TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/21 pp1053–1068
DOI: 10.26599/TST.2023.9010079
Volume 29, Number 4, August 2024

SemID: Blind Image Inpainting with Semantic Inconsistency Detection

Xin Li, Zhikuan Wang*, Chenglizhao Chen, Chunfeng Tao, Yuanbo Qiu, Junde Liu, and Baile Sun

Abstract: Most existing image inpainting methods aim to fill in the missing content in the inside-hole region of the target image. However, the areas to be restored in realistically degraded images are unspecified. Previous studies have failed to recover the degradations due to the absence of the explicit mask indication. Meanwhile, inconsistent patterns are blended complexly with the image content. Therefore, estimating whether certain pixels are out of distribution and considering whether the object is consistent with the context is necessary. Motivated by these observations, a two-stage blind image inpainting network, which utilizes global semantic features of the image to locate semantically inconsistent regions and then generates reasonable content in the areas, is proposed. Specifically, the representation differences between inconsistent and available content are first amplified, iteratively predicting the region to be restored from coarse to fine. A confidence-driven inpainting network based on prediction masks is then used to estimate the information regarding missing regions. Furthermore, a multiscale contextual aggregation module is introduced for spatial feature transfer to refine the generated contents. Extensive experiments over multiple datasets demonstrate that the proposed method can generate visually plausible and structurally complete results that are particularly effective in recovering diverse degraded images.

Key words: blind image inpainting; inconsistent pattern; representation difference; contextual aggregation

1 Introduction

As an important carrier of information preservation, image quality directly affects the preservation of information. However, images may be deteriorated during preservation, such as cracking and fading of artifacts due to natural erosion[1], or accidental scratches during recording, such as capturing occlusion[2] or object surface reflections[3]. The aforementioned phenomena can lead to a considerable loss in the expression of image content. These accidental elements are complex and diverse, with random spatial distributions. The degradations and damages can be repaired by manual retouching; however, repairing is usually time-consuming and laborious. Therefore, designing a robust blind image inpainting model that can automatically perform batch reconstruction of contaminated images has practical applications.

Most previous image inpainting methods[4–8] aim to investigate the recovery of coherent textures and structures for inside-hole regions by learning information from the outside-hole regions. They

• Xin Li, Zhikuan Wang, Chenglizhao Chen, Yuanbo Qiu, Junde Liu, and Baile Sun are with College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China. E-mail: lix@upc.edu.cn; 1607040219@s.upc.edu.cn; cclz123@163.com; s22070055@s.upc.edu.cn; 2007040220@s.upc.edu.cn; 2007010218@s.upc.edu.cn.
• Chunfeng Tao is with CNPC Oriental Geophysical Exploration Co. Ltd., Baoding 072751, China. E-mail: taochunfeng@cnpc.com.cn.
* To whom correspondence should be addressed.
Manuscript received: 2023-03-10 ; revised: 2023-07-04; accepted: 2023-07-27

typically take input as the corrupted image and the corresponding binary mask that calibrates the missing region. However, the previously complete image is not readily available for supervision in practice, and the region to be repaired in the real degraded image is normally unknown. Directly applying conventional inpainting methods is difficult in reality because they require the user to carefully locate the damaged area manually, which is a tedious task. In addition, the positioning accuracy of the mask influences the image restoration quality, and an incorrect mask can lead to poor results.

Blind image inpainting means that only one corrupted image is taken as input. As shown in Fig. 1, the crucial idea of the blind image inpainting task is to automatically identify and restore the damaged areas of an image. Compared to conventional inpainting research, blind image inpainting is highly relevant to real-life needs[9]. The existing blind inpainting methods[10, 11] use constant values or Gaussian noise to define image contaminations, such as simulating ink marks with black pixels. However, the contaminations assumed are oversimplified and can be detected with a single CNN. VCNet[12] uses image stitching and adding graffiti strokes as damages for training to improve generalization. Unreasonable pixels are detected based on image context information, which avoids the distinction of certain patterns of fixation as features.

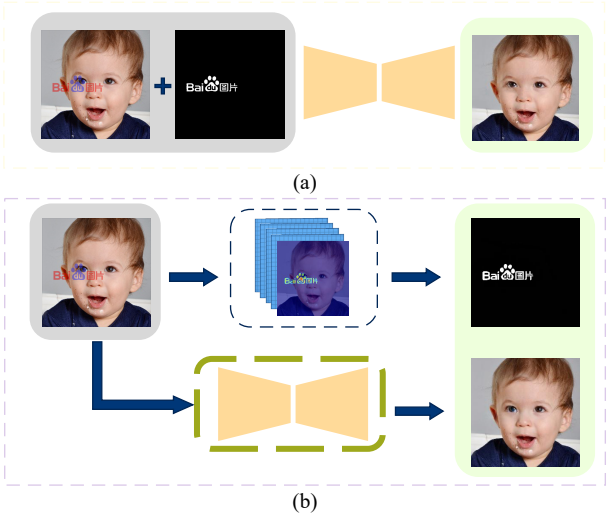


Fig. 1 Comparison of the conventional and blind image inpainting. In the case of a watermark removal, the conventional inpainting method requires a mask to be provided as input to mark the location of the watermark (a), while the blind inpainting method requires the model to automatically identify and fill the damaged area (b).

VCNet can remove contaminated pixels and noise but still struggles to identify content incongruent with the image.

Compared to the simulated degradation above, real inconsistent patterns (e.g., scratches, speckles, and occlusions) are intricately blended with the background, which not only has uncontrollably disparate pixel distributions, but also considerably differs in the deep semantics. Thus, we address this issue by comprehensively considering the deep semantics of the input image, detecting more semantically meaningful inconsistencies based on the structural context in contrast to previous methods.

In this work, we propose an end-to-end robust blind inpainting network that divides the blind repair task into two subtasks: the identification of inconsistent patterns and the reconstruction of missing content. Inevitably, another major issue in this process is nontrivial, that is, addressing the possible degradation of the generation effect caused by certain mask estimation errors. We design an Inconsistency Detection Network (IDN) due to these considerations to locate inconsistent patterns with high accuracy in a coarse-to-fine manner and an Image Reconstruction Network (IRN) to dynamically transfer the statistical features and contextual information of the valid region to the region to be repaired. These networks encourage semantic consistency and continuity of the data distribution. Specifically, IDN stacked Ringed Residual Blocks[13] (RResBlk) could amplify the representational difference between degradation and surroundings to determine the initial estimated mask. A Mask Refinement module (MRf) is then used to extract the key features of the inconsistent patterns and refine the detailed locations using key-value queries. Correspondingly, the Gated Residual Unit (GRU) is designed as the basic unit of the IRN, which can dynamically select features from different regions for information aggregation based on mask prediction probabilities. Moreover, we use the MultiScale Contextual Aggregation Module (MSCAM) at multiple abstraction layers, which enables the transfer of context information at multiple scales with limited resources to encourage coherency in the reconstruction structure.

We conduct extensive experiments on various datasets to evaluate the performance of the proposed model. Furthermore, we perform model analysis and ablation studies to validate our modifications. The main contributions of the proposed method can be

summarized as follows:

- We propose an end-to-end two-stage network for blind image inpainting that decouples the blind inpainting task into inconsistent pattern detection and image content restoration, thereby robustly complementing the image to recover reasonable contextual semantics.
- We introduce RResBlk, which focuses on amplifying the representation differences in each region to locate inconsistent patterns, and a mask refinement module to improve prediction accuracy by querying key features in similar regions.
- Extensive experiments have demonstrated that the proposed approach is more effective than previous works in reconstructing degraded images with diverse noise and corruption, especially in removing inconsistent objects from realistic images.

2 Related Work

2.1 General image inpainting

General image inpainting studies focus on filling in missing regions (usually described by white pixels) with reasonable content[14]. Traditional methods, such as diffusion-based[15] or batch-based[16, 17], borrow image-level patches from the source image to fill in masked areas. These methods leverage internal information to achieve local texture consistency but fail to generate semantic-level content. With the considerable advances[18, 19] in the feature learning capabilities of conditional generation models, filling in large gaps in images has recently become possible. These methods[4–6, 20, 21] are based on the GAN structure to learn image reconstruction under the supervision of the ground truth through reconstruction and adversarial losses. However, the above approaches lack the capability to capture the global structure. Thus, other approaches explicitly incorporate the auxiliary information as an important prerequisite for modeling the structure, such as edges[22, 23], contour[24], sketch[25], gradient information[26], and semantic segmentation[27]. Inspired by patch-based inpainting methods[16, 17], patch-borrowing operations have been integrated into generation models[7, 28, 29], attempting to reduce uncertainty by specifying known regions as references for filling in the missing regions. In addition, some approaches[8, 21, 30] investigate the irregular holes using a special convolution. They differentiate between the inside and outside regions of the hole to ensure that the convolution operates upon valid pixels, and the missing regions are gradually filled in as the network layers progress. These methods achieve visually plausible restorations but still require the mask to indicate the location of the hole explicitly.

2.2 Blind image inpainting

Unlike general inpainting studies that fill in gaps at fixed locations, blind image inpainting addresses the issue of mixed degradation. Liu et al.[11] were motivated by the residual learning algorithm, which aims to learn the missing information in corrupted regions. Cai et al.[10] proposed a fully convolutional neural network named BICNN to automatically identify and remove damaged areas directly. Wu et al.[31] presented a blind face inpainting framework to learn an effective nonlinear mapping between corrupted and clean ID card image pairs. However, they assumed that the contents of the image contaminations are simple data distribution patterns, such as pixels filled with constant values or Gaussian noise. These ideal assumptions increase the possibility of identification even when using uncomplicated models by treating them as specific characteristics without requiring a high-level understanding of the input image. Thus, Wang et al.[12] relaxed the assumption that degradation patterns are semantically incoherent with the background. They first simulate multiple degradation patterns (e.g., graffiti and image stitching) and design a robust model that can identify degraded regions from semantic differences between contamination and surroundings before restoration. Zhao et al.[32] proposed a one-stage hybrid autoencoder architecture, TransCNN-HAE, which exploits the powerful long-range context modeling capabilities of the transformer to avoid the possible degradation of inpainting performance from mask prediction errors. By contrast, the generalization capability of the model is considerably improved. In this work, we exploit contextual semantic inconsistencies to further explore the restoration of highly realistic degraded images.

2.3 Other degraded image restoration

Various existing scene removal tasks boil down to the same path as blind restoration, such as removing raindrops[33] and snow curtains[34] from images of natural scenes or eliminating moles, acnes, and wrinkles to beautify face photos[35]. These tasks share similar assumptions that the acquired images are

immediate and unique, differing in that the feature statistics of noisy regions are influenced by some strong priors.

3 Approach

For this task, we assume that the input to the model contains only degraded images I_{mask} and is formulated as

$$I_{\text{mask}} = I_{\text{gt}} \odot (1 - M) + \text{NOI} \odot M \tag{1}$$

where $I_{\mathbf{gt}}$ represents ground truth, M denotes the binary mask (0 for valid pixels and 1 for others), NOI represents the noise source, and $\odot$ is the Hadamard product. Given a degraded image I_{mask}, we expect the model to remove the contamination accurately and produce a visually complete result I_{comp}.

Inspired by VCNet[12] and image manipulation detection[13], we split the blind image inpainting task into the following two subtasks: detection of global semantic inconsistent patterns of the entire image and reasonable completions for missing regions. Correspondingly, the proposed network is designed into a two-stage structure, as shown in Fig. 2, the first subnetwork IDN utilizes the representational differences to position inconsistent patterns and generate prediction masks, while the second subnetwork IRN completes the structure and content of the suspected regions with the information from the valid regions. Notably, the two subnetworks are interrelated. IDN predicts an estimated mask $\hat{M}$, where $\hat{M} \in [0, 1]$ helps IRN locate inconsistent regions. By contrast, IRN largely standardizes IDN by using local and global semantic contexts to focus on detecting semantically inconsistent regions rather than simply fitting the generated data.

3.1 IDN

IDN is designed to learn a mapping P_{IDN}, where $\boldsymbol{P}_{\text{IDN}}(I_{\text{mask}}) \rightarrow \hat{M}$, which aims to complete the separation of damaged parts from the input image. The inconsistent patterns within the degraded regions are generally considerably different from the valid content in terms of deep semantics and visual texture. With this analysis, we attempt to roughly predict degraded regions by amplifying the differences in intrinsic properties between corruption and background and further refining the detail location using texture continuity or similarity within the same category of regions. Hence, IDN adopts a coarse-to-fine design principle with an encoder–decoder structure comprising RResBlk to acquire feature differences, and an MRf is then used to extract the key features of the degraded region for the relocation of the predicted region.

RResBlk: The differences in properties between degradation and valid content are a considerable basis for detecting inconsistent patterns in images. The residual structure[36] solves the gradient degradation problem. However, the discrimination of the essential attribute features of the image will be weak through the direct multilayer structure. Therefore, we introduce the RResBlk[13] to focus on the representation differences between the corrupted content and the background to improve the detection performance and address the challenge. The core idea of the RResBlk is inspired by the recall and consolidation mechanisms of the human brain and is implemented by the propagation and feedback processes of the residual in CNN. The residual propagation uses skip connections to solve the gradient degradation problem in the deep network. The residual feedback learns nonlinear relationships between

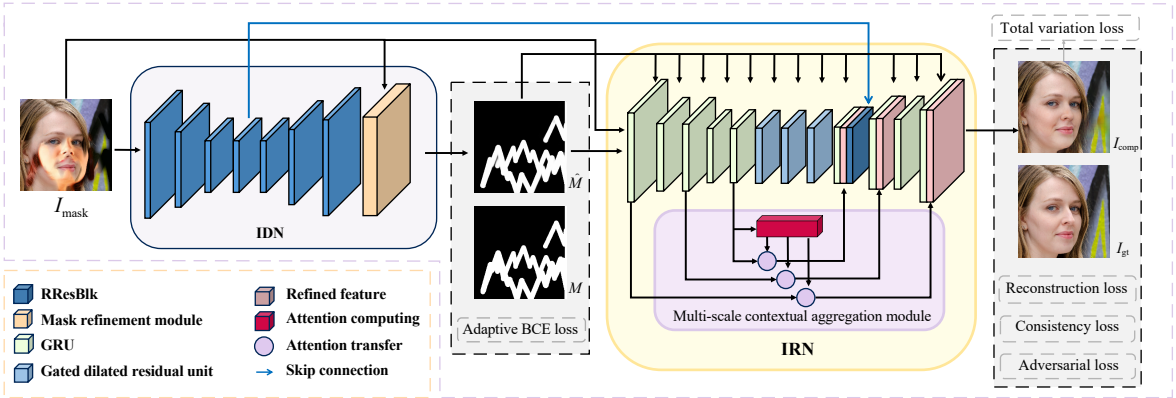


Fig. 2 Overview of the proposed network. IDN locates regions of inconsistent patterns, IRN later generates visually plausible results utilizing the estimated mask (BCE denotes binary cross entropy).

discriminable feature channels by using a simple gating mechanism to access a new understanding of features. The two processes can be defined as

$$y_f = F(x) + x \tag{2}$$

$$y_b = (\sigma(B_{y_f} + 1))\, x \tag{3}$$

where x is the input feature map, y_f is the output of residual propagation, $F()$ represents the residual mapping to be learned; y_b is the output of the residual feedback, B_{y_f} is a linear projection to change the dimensions of y_f, and the function σ represents the sigmoid activation function in this study.

The final RResBlk is structured in a ring by running the residual propagation twice and the residual feedback once, as shown in Fig. 3. This ring structure effectively enhances the learning effect of CNN and avoids gradient degradation as the network deepens. The structure also notably distinguishes essential attribute features between different categories and further reinforces the discrimination between inconsistent patterns and backgrounds.

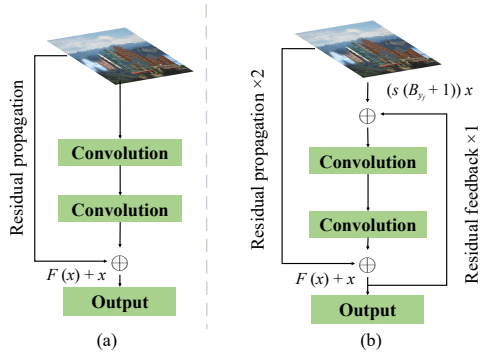


Fig. 3 Illustration of (a) residual block and (b) ringed residual block.

The main framework of IDN comprises seven RResBlks, which are independent of the human visual system and effectively reduce false predictions by maximizing contextual spatial information. Visualization of the feature map of the last RResBlk output x_f with a 1×1 convolution obtains the coarsely estimated mask as shown in Fig. 4.

The inconsistent patterns and backgrounds are fundamentally different at the deep representation level. Thus, their textural characteristics are also considerably distinct, while the textures are similar in their respective regions. Therefore, we jointly input high-level semantic information (feature map x_f) and low-level visual information (degraded image I_{mask}) into the MRf to obtain fine localization of degraded regions (Fig. 4). $x_f \in \mathbf{R}^{m\times n}$ is initially mapped to the initial segmentation prediction $x_{\text{sem}} \in \mathbf{R}^{c\times n}$ through a convolution layer with bias and a softmax layer, where m denotes m-dimensional channels in the feature map, c denotes the number of classes, and n is the number of pixels in a single channel. Afterward, following the nonlocal operation[37], we focus on similar textures on the entire image to distinguish regions of different classes. The low-confidence regions should be revised if they share similar textures to the high-confidence regions predicted in the same class. To achieve this goal, we first need to extract the key feature vector of the high-confidence region of each class. Specifically, we calculate the cosine similarity on the initial prediction x_{sem}, and the output is used as a new bias. The cosine similarity is calculated as follows:

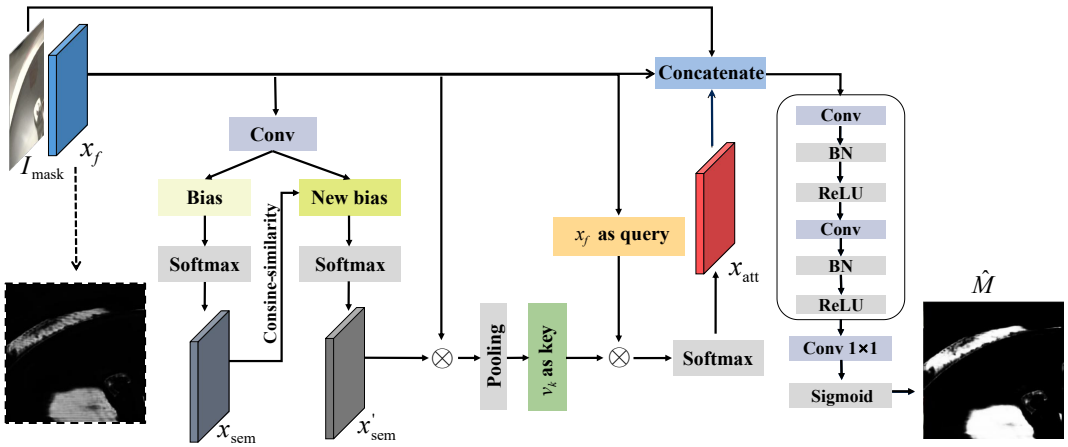


Fig. 4 Illustration of mask refinement module. The output of the last ringed residual block x_f and the degraded image I_{mask} are input into the module (BN is the abbreviation for batch normalization).

$$X_{ij}=\begin{cases}\dfrac{x_i x_j^{\mathrm{T}}}{\|x_i\|\cdot\|x_j\|}, & i\neq j;\\ 0, & i=j,\end{cases}$$

$$x_i=x_{\mathrm{sem}}^{(i,)}\in\mathbf{R}^{1\times n} \tag{4}$$

where $x_{\mathrm{sem}}^{(i,)}$ is the i-th channel of x_{sem}, that is, the predicted score map on class $i\in\{1, 2, \ldots, c\}$. In our case, $\|c\|=2$, corresponding to inconsistent patterns and background. The cosine similarity X_{ij} is then used as new biases, and a 1×1 convolution and softmax layer are repeated to obtain the new segmentation prediction $x'_{\mathrm{sem}}\in\mathbf{R}^{c\times n}$. By using zero bias in the same class and biases proportional to X_{ij} between different classes, equivalent to decreasing the confidence scores on class i, the remaining highly activated regions are considered sufficiently confident. The final key features are extracted from the high-confidence regions as global visual features for each class,

$$v_i=\frac{x_f\cdot\left(x_{\mathrm{sem}}^{\prime(i,)}\right)^{\mathrm{T}}}{\left\|x_{\mathrm{sem}}^{\prime(i,)}\right\|},\ v=\{v_i\mid i=1, 2, \ldots, c\} \tag{5}$$

where $v_i\in\mathbf{R}^{m\times1}$ denotes the pooled key vector for the i-th class, and $v\in\mathbf{R}^{m\times c}$ is the concatenated matrix from v_i. We then take v as key and x_f as query, and compute the query-key similarity $x_{\mathrm{att}}\in\mathbf{R}^{c\times n}$ through dot-product followed by softmax, highlighting those degenerate regions that may be ignored due to low-confidence in x_{sem}. We eventually fuse x_{att} with backbone feature x_f and degradation image I_{mask} through several extra convolutional layers, dynamically activating low-confidence regions with global similarity to the high-confidence regions to obtain the refined estimated mask $\hat{M}\in\mathbf{R}^{h\times w\times1}$, where h and w represent the height and width of degraded image I_{mask}.

Notably, the blind inpainting is sensitive to the prediction of contaminated areas. If the clean regions are incorrectly predicted, or the degraded regions are incompletely detected, then the presentation of the final results will suffer directly. The $\hat{M}$ is restricted to [0, 1] by the sigmoid activation function instead of the binary version to avoid possible misguidance by inaccurate prediction masks for the next content restoration. The benefits of a softness mask are as follows: (1) The softness of $\hat{M}$ enables the differentiability of the entire network. (2) Damaged pixels are incompletely discarded, which reduces the accumulation of errors caused by pixel misclassification. (3) Partial contaminations are allowed to blend naturally into the restoration process, generating a stylistic and artistic final result.

3.2 IRN

The estimated mask $\hat{M}$ of the inconsistent patterns localized by the IDN is used as regional prior guidance to restore the corrupted region. The IRN is designed to learn the mapping P_{IRN} that $P_{\mathrm{IRN}}(I_{\mathrm{mask}}|\hat{M})\rightarrow I_{\mathrm{comp}}$. However, considering that incorrect mask predictions may misguide inpainting, we explicitly use the predictions and repeatedly consider their confidence. The framework of IRN is an encoder-decoder structure comprising GRUs that explicitly consider prediction masks (Fig. 5). In addition, the Contextual Aggregation Module (CAM) is introduced to maintain the consistency of the global structure and improve the reconstruction quality of the final result. The CAM can capture the similarity between contextual patches in the feature space and fill in corrupted regions at multiple scales.

GRU: Despite the predicted mask acquired by IDN, we still remain skeptical regarding its accuracy. Instead of being concatenated with the degradation image and directly entered into the network initially, the prediction mask is used as partial input for each GRU. Notably, the mask regenerates the feature statistics of inconsistent areas by leveraging Probabilistic Context Normalization (PCN)[12], which can minimize the propagation of generated errors. Moreover, Gated Convolution (GC)[8] enables networks to learn a dynamic feature selection mechanism and is, therefore, mostly used to solve irregular hole inpainting. Thus, we adopt the gated mechanism to mask invalid features, which allows the extraction of meaningful features from degraded images mixed with irregularly corrupted regions. Consequently, the GRU is a variant of residual structure that combines GC and residual connection[36], and can be formulated as

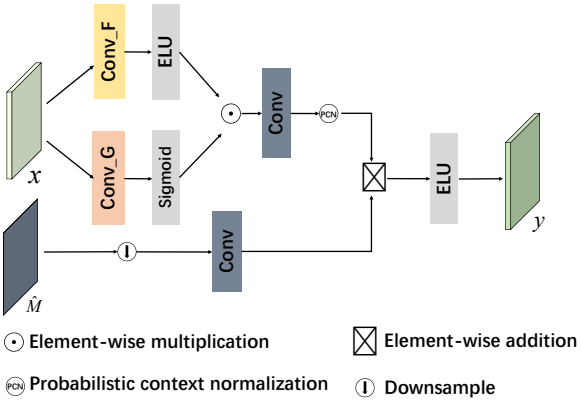


Fig. 5 Illustration of GRU.

$$y = W\left(\phi\left(W_f \cdot x\right) \odot \sigma\left(W_g \cdot x\right)\right) + x \tag{6}$$

where ϕ can be other activation functions (herein set to Exponential Linear Unit (ELU) activation). W, W_f, and W_g are different convolutional filters.

Various previous studies[38, 39] argue that feature mean is correlated with global semantic information, while variance is highly related to local patterns such as texture. PCN is a learnable convex combination of feature statistics from known and unknown areas. Feature statistics comprise the context feature transfer and preservation terms, which aim to facilitate the convergence of the distribution of unknown to known regions while maintaining the features of the known regions. The formulation of PCN is as follows:

$$\begin{aligned}\mathrm{PCN}(x, \hat{M}^l) =& x \odot (1 - \hat{M}^l) + \\ & \left[\omega \cdot \tau(x, \hat{M}^l) + (1 - \omega) x \odot \hat{M}^l\right] \odot \hat{M}^l\end{aligned} \tag{7}$$

where $\tau(\cdot)$ is the statistical information transfer operation of the feature map,

$$\tau(x, \hat{M}^l) = \frac{x_p - \mu(x_p)}{\sigma(x_p)} \cdot \sigma(x_q) + \mu(x_q) \tag{8}$$

where $\hat{M}^l$ represents the l-th feature map, it is downsampled from $\hat{M}$ to the same size as x, $x_p = x \odot \hat{M}^l$ and $x_q = x \odot (1 - \hat{M}^l)$ denote the features of the areas predicted as polluted and clean, respectively. ω is a learnable channel-wise weight computed from x, which is calculated using the ECA[40] instead of the squeeze-and-excitation module[41] in the original method to reduce the negative impact of the dimensionality reduction on the prediction of channel attention.

We also adopt dilated convolutions with different dilation rates in the deep GRUs to further expand the size of the receptive field. Therefore, our model explicitly considers the uncertainty of the prediction mask and aggregates contextual information with minimal prediction error accumulation driven by mask confidence.

MSCAM: GRU achieves overall continuity of information via the propagation of statistical features inside and outside the prediction areas. Furthermore, previous works[6–8, 42, 43] demonstrate that contextual attention enables the inpainting model to capture long-range spatial dependencies among image patches, which is crucial to learning patterns that cross spatial locations in image inpainting. We introduce an MSCAM[6] to enhance the correlation between local patterns, as shown in Fig. 2, which improves the consistency of features in the spatial dimension. Concretely, the cosine similarity between patches inside and outside predicted regions is first calculated on a high-level feature map in IRN,

$$c_{i,j} = \left\langle \frac{p_i}{\|p_i\|}, \frac{p_j}{\|p_j\|} \right\rangle \tag{9}$$

where p_i and p_j denote the i-th and j-th patches extracted from the high-level feature map x outside and inside the prediction mask, respectively. Then softmax is then applied on cosine-similarity to obtain the attention scores between patches,

$$s_{i,j} = \frac{\mathrm{e}^{c_{i,j}}}{\sum_{\mathrm{e}^{c_{i,j}} \in \text{all patches}} \mathrm{e}^{c_{i,j}}} \tag{10}$$

We replicate and aggregate the extra-hole patches with high affinity after obtaining the attention score to fill in the intra-hole regions at different levels. In contrast to FPN[28], which calculates attention and feature reconstruction separately at multiple layers, the design of attention transfer[6] is introduced in our framework. The attention scores $s_{i,j}$ are computed only once on a 32×32 feature map, and the corresponding gaps in the low-level feature maps at different scales are to be filled with contextual patches weighted by a shared attention score,

$$x_j^l = \sum_{i=1}^{N} L\left(s_{i,j}\right) x_i^l \tag{11}$$

where $l \in \{1, 2, 3\}$ denotes three different levels, x_i^l is the i-th patch extracted from outside masked regions of x^l, while x_j^l is the j-th patch to be filled inside. $L(\cdot)$ is used to resize $s_{i,j}$ into the same size as x^l. We apply attention transfer multiple times with the same set of attention scores in the structure, which leads to superior efficiency while maintaining the continuity of semantic features.

Notably, another special design in IRN is to connect the MPN bottleneck to the IDN bottleneck for feature fusion, which allows for joint optimization of the model. The aggregation of information from different directions and the introduction of potential spatial information not only enrich the feature information to produce natural results, but also enhance the discriminative learning of gradient-based localization problems in the generation process.

3.3 Loss functions

Given an input image I_{mask} with inconsistent patterns, our network predicts inconsistent regions $\hat{M}$ and reconstructs results I_{comp} from the IDN and IRN, respectively. We exploit binary cross entropy loss between M and $\hat{M}$ as the optimization goal to train IDN. Valid pixels outnumber damaged ones in most cases. Therefore, a self-adaptive loss that can handle the imbalance between positive and negative sample classifications of pixels is adopted to stabilize the training,

$$\mathcal{L}_{\text{mask}} = -\tau \sum_{\text{clean}} M_{\text{clean}} \cdot \log\left(\hat{M}_{\text{clean}}\right) - (1-\tau) \sum_{\text{dam}} M_{\text{dam}} \cdot \log\left(\hat{M}_{\text{dam}}\right) \tag{12}$$

where $\text{clean} \in \{\text{clean} \mid M_{\text{clean}} = 1\}$ represents clear pixel, and $\text{dam} \in \{\text{dam} \mid M_{\text{dam}} = 0\}$ is damaged pixel. $\tau = |\text{sum}(M_{\text{clean}})|/(h \times w)$ denotes the ratio of the valid region.

A combination of losses comprising reconstruction, consistency, adversarial and total variation losses is used for the training of IRN.

We use the L1 loss on the final output for the hole and the nonhole pixels to force the pixel-level consistency. We define reconstruction loss as

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{hole}} \left\| M \odot (I_{\text{gt}} - \text{I}_{\text{comp}}) \right\|_1 + \lambda_{\text{valid}} \left\| (1-M) \odot (I_{\text{gt}} - \text{I}_{\text{comp}}) \right\|_1 \tag{13}$$

where λ_{hole} is the weight for reconstructing the predicted damaged pixels, and λ_{valid} is the weight for reconstructing the nonpixels ($\lambda_{\text{hole}} = 2.0$ and $\lambda_{\text{valid}} = 1.0$).

ID-MRF loss[44] is employed as our consistency loss to minimize the difference between generated content and corresponding nearest-neighbors from ground truth in the feature space and facilitate effective recovery of the subtle details. This approach adopts a relative distance measure to model the relation between local and target features, which not only finds corresponding candidates from ground truth for each patch of inpainting result but also obtains the feature distribution to help capture variations in complicated textures.

Let I^l_{comp} and I^l_{gt} represent the l-th feature maps generated by a pretrained deep model from the restored image and ground truth, respectively, and the relative distance metric can be expressed as

$$d(v,s) = \exp\left(\frac{\mu(v,s)}{h \cdot (\max \mu(v,r) + \epsilon)}\right) \tag{14}$$

where $v \in \left\{I^l_{\text{comp}} \odot M\right\}$ denotes the patch extracted from the restored content of the missing region, s denotes the patch extracted from I^l_{gt}, and $r \in \rho_s\left(I^l_{\text{gt}}\right)$ means r belongs to I^l_{gt}, excluding s. $\mu(\cdot,\cdot)$ is the cosine similarity, and h and ε are two positive constants. After normalizing $d(\cdot)$, the ID-MRF loss calculation between I_{gt} and I_{comp} is expressed as follows:

$$\mathcal{L}_{\text{IDN}} = -\frac{1}{Z}\left(\sum_{s \in I^l_{\text{gt}}} \max_{v \in I^l_{\text{comp}}}\left(\frac{d(v,s)}{\sum_{r \in \rho_s\left(I^l_{\text{gt}}\right)} d(v,r)}\right)\right) \tag{15}$$

where Z is a normalization factor. Following previous practice[38], final consistency loss is computed on several feature layers from pretrained VGG-19,

$$\mathcal{L}_{\text{cs}} = \mathcal{L}_{\text{IDN}}\left(\phi_s(I_{\text{gt}}),\, \phi_s(I_{\text{comp}})\right) + \sum_{t=3}^{4} \mathcal{L}_{\text{IDN}}\left(\phi_t(I_{\text{gt}}),\, \phi_t(I_{\text{comp}})\right) \tag{16}$$

where the former uses the activation map of layer con4_2 to describe the image structure, and the latter outputs of layers con3_2 and con4_2 are utilized to describe the image texture.

For the adversarial term, WGAN-GP[18, 45] is adopted as our adversarial loss, which enforces the global consistency of final results. The adversarial loss term $\mathcal{L}_{\text{adv}}$ for the generator can be formulated as

$$\mathcal{L}_{\text{adv}} = -E_{I_{\text{comp}} \in P_{I_{\text{comp}}}}(D(I_{\text{comp}})) \tag{17}$$

where P indicates the distribution of the dataset, and $D(\,)$ is the discriminator. Correspondingly, the loss function $\mathcal{L}_D$ for the discriminator is defined as

$$\mathcal{L}_D = E_{I_{\text{comp}} \in \boldsymbol{P}_{I_{\text{comp}}}}(D(I_{\text{comp}})) - E_{I_{\text{gt}} \in \boldsymbol{P}_{I_{\text{gt}}}}(D(I_{\text{gt}})) + \lambda_{\text{gp}} E_{\hat{I} \in \boldsymbol{P}_{\hat{I}}}\left[\left\|\nabla_{\hat{I}} D(\hat{I})\right\|_2 - 1\right]^2 \tag{18}$$

where $\hat{I} = \alpha I_{\text{gt}} + (1-\alpha) I_{\text{comp}}$, $\alpha \in [0, 1]$ is interpolation between real images and generated results, and λ_{gp} is the gradient penalty that is set to 10 for stabilizing the adversarial.

The last loss term is the Total Variation (TV) loss, which is used to obtain naturally smooth results,

$$\mathcal{L}_{\text{tv}} = \sum\nolimits_{i,j} \sqrt{\left[\left(I^{i,j+1}_{\text{comp}} - I^{i,j}_{\text{comp}}\right)^2 + \left(I^{i+1,j}_{\text{comp}} - I^{i,j}_{\text{comp}}\right)^2\right]} \tag{19}$$

The total loss function $\mathcal{L}_{\text{IRN}}$ of the IRN is the combination of all the above loss functions,

$$\mathcal{L}_{\text{IRN}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cs}}\mathcal{L}_{\text{cs}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} \quad (20)$$

where λ_{rec}, λ_{cs}, λ_{adv}, and λ_{tv} are the weights to balance the reconstruction, consistency, TV, and adversarial loss, and are set to 1, 1×10^{-3}, 1×10^{-4}, and 1×10^{-3} in the experiments, respectively.

4 Experiment

4.1 Training datasets

We prove the advantages of the proposed method by showing the results of diverse image inpainting on the following four standard datasets:

•FFHQ[46]: It is a high-quality image dataset of human faces, which comprises 70 000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity, and image background. We randomly select 20 000 images as the training set and 4000 images as the test set.

•CelebA[47]: A dataset focusing on human face images and containing over 180 000 training images.

•Paris StreetView[48]: This dataset contains 14 900 training images and 100 test images collected from street views of Paris.

•Places2[49]: A dataset released by MIT containing over 8 000 000 images from over 365 scenes. In our implementation, we randomly sample and split them into a group of architectural scenes (including but not limited to hotels approximately 20 000 images) and a group of natural scenes (encoinpassing features such as mountain, roughly 30 000 images).

The generation of training data is given in Eq. (1), where the binary mask M is produced using random brush strokes. Notably, the binarization M is extended to a soft version by Gaussian smoothing to realize a naturally smooth fusion and eliminate the noticeable edges produced by direct fusion.

Regarding the definition of the noise signal NOI, we intercept the real image patch instead of a constant value or a certain kind of noise. Such an interception ensures that the noise is indistinguishable from the content in the local patch, forcing the model to draw an inference from contextual information and eventually improving the generalization for real-world data. The corresponding noise sources for FFHQ are drawn from CelebA-HQ during training. The artificial images of Places2 are used as noise sources for the training of Paris StreetView and the natural scenes of Plcaes2. We also simulate various damage modes, such as graffiti, target occlusion, and watermarks, to enhance the robustness of the model. In addition, SHIQ[3], a real specular highlight dataset, is utilized to participate in model training and test its capability to solve blind repairs in a real case. All the images and corresponding masks are resized to 256 pixel$\times$256 pixel.

4.2 Implementation

We implement our model using PyTorch 1.5.0. Training is launched on two RTX 2080Ti GPUs with a batch size of four. Two training stages are available. In the first stage, IDN and IRN are separately trained using an Adam optimizer ($\beta_1 = 0.5$ and $\beta_2 = 0.9$) with both learning rates of 1×10^{-4}. We jointly optimize $\min\{\lambda_m\mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{IRN}}\}$ ($\lambda_m = 2$) after the convergence of both networks using the same setting of the Adam optimizer. However, the learning rate is adjusted to 1×10^{-5}. Training on the FFHQ and SHIQ datasets takes approximately 20 000 iterations to converge (8000 iterations for the first stage). Meanwhile, training on Paris StreetView and Places2 takes approximately 30 000 iterations in total (the first stage costs 10 000 iterations).

4.3 Baselines

We compare our method against some advanced conventional inpainting methods and state-of-the-art blind inpainting approaches, including CA[7], GC[8], HiFill[6], VCNet[12], and TransCNN-HAE[32]. For a fair comparison, CA, GC, and HiFill are all equipped with our IDN via sequential connections. These methods are trained in the same way as ours (as described in the previous section) while using the prediction mask of IDN as part of their input. TransCNN-HAE is a one-stage blind inpainting framework without prediction mask guidance. Thus, we make no evaluation of the accuracy of its prediction mask during the experiment. Notably, the number of parameters in our SemID (7.43×10^6) is slightly higher compared to most baseline models, such as CA, GC, and HiFill, which are 6.68×10^6, 18.42×10^6, and 5.92×10^6, respectively. All these numbers include the complexity of our IDN, which is 3.10×10^6. Compared to VCNet (3.88×10^6) and TransCNN-HAE (2.76×10^6), which maintain the original model structure, the elevation in complexity is highly considerable. The training environment for the above models remains the same, and all outputs are not performed in any specific post-processing operations.

5 Result

5.1 Qualitative comparisons

Figures 6–8 show the qualitative comparisons of our method with all the baselines on synthetic data (generated based on FFHQ[46], Paris StreetView[48], and Places2[49] datasets, respectively). From the visual presentation, the results of regular baselines of splicing IDN in front of their inputs demonstrate artifacts that are markedly disturbed by noise content. The results produced by CA, GC, and HiFill tend to contain blurry, distorted content or artifacts. In addition, these tandem structures affect the mask estimation performance. Therefore, simply concatenating the estimated masks with the input images instead of performing a fusion of features is an ineffective way to address the blind inpainting task. VCNet can generate plausible results because it uses an efficient end-to-end network to jointly train the prediction of damaged areas and the reconstruction of content. However, VCNet fails to adequately consider contextual information integrity, resulting in discontinuities in the structure of the generated results. TransCNN-HAE captures global contextual information with the help of a transformer

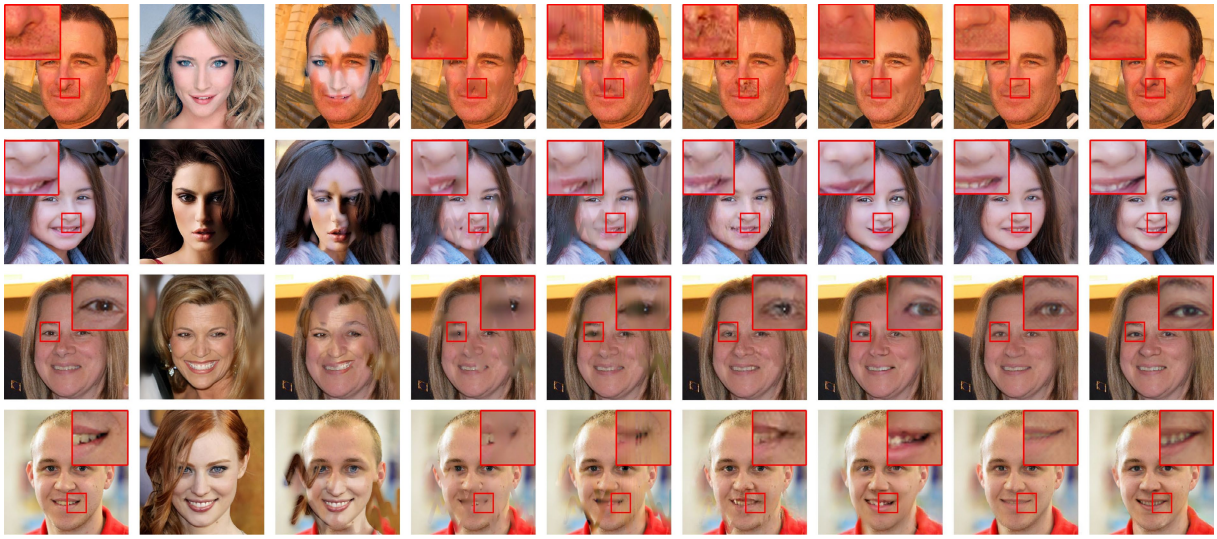


Fig. 6 Qualitative results on FFHQ (the ground truth masks and the estimated ones are shown in the corner of each image).

Fig. 7 Qualitative results on Paris StreetView (the ground truth masks and the estimated ones are shown in the corner of each image).

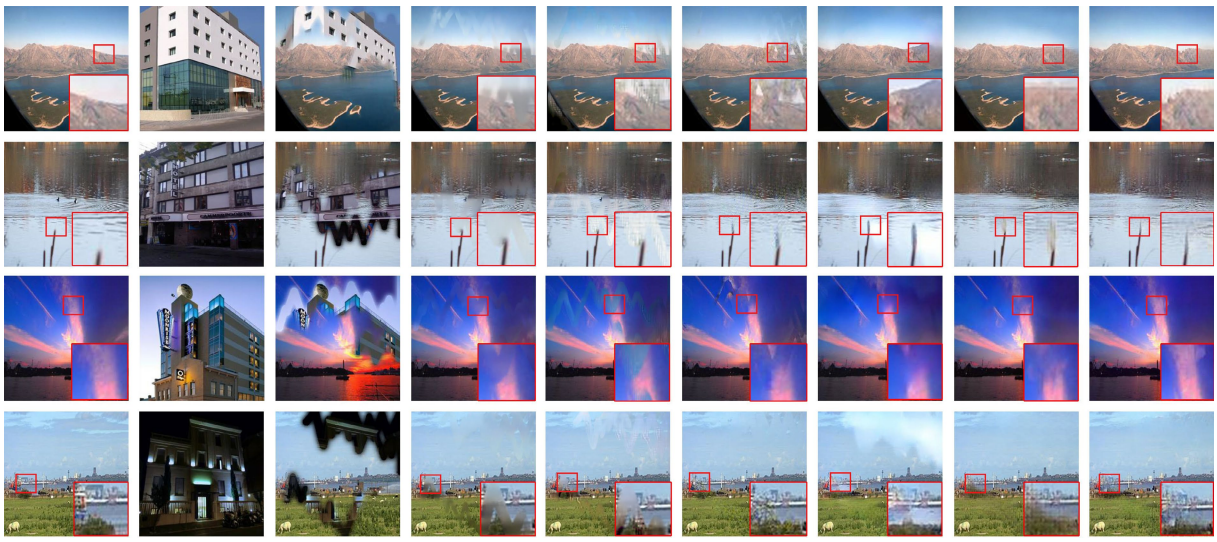


Fig. 8 Qualitative results on Places2 (the ground truth masks and the estimated ones are shown in the corner of each image).

to generate reasonable content. However, without optimizing the opponent's loss, the results are slightly blurry and visually unrealistic. In this paper, our model uses a multiscale attention-shifting network to capture the long-range contextual similarity to reconstruct the feature map and the encoded feature map into a feature map with complete semantic information and texture details. Ultimately, our approach can generate reasonable results with consistent, contextual structures and clear textures.

We also validate the restoration of real data on the specular highlight dataset SHIQ[3]. As shown in Fig. 9, VCNet fails to effectively remove highlighted areas because it fails to distinguish these areas from the global semantic plausibility. Moreover, the results of VCNet are generally dim and produce color/shading distortion. Benefiting from the ringed residual structure to amplify representation differences and prediction category confidence-driven refinement of the prediction mask, the highlights are accurately located and removed by our method. Moreover, the results effectively maintain the color saturation of the background portion. Considering image quality, our results are comparable to those of SHDR[3], which are considerably similar to ground truths. The results demonstrate that the visual consistency capability learned for our model is generalized despite different categories of degraded images.

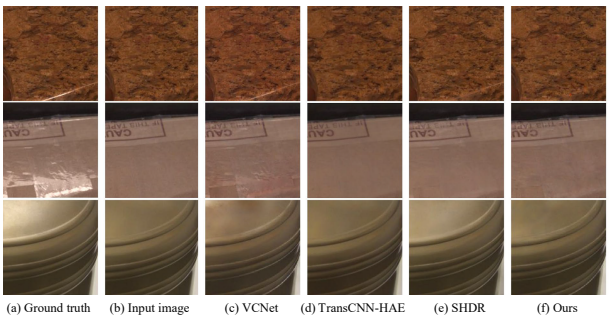


Fig. 9 Visual comparison on SHIQ.

5.2 Quantitative comparisons

The mask prediction quality of all used methods is evaluated with mIoU. We adopt three common evaluation metrics for the quantification of final performances: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Frechet Inception Distance (FID). The evaluation results on FFHQ[46], Paris StreetView[48], and Places2[49] are shown in Table 1. Undeniably, Trans CNN-HAE slightly outperforms the proposed method in SSIM and FID on Paris StreetView and Places2. This phenomenon may be attributed to the powerful global information learning capability of the transformer structure, which facilitates the easy capture of global features of contaminated and uncontaminated areas. Overall, for most of the metrics used for evaluation, our method performs better than other approaches. This finding indicates that the proposed method outperforms other approaches in terms of the recognition accuracy of degraded regions and the image restoration quality. Such a performance confirms that our training model is highly effective for the blind inpainting task.

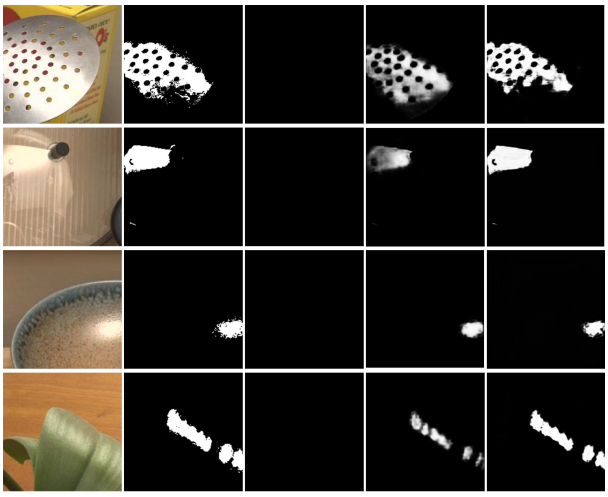
Table 1 Quantitative comparison of our method with others on FFHQ, Paris StreetView, and Places2 (the best and the second-best results are shown in red and blue, respectively. "↑" means large is superior, and ↓ means low is superior).

Method	FFHQ				Paris StreetView				Places2			
	mIoU↑	PSNR↑	SSIM↑	FID↓	mIoU↑	PSNR↑	SSIM↑	FID↓	mIoU↑	PSNR↑	SSIM↑	FID↓
CA[7]	0.9104	16.57	0.6210	27.09	0.9137	18.16	0.6387	78.65	0.7029	14.35	0.6281	33.82
GC[8]	0.8336	14.80	0.4161	22.52	0.7909	17.06	0.5349	66.17	0.7248	13.34	0.3497	31.85
HiFill[6]	0.9456	21.06	0.6076	28.60	0.9589	21.53	0.5843	63.36	0.9007	18.65	0.4900	11.24
VCNet[12]	0.9566	24.17	0.6691	16.88	0.9648	24.06	0.7272	55.93	0.9737	23.05	0.6854	14.65
TransCNN-HAE[32]	–	23.29	0.6945	9.74	–	24.48	0.7469	43.19	–	23.47	0.7115	18.23
Ours	0.9662	25.01	0.7317	9.22	0.9669	24.64	0.7442	50.45	0.9675	24.40	0.7144	13.94

5.3 Ablation studies

Residual Block (ResBlock) v.s. RResBlk: As we have verified on the highlight dataset SHIQ shown in Fig. 10, the conventional residual structure fails to identify natural degradation patterns in real-world images, while the ringed residual structure can roughly locate degradation regions by amplifying the inherent property differences between different classes of pixels. This finding indicates that the amplification of representational differences to identify inconsistent patterns is effective.

Effectiveness of MRf: The MRf is introduced to refine the localization of the prediction mask. In the manner of key-value queries, regions similar to the key features are rediscovered and classified into the same category. The results are shown in Fig. 10, where the mask positions predicted as highlight regions are complete and accurate. The quantitative results in Table 2 also validate its necessity. The recognition accuracy of the highlight region on SHIQ is improved by 0.0523 (an improvement of 13.7%) despite only 145 000 parameters (4.7% of the IDN total parameters) in the MRf. The MRf is designed with the lightest possible parameters to achieve a considerable improvement in the recognition of degraded regions compared to established segmentation algorithms.



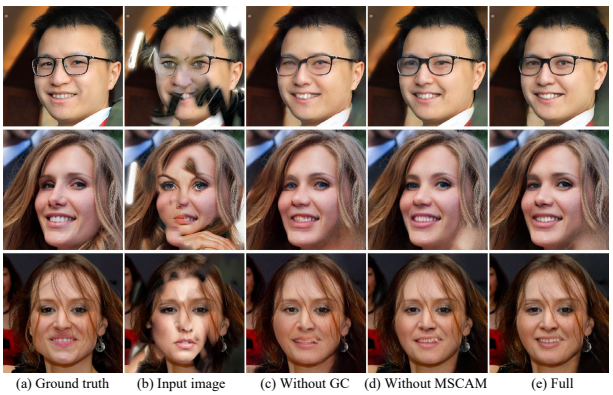
(a) Input image (b) Ground truth (c) ResBlock (d) RResBlk (e) RResBlk+MRf

Fig. 10 Ablation studies on the ringed residual block and mask refinement module in SHIQ. (a) Input image, (b) ground truth mask, (c) mask predicted with residual block, (d) mask predicted with ringed residual block, and (e) mask predicted by the full model.

Effectiveness of GC and MSCAM: We further perform experiments on the CelebA-HQ dataset to validate the effects of different components of the introduced methodology. In Fig. 11, we show the comp arison between different variants of our method, including replacing GC with vanilla convolution and without the MSCAM. Compared to the full baseline, the artifacts are observed in Fig. 11c, replacing GC with vanilla convolutions. In the absence of MSCAM to

Table 2 Quantitative ablation studies on SHIQ (high values indicate superiority).

Module	mIoU
ResBlock	–
RResBlk	0.3814
RResBlk+MRf	**0.4337**



(a) Ground truth (b) Input image (c) Without GC (d) Without MSCAM (e) Full

Fig. 11 Visualization of ablation studies of image reconstruction network on FFHQ.

model the correlation between local features, a considerable visual discontinuity exists between image patches (Fig. 11d). Quantitative results are given in Table 3 to provide specific comparisons.

5.4 Robustness testing and real removal cases

Figure 12 shows the repair results of directly processing graffiti simulated with constant color fills to demonstrate the robustness of the proposed model for filling content. This finding also illustrates the capability of our training scheme to actively learn semantic inconsistencies rather than fit the expected data distribution from the training phase.

We evaluate the generality of removing the task in several common scenarios by retraining the model with a few target data. Figure 13 presents the results of real face mask removal, wherein the model is fine-tuned (pretrained on FFHQ with random strokes) using a set of simulated masks generated on the CelebA dataset (approximately 1000 images). We immediately demonstrate the generality of the model to remove watermarks from images. We generate a set of watermarked images for training by combining randomly selected images from the FFHQ and Paris StreetView datasets with the watermarked image CLWD[50] collected from websites of open-sourced logo images. When making a watermarked image, the size, position, and rotation angle of each watermark are set randomly across the different images. Figure 14 shows the comparison results of our method with SLBR[50] (a state-of-the-art watermark removal model). In the same test environment, the PSNR of SLBR is 43.26, while ours is 39.75, achieving a satisfactory result. Our models are suitable for numerous practical applications due to their outstanding capability to differentiate between representational differences.

Table 3 Quantitative ablation studies in FFHQ. "↑" means large is superior and ↓ means low is superior.

Module	PSNR↑	SSIM↑	FID↓
Without GC	24.60	0.7095	10.87
Without MSCAM	25.09	0.7311	10.37
Full	**25.19**	**0.7378**	**9.221**

Fig. 12 Visual evaluations on FFHQ with random scratches filled with graffiti (the first row is the input, the second row is the corresponding restored result).

Fig. 13 Blind inpainting on the real occluded faces.

5.5 Limitation and failure cases

If the contaminated area blends too naturally with the image background, the accuracy of locating the damaged area is affected considerably to some extent. As shown in the first row of Fig. 15, the direct confidence level of our predicted soft mask is considerably reduced when the damaged areas are too close to the colour and

Fig. 14 Visual evaluation of the watermark removal.

Fig. 15 Failure cases.

image features of the background, making it impossible to remove the predicted areas cleanly. In addition, as shown in the second row of Fig. 15, for some images with rigorous structure (e.g., faces), our generated results may have difficulty maintaining their rigorous structure. The issue could be limited in the future by considering the introduction of a priori information about the target domain, e.g. adding information about the key points of the face in the mask removal task is more helpful in recovering a natural face.

6 Conclusion

We jointly model mask prediction and content reconstruction, and propose an end-to-end blind inpainting network that can robustly accomplish the restoration of multiple degraded images in the real world. Particularly when obtaining manually labeled masks is difficult, the proposed method can automatically identify and reconstruct images. The model first identifies the semantically inconsistent content by amplifying the representation differences between the inconsistent patterns and the valid content with a coarse-to-fine strategy. The inpainting process is then driven by the confidence of the estimated mask, which enables the efficient transfer of contextual information while avoiding the accumulation of prediction errors. Experiments demonstrate that the proposed approach can inherently learn to achieve a visually reasonable and semantically consistent recovery of degraded images. In the future, we plan to simplify the model framework to further reduce the degradation of results due to intermediate prediction mask errors.

Acknowledgment

This work was supported by the Natural Science Foundation of Shandong Province of China (No. ZR2020MF140), the Major Scientific and Technological Projects of CNPC (No. ZD2019-183-004), and the Fundamental Research Funds for the Central Universities (No. 20CX05019A).

References

[1] Z. Zou, P. Zhao, and X. Zhao, Automatic segmentation, inpainting, and classification of defective patterns on ancient architecture using multiple deep learning algorithms, *Struct. Control. Health Monit.*, vol. 28, no. 7, p. e2742, 2021.

[2] G. Tudavekar, S. R. Patil, and S. S. Saraf, Dual-tree complex wavelet transform and super-resolution based video inpainting application to object removal and error concealment, *CAAI Trans. Intell. Technol.*, vol. 5, no. 4, pp. 314–319, 2020.

[3] G. Fu, Q. Zhang, L. Zhu, P. Li, and C. Xiao, A multi-task network for joint specular highlight detection and removal, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 7752–7761.

[4] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, Context encoders: Feature learning by inpainting, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2536–2544.

[5] D. Wang, C. Xie, S. Liu, Z. Niu, and W. Zuo, Image inpainting with edge-guided learnable bidirectional attention maps, arXiv preprint arXiv: 2104.12087, 2021.

[6] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, Contextual residual aggregation for ultra high-resolution image inpainting, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 7508–7517.

[7] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, Generative image inpainting with contextual attention, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5505–5514.

[8] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, Free-form image inpainting with gated convolution, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 4471–4480.

[9] Y. Guo and H. Ma, Image blind deblurring using an adaptive patch prior, *Tsinghua Science and Technology*, vol. 24, no. 2, pp. 238–248, 2019.

[10] N. Cai, Z. Su, Z. Lin, H. Wang, Z. Yang, and B. W. K. Ling, Blind inpainting using the fully convolutional neural network, *Vis. Comput.*, vol. 33, no. 2, pp. 249–261, 2017.

[11] Y. Liu, J. Pan, and Z. Su, Deep blind image inpainting, in *Proc. 9th Int. Conf. Intelligent Science and Big Data Engineering*, Nanjing, China, 2019, pp. 128–141.

[12] Y. Wang, Y. C. Chen, X. Tao, and J. Jia, VCNet: A robust approach to blind image inpainting, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 752–768.

[13] X. Bi, Y. Wei, B. Xiao, and W. Li, RRU-Net: The ringed residual U-net for image splicing forgery detection, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, 2019, pp. 30–39.

[14] X. Wu, K. Xu, and P. Hall, A survey of image synthesis and editing with generative adversarial networks, *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 660–674, 2017.

[15] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, Filling-in by joint interpolation of vector fields and gray levels, *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1200–1211, 2001.

[16] C. Barnes, E. Shechtman, A. Finkelstein, and D. B.

Goldman, PatchMatch: A randomized correspondence algorithm for structural image editing, *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.
[17] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, Image melding: Combining inconsistent images using patch-based synthesis, *ACM Trans. Graph.*, vol. 31, no. 4, p. 82, 2012.
[18] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein generative adversarial networks, in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 214–223.
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672–2680.
[20] S. Iizuka, E. Simo-Serra, and H. Ishikawa, Globally and locally consistent image completion, *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, 2017.
[21] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 85–100.
[22] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, EdgeConnect: Structure guided image inpainting using edge prediction, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision Workshops*, Seoul, Republic of Korea, 2019, pp. 3265–3274.
[23] X. Li, H. Zhang, L. Feng, J. Hu, R. Zhang, and Q. Qiao, Edge-aware image outpainting with attentional generative adversarial networks, *IET Image Process.*, vol. 16, no. 7, pp. 1807–1821, 2022.
[24] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, Foreground-aware image inpainting, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 5840–5848.
[25] C. Cao and Y. Fu, Learning a sketch tensor space for image inpainting of man-made scenes, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 14509–14518.
[26] J. Zhang, L. Niu, D. Yang, L. Kang, Y. Li, W. Zhao, and L. Zhang, GAIN: Gradient augmented inpainting network for irregular holes, in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 1870–1878.
[27] L. Liao, J. Xiao, Z. Wang, C. W. Lin, and S. Satoh, Image inpainting guided by coherence priors of semantics and textures, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 6539–6548.
[28] Y. Zeng, J. Fu, H. Chao, and B. Guo, Learning pyramid-context encoder network for high-quality image inpainting, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 1486–1494.
[29] Y. Zeng, Z. Lin, H. Lu, and V. M. Patel, CR-Fill: Generative image inpainting with auxiliary contextual reconstruction, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 14164–14173.
[30] S. Navasardyan and M. Ohanyan, Image inpainting with onion convolutions, in *Proc. 15th Asian Conf. Computer Vision*, Kyoto, Japan, 2020, pp. 3–19.
[31] F. Wu, Y. Kong, W. Dong, and Y. Wu, Gradient-aware blind face inpainting for deep face verification, *Neurocomputing*, vol. 331, pp. 301–311, 2019.
[32] H. Zhao, Z. Gu, B. Zheng, and H. Zheng, TransCNN-HAE: Transformer-CNN hybrid AutoEncoder for blind image inpainting, in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 6813–6821.
[33] M. Hu, J. Yang, N. Ling, Y. Liu, and J. Fan, Lightweight single image deraining algorithm incorporating visual saliency, *IET Image Process.*, vol. 16, no. 12, pp. 3190–3200, 2022.
[34] P. Li, M. Yun, J. Tian, Y. Tang, G. Wang, and C. Wu, Stacked dense networks for single-image snow removal, *Neurocomputing*, vol. 367, pp. 152–163, 2019.
[35] X. Li, X. Li, X. Zhang, Y. Liu, J. Liang, Z. Guo, and K. Zhai, A method of inpainting moles and acne on the high-resolution face photos, *IET Image Process.*, vol. 15, no. 3, pp. 833–844, 2021.
[36] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.
[37] X. Wang, R. Girshick, A. Gupta, and K. He, Non-local neural networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.
[38] L. A. Gatys, A. S. Ecker, and M. Bethge, Image style transfer using convolutional neural networks, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2414–2423.
[39] J. Johnson, A. Alahi, and L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, the Netherlands, 2016, pp. 694–711.
[40] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 11531–11539.
[41] J. Hu, L. Shen, and G. Sun, Squeeze-and-excitation networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.
[42] J. Xu, N. Wang, and Y. Wang, Multi-pyramid image spatial structure based on coarse-to-fine pyramid and scale space, *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 228–234, 2018.
[43] Q. Hua, L. Chen, P. Li, S. Zhao, and Y. Li, A pixel-channel hybrid attention model for image processing, *Tsinghua Science and Technology*, vol. 27, no. 5, pp. 804–816, 2022.
[44] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, Image inpainting via generative multi-column convolutional neural networks, in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Montréal, Canada, 2018,

pp. 329–338.
[45] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, Improved training of wasserstein GANs, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 5769–5779.
[46] T. Karras, S. Laine, and T. Aila, A style-based generator architecture for generative adversarial networks, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 4401–4410.
[47] T. Karras, T. Aila, S. Laine, and J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in *Proc. the 6th Int. Conf. Learning Representations*, arXiv preprint arXiv: 1710.10196, 2017.
[48] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, What makes Paris look like Paris? *ACM Trans. Graph.*, vol. 31, no. 4, p. 101, 2012.
[49] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, Places: A 10 million image database for scene recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018.
[50] J. Liang, L. Niu, F. Guo, T. Long, and L. Zhang, Visible watermark removal via self-calibrated localization and background refinement, in *Proc. 29th ACM Int. Conf. Multimedia*, Chengdu, China, 2021, pp. 4426–4434.

Xin Li received the BEng degree from China University of Petroleum (East China) in 2004, and the PhD degree from Zhejiang University, China in 2014. He is currently an associate professor at College of Computer Science and Technology, China University of Petroleum (East China). His current research interests include computer vision and visual analytics.

Chenglizhao Chen received the PhD degree from Beihang University, China in 2017. Before that, he was also a joint PhD candidate at Stony Brook University (2015–2016). After graduation, he joined Qingdao University as an assistant professor (2017–2019), an associate professor, a tenure-track professor, and the vice director of Computer Vision Laboratory (2019–2021). In Nov. 2021, he joined College of Computer Science and Technology, China University of Petroleum (East China) as a professor. His current research interests include virtual reality, computer vision, deep learning, and data mining.

Yuanbo Qiu received the BEng degree from Qingdao University of Technology, China in 2022. He is currently a master student at College of Computer Science and Technology, China University of Petroleum (East China). His main research interest is image inpainting and segmentation.

Baile Sun is an undergraduate student at student China University of Petroleum (East China). His research interests include computer vision and data visualization.

Zhikuan Wang received the BEng degree from China University of Petroleum (East China) in 2016. He is currently a master student at College of Computer Science and Technology, China University of Petroleum (East China). His main research interest is computer vision.

Chunfeng Tao received the BEng degree in computer application from China University of Petroleum in 2001. He joined CNPC Oriental Geophysical Exploration Co. Ltd. in 2001, and he has more than 20 years of software development experience. His current research interests include computer vision, visual analytics, and seismic interpretation software.

Junde Liu is an undergraduate student at China University of Petroleum (East China). His research interest covers areas of computer vision and machine learning.