# Hierarchical damage correlations for old photo restoration

Weiwei Cai [a], Xuemiao Xu [a,*], Jiajia Xu [a], Huaidong Zhang [b], Haoxin Yang [a], Kun Zhang [c], Shengfeng He [d,*]

a School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China
b School of Future Technology, South China University of Technology, Guangzhou, 510641, China
c Philosophy Department, Carnegie Mellon University, 15213, Pittsburgh, The United States of America
d School of Computing and Information Systems, Singapore Management University, 178902, Singapore

Published in Information Fusion (2024) 102340. DOI: 10.1016/j.inffus.2024.102340

**Abstract**

Restoring old photographs can preserve cherished memories. Previous methods handled diverse damages within the same network structure, which proved impractical. In addition, these methods cannot exploit correlations among artifacts, especially in scratches versus patch-misses issues. Hence, a tailored network is particularly crucial. In light of this, we propose a unified framework consisting of two key components: ScratchNet and PatchNet. In detail, ScratchNet employs the parallel Multi-scale Partial Convolution Module to effectively repair scratches, learning from multi-scale local receptive fields. In contrast, the patch-misses necessitate the network to emphasize global information. To this end, we incorporate a transformer-based encoder and decoder architecture. In the encoder phase, we introduce a Non-local Inpainting Attention Module, replacing the multi-head attention, to facilitate holistic context inpainting. In the decoder phase, the Mask-aware Instance Norm Module replaces the Layer Normalization, ensuring style consistency between foreground and background. Finally, the outcomes of ScratchNet are integrated into the PatchNet pipeline to supplement contextual information hierarchically. Mining damage correlations assists in training the network in an easy-to-hard manner. Extensive experiments demonstrate the superiority of our method over state-of-the-art approaches. The code is available at https://github.com/cwyyt/Hierarchical-Damage-Correlations-for-OldPhoto-Restoration.

**Keywords**: Image inpainting, Old photo restoration, Transformer

## 1. Introduction

Photos can record some unforgettable moments like weddings, birthdays, and other memorable events. Unfortunately, these photos are gradually becoming full of flaws, such as scratches and patch-misses, due to improper preservation. Moreover, seeking assistance from professional photo restoration services is both financially burdensome and inefficient. To address this issue, a series of deep learning-based methods [1], [2], [3] have emerged in recent years. These algorithms exhibit satisfactory performance under certain conditions. While their results tend to deteriorate when confronted with intricate scenarios significantly. To describe our method explicitly, we present Fig. 1 in this section. We can observe that the results of OPBL [1] are blurry and missing details in the scratched regions. The reason is that they stack excessive ResNet blocks for extracting feature embeddings, which is bad for scratched restoration due to missing details and computing unrelated pixels. On the contrary, CSI [3] cannot restore the patch-misses perfectly due to their network being designed for learning local perceptive fields. Our method overcomes these shortcomings while promoting their respective strengths. It not only can restore thin scratches with rich details but also repair patch-misses with contextual consistency. From Fig. 1, our results are more visually pleasing than the other two methods.

Overall, they have a similar limitation, employing one network to address different damages. However, note that different damages own distinct characteristics. For instance, scratches need the network to attend to the local receptive field while patch-misses require learning global contextual information. Consequently, the network is imperative to consider the diverse damaged nature. Beyond this, prior methods neglect to exploit the potential correlations between scratches and patches. Notably, the scratches are less intricate than the patches. The results of scratch restoration can serve as supplementary knowledge for patch-wise repairing. Through these observations, we adopt an easy-to-hard learning strategy which is first to restore simple scratches and then to restore complex patch-misses.

Within our framework, it mainly includes ScratchNet and PatchNet. We first train a Multi-task Artifact Detection Network that generates distinct damaged masks (scratch mask, patch mask, overall mask). They are indispensable for subsequent restoration tasks. In our

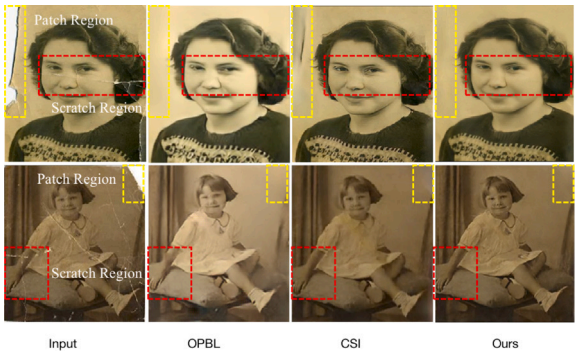


**Fig. 1. Compared with the state-of-the-art patch-wise and scratch-wise old photo restoration method** [1,3], we design a hierarchical damage-aware restoration network, which can faithfully recover the corrupted region. Zoom in for better visualization.

restoration network, we have two data flows containing scratch-wise and patch-wise inpainting pipelines. Through the above analysis, we prioritize scratches before patch-misses. In detail, the input image multiples with a patch mask to cover patched damage and concatenates scratch mask to retain underlying contextual information. Then we input these into a Siamese Encoder to obtain the scratched feature embeddings. Next, they are fed into ScratchNet. To avoid introducing much computational complexity and numerous unrelated pixels, we leverage Multi-scale Partial Convolution Module [4] for local-aware learning. Due to multi-scale scratches dispersing, we concatenate $3\times3, 5\times5$, and $7 \times 7$ partial convolutional kernels for multi-scale restoration. Finally, the restored results are regarded as supplementary assistant content to aid in following patch-wise restoration.

In PatchNet, the network requires learning holistic context to make patch-wise inpainting feasible. Consequently, we introduce transformer-based methods on account of their global learning ability. Similar to the workflow of ScratchNet, we employ overall masks to cover broken areas. Subsequently, we feed these into another Siamese Encoder to obtain the feature embeddings. Note that the two Siamese Encoders are shared parameters that mutually benefit each other. We then input these feature embeddings into the encoder of PatchNet. Specifically, we name each block as Transfill in the encoder phase and Transhomon in the decoder phase. Within each Transfill block, we introduce a Non-local Inpainting Attention Module. This module serves as a replacement for the Multi-head Attention Module and is tailored specifically for patch-wise inpainting. Considering that the restored foreground should exhibit the same style as the background, we introduce a Mask-aware Instance Norm Module within Transhomon blocks. This module replaces the Layer Norm in the transformer blocks and ensures that the entire image maintains a harmonious style. Furthermore, we establish a hierarchical fusion mechanism between the ScratchNet and PatchNet. This fusion approach enriches the contextual information, contributing to more accurate and coherent results. Overall, the main contributions of this paper are summarized as follows:

- We introduce a damage-aware hierarchical old photo restoration network that leverages ScratchNet and PatchNet to restore various damages. Furthermore, we exploit the correlations among damages by training the network in an easy-to-hard strategy.
- In our proposed framework, we introduce a Non-local Inpainting Attention Module to address the patch-misses during the encoder phase of PatchNet, incorporating global semantic context information for patch-wise restoration.
- To ensure style consistency between the damaged foreground and the background, we introduce a Mask-aware Instance Norm Module. This module replaces the Layer Norm in the decoder phase of PatchNet.
- Through comprehensive evaluation and comparison, our approach outperforms existing techniques in terms of restoration quality. Our proposed model successfully addresses various damages, yielding impressive restoration results.

## 2. Related work

### 2.1. Old photo restoration

There are some traditional algorithms for photo restoration in prior years. Yamauchi et al. [5] combine techniques from texture synthesis and image inpainting. Criminisi et al. [6] proposes the confidence of the synthesized pixel in texture synthesis, propagating like the information in inpainting. With the development of deep learning [7–12], recently, most old photo restoration methods are based on deep learning. Wan et al. [1] first propose the deep old photo restoration method, they restore the damaged region in latent space with a Non-local network [13]. The Non-local network is global attention which is appropriate for patch-missed artifacts, not scratches, because scratches generally require local context. Global semantic learning will bring redundancy pixels. Liu et al. [2] design the class-attribute guided generative adversarial network for old photo restoration. They use multiple encoders composed of identical structures. While they have no special designs for different damages. Cai et al. [3] propose a contextual-assisted photo restoration method. Nevertheless, they mainly focus on how to repair scratches rather than mixed scratches and patch-missed issues. Xu et al. [14] present a novel reference-based end-to-end learning framework that can both repair and colorize old photos. However, they mainly focus on the colorization task for black-and-white photos.

### 2.2. Image inpainting

Some image inpainting methods are relevant to our tasks. These methods can be classified into two categories: Random mask image inpainting and regular mask image inpainting, based on the shape of the inpainting mask. For the random image inpainting methods [4,15–19], they concentrate on the random mask to design the module structure. Generally, these methods are primarily based on convolutional neural networks. However, other inpainting methods [13,20–25] are developed to deal with regular masks. These methods usually restore huge holes with global perceptive fields. Consequently, both conventional photo restoration and image inpainting algorithms exhibit a shared limitation, as they lack a comprehensive framework to address mixed forms of damage. Furthermore, they tend to ignore exploiting the correlation among artifacts.

### 2.3. Vision transformer

The Transformer has garnered remarkable success in various natural language processing [26–29] tasks, primarily owing to the ability of self-attention mechanism to effectively capture long-range dependencies. Recently, several works [30–37] attempt to introduce transformers into the computer vision community. Dosovitskiy et al. [38] develop a new vision transformer (VIT) for image classification. Their approach involves splitting each image into fixed-size patches and feeding them into a standard transformer network. Based on VIT, Li et al. [32] propose a plain vision transformer for object detection. In image fusion, a two-branch network [39] with transformers is presented to fusion hyperspectral and multispectral images. However, the self-attention layer in these transformer-based methods heavily relies on learning the relationship between patches, which leads to neglecting the local detail information within these patches. To address this problem, Yang et al. [40] develop a dual-encoder unit comprising a transformer block and a pre-trained ResNet34 [41] network to extract global and local context features. A novel CNN-Transformer Blender [42] model is

**Table 1**
Detailed architecture of our network.

| Module | Block | Input Size/Chan. | Output Size/Chan. |
|---|---|---|---|
| ScratchNet | RCAB | 256 × 256/3 | 256 × 256/48 |
| | MultiPconv | 256 × 256/48 | 256 × 256/48 |
| | MultiPconv | 256 × 256/48 | 128 × 128/96 |
| | MultiPconv | 128 × 128/96 | 64 × 64/192 |
| | MultiPconv | 64 × 64/192 | 32 × 32/384 |
| PatchNet | RCAB | 256 × 256/3 | 256 × 256/48 |
| | Transfill | 256 × 256/48 | 256 × 256/48 |
| | Transfill | 256 × 256/48 | 128 × 128/96 |
| | Transfill | 128 × 128/96 | 64 × 64/192 |
| | Transfill | 64 × 64/192 | 32 × 32/384 |
| | Transhomon | 32 × 32/384 | 32 × 32/384 |
| | Transhomon | 32 × 32/384 | 64 × 64/192 |
| | Transhomon | 64 × 64/192 | 128 × 128/96 |
| | Transhomon | 128 × 128/96 | 256 × 256/48 |
| | Conv | 256 × 256/48 | 256 × 256/3 |

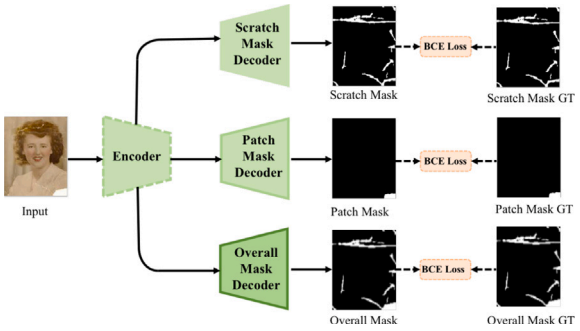


**Fig. 2. The structure of our Multi-task Artifact Detection model.** It consists of one encoder and three decoders. It can generate three different masks for downstream restoration tasks.

used for domain adaptive object detection. To learn discriminative global features and long-range dependencies, Wu et al. [43] propose a FAT-Net model for skin lesion semantic segmentation. Unlike the above methods, our method aims to effectively integrate convolution and transformer techniques to handle the complexities of old photo restoration tasks.

## 3. Method

### 3.1. Overview

#### 3.1.1. The network of multi-task artifact detection

We construct a Multi-task Artifact Detection Model to obtain three damaged masks: scratch mask, patch mask, and overall mask. They are necessary guidance to assist the old photo restoration. The architectural details are illustrated in Fig. 2. The network architecture consists of an encoder and three distinct decoders. Each decoder is assigned a different annotated ground truth mask and a BCE loss supervises every decoder. Generally, the BCE loss is used to learn binary pixel segmentation. Our detection loss contains the above three BCE losses equally. Note that we train this detection network in an end-to-end manner. The basic component is Residual Channel Attention Block [44]. It effectively compresses the number of parameters while maintaining essential features. To ensure the preservation of detailed information, we establish connections between the encoder and each decoder using a U-Net way. We also show the three distinct detected results in Fig. 3. From Fig. 3, our network successfully generates satisfactory masks, which are vital to guide the subsequent restoration process.

#### 3.1.2. The network of old photo restoration

**ScratchNet**: We propose a dedicated network called ScratchNet to address the various characteristic scratches. As shown in Fig. 4, in

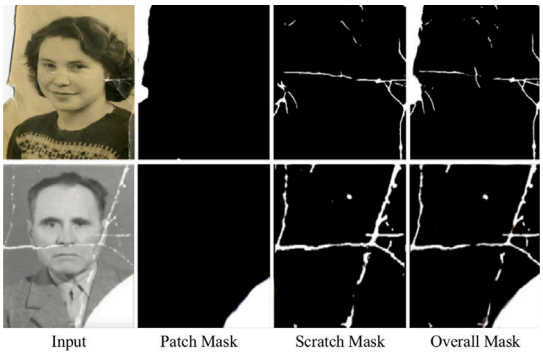


**Fig. 3. The detected results of different damage regions.** Zoom in for better visualization.

the feature fusion module, we employ input images to multiply patch masks to cover the patch-missed regions. Subsequently, these inputs are concatenated with the scratched masks to guide scratched restoration, retaining the underlying contextual information. Next, we input these into the Siamese Encoder which incorporates the Residual Channel Attention Block [45] and a 3 × 3 convolution model, to obtain the feature embeddings. It can extract high-level feature representations as well as reduce detail loss. The features are then passed through ScratchNet. Universal global learning is unreasonable for scratch restoration because it wastes computational resources and imports some unrelated pixels. Therefore, we use the Partial Convolution Module to focus on how to attend the local perceptive field. Furthermore, to effectively restore various scratches, we employ parallel Multi-scale Partial Convolution kernels with 3×3, 5×5, and 7 × 7. Additionally, we regard restored results as middle auxiliary content and feed them into PatchNet for patch-aware restoration.

**PatchNet**: In contrast to scratches, restoring patch-missed artifacts requires the network to comprehensively analyze and incorporate global information. To accomplish this, we utilize a transformer-based structure which is well-suited for capturing long-range dependencies. Additionally, we use the channel-wise transformer [46] which focuses on how to enhance the restoration quality as well as compress the computational complexity, to replace the normal spatial-wise transformer. Every block in the encoder is called Transfill in Fig. 4. It is specifically dedicated to repairing the patch-missed region. To catch more referred contextual information, we hierarchically fuse the restored results from ScratchNet with PatchNet in addition. One notable component we have introduced is the Non-local Inpainting Attention Module in Fig. 5 responsible for global inpainting. This module focuses on capturing long-range dependencies and incorporating channel-wise information to ensure comprehensive restoration.

In the decoder phase, we call every block in the decoder Transhomon in Fig. 4. We incorporate the Mask-aware Instance Norm Module, which replaces the original Layer Norm. This module draws inspiration from Instance Norm techniques [19,47], as it helps maintain style consistency between the foreground and background elements of the restored image. Therefore, it is crucial to employ the Mask-aware Instance Norm Module in Fig. 6 within our structure. Both the encoder and decoder utilize skip connections to facilitate information flow between different layers. Every encoder and decoder is composed of four Transfill/Transhomon blocks. Additionally, a 3 × 3 convolutional layer is employed to map the extracted features to the final RGB results. The detailed network structure please see Table 1.

### 3.2. Non-local inpainting attention module

Inspired by this approach [46], a method known as MDTA was proposed. This method aims to mitigate the computational burden

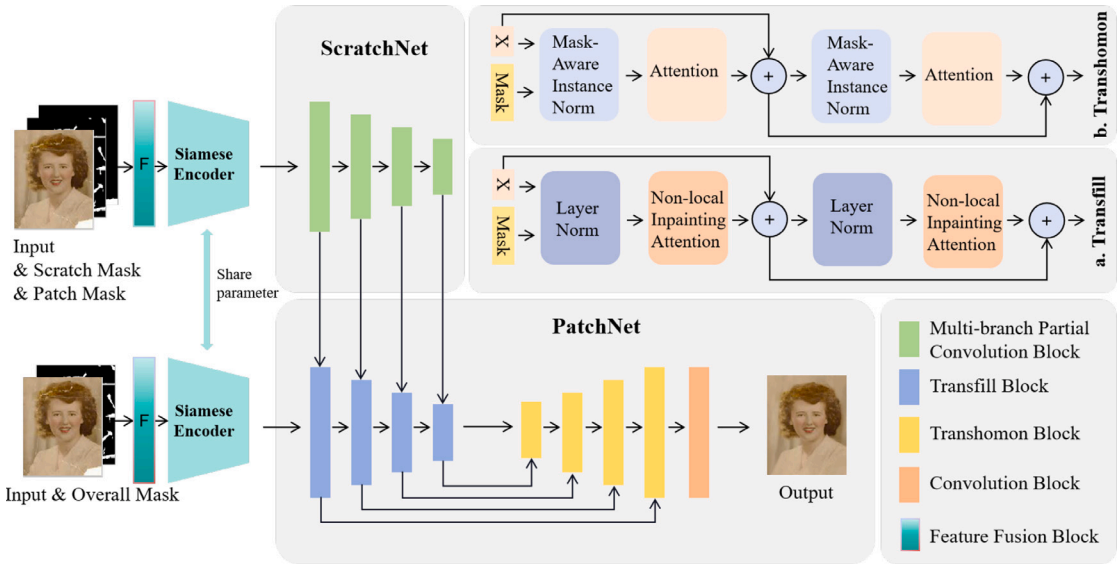


**Fig. 4. Overview of the proposed hierarchical damage-aware old photo restoration network.** The network is mainly composed of ScratchNet and PatchNet. ScratchNet is applied to restore the missed scratched content with the scratch mask, and PatchNet aims to learn to fuse the scratch-restored features to assist in the patch-missing restoration. Particularly, the PatchNet encoder consists of Transfill blocks and the PatchNet decoder consists of Transhomon blocks. Note that the overall mask indicates the mixed scratch mask and patch mask.

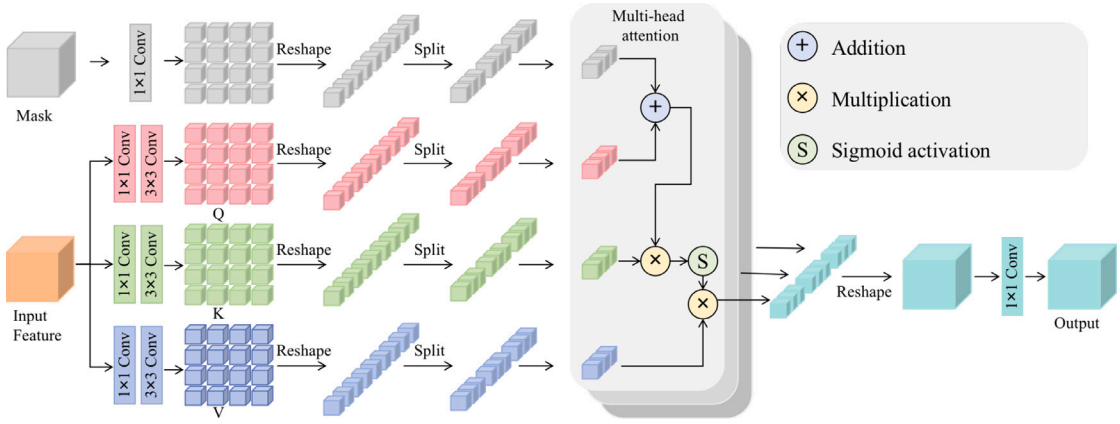


**Fig. 5. Network details of our proposed Non-local Inpainting Attention Module.**

associated with transformers. MDTA achieves this by implementing self-attention mechanisms that encompass both the channel and the spatial dimensions of the input data. Our method leverages masks to guide the network process of hole-filling. As depicted in Fig. 5, the input feature is sourced from the Siamese Encoder, and the mask is resized to match the dimensions of the input feature. Within this mask, the value of 1 represents damaged regions, while the value of 0 signifies intact areas. In the mask pipeline, the multiplication of the input feature by (1-mask) causes the attention (k×q) to become 0 in the damaged region. This leads to the damaged area not being repaired with the attention map. The mask pipeline resolves this issue by transforming 0 to 1, which is easy but effective.

To facilitate the aggregation of cross-channel context at the pixel level, $1 \times 1$ convolutions are employed. Additionally, $3 \times 3$ depth-wise convolutions are utilized to encode spatial context. The query, key, and value projections, along with the mask projection generate transposed attention maps through several interactions. This design effectively circumvents excessive large attention maps. We distribute the total number of channels into distinct heads, allowing for the simultaneous learning of separate attention maps. As illustrated in Fig. 5, we set the head as 4. Then we fuse the outputs from every head and reshape them into the original feature dimensions. Finally, the $1 \times 1$ convolution layer outputs the final projections. This approach ensures that the Non-local Inpainting Attention Module effectively leverages both channel-wise and spatial-wise information, ultimately enhancing its capacity to restore damaged regions.

### 3.3. Mask-aware instance norm module

Introduced by Yu et al. [19], they creatively proposed a spatial region-wise normalization named Region Normalization (RN) to resolve the image inpainting task. Ling et al. [47] developed the RAIN technique to harmonize foreground and background elements. We incorporate these insights into our transformer block to preserve style consistency throughout the photo. As depicted in Fig. 6, the process unfolds as follows: Initially, the input features multiply with foreground mask and background mask separately. Subsequently, the foreground features are normalized using the Instance Norm. This normalized representation is then subjected to an affine transformation, incorporating learned mean and variance parameters derived from the background features. In this way, we can harmonize the style between the restored damaged foreground and the background.

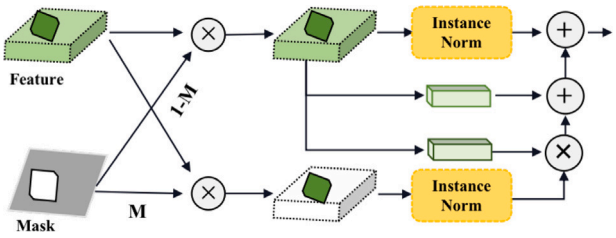


**Fig. 6. Network Details of our proposed Mask-aware Instance Norm Module.**

### 3.4. Loss function

In the old photo detection stage, we use three bce losses $\mathcal{L}_{bce}$ to train three decoder heads. In the old photo inpainting phase, we employ several loss functions which are combined with reconstruction loss $\mathcal{L}_{rec}$, perceptual loss $\mathcal{L}_{per}$, and style loss $\mathcal{L}_{sty}$ in special proportion. Moreover, we employ adversarial loss $\mathcal{L}_{adv}$ to conduct adversarial learning. We will describe these losses in the following part.

#### 3.4.1. BCE loss

To compute the pixel-wise discrepancy between the detected mask $M_{pre}$ and the ground truth mask $M_{gt}$, we utilize the binary cross entropy loss. This loss function is formulated as follows:

$$\mathcal{L}_{bce} = -(M_{gt} log(M_{pre}) + (1 - M_{pre}) log(1 - M_{pre})). \tag{1}$$

#### 3.4.2. Detection loss

To supervise the detection network, we leverage the detection loss. This loss function is formulated as follows:

$$\mathcal{L}_{det} = \lambda_s \mathcal{L}_{bce} + \lambda_p \mathcal{L}_{bce} + \lambda_o \mathcal{L}_{bce}, \tag{2}$$

where the empirical coefficients $\lambda_s = 1$, $\lambda_p = 1$, $\lambda_o = 1$.

#### 3.4.3. Reconstruction loss

To quantify the pixel-wise distance between the predicted image $I_{pre}$ and the ground truth image $I_{gt}$, we utilize the reconstruction loss. This loss function is formulated as follows:

$$\mathcal{L}_{rec} = \|(I_{pre} - I_{gt})\|_1. \tag{3}$$

#### 3.4.4. Perceptual loss

We employ perceptual loss $\mathcal{L}_{per}$ to generate high-quality results according to human perception standards. Specifically, it is captured by the pre-trained VGG-16 backbone and can be described as follows:

$$\mathcal{L}_{per} = \sum_i \frac{1}{N^i} \|\Phi_i(I_{pre}) - \Phi_i(I_{gt})\|_1, \tag{4}$$

where $\Phi_i$ denotes the $i$th layer feature map of VGG-16 network, and $N^i$ indicates the total number of activations.

#### 3.4.5. Style loss

The photo style should be consistent with the ground truth. We utilize the style loss, and we compute this loss as follows:

$$\mathcal{L}_{sty} = \|G_j^{\Phi}(I_{pre}) - G_j^{\Phi}(I_{gt})\|_1. \tag{5}$$

The dimension of the feature activation map is $C_j \times H_j \times W_j$, where $G_j^{\phi}$ is a $C_j \times C_j$ Gram matrix computed by each activation map. These activation maps are the same as those adopted in perceptual loss.

#### 3.4.6. Adversarial loss

The adversarial loss is defined as

$$\mathcal{L}_{adv} = -\mathbb{E}_{x_r}\left[log(1 - D(x_r, x_f))\right] - \mathbb{E}_{x_f}\left[log(D(x_f, x_r))\right], \tag{6}$$

where $D(x_r, x_f)$ denotes the local or global discriminator. $x_r$ and $x_f$ denote the predicted ground truth photo and scratch-free photo sampled from the training set, respectively.

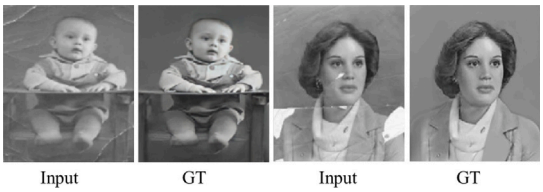


**Fig. 7. Two cases come from our dataset.**

**Table 2**
Quantitative comparison of our and other state-of-the-art methods on our dataset. The dataset includes various degradations such as scratches and patch-misses. All results are measured in terms of PSNR, SSIM, LPIPS, and FID. The best results are shown in boldface.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| MFE [20] | 28.5920 | 0.8796 | 0.1712 | 112.3100 |
| OPBL [1] | 22.4579 | 0.8422 | 0.1975 | 115.2023 |
| OPBL-Re [1] | 22.4571 | 0.8363 | 0.2159 | 105.2039 |
| HGA [49] | 26.8991 | 0.8881 | 0.1596 | 98.3006 |
| CSI [3] | 29.0911 | 0.9009 | 0.1392 | 86.7571 |
| Ours | **29.4469** | **0.9010** | **0.1223** | **72.8066** |

#### 3.4.7. Total loss

The final training objective is the weighted sum of $\mathcal{L}_{rec}$, $\mathcal{L}_{per}$, $\mathcal{L}_{sty}$ and $\mathcal{L}_{adv}$ :

$$\mathcal{L}_{su} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{sty}\mathcal{L}_{sty} + \lambda_{adv}\mathcal{L}_{adv}, \tag{7}$$

where the empirical coefficients $\lambda_{rec} = 1$, $\lambda_{per} = 0.2$, $\lambda_{sty} = 250$, $\lambda_{adv} = 0.2$.

## 4. Experiment

### 4.1. Implementation details

Our proposed network is implemented using PyTorch and all the comparative models are trained and tested on a single Nvidia RTX 3090Ti-24G GPU. We set the transformer heads in Transfill and Transhomon blocks as 1, 2, 4, and 8 separately. The attention heads are a crucial part of the transformer model that enables it to capture dependencies and relationships between different parts, contributing to its ability to handle global information effectively. We employed the Adam [48] optimizer with a learning rate of $2 \times 10^{-4}$. The masks and images for training and testing were resized to $256 \times 256$.

### 4.2. Dataset

We collect the old photos from the public websites. Then we invited an expert to annotate the masks on the image. They are constructed to train the detection network. Furthermore, we invite another expert to restore the photos and obtain the ground truth for old photo restoration. We randomly selected 176 photos for training and 68 photos for testing in all quantitative experiments. We employed various data augmentation techniques, such as horizontal/vertical flips, random cropping, color jitter, and image transpose. As a result, our training dataset was expanded to a total of 2223 photos. We illustrate two samples of the dataset as Fig. 7. For fairness, we further testify the comparison experiments with the data of PixFix [14]. Their dataset includes 200 pairs which are mainly created for old photo restoration and colorization tasks.

### 4.3. Comparisons with the state-of-the-arts

We compare the proposed approach with state-of-the-art methods, including MFE [20], OPBL [1], HGA [49], and CSI [3]. MFE and HGA are pure image inpainting methods, while OPBL and CSI are special


**Fig. 8. Compared with the state-of-the-art methods in patch-wise restoration.** The sixth column presents our results. Zoom in for better visualization.

**Table 3**
Quantitative comparison of our and other state-of-the-art methods on PixFix dataset. All results are measured in terms of NIQE [50] and BRISQUE [51].

| Method | MFE | OPBL-Re | HGA | CSI | Ours |
|---|---|---|---|---|---|
| NIQE ↓ | 5.5818 | 6.6768 | 5.6335 | 5.6574 | **5.3590** |
| BRISQUE ↓ | 22.7138 | 27.7693 | 23.4790 | 24.3912 | **19.5255** |

**Table 4**
The parameters and reference time comparison of our and other methods.

| Method | Param. (M) | Refer. (s) |
|---|---|---|
| MFE [20] | 135.83 | 0.10 |
| OPBL-Re [1] | 53.09 | 0.14 |
| HGA [49] | 22.21 | 0.19 |
| CSI [3] | 6.45 | 0.07 |
| Ours | 80.61 | 0.13 |

designs for old photo restoration. Note that the OPBL is more apt to restore patch-wise old photos, whereas CSI tends to restore scratch-wise ones. Importantly, we also present the results of OPBL-Re, training their middle mapping network with our data. To ensure fairness, all the methods, including MFE, HGA, OPBL, OPBL-Re, and CSI, are trained using our datasets with their default parameters. We test the comparable methods with our dataset and PixFix [14]. Most importantly, we provide the damaged masks for all the comparable methods. It can ensure we can fairly compare their restoration networks. We also provide a comparison of the network parameters and reference time in this section.

#### 4.3.1. *Quantitative comparison*

As the results in Table 2, we have the following observations: (1). Our method achieves better PSNR, SSIM, LPIPS, and FID scores than other competitive methods. Compared to CSI and MFE, we achieve 1.22% and 2.99% relatively higher PSNR. (2). Considering the LPIPS, our method gets 12.14% and 23.37% relatively lower LPIPS than CSI and HGA. For the FID metric, our approach archives 16.08% and 25.93% desperately lower FID than CSI and HGA. (3). Through comprehensive observation, the CSI is the second-best method. It is well in the scratched restoration rather than mixed restoration. The HGA is the third-best method. The transformers are the primary components of their network which focus on huge hole restoration, not scratches. Overall, These comparable results demonstrate the impressive superiority of our method.

We also expand the comparison experiments on the dataset of PixFix [14]. This dataset concentrates on restoration and colorization tasks for old photos. Their ground truth images are unaligned and colorized. Therefore, we test the damaged ones with NIQE [50] and BRISQUE [51] which are the metrics of blind quality assessments. From Table 3, our results still outperform others. Furthermore, we have a comparison experiment about parameters and reference time. From Table Table 4, although we achieve excellent results, there is much room for our method to be promoted. We explore how to balance efficiency and effectiveness in the future.

#### 4.3.2. *Qualitative comparison*

For fairness, we provide the visualization results on patch-wise restoration in Fig. 8 and scratch-wise restoration in Fig. 9 respectively. Our method generates more natural and high-fidelity images significantly in the above two conditions. In Fig. 8 and Fig. 9, we find


**Fig. 9. Compared with the state-of-the-art methods in scratch-wise restoration.** The sixth column presents our results. Zoom in for better visualization.

**Table 5**
Quantitative ablation study for ScratchNet and PatchNet. The best results are shown in boldface.

| Model | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| w/o ScratchNet | 29.2102 | **0.9022** | 0.1224 | 76.9057 |
| w/o NIAM-PatchNet | 29.0260 | 0.8976 | 0.1259 | 78.9264 |
| w/o MINM-PatchNet | 29.2235 | 0.8993 | 0.1239 | 77.9329 |
| Full (ours) | **29.4469** | 0.9010 | **0.1223** | **72.8066** |

that: (1). MFE, HGA, and CSI results have distinct artifacts (distortions and blur) in the patched and scratched region. Because MFE and HGA are designed for image inpainting. They have no special design for old photo issues. Although the CSI method is a specialized scratch photo network, it still cannot perform well in patch regions due to their local perceptive field. (2). Results of the OPBL method are excellent as well as ours in the patch area. However, their results have a global blur and details are missing especially in scratched areas. Because they stack excessive ResNet blocks to extract the feature embeddings before restoration. (3). Distinguishing performances exist between the first two rows and the last two rows, attributed to the inherently lower resolution and increased challenges in the inputs of the first two rows. The intricacies are inevitably affected by the network transmission process, particularly when dealing with lower-resolution input images. Despite these challenges, our results outperform those of other methods. In conclusion, our hierarchical damage-aware old photo restoration architecture always provides fine-grained structures and details.

### *4.4. Ablation study*

#### *4.4.1. Quantitative comparison*

To validate the effectiveness of our network, we consider four variants: w/o ScratchNet, w/o NIAM (Non-local Inpainting Attention Module), w/o MINM (Mask-aware Instance Norm Module), and full components. In Table 5, we observe that: (1). Combining the ScratchNet and PatchNet, yields better scores on the majority of metrics than w/o ScratchNet. Because the PatchNet employs the global perceptive field which is not appropriate for scratches restoration. Furthermore, it abandons the hierarchical fusion manner between ScratchNet and PatchNet, losing much contextual information. (2). w/o NIAM indicates we maintain other structures unchanged while eliminating the mask pipeline. Due to input features multiplied by (1-masks) at the beginning of the overview, the attention values (k×q) become 0 in the damaged regions if the network lacks the mask pipeline. This results in the inability to effectively restore the damaged areas using global attention. The mask pipeline can rectify this by converting 0 to 1, thus obtaining a reasonable attention value to repaint the artifacts. (3). w/o MINM indicates we use the Layer Norm instead of mask-wise Instance Norm. The results are also influenced without MINM, owing to the inability to maintain style consistency between foreground and background.

#### *4.4.2. Qualitative comparison*

We visualize the results of our ablation experiments in Fig. 10. From Fig. 10, it becomes evident that when the ScratchNet is omitted, along with the absence of the NIAM or the MINM. There are noticeable artifacts present within the damaged regions, as indicated by the red and yellow boxes. In contrast, our proposed network showcases the capability to achieve visually pleasing and semantically consistent

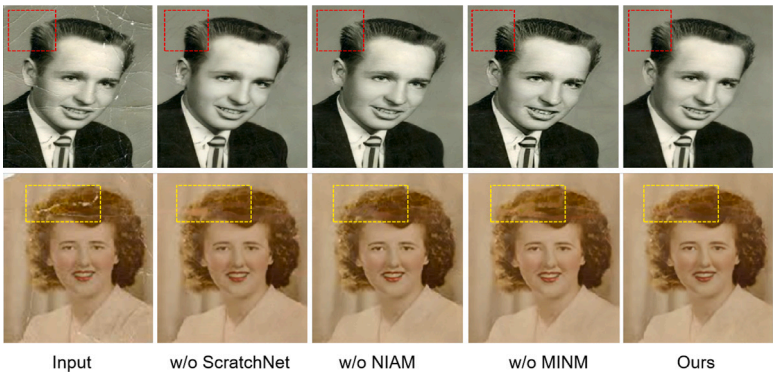


**Fig. 10. The ablation visual results are shown.** w/o ScratchNet: we only use PatchNet to restore whole artifacts with global masks. w/o NIAM: We use normal Multi-head attention instead of our Non-local Inpainting Attention Module. w/o MIHM: We employ Layer Norm to replace our proposed Mask-aware Instance Norm Module. Zoom in for better visualization.

**Fig. 11. The results for the real scenario are shown**.

restoration outcomes. This success can be attributed to exploiting the hierarchical damage correlations with ScratchNet and PatchNet and maintaining style consistency between foreground and background.

### 4.5. In-the-wild results

To demonstrate the practical value of the method further, we illustrate several results stemming from the real scenarios. In Fig. 11, the top row represents the damaged old photos and the bottom row indicates our restored results. Through observation, they perform well not only with scratches and patch-misses problems in Fig. 8 and Fig. 9 but also with spots and creases. These results testify to the generalization ability of our network. In the future, we will explore more for old photo restoration, like colorization and photo enhancement.

### 4.6. Discussion

We extend external experiments to investigate the influence of the network components. Specifically, we delve into exploring the significance of the detection model. Furthermore, we explore the effects of different fusion methods between ScratchNet and PatchNet. Additionally, we examine various processing techniques for scratched inputs. Next, we discuss the impact of the number of multi-branch partial convolution blocks in ScratchNet. It is noteworthy to mention that we conduct these experiments under the condition of having only one transformer block in every Transfill or Transhomon for fairness and effectiveness. Lastly, we also discuss the impact of different numbers of transformer blocks.

#### 4.6.1. The impact of the multi-task detection model

To assess the significance of the Multi-task Detection Model, we conduct a set of experiments. We provide the validation experiments in Table 6. Besides, we conduct the ablation experiments, encompassing three distinct settings: without the overall mask (O_mask), without the scratch mask (S_mask), and the patch mask (P_mask). It is worth noting that the overall mask is mixed with the scratch mask and patch mask. The results, as presented in Table 7, the outcomes indicate that the inclusion of these masks is crucial for the effectiveness of the restoration process.

**Table 6**
The detection results of different masks.

| Model | Recall ↑ | FPR↓ |
|---|---|---|
| O_mask | 0.8299 | 0.0324 |
| S_mask | 0.8169 | 0.0227 |
| P_mask | 0.7592 | 0.0132 |

**Table 7**
The results of different mask fusion strategies between the detection stage and restoration stage. The best results are shown in boldface.

| Model | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| w/o O_mask | 28.2977 | 0.8905 | 0.1384 | 95.5513 |
| w/o S_mask | 28.8712 | 0.9001 | 0.1286 | 82.1801 |
| w/o P_mask | 28.4819 | 0.8937 | 0.1341 | 91.6486 |
| Ours | **29.5119** | **0.9041** | **0.1238** | **74.9684** |

**Table 8**
The results of different fusion strategies between ScratchNet and PatchNet. The best results are shown in boldface.

| Model | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| Concat. | 29.4791 | 0.9023 | 0.1239 | 76.9149 |
| Add. | **29.5119** | **0.9041** | **0.1238** | **74.9684** |

#### 4.6.2. The impact of different fusion manners

We examine the influence of different fusion methods between ScratchNet and PatchNet. The results of these experiments are presented in Table 8. We can conclude that the addition operation performs better than the concatenation for the fusion manner. The superiority of the addition operation can be attributed to its ability to preserve and enhance important information from both networks. By adding the restored results at different stages of the restoration process, the network benefits from the complementary nature of ScratchNet and PatchNet, leading to more effective restoration results.

#### 4.6.3. The impact of different pre-processed methods

To discuss the influence of different pre-processed methods for scratched inputs, we conduct experiments using two distinct settings: multiplication and concatenation. These settings determine how the scratched inputs are combined with their corresponding masks. The


**Fig. 12. Some results from different methods are illustrated in failure cases.** Zoom in for better visualization.

**Table 9**
The results of different pre-processed methods for scratched inputs. The best results are shown in boldface.

| Model | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| Multi. | 29.2075 | 0.9024 | 0.1242 | 76.2010 |
| Concat. | **29.5119** | **0.9041** | **0.1238** | **74.9684** |

**Table 10**
The results of different numbers of multi-branch partial convolution block in ScratchNet. The best results are shown in boldface.

| Model | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| MPC-1 | 29.2368 | 0.8977 | 0.1299 | 79.3149 |
| MPC-2 | 29.3274 | 0.9017 | 0.1250 | 75.1267 |
| MPC-3 | **29.5119** | **0.9041** | **0.1238** | **74.9684** |
| MPC-4 | 29.3745 | 0.9000 | 0.1276 | 75.1266 |

results of these experiments are presented in Table 9. The concatenation method yields more effective results compared to the multiplication method. The reason for this preference lies in the characteristics of the two approaches. Multiplication strategy results in the loss of underlying contextual information within the scratched region. Consequently, it may lead to semantic inconsistency in scratched results and impact patch-wise restoration further.

#### *4.6.4. The impact of the number of multi-branch partial convolution*

We conducted experiments with different numbers of stacked partial convolution blocks and evaluated their performance. The results are presented in Table 10. We can observe that the results improve as the number of stacked partial convolution blocks increases. With three stacked blocks, the network gains the best ability to capture and restore finer details and textures.

#### *4.6.5. The impact of the number of transformer blocks*

We conduct experiments using different configurations of the Transfill and Transhomon blocks. Specifically, we evaluate the performance of 1, 2, 3, and 4 transformer blocks in each Transfill and Transhomon. The results of these experiments are presented in Table 11. The performance is declining with many transformer blocks. This is attributed to the stacked excessive networks that will lose more context. To comprehensively assess the image quality and computational burden, we set the number of transformer blocks to two. The network is capable of exploiting more complex relationships, leading to enhanced restoration performance.

**Table 11**
The results of different numbers of transformer blocks in the PatchNet. The best results are shown in boldface.

| Model | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| Trans-1 | 29.5119 | **0.9041** | 0.1238 | 74.9684 |
| Trans-2 | 29.4469 | 0.9010 | **0.1223** | **72.8066** |
| Trans-3 | 29.4312 | 0.9016 | 0.1260 | 73.7134 |
| Trans-4 | **29.6590** | 0.9014 | 0.1267 | 74.9213 |

### *4.7. User study*

We engage users to assess and score the restoration results achieved by the compared methods. For this evaluation, we randomly selected 20 photos from the test set and instructed the users to rank the restoration outcomes subjectively. The participants were asked to sort the results, assigning a higher rank to better-quality restorations. We collected the responses from 38 volunteers and analyzed the rankings. The Top $k$ metric is employed, which $k$ ranged from 1 to 4, indicating the percentage of times each method was selected within the top $k$ rankings. The results of this user study are presented in Table 12. From the table, our proposed method exhibits a significant advantage. It achieves a probability of 64.08% to be selected as the top 1. This demonstrates the superior quality of our algorithm in comparison with the other methods.

**Table 12**
User study results. The Top $k$ denotes the percentage(%) of every image restoration algorithm by users. The best results are shown in boldface.

| Method | Top 1 | Top 2 | Top 3 | Top 4 |
|---|---|---|---|---|
| MFE [20] | 2.09 | 8.74 | 32.74 | 87.6 |
| OPBL-Re [1] | 0.26 | 2.10 | 4.83 | 20.14 |
| HGA [49] | 5.93 | 21.47 | 68.96 | 94.97 |
| CSI [3] | 27.62 | 76.36 | 94.13 | 97.25 |
| Ours | **64.08** | **91.31** | **99.31** | **99.99** |

### 4.8. Failure case

Despite the overall success of our approach, certain failure cases occur under specific conditions. As shown in Fig. 12, we can observe that the restoration results are imperfect when dealing with both low-resolution and highly intricate. We also illustrated the performance of other methods on our identified failure cases. These results show that other methods tend to either lose details or introduce artifacts. To address this limitation, we plan to collect a larger dataset of low-resolution photos with more diverse and complicated damages. By incorporating these challenging samples into our training set, we aim to handle complex artifacts in low-resolution old photo restoration tasks.

## 5. Conclusion

We propose a novel approach specifically designed to restore various damages in old photos and exploit the corrections between damages. The network is composed of two main components: ScratchNet and PatchNet. ScratchNet is responsible for restoring scratches. It consists of parallel Multi-scale Partial Convolution Modules to restore the scratches with different scales. PatchNet focuses on restoring patch-misses. It is composed of a transformer encoder and a transformer decoder. Particularly, the Non-local Inpainting Attention in the encoder is responsible for inpainting the damaged holes by learning the global semantic context. The Mask-aware Instance Norm Module in the decoder aims to achieve harmonization between the foreground and background, ensuring consistency in style and appearance. Last but important, restored results from ScratchNet are hierarchically fused with the PatchNet pipeline, providing supplement assistant contextual information hierarchically. Overall, our method can address complex damaged old photos and obtain pretty good results.

## CRediT authorship contribution statement

**Weiwei Cai:** Writing – original draft, Software, Methodology. **Xuemiao Xu:** Writing – review & editing, Supervision, Funding acquisition. **Jiajia Xu:** Writing – review & editing, Project administration, Conceptualization. **Huaidong Zhang:** Visualization, Software. **Haoxin Yang:** Visualization, Investigation. **Kun Zhang:** Writing – review & editing, Resources, Formal analysis. **Shengfeng He:** Writing – review & editing, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

The work is supported by the Guangdong International Technology Cooperation Project (No. 2022A0505050009), China National Key R&D Program (Grant No. 2023YFE0202700), Key-Area Research and Development Program of Guangzhou City (No. 2023B01J0022), Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097), Singapore MOE Tier 1 Funds (MSS23C002), and the National Research Foundation Singapore under the AI Singapore Programme (No: AISG3-GV-2023-011). Xuemiao Xu is with the School of Computer Science and Engineering at South China University of Technology, and also with the State Key Laboratory of Subtropical Building Science, Ministry of Education Key Laboratory of Big Data and Intelligent Robot, and Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information.

## References

[1] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, F. Wen, Bringing old photos back to life, in: Computer Vision and Pattern Recognition, 2020, pp. 2747–2757.
[2] J. Liu, R. Chen, S. An, H. Zhang, CG-GAN: Class-attribute guided generative adversarial network for old photo restoration, in: ACM International Conference on Multimedia, 2021, pp. 5391–5399.
[3] W. Cai, H. Zhang, X. Xu, S. He, K. Zhang, J. Qin, Contextual-assisted scratched photo restoration, IEEE Trans. Circuits Syst. Video Technol. (2023).
[4] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: European Conference on Computer Vision, 2018, pp. 85–100.
[5] H. Yamauchi, J. Haber, H.-P. Seidel, Image restoration using multiresolution texture synthesis and image inpainting, in: Proceedings Computer Graphics International 2003, IEEE, 2003, pp. 120–125.
[6] A. Criminisi, P. Pérez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Trans. Image Process. 13 (9) (2004) 1200–1212.
[7] Z. Hu, X. Liu, X. Wang, Y.-m. Cheung, N. Wang, Y. Chen, Triplet fusion network hashing for unpaired cross-modal retrieval, in: Proceedings of the ACM International Conference on Multimedia Retrieval, 2019, pp. 141–149.
[8] Y. Xie, H. Zhang, X. Xu, J. Zhu, S. He, Towards a smaller student: Capacity dynamic distillation for efficient image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 16006–16015.
[9] Z. Zhou, J. Li, Y. Quan, R. Xu, Image quality assessment using kernel sparse coding, IEEE Trans. Multimed. 23 (2021) 1592–1604, http://dx.doi.org/10.1109/TMM.2020.3001472.
[10] M. Li, Y.-m. Cheung, Z. Hu, Key point sensitive loss for long-tailed visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2023) 4812–4825.
[11] Q. Zhou, K. Sheng, X. Zheng, K. Li, X. Sun, Y. Tian, J. Chen, R. Ji, Training-free transformer architecture search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10894–10903.
[12] W. Ding, H. Wang, J. Huang, H. Ju, Y. Geng, C.-T. Lin, W. Pedrycz, FTransCNN: Fusing transformer and a CNN based on fuzzy logic for uncertain medical image segmentation, Inf. Fusion (2023) 101880.
[13] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[14] R. Xu, Z. Tu, Y. Du, X. Dong, J. Li, Z. Meng, J. Ma, A. Bovik, H. Yu, Pik-fix: Restoring and colorizing old photos, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1724–1734.
[15] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Trans. Graph. 36 (4) (2017) 1–14.
[16] Z. Yan, X. Li, M. Li, W. Zuo, S. Shan, Shift-Net: Image inpainting via deep feature rearrangement, in: European Conference on Computer Vision, 2018, pp. 1–17.
[17] Y. Zeng, J. Fu, H. Chao, B. Guo, Learning pyramid-context encoder network for high-quality image inpainting, in: Computer Vision and Pattern Recognition, 2019, pp. 1486–1494.
[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Free-form image inpainting with gated convolution, in: Computer Vision and Pattern Recognition, 2019, pp. 4471–4480.
[19] T. Yu, Z. Guo, X. Jin, S. Wu, Z. Chen, W. Li, Z. Zhang, S. Liu, Region normalization for image inpainting, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, 2020, pp. 12733–12740.
[20] H. Liu, B. Jiang, Y. Song, W. Huang, C. Yang, Rethinking image inpainting via a mutual encoder-decoder with feature equalizations, in: European Conference on Computer Vision, 2020, pp. 725–741.
[21] J. Li, N. Wang, L. Zhang, B. Du, D. Tao, Recurrent feature reasoning for image inpainting, in: Computer Vision and Pattern Recognition, 2020, pp. 7760–7768.

[22] W. Zhang, J. Zhu, Y. Tai, Y. Wang, W. Chu, B. Ni, C. Wang, X. Yang, Context-aware image inpainting with learned semantic priors, in: International Joint Conference on Artificial Intelligence, 2021, pp. 1323–1329.
[23] N. Wang, J. Li, L. Zhang, B. Du, MUSICAL: Multi-scale image contextual attention learning for inpainting, in: International Joint Conference on Artificial Intelligence, 2019, pp. 3748–3754.
[24] H. Liu, B. Jiang, Y. Xiao, C. Yang, Coherent semantic attention for image inpainting, in: International Conference on Computer Vision, 2019, pp. 4170–4179.
[25] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, E. Ding, Image inpainting with learnable bidirectional attention maps, in: International Conference on Computer Vision, 2019, pp. 8858–8867.
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[27] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving Language Understanding by Generative Pre-Training, OpenAI, 2018.
[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[30] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1833–1844.
[31] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, W. Liu, You only look at one sequence: Rethinking transformer in vision through object detection, Adv. Neural Inf. Process. Syst. 34 (2021) 26183–26197.
[32] Y. Li, H. Mao, R. Girshick, K. He, Exploring plain vision transformer backbones for object detection, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, Springer, 2022, pp. 280–296.
[33] G. Han, J. Ma, S. Huang, L. Chen, S.-F. Chang, Few-shot object detection with fully cross-transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5321–5330.
[34] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, C. Feichtenhofer, Mvitv2: Improved multiscale vision transformers for classification and detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4804–4814.
[35] M. Yang, S. Xu, A novel degraded document binarization model through vision transformer network, Inf. Fusion 93 (2023) 159–173.
[36] Z. Chang, S. Yang, Z. Feng, Q. Gao, S. Wang, Y. Cui, Semantic-relation transformer for visible and infrared fused image quality assessment, Inf. Fusion 95 (2023) 454–470.
[37] M. Ma, W. Ma, L. Jiao, X. Liu, L. Li, Z. Feng, S. Yang, et al., A multimodal hyper-fusion transformer for remote sensing image classification, Inf. Fusion 96 (2023) 66–79.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[39] S. Jia, Z. Min, X. Fu, Multiscale spatial-spectral transformer network for hyperspectral and multispectral image fusion, Inf. Fusion 96 (2023) 117–129.
[40] L. Yang, Y. Gu, G. Bian, Y. Liu, TMF-Net: A transformer-based multiscale fusion network for surgical instrument segmentation from endoscopic images, IEEE Trans. Instrum. Meas. 72 (2022) 1–15.
[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[42] J. Zhang, J. Huang, Z. Luo, G. Zhang, X. Zhang, S. Lu, DA-DETR: Domain adaptive detection transformer with information fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23787–23798.
[43] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, Z. Wen, FAT-net: Feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327.
[44] A. Mehri, P.B. Ardakani, A.D. Sappa, MPRNet: Multi-path residual network for lightweight image super resolution, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2704–2713.
[45] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, L. Shao, Multi-stage progressive image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14821–14831.
[46] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
[47] J. Ling, H. Xue, L. Song, R. Xie, X. Gu, Region-aware adaptive instance normalization for image harmonization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9361–9370.
[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[49] Y. Deng, S. Hui, R. Meng, S. Zhou, J. Wang, Hourglass attention network for image inpainting, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII, Springer, 2022, pp. 483–501.
[50] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2012) 209–212.
[51] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.