al., "GAIN: gradient augmented inpainting network for irregular holes," in *Proceedings of the 27th ACM International Conference. ACM, 2019*, pp. 1870–1878, Nice, France, October 2019.

[17] W. Xiong, J. Yu, Z. Lin, and J. Yang, "Foreground-aware image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019*, vol. 3, Long Beach, CA, USA, June 2019.

[18] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: structure guided image inpainting using edge prediction," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, October 2019.

[19] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020*, pp. 7760–7768, Seattle, WA, USA, June 2020.

[20] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, "Semantic image inpainting with progressive generative networks," in *Proceedings of the 26th ACM international conference on Multimedia*, vol. 2, no. 3, pp. 770–778, Seoul, Korea, October 2018.

[21] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, "Progressive image inpainting with full-resolution residual network," in *Proceedings of the 27th ACM International Conference on Multimedia*, vol. 2, no. 3, pp. 85–100, Nice, France, October 2019.

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5505–5514, Salt Lake City, UT, USA, June 2018.

[23] J. Yu, Z. Lin, and J. Yang, "Contextual attention image inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2018*, Salt Lake City, UT, USA, June 2018.

[24] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE International Conference on Computer Vision. 2019*, pp. 4471–4480, Seoul, Korea, November 2019.

[25] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

[26] C. Zheng, T. J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Y. Yu, F. Zhan, R. Wu et al., "Diverse image inpainting with bidirectional and autoregressive transformers," in *Proceedings*

of the 29th ACM International Conference on Multimedia 2021, pp. 69–78, Chengdu, China, October 2021.

[28] C. Zheng, G. Song, and T. J. Cham, "High-quality pluralistic image completion via code shared VQGAN," 2022, https://arxiv.org/abs/2204.01931.

[29] Z. Wan, B. Zhang, D. Chen, J. Liao, and F. Wen, "Bringing old photos back to life," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020*, June 2020.

[30] Y. Zeng, J. Fu, and H. Chao, "Aggregated contextual transformations for high-resolution image inpainting," in *Proceedings of the IEEE Transactions on Visualization and Computer Graphics*, Piscataway, NJ, USA, February 2021.

[31] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "PD-GAN: probabilistic diverse gan for image inpainting," in *Proceedings of the Computer Vision and Pattern Recognition 2021*, Nashville, TN, USA, June 2021.

[32] Y. Zeng, Z. Lin, H. Lu, and V. M. Patel, "CR-fill: generative image inpainting with auxiliary contexutal reconstruction," in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, June 2021.

[33] C. Zheng, T. J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, June 2022.

[34] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) 2015*, pp. 5197–5206, Boston, MA, USA, June 2015.

[35] L. Sharan, R. Rosenholtz, and E. H. Adelson, "Accuracy and speed of material categorization in real-world images," *Journal of Vision*, vol. 14, no. 9, pp. 12–24, 2014.

[36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, Santiago, Chile, December 2015.

[37] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800-801, 2008.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.