# 利用纹理与结构双流模型的图像补全的方法进行人脸表情编辑

邓庆新

**摘要**

图像补全要求算法根据图像自身或图像库信息来补全待修复的图像的缺失区域使得修复后的图像看起来十分自然。图像补全在许多场景中都有着广泛的应用，比如无关物体的移除，缺失区域的补全以及老照片的修复等等。本文创新性地将用于图像补全的模型应用到人脸表情编辑上。主要思路是先对最新的用于图像补全的模型在大规模的人脸数据集上进行预训练，使该模型能够得到进行人脸补全的能力。然后，人工选取带笑脸的小规模人脸图像，并利用人脸关键点监测数据生成嘴部掩码，对预训练好的模型进行微调训练，使模型能够有在嘴部区域生成笑脸能力。如此一来，输入一张不笑的照片，以及其嘴部掩码，便能够得到相应的露齿笑的照片，从而实现了人脸表情编辑的目的

**关键词：**生成对抗网络；图像补全；表情编辑；深度学习

## 1 引言

图像补全是指在保持图像整体一致性的同时重建图像受损区域的过程，是一种典型的计算机视觉任务，具有许多实际应用，如照片编辑、无关物体的移除、恢复损坏部分等。和许多计算机视觉任务一样，如今图像补全算法大多数为现在流行的深度学习生成算法。在此之前就有许多深度学习模型用于图像补全，经过不断地改进，如今用于图像补全的深度学习模型多是融合图像的纹理信息和结构信息。比如 PRVS(Progressive Reconstruction of Visual Structure)[1]，以及 MED(Mutual Encoder-Decoder)[2]，尽管这两个模型都取得了一些效果上的提升，但是它们还没有充分地利用好图像结构和纹理之间的关系，尤其是没有充分利用好纹理与结构之间的关联性，使其传递整体的互补信息来辅助对方。针对这个问题，作者提出了一种双流的网络结构，即结构约束下的纹理合成和纹理引导下的结构重建，并以生成对抗网络为总体框架，并且在其中融合了双向门控特征融合模块 (Bi-GFF) 来融合重建的纹理特征和结构特征，以增加它们在图像中的相容性，还引入了背景特征聚合模块 (CFA) 来融合背景信息。该 CTSDG 模型在图像补全方面取得了较好的效果。人脸编辑是指在不改变其他特征或者区域的情况下，对输入的人脸图像进行特定特征和区域的编辑。得益于生成对抗网络的快速发展，许多人脸编辑的工作都采用了生成对抗模型，并取得了很好的效果，但是大多数用于人脸编辑的深度学习模型在训练的时候都需要大量的成对的样本，因为它们大多都采用 cycle consistency 来保持其他区域的不变性，然而在实际情况下成对的大规模样本很难得到。图像补全的思想给了我新的思路，如果通过训练使得模型能够根据已有的知识对缺失的区域进行补全，那么根据人脸编辑的目标，对希望编辑的区域添加掩膜，再利用经过训练的图像补全模型对其进行预想中的补全，就可以在不改变其他区域的情况下对特定的区域实现编辑的目的了。基于这个想法，我对上面提到的最新的 CTSDG 模型进行了重头的预训练，再针对特定区域进行了微调训练，得到了可用于人脸笑脸编辑的模型。

# 2 相关工作

此部分对课题内容相关的工作进行简要的分类概括与描述，二级标题中的内容为示意，可按照行文内容进行增删与更改，若二级标题无法对描述内容进行概括，可自行增加三级标题，后面内容同样如此，引文的 bib 文件统一粘贴到 Ref 文件夹下的 Collection.bib 中并采用如下引用方式[3]。

## 2.1 图像补全的传统方法

图像补全的传统方法一般是基于计算机图形学的方法，可以分为两大类：基于扩散模型和基于像素块模型。扩散模型通过参考邻近区域的信息来对缺失区域进行渲染[4][5]，基于像素块的模型[6][7]通过搜索图像未缺失区域中与缺失区域最相似的像素块来填补缺失区域，这种方式很好地利用了图像的远距离像素信息。上述基于计算机图形学的方法虽然也能取得不错的效果但是耗时长，去需要大量的计算资源。

## 2.2 基于深度学习的方法

基于深度学习的生成方法现在是图像补全领域的主流方法，基于深度学习的方法的核心是利用深度神经网络的强大的特征学习能力，经过训练的深度神经网络从有缺失的图像中提取出高级的语义信息并对缺失区域进行合理地填充。近些年有许多基于深度学习的生成的方法，主要是基于生成对抗网络的，但他们中的大多数都只利用了图像的纹理信息或者图像的轮廓结构信息，值得一提的是，刘的团队[2]提出了一种相互作用的编码器-解码器结构，在神经网络的不同层中同时学习图像的结构信息和纹理信息。但是这项工作并没有给出如何让纹理信息和结构信息作为相互补充的信息来相辅相成地进行图像补全的答案。

# 3 本文方法

## 3.1 本文方法概述

本文复现了原文提出的生成模型 CTSDG，该模型的整体架构是一个生成对抗网络。对于生成器，作者提出了一种新颖的双流结构，这种双流结构由两个类似于 U-NET 的模块组成，分别是结构约束下的纹理合成模块以及纹理引导下的结构重建模块。经过该双流结构后，会得到纹理特征图以及结构特征图，得到的特征图在生成器中会进一步经过双向门控特征融合模块以及背景特征聚合模块从而取得更好的结果。在判别器中，有纹理分支以及结构分支，分别判定纹理合成以及结构重建的质量和一致性。该模型的总体框架如图 1所示：

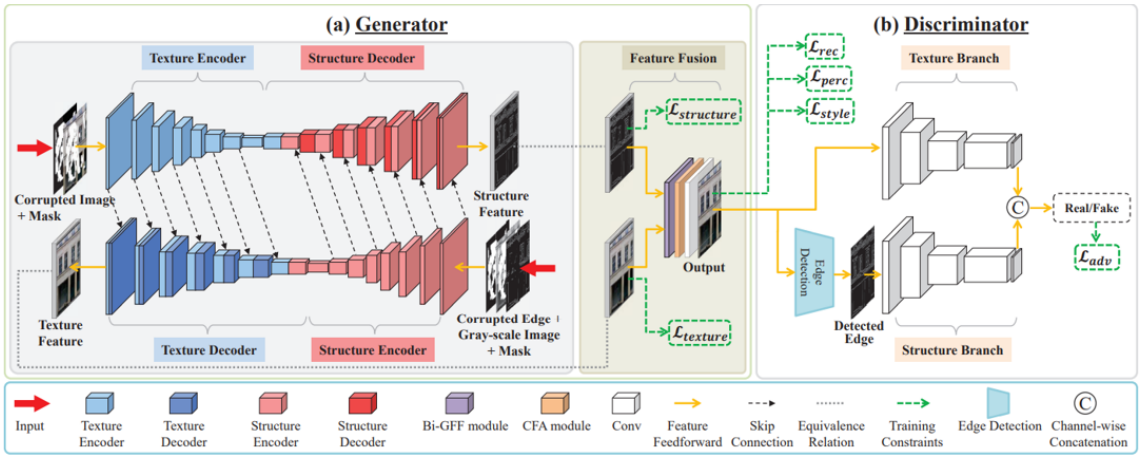


图 1: 模型概述图

### 3.2 双流结构

生成器是一个双流架构，整个双流的模型是以 U-NET 为骨架。在编码阶段，有缺失的图像和其对应的结构图分别映射到隐空间；在解码阶段，纹理解码器通过融合结构编码器得到的结构特征来进行结构约束下的纹理合成，而结构解码器通过融合纹理编码器得到的纹理特征来进行纹理引导下的结构重建。通过这样的双重生成过程，图像的纹理信息和结构信息很好地互补，从而使图像补全取得了更好的效果。

### 3.3 双向门控特征融合模块

该模块是为了进一步结合经过解码的纹理特征和结构特征，它的作用是交换融合纹理和结构两种特征信息。该模块的框架图如下所示 2所示：

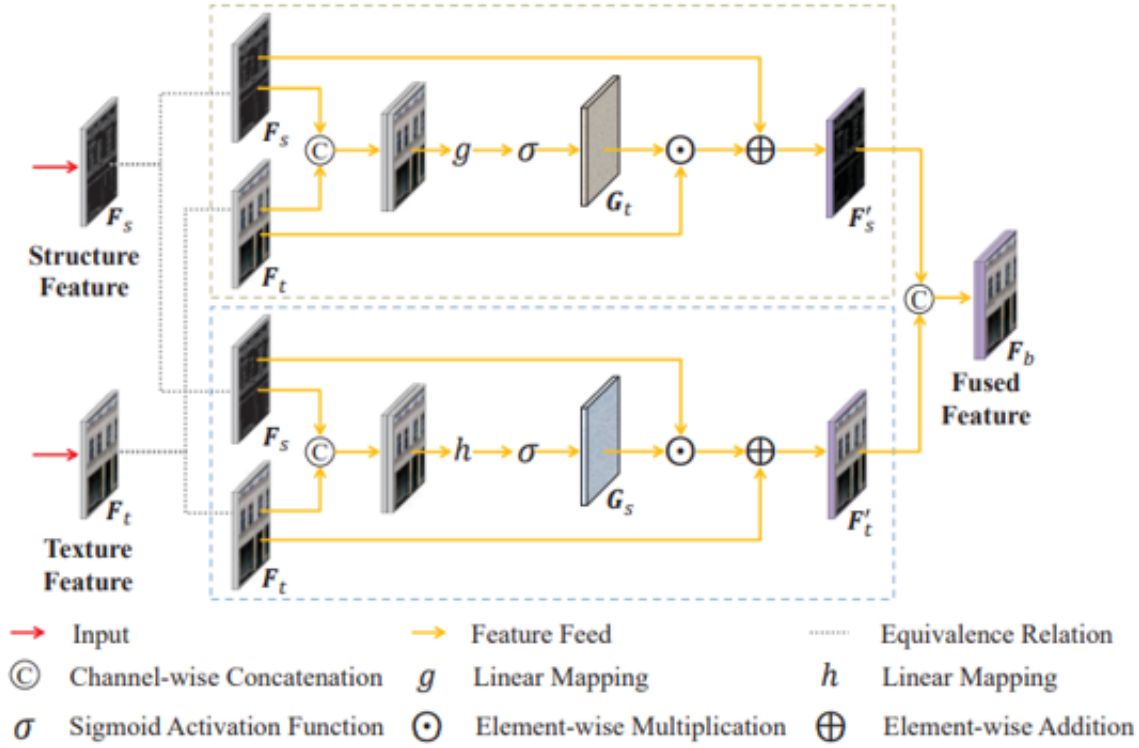


图 2: 双向门控特征融合模块

以对纹理特征图的处理为例：

$$G_t = \sigma(g(Concat(F_t, F_s))), \tag{1}$$

其中 Concat 代表通道级的拼接，$F_t$，$F_s$ 分别表示纹理特征图以及结构特征图，g() 是通过卷积层实现的一个映射函数，$\sigma$ 则是激活函数。上述公式实际上得到了门控值 $G_t$，表示了纹理特征会以多大的程度与结构特征相融合。

$$F_s' = \alpha(G_t \otimes F_t) \oplus F_s \tag{2}$$

上式中的，经过上式，结构特征图与经过门控函数的纹理特征图结合得到最终的结构特征图，而最终的纹理特征图也是同理。最终，互相融合的结构特征图将与纹理特征图拼接，得到融合的特征图。

$$F_b = Concat(F_s', F_t') \tag{3}$$

$$X'_f = \theta_{0x} + \theta_X \cdot X_f^{ARIMA} + \sum_{i=1}^{N_{obj}} \theta_{ix} \cdot O_{Xif}$$

$$Y'_f = \theta_{0y} + \theta_Y \cdot Y_f^{ARIMA} + \sum_{i=1}^{N_{obj}} \theta_{iy} \cdot O_{Yif}$$

### 3.4 背景特征融合模块

在经过双向门控特征融合模块后，还要经过背景特征融合模块。这个模块的设计是为了更好地利用图片中的未缺失区域来填补缺失区域，以提高局部区域与全局区域的相容性。该模块的框架图如下所示 3

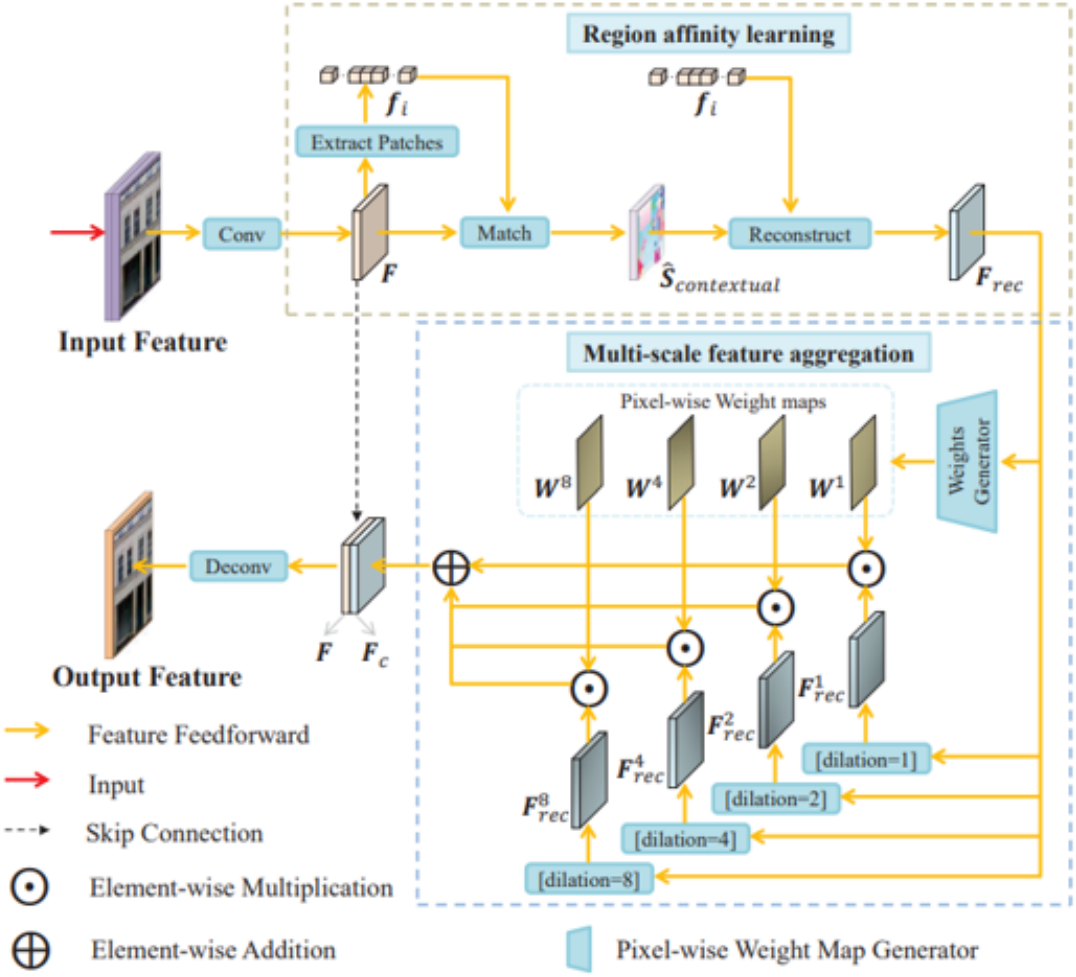


图 3: 背景特征融合模块

可以看到，在该模块中，首先将特征图提取为许多 3*3 的块，再计算块与块之间的余弦相似度，然后再经过 softmax 函数得到每个块的注意力分数，最后根据计算得到的注意力分数图得到重构的特征图。该模块也用到了多尺度空间，更好地融合了语义特征。

### 3.5 判别器

受到全局与局部 GAN[8]，门卷积[9]，马尔可夫 GAN[10]的启发，作者设计了一种双流的判别器由结构分支和纹理分支组成，通过估计纹理和结构的特征估计量来区分真实图像和生成图像。

### 3.6 损失函数定义

原文中的模型采用了联合的损失函数进行训练，这其中包括重构损失，感知损失，风格损失和对抗损失，以此来取得视觉上的逼真效果以及语义信息的合理性。其中重构损失表达式如下：

$$\mathcal{L}_{rec} = \varepsilon \left[ \| \boldsymbol{I}_{\boldsymbol{out}} - \boldsymbol{I}_{\boldsymbol{gt}} \|_1 \right] \tag{4}$$

其中 $I_{gt}$ 表示真实的图像，$I_{out}$ 表示生成器生成的图像，重构损失以它们之间的 l1 范数作为损失函数接下来的感知损失表达式如下：

$$\mathcal{L}_{perc} = \varepsilon \left[ \sum_i \| \phi_i(\boldsymbol{I}_{\boldsymbol{out}}) - \phi_i(\boldsymbol{I}_{\boldsymbol{out}}) \|_1 \right] \tag{5}$$

其中 $\phi_i()$ 表示第 i 个池化层的激活映射。风格损失函数用来保证风格的一致性，其表达式如下：

$$\mathcal{L}_{style} = \varepsilon \left[ \sum_i \| \psi_i(\boldsymbol{I}_{\boldsymbol{out}}) - \psi_i(\boldsymbol{I}_{\boldsymbol{out}}) \|_1 \right] \tag{6}$$

其中 $\psi_i()$ 表示激活映射的格拉姆矩阵。加上保证补全的图像的视觉真实性的对抗损失

$$\mathcal{L}_{adv}$$

，以及过程中监督纹理特征图以及结构特征图的中间损失

$$\mathcal{L}_{inter}$$

。最终的联合损失函数为上述损失函数的加权和：

$$\mathcal{L}_{joint} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{inter}\mathcal{L}_{inter} \tag{7}$$

## 4 复现细节

### 4.1 与已有开源代码对比

我所复现的论文为开源的工作，原文给出了源代码，地址为：https://github.com/Xiefan-Guo/CTSDG. 我的工作使用了文中提出的模型，因此使用了原文给出的全部代码，包括生成器部分和判别器部分。我的改进工作是利用这个图像补全的模型，通过数据预处理和微调训练，将其应用到人脸编辑领域。我的具体工作内容如下：1. 利用原文的用于图像补全的模型使用大规模的人脸数据集名人数据集（内含 202599 张人脸照片）以及 Irregular masks 数据集；对它进行重头训练，使该模型能够学习补全人脸特征；2. 利用 dlib 库的人脸关键点检测技术，针对每张人脸生成只遮盖嘴部的特定区域的掩码（如下图所示 4）；3. 之后使用特定的样本数量较少的数据集（3019 张笑脸人脸图像），以及制作的嘴部区域的掩码在第二步得到的模型上进行微调训练，从而使模型学习得到生成笑脸的能力；

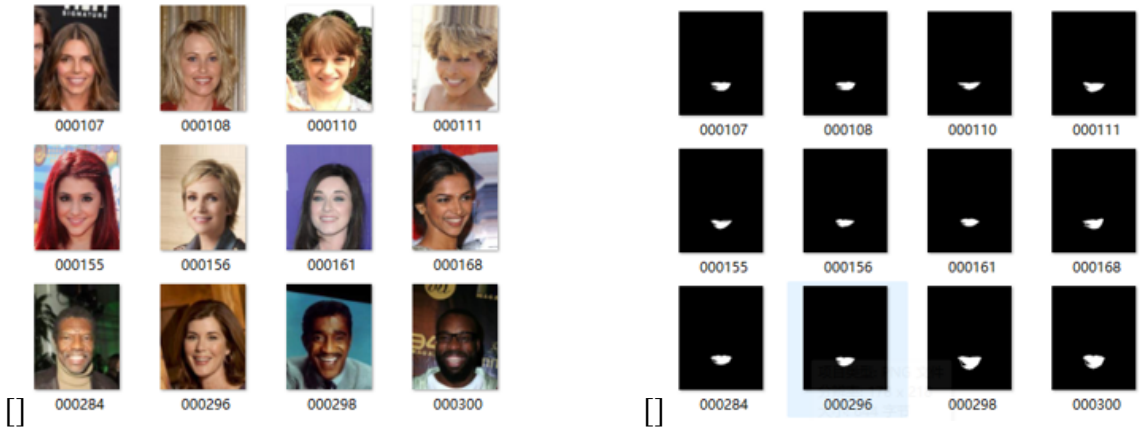


图 4: 用于微调训练的笑脸数据集及其嘴部掩码（部分）(a) 笑脸数据集 (b) 通过人脸关键点检测技术生成的嘴部掩码

此部分为必填内容。如果没有参考任何相关源代码，请在此明确申明。如果复现过程中引用参考了任何其他人发布的代码，请列出所有引用代码并详细描述使用情况。同时应在此部分突出你自己的工作，包括创新增量、显著改进或者新功能等，应该有足够差异和优势来证明你的工作量与技术贡献。

### 4.2 实验环境搭建

本次实验使用神经网络训练使用 python3.7.4，pytorch1.10.2；使用了 dlib 库的 68 点人脸关键点检测技术。

### 4.3 嘴部掩码的生成

**Procedure 1** 对人脸图像生成嘴部掩码

**Input:** images $X$, face detector $D$, predictor $P$
**Output:** masks $M$
**for** $l$ ***in*** *reference images indices* **do**
$face_l = \mathrm{D}(X_l)$ $keypoints_l = \mathrm{P}(face_l)$ $ROI_l = \text{Indices of mouth region}(\mathrm{FFM}(\text{keypoints}_l))$
$M_l = \text{FillConvexPoly}(\text{ROI}_l)$
**end**

### 4.4 创新点

原文提出的模型是用于图像补全的，在原文的实验中训练所用的掩膜都是根据覆盖率随机生成的掩码。之前也有许多深度学习的生成模型用来进行人脸编辑，但是这其中很多模型都需要成对的样本来进行训练。根据人脸编辑的目标：在保持其他区域不变的情况下，改变人脸的特定区域。于是对数据集进行筛选和预处理，创新性地将这个效果非常好的图像补全模型用于人脸编辑，在不需要很大数据量的成对的数据的情况下取得了不错的效果。

## 5 实验结果分析

对微调改进后的模型进行测试的结果如下图所示：

图 5: 测试的可视化结果

## 6 总结与展望

本文所复现的论文提出了一种新的双流图像修复方法，该方法通过同时建模结构约束的纹理合成和纹理引导的结构重建来恢复损坏的图像。通过这种方式，两个子任务交换有用的信息，从而相互促进。此外，引入了一个双向门控特征融合模块，然后引入了一个背景特征聚合模块，以优化结果，具有语义合理的结构和丰富的细节纹理。实验结果表明，该模型能够较好地解决这一问题，并优于目前最先进的模型。我利用该模型，先进行重头训练得到预训练模型，对训练数据集进行预处理然后对预训练模型进行微调训练，使其可以用于人脸编辑领域，这样的方法避免了人脸编辑模型需要大量成对数据集的问题。但是也存在一些问题，比如说，在数据预处理阶段，根据嘴部的关键点生成嘴部掩码，之后的编辑只在这个区域进行，变化的可拓展性很小，最显著的问题就在于如果嘴部太小的话，效果会十分不理想，而且笑脸涉及到的其他脸部肌肉不变化也会造成不协调。第二个问题在于，这样的方法一次训练只能改变一种表情或区域，单一模型不能实现脸部表情的连续变化。上述的两个问题是我们将来需要解决的，解决的可能思路是将脸部表情作为编码嵌入到生成器中进行训练。

## 参考文献


[1] LI J, HE F, ZHANG L, et al. Progressive Reconstruction of Visual Structure for Image Inpainting[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 5961-5970. DOI: 10.1109/ICCV.2019.00606.

[2] LIU H, JIANG B, SONG Y, et al. Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations[EB/OL]. arXiv. 2020. https://arxiv.org/abs/2007.06929.

[3] 佘嘉博. Research on Runge Phenomenon[J]. Advances in Applied Mathematics, 2019, 08(08): 1500-1510.

[4] BALLESTER C, BERTALMIO M, CASELLES V, et al. Filling-in by joint interpolation of vector fields and gray levels[J]. IEEE Transactions on Image Processing, 2001, 10(8): 1200-1211. DOI: 10.1109/83.935036.

[5] BERTALMIO M, SAPIRO G, CASELLES V, et al. Image Inpainting[C/OL]//SIGGRAPH '00: Pro-

ceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. USA: ACM Press/Addison-Wesley Publishing Co., 2000: 417-424. https://doi.org/10.1145/344779.344972. DOI: 10.1145/344779.344972.

[6] XU Z, SUN J. Image Inpainting by Patch Propagation Using Patch Sparsity[J]. IEEE Transactions on Image Processing, 2010, 19(5): 1153-1165. DOI: 10.1109/TIP.2010.2042098.

[7] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing[J/OL]. ACM Trans. Graph., 2009, 28(3). https://doi.org/10.1145/1531326.1531330. DOI: 10.1145/1531326.1531330.

[8] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and Locally Consistent Image Completion [J/OL]. ACM Trans. Graph., 2017, 36(4). https://doi.org/10.1145/3072959.3073659. DOI: 10.1145/3072959.3073659.

[9] YU J, LIN Z, YANG J, et al. Free-Form Image Inpainting With Gated Convolution[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 4470-4479. DOI: 10.1109/ICCV.2019.00457.

[10] LI C, WAND M. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks[C]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 702-716.