Expert Systems With Applications 235 (2024) 121148

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Uncertainty-aware image inpainting with adaptive feedback network

Xin Ma [a], Xiaoqiang Zhou [d,c], Huaibo Huang [c], Gengyun Jia [b], Yaohui Wang [e], Xinyuan Chen [e], Cunjian Chen [a,*]

[a] *Monash University, Australia*
[b] *School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, China*
[c] *Institute of Automation, Chinese Academy of Sciences, China*
[d] *University of Science and Technology of China, China*
[e] *Shanghai Artificial Intelligence Laboratory, China*

ARTICLE INFO

Keywords:
Image inpainting
Uncertainty estimation
Feedback mechanism

ABSTRACT

While most image inpainting methods perform well on small image defects, they still struggle to deliver satisfactory results on large holes due to insufficient image guidance. To address this challenge, this paper proposes an uncertainty-aware adaptive feedback network (U2AFN), which incorporates an adaptive feedback mechanism to refine inpainting regions progressively. U2AFN predicts both an uncertainty map and an inpainting result simultaneously. During each iteration, the adaptive integration feedback block utilizes inpainting pixels with low uncertainty to guide the subsequent learning iteration. This process leads to a gradual reduction in uncertainty and produces more reliable inpainting outcomes. Our approach is extensively evaluated and compared on multiple datasets, demonstrating its superior performance over existing methods. The code is available at: https://codeocean.com/capsule/1901983/tree.

1. Introduction

The objective of image inpainting is to restore missing areas in corrupted images (Ma et al., 2022). This problem is fundamental in computer vision and has practical applications in various fields, such as object removal and photo restoration as shown in Fig. 1 (Li et al., 2022). The primary goal of image inpainting is to generate visually realistic and semantically plausible alternative content in the missing regions that is coherent with the known available content (Jam et al., 2021; Xiang et al., 2023; Zhang et al., 2022).

Image inpainting has been a topic of study for several decades, and various solutions have been proposed. In the early stages, traditional methods were proposed to address this problem by propagating information from known regions to unknown ones, a.k.a., diffusion-based methods (Bertalmio et al., 2003; Levin et al., 2003). Alternatively, other methods attempted to recover missing areas by copying pixels or patches from known regions, termed example-based methods (Barnes et al., 2009; Xu & Sun, 2010). Though both diffusion-based and example-based methods have been widely used for image inpainting, they have limitations when dealing with non-repetitive and complex scenes, as they may lack the ability to capture high-level semantics.

In recent years, great progress has been made in image completion through the development of deep learning (Liu et al., 2018), with impressive results achieved by using joint training with generative adversarial networks (GANs) (Goodfellow et al., 2014). However, these methods tend to generate boundary artifacts and inconsistent structures when dealing with large holes, which are particularly challenging to fill due to the absence of sufficient constraints (Li et al., 2020). Inspired by the progress of cortical physiology (Hupé et al., 1998), researchers have explored the use of feedback connections between higher- and lower-order visual areas, and effectively reformulated image inpainting as a curriculum learning problem (Guo et al., 2019; Li et al., 2020; Zhang et al., 2018). Nevertheless, these methods have limitations, including the high computational cost of performing progressive inpainting at the pixel level and insufficient utilization of the feedback information. While some researchers have utilized transformer-based models to address the issue of large holes (Dong et al., 2022; Li et al., 2022), these models often require a significant amount of data as they lack prior inductive bias, as well as more computational resources.

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

[*] Corresponding author.

E-mail addresses: xin.ma1@monash.edu (X. Ma), xq525@mail.ustc.edu.cn (X. Zhou), huaibo.huang@cripac.ia.ac.cn (H. Huang), gengyun.jia@njupt.edu.cn (G. Jia), wangyaohui@pjlab.org.cn (Y. Wang), chenxinyuan@pjlab.org.cn (X. Chen), cunjian.chen@monash.edu (C. Chen).

https://doi.org/10.1016/j.eswa.2023.121148
Received 12 April 2023; Received in revised form 7 August 2023; Accepted 7 August 2023
Available online 10 August 2023

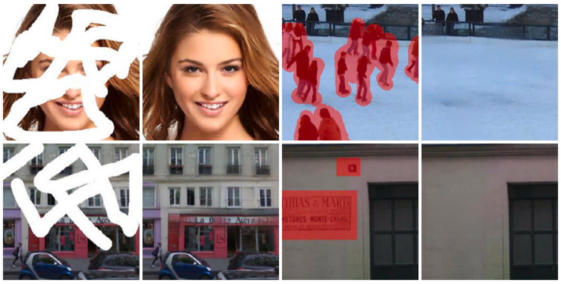

Fig. 1. Illustration of the image inpainting task with face and natural scenes.

In this regard, we present a new architecture for image inpainting called U2AFN. The key components are the adaptive integration feedback block in the latent space and uncertainty estimation. The adaptive integration feedback block integrates both low-level and high-level information in an adaptive manner (Zamir et al., 2017). It refines low-level information with high-level information to enhance the model's ability to reconstruct images. To extract trustworthy feedback information, an uncertainty map illustrating the model's aleatoric uncertainty is estimated alongside the corresponding inpainting results (Kendall & Gal, 2017). These maps are used to guide the generation process by enabling the model to identify successfully restored regions and treat remaining portions as new missing areas. By gradually strengthening the constraints that determine the internal content, our method generates semantically consistent results. Overall, our approach demonstrates promising results in addressing the limitations of previous methods, as well as improving image inpainting performance.

The main contributions of this work are summarized here:

- We propose an adaptive integration feedback block that adaptively integrates low-level and high-level information.
- We propose to estimate uncertainty maps to identify the reliable inpainting areas, using them as guidance to refine the high-uncertainty regions in the next iteration.
- We integrate uncertainty estimation with the proposed adaptive integration feedback block, which enables training under a curriculum learning strategy.
- We evaluate the performance of our model through both qualitative and quantitative experiments and demonstrate the superiority of our proposed method against other state-of-the-art approaches on benchmark datasets.

2. Related work

In this section, we provide a brief discussion of existing literature related to our proposed method.

Image inpainting. Image inpainting techniques can be broadly classified into two categories: traditional and deep learning-based approaches. Traditional approaches encompass diffusion-based and example-based methods. Diffusion-based methods rely on propagating information from neighboring pixels to the missing regions. However, they only consider surrounding known pixels of holes, which renders them unsuitable for background inpainting tasks and generating semantically meaningful content (Bertalmio et al., 2000, 2003; Levin et al., 2003, 2003). In contrast, example-based methods transfer similar and relevant patches from known regions to unknown regions. However, due to the searching and optimization processes involved, they can be computationally expensive for high-resolution images (Barnes et al., 2009; Darabi et al., 2012; Drori et al., 2003; Xu & Sun, 2010).

In recent years, deep learning-based image inpainting methods have emerged as a promising solution to the above challenges. Pathak et al. (2016) were the first to train a convolutional encoder–decoder network using the adversarial training strategy, demonstrating the potential of CNNs for image inpainting tasks. Iizuka et al. (2017) proposed to enforce image coherency by using global and local context discriminators, along with preserving rich high-frequency information through Poisson blending. Yan et al. (2018), Yu et al. (2018) introduced methods that are capable of allowing the model to copy or borrow information from distant spatial locations of the images through feature shift and contextual attention operations, respectively. Moreover, Liu et al. (2018) addressed the issue of irregular masked images using a partial convolutional layer and a mask-update operation. Hong et al. (2019) proposed a fusion block to generate a flexible alpha composition map for combining known and unknown regions for the purpose of harmonically blending the restored image into existing content. Xie et al. (2019) proposed a more effective learnable attention map module for adapting arbitrary irregular holes and convolution layer propagation. At the same time, Yu, Lin, et al. (2019) proposed gated convolutions to solve the problem of vanilla convolution treating all input pixels as valid ones. Li et al. (2020) proposed a recurrent feature reasoning network that iteratively deduces the empty boundary of the convolutional feature map and uses it as a clue for further inferences. Zheng et al. (2021) proposed a global context modeling network to capture the global contextual information effectively for recovering images with heavy corruption. They also proposed a novel deep multi-resolution mutual learning strategy, which can explore the image information from different image resolutions and guide the image inpainting process (Zheng et al., 2022). Liu et al. (2023) proposed a novel framework for efficient high-resolution image inpainting, which is based on parameterized coordinate querying and enables automatic focusing on the masked region.

Furthermore, researchers have also explored the use of transformer-based models to fill missing regions of images, leveraging their ability to sense long-term dependencies. Li et al. (2022) efficiently processed the high-resolution images by unifying the individual strengths of transformers and CNNs. Dong et al. (2022) developed a powerful attention-based transformer model that restored holistic image structures in a fixed low-resolution sketch space. While transformer-based models lack artificial pre-set prior knowledge, such as inductive bias, and require large-scale data and more computational resources to achieve performance comparable to CNN-based models, our study demonstrates that pure CNN models driven by uncertainty estimation can outperform transformer models.

The differences and advantages of our proposed method compared to the above image inpainting methods are mainly reflected in the following aspects. (1) While predicting the completion results, we also predict uncertainty maps simultaneously. We design a new loss function based on these uncertainty maps. This loss function can be used to strengthen the constraints on the completion results. (2) We introduce a feedback mechanism to address the learning difficulties caused by large missing areas in the image. At the same time, we cleverly combine the use of uncertainty estimation mentioned in the first point, allowing the model to pay more attention to the areas with higher uncertainty values in each iteration. (3) By combining the above two points, as the number of feedback iterations increases, the completion results gradually become explicit, making it easier for the model to learn how to inpaint the image.

Feedback mechanism. Feedback mechanisms have been widely deployed in computer vision tasks to enable networks to use output information to rectify previous states (Li et al., 2020, 2019; Yiasemis et al., 2022; Zamir et al., 2017). As a result, some researchers have explored the use of feedback or analogous concepts in the context of image inpainting. For instance, Zhang et al. (2018) proposed the progressive generative networks (PGN), which breaks down the hole-filling process into several phases and integrates them using an LSTM framework. Guo et al. (2019) developed a full-resolution residual network (FRRN), which uses a well-designed residual architecture to progressively fill a hole. Oh et al. (2019) proposed an Onion-Peel network, which can achieve richer contextual information by progressively

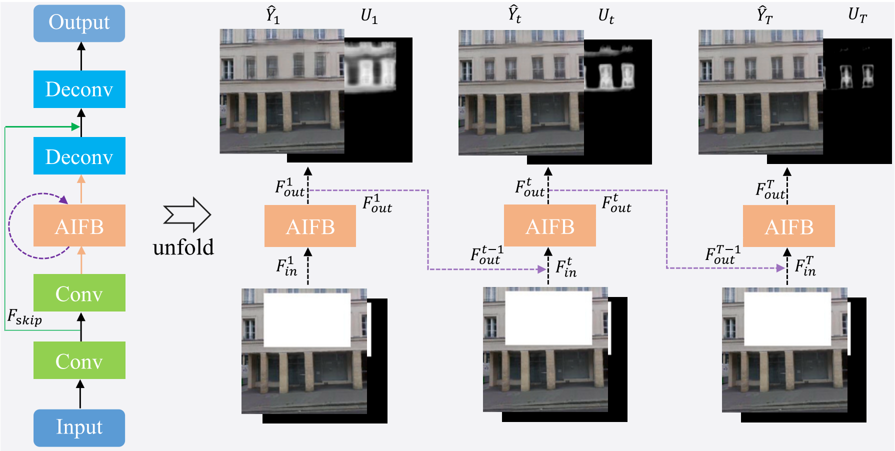

Fig. 2. The architecture of our proposed method. The green arrows represent the local skip connections, while the purple arrows represent the feedback paths.

filling in the hole. Similarly, Li et al. (2020) introduced a recurrent feature reasoning (RFR) method, which can recurrently predict the hole boundaries of convolutional feature maps. However, these approaches possess certain limitations, as previously described.

Uncertainty Estimation. There are two major types of uncertainty that we can model: aleatoric uncertainty and epistemic uncertainty (Kendall & Gal, 2017). Some works have studied uncertainty estimation in terms of the input data, labels, model weights, and so on Choi et al. (2019), Tang et al. (2020), Wu et al. (2020). Kendall et al., proposed a Bayesian SegNet that can predict pixel-wise class labels with a measure of model uncertainty. Yu, Li, et al. (2019) proposed a method that models the representation of each person image as a Gaussian distribution as well as a predicted variance that depicts the uncertainty of the extracted features. Tang et al. (2020) proposed an uncertainty-aware score distribution learning approach, where the action was treated as an instance related to a score distribution. Kundu et al. (2022) proposed a multi-representation pose network (MRP-Net) for 3D human pose estimation. The adaptation process of the MRP-Net aims to minimize the uncertainty for the unlabeled target images while maximizing it for an extreme out-of-distribution dataset. In this work, our goal is to develop a model that can not only predict an inpainting output but also estimate an uncertainty map that pertains to the predicted result. In contrast to prior techniques, we refrain from utilizing additional modules to predict the uncertainty map, thus avoiding a rise in the computational cost. Additionally, we deliberately incorporate the estimation of uncertainty during the training process by designing a loss function that enables the model to self-calibrate and learn an explainable uncertainty map.

3. Methodology

In this section, we will first introduce our network framework and the adaptive integration feedback block. Then, we will discuss the details of our iterative inpainting process and the corresponding uncertainty map. Lastly, we will outline the loss functions utilized in our approach.

3.1. Network framework

As illustrated in Fig. 2, our proposed method involves unfolding into T iterations, during which the feedback mechanism operates in the latent space, and the predicted uncertainty map depicts the aleatoric uncertainty. Each unfolded network comprises three fundamental components: a shallow feature extraction block, an adaptive integration feedback block, and a reconstruction module. At the tth iteration, the input image, corresponding mask, and output image are represented by X, M, and $\hat{Y}_t$, respectively. To extract shallow features containing information from the input image, we apply two convolutional layers, resulting in F_{in}^t.

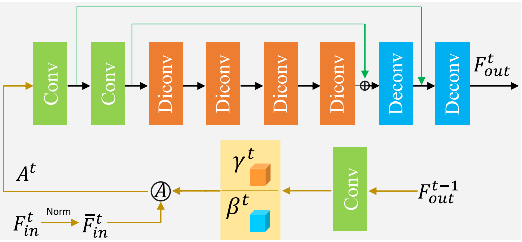

Fig. 3. The architecture of the proposed adaptive integration feedback block. *Diconv* denotes the dilated convolution.

During the tth iteration, we feed the hidden state from the previous iteration, F_{out}^{t-1}, and the shallow feature, F_{in}^t, into the adaptive integration feedback block. The resulting output of the adaptive integration feedback block is denoted as F_{out}^t, which can be expressed as:

$$F_{out}^t = H_{AIFB}(F_{out}^{t-1}, F_{in}^t), \tag{1}$$

where, $H_{AIFB}(\cdot)$ represents our proposed adaptive integration feedback process.

The reconstruction module consists of two transposed convolutional layers, responsible for producing the inpainting results and corresponding uncertainty maps:

$$\hat{Y}^t, U^t = H_{RP}(F_{out}^t, F_{skip}), \tag{2}$$

where, $H_{RP}(\cdot)$ denotes the operation of the reconstruction module and U^t depicts the generated uncertainty map. The shallow feature, F_{skip}, is obtained through the local-source skip connection in the shallow feature extraction block. Upon completing T iterations, we obtain T inpainting results and their corresponding uncertainty maps (i.e., $\{\hat{Y}^1$, $\hat{Y}^2, \dots, \hat{Y}^T\}$ and $\{U^1, U^2, \dots, U^T\}$).

3.2. Adaptive integration feedback block

To refine the low-level representation F_{in}^t through the feedback path, we utilize the hidden state F_{out}^{t-1}. However, instead of using a simple concatenation approach, we adopt a more adaptive method to integrate information.

As depicted in Fig. 3, to integrate F_{out}^{t-1} and F_{in}^t, We first apply instance normalization to F_{in}^t:

$$\bar{F}_{in}^t = \frac{F_{in}^t - \mu_c}{\sigma_c}, \tag{3}$$

where, μ_c and σ_c denote the mean and standard deviation of F^t_{in} in channel c, respectively. Then, we use an adaptive integration feedback process that involves denormalizing the normalized feature map $\bar{F}^t_{in}$ based on the feedback information from the previous iteration. Specifically, we compute a refined version of F^t_{in}, denoted as A^t, as follows:

$$A^t = H_A(\gamma^t, \beta^t, \bar{F}^t_{in}) = (1 + \gamma^t) \otimes \bar{F}^t_{in} + \beta^t, \tag{4}$$

where $H_A(\cdot)$ denotes the fusion operation function. γ^t and β^t are scale and shift factors that are derived after convolving the feature map F^{t-1}_{out} at the tth iteration. The dimensions of γ^t and β^t are the same as that of $\bar{F}^t_{in}$. Finally, A^t is fed into U-Net (Ronneberger et al., 2015) (black color flows) to obtain F^t_{out}.

3.3. Iterative inpainting with uncertainty estimation

Our model generates both an inpainting result and a corresponding uncertainty map at each iteration. These outputs are then utilized to guide the whole image inpainting process through a novel loss function:

$$\mathcal{L}^t_{unc} = -\frac{1}{|\Omega|} \sum_{\mu\nu\in\Omega} ln \frac{1}{\sqrt{2}U^t_{\mu\nu}} exp^{-\frac{\sqrt{2}\mathcal{L}^t_{rec,\mu\nu}}{U^t_{\mu\nu}}}, \tag{5}$$

where Ω denotes the image pixel coordinate. $\mathcal{L}_{rec,\mu\nu}$ represents the $\mathcal{L}_1$ loss between the pixel intensities at location $\mu\nu$:

$$\mathcal{L}^t_{rec,\mu\nu} = \| \hat{Y}^t_{\mu\nu} - Y_{\mu\nu} \|_1, \tag{6}$$

where Y denotes the ground truth image.

Modeling uncertainty is crucial for effectively addressing the underdetermined inverse problem that arises in image inpainting (Zhao et al., 2020). We accomplish this by creating a self-calibrated model and learning a meaningful uncertainty map (Kendall & Gal, 2017; Wu et al., 2020), which enables us to minimize the loss function $\mathcal{L}_{unc}$. By providing information about which parts of the image hole are successfully filled, the uncertainty map enables the model to learn more effectively and ultimately improve the quality of the inpainted image. Eq. (5) can be rephrased as:

$$\mathcal{L}^t_{unc} = \frac{\sqrt{2}}{|\Omega|} \sum_{\mu\nu\in\Omega} [\frac{\mathcal{L}^t_{rec,\mu\nu}}{U^t_{\mu\nu}} + \lambda^t_U U^t_{\mu\nu}]. \tag{7}$$

The model prioritizes processing low-uncertainty pixels during each iteration while postponing the processing of high-uncertainty pixels (i.e., those with uncertain predictions) until the next iteration. We apply a penalty to the estimated uncertainty map using values of $\lambda^t_U = \{1, 2, 4, 8\}$.

Fig. 4 illustrates the inpainting results and corresponding uncertainty maps at different iterations. It is evident that the uncertainty of the inpainting result decreases progressively as t increases. Additionally, the quality of the image inpainting results also gradually improves as t increases, providing compelling evidence for the effectiveness of our proposed method.

3.4. Loss function

Perceptual Loss. To align with human perception of image quality, we introduce the concept of perceptual loss:

$$\mathcal{L}^t_{per} = \frac{1}{N} \sum_{i=1}^{N} \| \Phi^i(Y) - \Phi^i(\hat{Y}^t) \|_1, \tag{8}$$

where, Φ is a VGG-16 network pre-trained on ImageNet (Deng et al., 2009; Simonyan & Zisserman, 2015). $\Phi^i(\cdot)$ outputs feature maps of the ith pooling layer. We adopt *pool*-1, *pool*-2, and *pool*-3 layers of the pre-trained VGG-16 in this work.

Style Loss. To facilitate the recovery of image textures, we also incorporate the style loss. This loss is computed as the $\mathcal{L}_1$-norm between the Gram matrices of feature maps generated by VGG-16 (Liu et al., 2018; Xie et al., 2019).

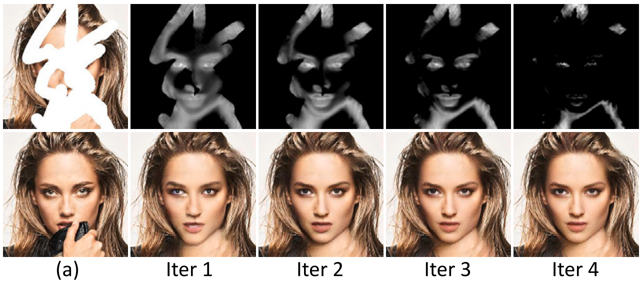

Fig. 4. Illustration of inpainting results and their corresponding uncertainty maps at different iterations. "Iter t" denotes the tth iteration.

$$\mathcal{L}^t_{style} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C_i \cdot C_i} \| \Phi^i(Y)(\Phi^i(Y))^T - \Phi^i(\hat{Y}^t)(\Phi^i(\hat{Y}^t))^T \|_1, \tag{9}$$

where, C_i represents the channel number of the feature map at the ith layer in the pre-trained VGG-16.

Total Variation Loss. It is computed on the region surrounding the hole, which includes an additional 1-pixel dilation and enforces visual coherence among the image pixels.

$$\mathcal{L}^t_{tv} = \frac{1}{N} \sum_{(i,j),(i,j+1)\in\Theta} \| \hat{Y}^t_{i,j+1} - \hat{Y}^t_{i,j} \|_1 + \frac{1}{N} \sum_{(i,j),(i+1,j)\in\Theta} \| \hat{Y}^t_{i+1,j} - \hat{Y}^t_{i,j} \|_1, \tag{10}$$

where Θ indicates the unknown regions.

Adversarial Loss. To improve the visual quality of inpainting results, we introduce the Wasserstein distance (Gulrajani et al., 2017) as the GAN loss:

$$\mathcal{L}^t_{adv} = \min_G \max_D \mathbb{E}_{Y\sim P_Y} D(Y) - \mathbb{E}_{\hat{Y}^t\sim P_{\hat{Y}^t}} D(\hat{Y}^t) + \lambda \mathbb{E}_{Y'\sim P_{Y'}} ((\| \nabla_{Y'} D(Y') \|)^2) - 1)^2, \tag{11}$$

where $D(\cdot)$ means the discriminator and Y' is a resized version with a random scale factor that is sampled from $\hat{Y}^t$ and Y. We set λ as 10 in this work.

Model Objective. The model objective of our method consists of the aforementioned loss functions:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \omega_t (\lambda_{unc}\mathcal{L}^t_{unc} + \lambda_{per}\mathcal{L}^t_{per} + \lambda_{style}\mathcal{L}^t_{style} + \lambda_{tv}\mathcal{L}^t_{tv} + \lambda_{adv}\mathcal{L}^t_{adv}), \tag{12}$$

where, $\lambda_{unc}, \lambda_{per}, \lambda_{style}, \lambda_{tv}, \lambda_{adv}$ are the hyperparameters of the loss functions and set as 1, 0.1, 240, 0.1 and 0.01 (Nazeri et al., 2019). ω_t is a constant factor, which indicates the importance of the output at the tth iteration. We set it as 1 for each iteration (Zamir et al., 2017).

4. Experiments

In this section, we will provide a detailed explanation of our experimental setup and present the results of extensive experimental evaluations. To demonstrate the effectiveness of our proposed method, we conducted both quantitative and qualitative experiments over multiple state-of-the-art image inpainting approaches. The compared methods are CA (Yu et al., 2018), DFNet (Hong et al., 2019), CatedConv (Yu, Lin, et al., 2019), LBAM (Xie et al., 2019), RFR (Li et al., 2020) and CoordFill (Liu et al., 2023). To ensure a fair evaluation, we trained these models until convergence under the same experimental setting as ours.

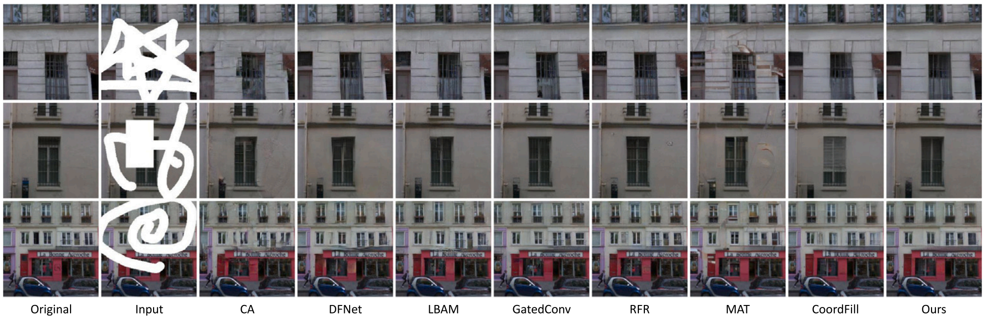

Fig. 5. Qualitative comparison of the proposed U2AFN method against other state-of-the-art inpainting methods on the Paris StreetView dataset.

Table 1
The quantitative evaluation results of CA (Yu et al., 2018), DFNet (Hong et al., 2019), LBAM (Xie et al., 2019), GatedConv (Yu, Lin, et al., 2019), RFR (Li et al., 2020), MAT (Li et al., 2022), CoordFill (Liu et al., 2023) and our proposed U2AFN on CelebA-HQ and Places2 datasets.

Dataset		CelebA-HQ				Places2			
Mask Ratio		10%–20%	20%–30%	30%–40%	40%–50%	10%–20%	20%–30%	30%–40%	40%–50%
CA	mean l_1[a]	0.0183	0.0329	0.0458	0.0718	0.0204	0.0371	0.0529	0.0745
DFNet		0.0158	0.0306	0.0456	0.0792	0.0172	0.0331	0.0483	0.0685
LBAM		0.0131	0.0249	0.0354	0.0580	0.0177	0.0341	0.0497	0.0705
GatedConv		0.0145	0.0268	0.0384	0.0584	0.0206	0.0383	0.0554	0.0793
RFR		0.0159	0.0314	0.0500	0.0876	0.0267	0.0518	0.0736	0.1030
MAT		0.0185	0.0325	0.0444	0.0460	0.0201	0.0342	0.0546	0.0714
CoordFill		0.0157	0.0323	0.0378	0.0522	0.0181	0.0381	0.0510	0.0712
Ours		**0.0113**	**0.0215**	**0.0308**	**0.0459**	**0.0169**	**0.0309**	**0.0447**	**0.0641**
CA	PSNR[b]	25.75	22.59	21.18	18.80	26.23	22.79	20.75	18.82
DFNet		26.75	23.33	21.44	18.39	27.31	23.72	21.65	19.74
LBAM		27.90	24.68	23.14	20.78	27.07	23.42	21.34	19.49
GatedConv		27.74	24.22	22.51	20.35	26.33	22.61	20.42	18.34
RFR		26.88	23.34	20.98	17.94	23.55	20.10	18.42	16.77
MAT		26.77	22.29	23.43	20.85	23.25	20.81	20.66	19.15
CoordFill		27.75	23.74	22.41	20.37	26.03	23.14	20.42	19.34
Ours		**29.20**	**25.82**	**24.13**	**22.10**	**27.44**	**24.35**	**22.32**	**20.36**
CA	SSIM[b]	0.9245	0.8620	0.7955	0.6610	0.8956	0.8072	0.7120	0.5731
DFNet		0.9348	0.8763	0.8106	0.6866	0.9040	0.8167	0.7230	0.5903
LBAM		0.9468	0.8990	0.8490	0.7453	0.9011	0.8105	0.7157	0.5792
GatedConv		0.9460	0.8948	0.8389	0.7365	0.8954	0.8056	0.7092	0.5688
RFR		0.9415	0.8831	0.8050	0.6448	0.8547	0.7397	0.6264	0.4610
MAT		0.8848	0.8434	0.8280	0.7612	0.8949	0.8185	0.7271	0.5987
CoordFill		0.9306	0.8841	0.8193	0.7264	0.8751	0.7967	0.7172	0.5864
Ours		**0.9575**	**0.9171**	**0.8745**	**0.8000**	**0.9052**	**0.8284**	**0.7422**	**0.6152**

[a] Lower is better.

[b] Higher is better.

4.1. Experiment settings

Dataset. We have conducted a series of experiments on three publicly available datasets, namely CelebA-HQ (Karras et al., 2018), Paris StreetView (Doersch et al., 2015), and Places2 (Zhou et al., 2017). CelebA-HQ consists of high-quality images of human faces, Paris StreetView comprises street images, and Places2 is the most demanding dataset among the three, containing over 10 million images spanning more than 365 scene categories. We used the original training and testing splits for both the Paris StreetView and Places2 datasets. Regarding CelebA-HQ, we randomly selected 28,000 images to create our training set, while using the remaining images for the testing set.

Implementation Details. During both the training and testing phases, we resized all the images to 256 × 256 pixels and applied data augmentation techniques such as flipping. Following the approach proposed by Liu et al. (2018), masks were generated automatically on-the-fly during the training. The masks were classified based on the relative size of the unknown regions to the whole image, such as 20%–30%. After further investigation, we have decided to set the number of iterations to 4 and use the Adam optimizer with a learning rate of 10^{-4} for training. Note that our model generates inpainting results directly without requiring any external post-processing.

4.2. Quantitative results

We performed quantitative experiments on three datasets with varying mask ratios, and evaluated our results using three different metrics, including mean l_1 error, structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR). All the aforementioned metrics are widely used to evaluate the quality of inpainted images. To elaborate, the mean l_1 error measures the average difference between the predicted and the ground truth images, with a lower value indicating better performance. SSIM measures the structural similarity between the predicted and the ground truth images, taking into consideration of luminance, contrast, and structure. A higher SSIM score denotes better similarity between the two images. PSNR measures the ratio between the maximum possible power of the signal and the power of corrupting noise, with a higher PSNR indicating less distortion in the image.

Our results, presented in Table 1, demonstrate that the proposed method outperforms state-of-the-art approaches on all three datasets in terms of PSNR, SSIM, and mean l_1 error. These findings suggest that our proposed approach can effectively fill in the missing regions in the images while enabling the structural and semantic information to be well preserved for the original image.

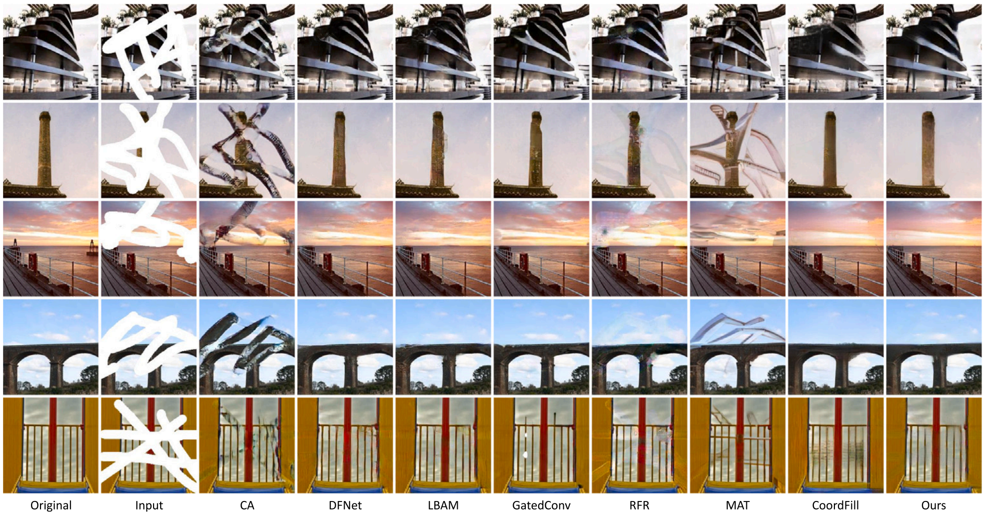

Fig. 6. Qualitative comparisons of the proposed U2AFN method against other state-of-the-art inpainting algorithms on Places2 dataset.

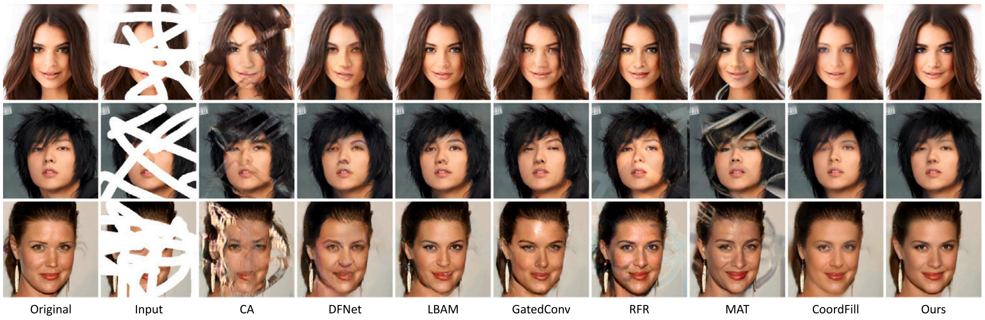

Fig. 7. Qualitative comparisons of the proposed U2AFN method against other state-of-the-art inpainting algorithms on the CelebA-HQ dataset.

Table 2
Analysis of the influence of hyper-parameters in the loss function.

Type	-Adv	-TV	-Style	-Perceptual	Full model
PSNR	21.66	21.78	21.92	21.78	22.32
SSIM	0.6924	0.6948	0.6998	0.6952	0.7422

4.3. Qualitative results

In this section, we also show the qualitative experiments on the three datasets in order to evaluate the visual and semantic coherence. To achieve this, we utilize irregular masks, commonly known as free-form masks, to corrupt the test images, as shown in Figs. 5, 6, and 7. Our analysis reveals that methods such as CA (Yu et al., 2018), DFNet (Hong et al., 2019), LBAM (Xie et al., 2019), GatedConv (Yu, Lin, et al., 2019), RFR (Li et al., 2020), MAT (Li et al., 2022) and CoordFill (Liu et al., 2023) produced inpainting results with noticeable artifacts and color discrepancies. In contrast, our proposed approach generates semantically reasonable inpainting results with rich texture details. These findings have corroborated that our method outperforms state-of-the-art techniques in the qualitative experiment setting.

4.4. Object removal

In this section, our objective is to eliminate distracting objects from Figs. 1 and 10. We present a comparison of our proposed approach with other methods including CA (Yu et al., 2018), DFNet (Hong et al., 2019), LBAM (Xie et al., 2019), and GatedConv (Yu, Lin, et al., 2019) in Fig. 10. Unlike the compared methods, our proposed approach can generate alternative contents that are both realistic and coherent by effectively utilizing both global semantics and local texture details.

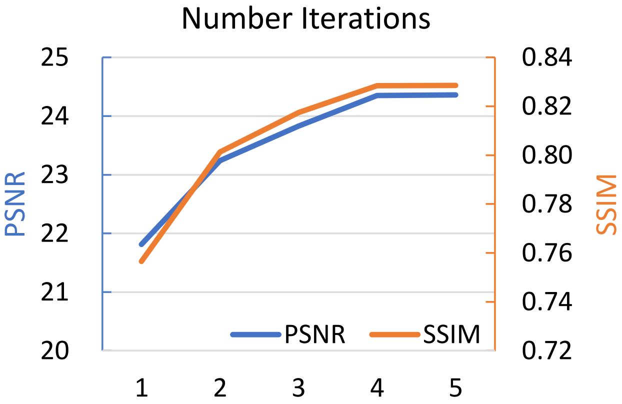

Fig. 8. Analysis of the impact of the number of iterations T on the Places2 dataset.

Table 3
Ablation study experiments on the Places2 dataset to analyze the effects of the uncertainty estimation and feedback path components of our method.

Mask Ratio	uncertainty feedback	✗ ✗	✓ ✗	✗ ✓	✓ ✓
20%-30%	PSNR SSIM	21.40 0.7535	21.81 0.7565	23.85 0.8162	**24.35** **0.8284**
30%-40%	PSNR SSIM	19.95 0.6507	20.30 0.6541	21.89 0.7253	**22.32** **0.7422**

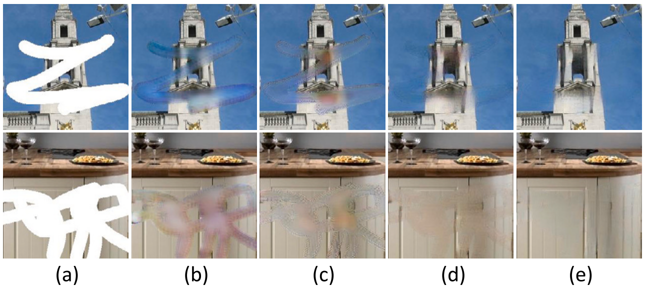

Fig. 9. Images generated by different variants of our proposed method. (a) displays the input images with irregular masks. (b), (c), and (d) correspond to results obtained using models without the uncertainty estimation and/or feedback connection components. (e) shows the inpainting result generated by our full model.

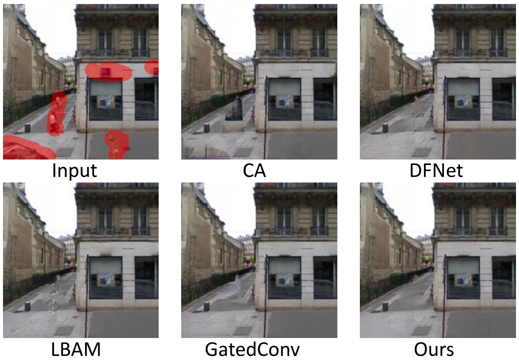

Fig. 10. Comparison of the proposed method against existing works on the object removal task.

Table 4
Analysis of the information fusion method in the adaptive integration feedback block.

Mask Ratio	20%-30%		30%-40%	
Fusion Type	Concatenation	Adaptive	Concatenation	Adaptive
PSRN	25.30	**25.82**	23.88	**24.13**
SSIM	0.8680	**0.9171**	0.8334	**0.8745**

Table 5
Analysis of model efficiency with respect to varying feedback iterations.

T	1	2	3	4	5
Time(s)	0.01186	0.02115	0.03054	0.03988	0.04937
MACs(G)	47.837	92.166	136.494	180.823	225.151

4.5. Model efficiency

To test the impact of feedback iterations, we measured the inference time (in seconds) and multiply-accumulate operations (MACs) to evaluate the inference efficiency on a single GPU. Table 5 presents the results, where the inference time increases as T becomes larger. For instance, when T equals 4, the inference time is 0.03988 s, which is still considered to be reasonable. Additionally, we calculated the MACs using an input size of $3 \times 256 \times 256$.

4.6. Ablation study

In this section, we will examine the impact of uncertainty estimation described in Section 3.3, feedback connections described in Section 3.2, loss functions, as well as the number of iterations T.

Effect of the components. We train variants of our proposed method, including ones without uncertainty estimation and/or feedback connections, on Places2 using a mask ratio of 20% to 30% and 30% to 40%. From Table 3, we see that the baseline model without uncertainty estimation and feedback connections, under a mask ratio of 20% to 30%, has PSNR and SSIM values of 21.40 and 0.7535, respectively. When we add uncertainty estimation and feedback connections to the baseline model, the performance of the model improves relative to the baseline. This indicates that both of these modules bring benefits to the task of image inpainting. We observe that adding these two modules to the baseline model provides the greatest improvement in model performance, resulting in a PSNR gain of 2.95 and an SSIM gain of 0.0749, respectively. Additionally, Fig. 9 shows that incomplete models often produce inpainting results with noticeable artifacts, while our full-fledged model effectively suppresses artifacts and color discrepancies.

Effect of the loss functions. To evaluate the effectiveness of individual loss functions, we conducted experiments in which we removed the perceptual loss, style loss, TV loss, and adversarial loss separately. We then tested the resulting model variants on the Places2 dataset, using a mask ratio of 30% to 40%. Our findings, presented in Table 2, indicate a decline in performance compared to the full model where all the loss functions were used. These results effectively demonstrate the significance of each loss function in achieving the best performance.

Effect of the information fusion methods. We conducted ablative experiments on the information fusion method in the Adaptive Integration Feedback Block on the CelebA-HQ dataset. From Table 4, it can be observed that under a mask ratio of 20%–30%, the values of PSNR and SSIM using the adaptive integration method are 25.82 and 0.9171, respectively, which are higher by 0.21 and 0.0482 compared to the Cat method. Under a mask ratio of 30%–40%, the adaptive integration information fusion method yields the highest performance gain for the model.

Study of T. Our investigation into the impact of the number of iterations, T, revealed the efficacy of employing a feedback connection in the network. As illustrated in Fig. 8, the reconstruction performance of the network with the feedback connection exhibits significant improvement compared to that without it (i.e., $T = 1$). Additionally, we observe a consistent boost in reconstruction performance as the number of iterations, T, increases. However, the performance seems to plateau when T reaches 5. Therefore, we set T to 4 to conserve computing resources while still achieving optimal inpainting results.

4.7. Limitations, application, and future works

Our method requires four iterations of feedback to achieve optimal performance, which inevitably increases the inference time. In practical applications, our work can be applied to scenarios such as watermark removal, face occlusion detection/recognition, and object removal. In the future, we will improve our model architecture by exploring the integration of CNN and Transformer to design a model structure that is suitable for image inpainting.

5. Conclusion

This paper introduces U2AFN, a novel approach for generating visually realistic and semantically plausible alternative content for image inpainting in missing regions. The proposed method leverages an adaptive integration feedback block to effectively combine low-level and high-level information, thereby enhancing the quality of the generated images. Additionally, U2AFN incorporates uncertainty maps to guide the entire generation process, ensuring more precise and dependable results. Extensive experimentation on CelebA-HQ, Places2, and Pairs StreetView datasets verifies the superior performance of our method compared to existing approaches. These findings underscore the potential of U2AFN as a promising solution for image inpainting tasks.

CRediT authorship contribution statement

Xin Ma: Conceptualization, Methodology, Writing – original draft. **Xiaoqiang Zhou:** Comparison experiment. **Huaibo Huang:** Methodology. **Gengyun Jia:** Validation, Writing – review & editing. **Yaohui Wang:** Validation, Investigation. **Xinyuan Chen:** Validation, Data curation. **Cunjian Chen:** Supervision, Writing – review & editing.

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

The datasets are public.

Acknowledgments

This research is based upon work supported in part by the Faculty Initiatives Research, Monash University, via Contract No. 2901912, and support from the NVIDIA Academic Hardware Grant Program; in part by the National Key R&D Program of China, via Contract No. 2022ZD0160100; and in part by the Shanghai Committee of Science and Technology, via Contract No. 21DZ1100100.

References

Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, *28*(3), 24.

Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. In *ACM SIGGRAPH* (pp. 417–424).

Bertalmio, M., Vese, L., Sapiro, G., & Osher, S. (2003). Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, *12*(8), 882–889.

Choi, J., Chun, D., Kim, H., & Lee, H.-J. (2019). Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *International conference on computer vision* (pp. 502–511).

Darabi, S., Shechtman, E., Barnes, C., Goldman, D. B., & Sen, P. (2012). Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics*, *31*(4), 1–10.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Computer vision and pattern recognition* (pp. 248–255).

Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. (2015). What makes paris look like paris? *Communications of the ACM*, 103–110.

Dong, Q., Cao, C., & Fu, Y. (2022). Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Computer vision and pattern recognition* (pp. 11358–11368).

Drori, I., Cohen-Or, D., & Yeshurun, H. (2003). Fragment-based image completion. In *ACM SIGGRAPH* (pp. 303–312).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Neural information processing systems* (pp. 2672–2680).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein GANs. In *Neural information processing systems* (pp. 5767–5777).

Guo, Z., Chen, Z., Yu, T., Chen, J., & Liu, S. (2019). Progressive image inpainting with full-resolution residual network. In *ACM multimedia* (pp. 2496–2504).

Hong, X., Xiong, P., Ji, R., & Fan, H. (2019). Deep fusion network for image completion. In *ACM multimedia* (pp. 2033–2042).

Hupé, J., James, A., Payne, B., Lomber, S., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, *394*(6695), 784–787.

Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics*, *36*(4), 1–14.

Jam, J., Kendrick, C., Walker, K., Drouard, V., Hsu, J. G.-S., & Yap, M. H. (2021). A comprehensive review of past and present image inpainting methods. *Computer Vision and Image Understanding*, *203*, Article 103147.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *International conference on learning representations*.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Neural information processing systems* (pp. 5574–5584).

Kundu, J. N., Seth, S., YM, P., Jampani, V., Chakraborty, A., & Babu, R. V. (2022). Uncertainty-aware adaptation for self-supervised 3D human pose estimation. In *Computer vision and pattern recognition* (pp. 20448–20459).

Levin, A., Zomet, A., & Weiss, Y. (2003). Learning how to inpaint from global image statistics. In *International conference on computer vision* (p. 305). IEEE.

Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., & Jia, J. (2022). MAT: Mask-aware transformer for large hole image inpainting. In *Computer vision and pattern recognition* (pp. 10758–10768).

Li, J., Wang, N., Zhang, L., Du, B., & Tao, D. (2020). Recurrent feature reasoning for image inpainting. In *Computer vision and pattern recognition* (pp. 7760–7768).

Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., & Wu, W. (2019). Feedback network for image super-resolution. In *Computer vision and pattern recognition* (pp. 3867–3876).

Liu, W., Cun, X., Pun, C.-M., Xia, M., Zhang, Y., & Wang, J. (2023). CoordFill: Efficient high-resolution image inpainting via parameterized coordinate querying. In *Association for the advancement of artificial intelligence*.

Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *European conference on computer vision* (pp. 85–100).

Ma, X., Zhou, X., Huang, H., Jia, G., Chai, Z., & Wei, X. (2022). Contrastive attention network with dense field estimation for face completion. *Pattern Recognition*, *124*, Article 108465.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (2019). EdgeConnect: Structure guided image inpainting using edge prediction. In *International conference on computer vision workshops*.

Oh, S. W., Lee, S., Lee, J.-Y., & Kim, S. J. (2019). Onion-peel networks for deep video completion. In *International conference on computer vision* (pp. 4403–4412).

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Computer vision and pattern recognition* (pp. 2536–2544).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer assisted intervention* (pp. 234–241).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., & Zhou, J. (2020). Uncertainty-aware score distribution learning for action quality assessment. In *Computer vision and pattern recognition* (pp. 9839–9848).

Wu, S., Rupprecht, C., & Vedaldi, A. (2020). Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *Computer vision and pattern recognition* (pp. 1–10).

Xiang, H., Zou, Q., Nawaz, M. A., Huang, X., Zhang, F., & Yu, H. (2023). Deep learning for image inpainting: A survey. *Pattern Recognition*, *134*, Article 109046.

Xie, C., Liu, S., Li, C., Cheng, M.-M., Zuo, W., Liu, X., Wen, S., & Ding, E. (2019). Image inpainting with learnable bidirectional attention maps. In *International conference on computer vision* (pp. 8858–8867).

Xu, Z., & Sun, J. (2010). Image inpainting by patch propagation using patch sparsity. *IEEE Transactions on Image Processing*, *19*(5), 1153–1165.

Yan, Z., Li, X., Li, M., Zuo, W., & Shan, S. (2018). Shift-net: Image inpainting via deep feature rearrangement. In *European conference on computer vision* (pp. 1–17).

Yiasemis, G., Sonke, J.-J., Sánchez, C., & Teuwen, J. (2022). Recurrent variational network: A deep learning inverse problem solver applied to the task of accelerated MRI reconstruction. In *Conference on computer vision and pattern recognition* (pp. 732–741).

Yu, T., Li, D., Yang, Y., Hospedales, T. M., & Xiang, T. (2019). Robust person re-identification by modelling feature uncertainty. In *International conference on computer vision* (pp. 552–561).

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Computer vision and pattern recognition* (pp. 5505–5514).

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *International conference on computer vision* (pp. 4471–4480).

Zamir, A. R., Wu, T.-L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., & Savarese, S. (2017). Feedback networks. In *Computer vision and pattern recognition* (pp. 1308–1317).

Zhang, H., Hu, Z., Luo, C., Zuo, W., & Wang, M. (2018). Semantic image inpainting with progressive generative networks. In *ACM multimedia* (pp. 1939–1947).

Zhang, X., Zhai, D., Li, T., Zhou, Y., & Lin, Y. (2022). Image inpainting based on deep learning: A review. *Information Fusion*.

Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., & Lu, D. (2020). UCTGAN: Diverse image inpainting based on unsupervised cross-space translation. In *Computer vision and pattern recognition* (pp. 5741–5750).

Zheng, H., Zhang, Z., Wang, Y., Zhang, Z., Xu, M., Yang, Y., & Wang, M. (2021). GCM-Net: Towards effective global context modeling for image inpainting. In *ACM international conference on multimedia* (pp. 2586–2594).

Zheng, H., Zhang, Z., Zhang, H., Yang, Y., Yan, S., & Wang, M. (2022). Deep multi-resolution mutual learning for image inpainting. In *ACM international conference on multimedia* (pp. 6359–6367).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.