图像处理

结课作业

Paper Reading

Title：《Generative Image Inpainting with Contextual Attention》

Author：Yu J , Lin Z , Yang J , et al

组号：24

姓名：李贺航 石云天

袁萌启 王寅合

日期：2022.12.24

目　录

一、要点引入

1.1 文章解决痛点

由于基于传统 CNN 的图像修复网络感受野十分有限，不能有效的建立起破损区域和距离破损区域较远区域之间的联系，导致了目标区域边界上结构的畸变、纹理的模糊和周围区域的不连贯这一系列问题的出现。

1.2 核心思想

为了解决这些问题，作者提出了一种基于深度生成模型的方法，该方法可以合成新颖的图像结构，而且可以在网络训练期间明确地利用周围地图像特征作为图像修复地参考，以此来获得更好地预测。

1.3 文章创新点

文章的创新点在于作者提出了一种语境注意力层（contextual attention layer），来从距离远的区域提取近似待修复区域的特征。简单的说就是，对一个待修补的区域，通过卷积的方法，从整个图像匹配出和待修补区域比较相似的信息，利用这些信息来重建待修补区域，以此来提升网络的远距离信息抓取能力。

二、方法途径

2.1 模型结构

从整体上看，本文提出了一个由两阶段组成的由粗到细的修复网络，该网络是一个前馈地完全卷积神经网络，可以修复任意位置形状缺失地图像。第一阶段是一个简单的卷积网络（粗修复网络），通过不断修复确实缺失区域来产生损失值 reconstruction loss，修复出一个比较模糊地结果；第二阶段是内容感知层的训练，其核心思想是：使用已知的图像斑块特征来作为卷积核来加工生成出来的斑块，对第一阶段模糊的修复结果精细化。总体网络结构呈现细而深的形状，这使得所需参数能减少，从而提高效率。

2.1.1 第一阶段：粗修复网络

2.1.1.1 功能结构

在功能方面，该粗修复网络修补出缺失的概貌，提供雏形，供后续精细化修复使用，具体效果与结构见下图 2-1：

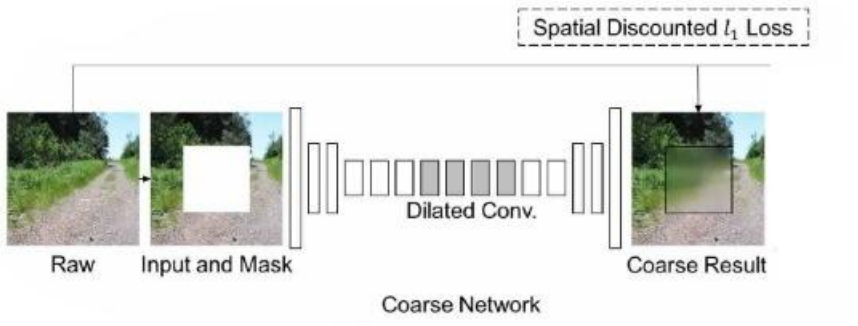

图 2-1 粗修复网络效果与结构[1]

在结构方面，作者借鉴了 Iizuka et al. 在论文《Globally and Locally Consistent Image Completion》[2]中提出的网络结构，具体要点如下：

1. 输入尺寸 256*256+一张用于标记空洞区域的二值 mask（1 表示源区域 0 表示目标区域）；

2. 网络的降采样和升采样分别由步长为 2 的卷积/转置卷积实现；

3. 网络中间灰色部分为空洞卷积，这使得使用同样参数的情况下输出像素能根据更大面积的输入进行计算得到，这对于填充非常重要，因为只有输入感受野足够大能够涵盖到空洞周围的信息，这样才能利用周围的信息计算目标区

域的像素值，这种扩大输入感受野的方式被称为 spatial support；

4. 网络的输出是与输入同样大小的三通道 RGB 图，且仅保留网络在填充区域的更改。

2.1.1.2 损失函数

Iizuka 的生成网络[2]的重构损失是生成图和真值图的l_2损失，而此处则是带空间衰减的l_1损失（Spatially discounted reconstruction loss）：直观上来说，在缺失区域的边界上修复的结果的歧义性（ambiguity），要远小于中心区域，即缺失区域越往内取值越不确定，所以在用真值图引导缺失区域重建时，离填充边缘不同距离的像素点应该被赋予不同的权重，离边缘越远权重应该越低，这样计算损失值时，不会因为中心结果和原始图像差距过大，从而误导训练过程。（此处默许不一定非要生成和真值图一模一样的内容，只要内容自然合理也是好算法，所以允许缺失区域中心区域“自由发挥”）。具体做法是使用一个带有权值的$mask\ M$，在M上，每一点的权值由γ^l来计算，其中γ被设定为 0.99，l是该点到最近的已知像素点的距离（街区距离）。在修复大面积缺失区域的时候，这种带衰减的损失值在提高修复质量上将更有效。

2.1.2 第二阶段：精修复网络

精修复阶段包含平行的两路编码器：上路分支是包含语义注意力层（contextual attention layer）的编码器，这是本文作者主要的创新点；下路分支的编码器就是类似 Iizuka 的常规的编码器[2]。两路编码器输出的特征图最后拼接在一起合成一个特征图，最后通过解码器生成修复结果，见下图 2-2：

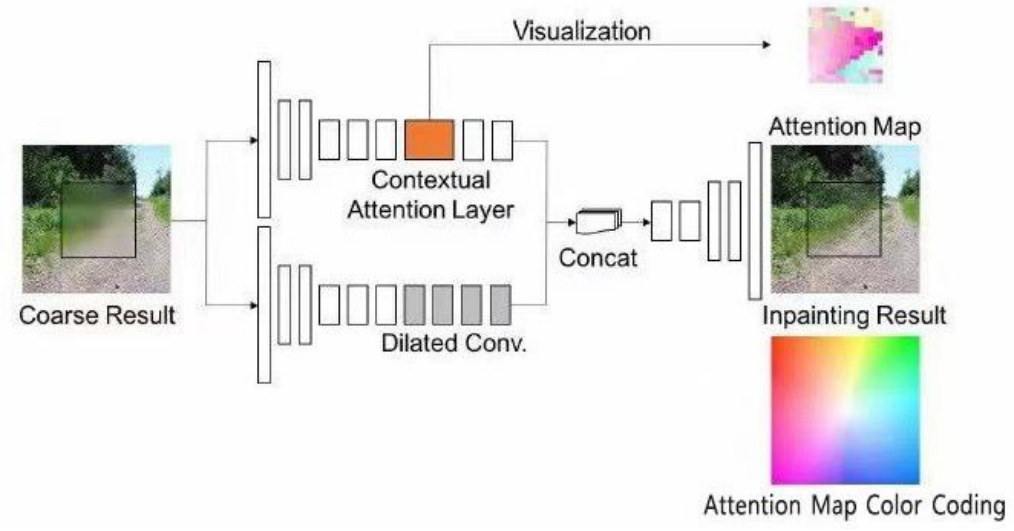

图 2-2 精修复网络效果与结构[1]

2.1.2.1 Encoder 上路分支

首先对于上路分支：包含语义注意力层的编码器，上文介绍过：传统卷积神经网络的缺陷是仅通过一层层的卷积，很难从从空间位置相隔较远的两个区域之间发生联系，所以为了克服这一限制，作者提出了语义注意力层（下图黄色块），它的功能是不受空间限制地从已知区域借鉴相似的特征信息，以此来生成缺失的信息，这是本文关键的创新点，我们对其重点讨论，其结构见下图 2-3：

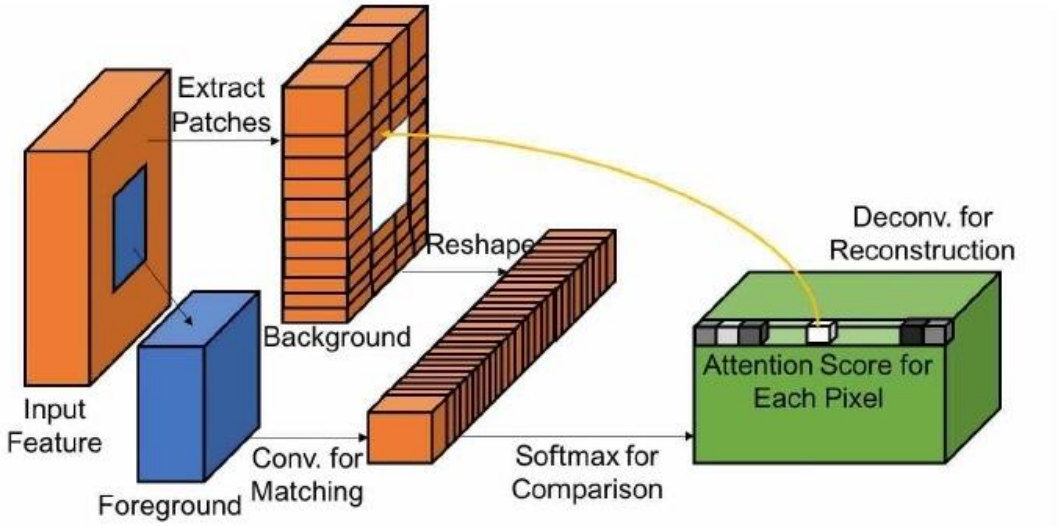

图 2-3 语义注意力层结构[1]

工作机制如下：输入的特征图包含前景（蓝色）和背景（黄色），前景是缺失区域对应的位置，背景是源区域对应的位置。（注意缺失区域内的内容不是空洞，在第一阶段粗修复里面已经填充了概貌，故其中有内容）。作者从整个背景区域提取出一堆 3x3 大小的 patches，然后将这些 patches 作为卷积核，对前景区域进行归一化卷积操作，这样就得到了图中绿色的输出特征图。记$f_{x,y}$是以(x,y)为中心的前景 patch，$b_{x'y'}$是以(x',y')为中心的背景 patch。显然，绿色特征图中每个像素(x,y)在不同通道的值就是以这个像素为中心的$f_{x,y}$和对应的$b_{x'y'}$的卷积结果。这个卷积的意义实际上就是$f_{x,y}$和$b_{x'y'}$的余弦相似度：

$$s_{x,y,x',y'} = < \frac{f_{x,y}}{\|f_{x,y}\|}, \frac{b_{x',y'}}{\|b_{x',y'}\|} >$$

式中尖括号为内积运算，内积运算先求对应元素乘积然后相加求和，与卷积是一致，因此通过卷积可以高效实现这种相似度匹配的过程，卷积的使用同时还使得这一过程可微分，因此可以反向求导进行更新参数。随后还需要对绿色特征图上每个像素在通道维度上进行 softmax，这样便得到了 attention score：

$$s^{*}_{x,y,x',y'} = soft\max_{x',y'}(\lambda s_{x,y,x',y'})$$

式中λ是一个常数，softmax 脚标x',y'代表该运算是按照x',y'方向进行的，其含义为按背景 patches 集合$\{b_{x'y'}\}$计算，即在绿色特征图中按通道维度计算，反映在代码里就是做一个 channel-wise softmax。而绿色的 attention score map 的物理意义为相似度图（互相关图），其中的每个数值都代表一个前景像素和一个背景像素之间的相似度。

最后，再次使用背景的 patches 集合$\{b_{x'y'}\}$作为卷积核，对得到的绿色的 attention score map 进行转置卷积操作（重合的像素取平均以避免棋盘效应），用于重建前景。通过理解转置卷积的计算过程，得到重建机理为：如果$f_{x,y}$和$b_{x'y'}$越相似，那么它们所对应的 attention score 的值$s^*_{x,y,x',y'}$越高，这使得$b_{x'y'}$对转置卷积在(x,y)位置输出的结果的贡献就越大，即前景（未知）区域向背景（已知）区域借鉴相似的纹理特征，来进一步帮助目标区域的重建。小结以上过程，见下图 2-4：

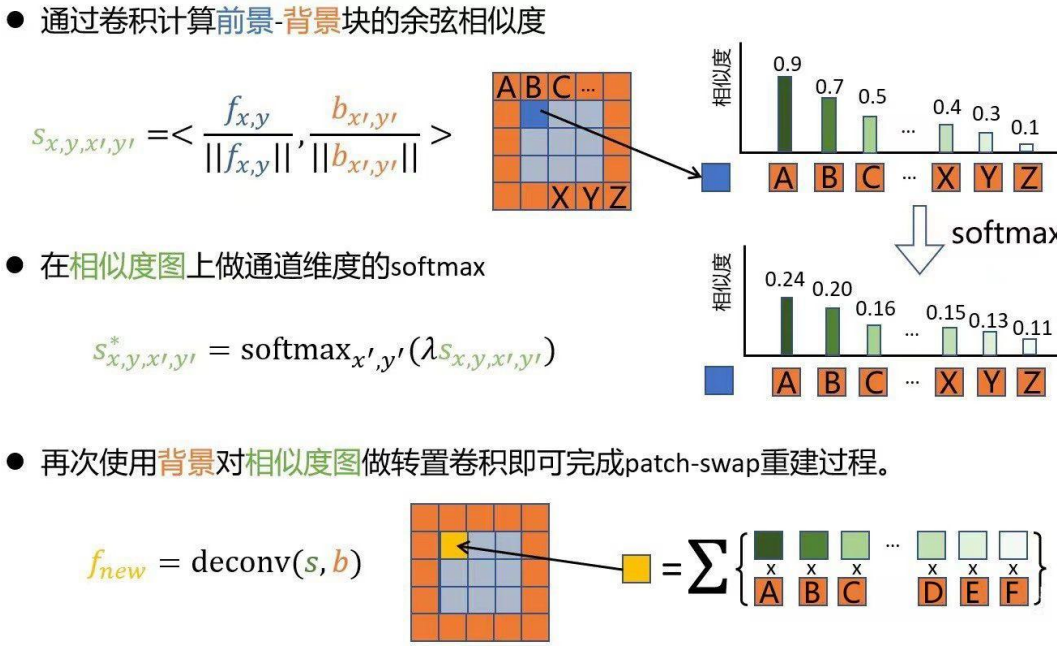

图 2-4 上路分支计算过程总结

2.1.2.2 Encoder 下路分支与解码器

下路分支结构较为简单常规：常规卷积+空洞卷积，见下图 2-5

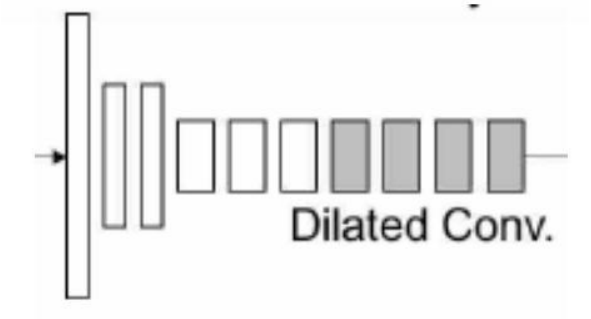

图 2-5 下路分支结构[1]

对于解码器部分，其通过拼接语义注意力支路的输出特征图与常规支路的输出特征图，进行解码，还原出精修复结果，送入双判别器内。

2.1.2.3 双判别器

判别器主要起对抗作用，提供对抗损失来更好地锻炼生成器。作者借鉴了 Iizuka 提出的 GL 模型里的全局+局部双判别器[2]，结构也基本相似，具体结构见下图 2-6：

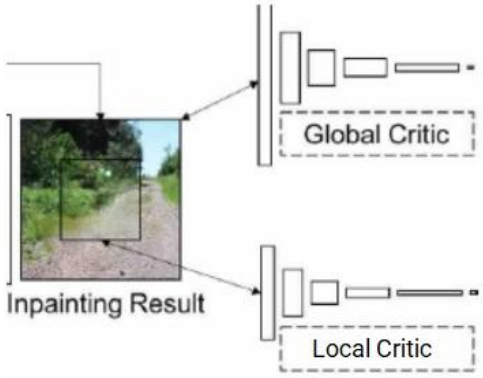

图 2-6 双判别器结构[1]

两者不同之处有如下几点：

1. 本文去掉了 batchnorm，作者认为其会破坏颜色；

2. 本文使用镜像 padding 代替了原来的 zero padding，同时使用 ELU 替代了原来的 Relu；

3. GL 模型把全局和局部两个判别器末端的全连接层拼在一起，最后只输出一个值，本文将两个判别器分开，两者各自产生判别值；

4. lizuka 的论文[2]中使用 Goodfellow 传统对抗损失，而本文中的判别器使用了$WGAN-GP\ loss$（具体说明见 2.2.4 节）。

2.1.2.4 损失函数

本阶段作者主要使用了两个损失函数：

1. 重建损失：生成图和真值图在目标区域差值的l_1范数（带有空间权重衰减，与第一阶段相同）；

2. 全局-局部$WGAN-GP\ loss$，其公式如下：

$$\min\max E_{x\sim P_r}[D(x)]-E_{\tilde{x}\sim P_g}[D(\tilde{x})]+\lambda E_{\hat{x}\sim P_{\hat{x}}}\left([\|\nabla_{\hat{x}}D(\hat{x})\|_2-1]\right)^2$$

式中G和D代表生成器和判别器，ζ是满足 Lipschitz 连续性的所有函数集合。x是服从真实分布P_r的真实样本，$\tilde{x}=G(z)$是服从生成（拟合）分布P_g的生成样

本（P_g是通过生成器函数G隐式定义的分布，一般没有显性的表达式），$z=x\odot m$是输入的缺失图像，m是$mask$，$\hat{x}$是x和$\tilde{x}$的线性组合。

2.1.3 模型训练过程伪代码

表 2-1 模型训练过程伪代码

Algorithm 1 Training of our proposed framework.[1]

while G has not coverged **do**
  for $i=1,...,5$ **do**
    Sample batch images x from training data;
    Generate random masks m for x;
    Construct inputs $z \leftarrow x \odot m$;
    Get predictions $\tilde{x} \leftarrow z + G(z,m) \odot (1-m)$
    Sample $t \sim U[0,1]$ and $\hat{x} \leftarrow (1-t)x + t\tilde{x}$;
    Update two critics with x, $\tilde{x}$ and $\hat{x}$;
  end for
  Sample batch images x from training data;
  Generate random masks m for x;
  Update inpainting network G with spatial dis-
  counted l_1 loss and two adversarial critic losses
end while

2.2 论文效果展示

利用数据集 CalebA、DXD 与 ImageNet 进行训练与测试，得到结果见下图 3-1：

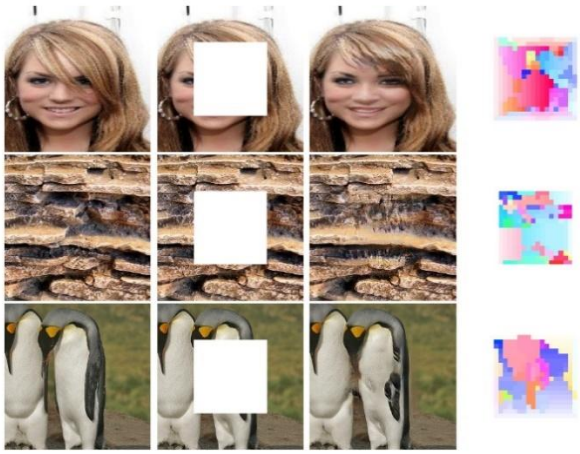

图 3-1 测试结果展示（所用数据集从上到下依次为：CalebA、DXD、ImageNet）[1]

2.3 创新拓展

2.3.1 聚焦人脸区域

本次实验聚焦人脸区域，降低了样本空间大小，同时降低了映射空间的大小，相当于进行了空间上的降维，降低了本次实验中对网络优化的要求，更容易优化到局部最优解的效果。比起原文在各种图像上进行生成，只关注人脸区域，可以提升其在该方面的效果。

本次实验中，使用论文作者开源的预训练模型。虽然作者将模型应用到了各种图像上，但由于其在训练过程中使用了 CelebA 作为训练集的一部分，所以将其预训练模型应用到人脸区域是可行的。以论文作者的预训练模型为基础，使用 CelebHQ 高清人脸数据集进行进一步训练，使得最终得到的训练结果聚焦于人脸区域。

2.3.2 固定生成范围

过去，人脸中容易被遮挡的部位大多为眼睛，被墨镜、眼罩之类的物体遮挡。但是，由于近年来疫情的蔓延与发展，口罩成为了人们出门在外必不可少的物品之一。因此，在现在及将来很长一段时间内，口罩都将是遮挡人脸的主要物品。而且口罩遮挡的区域最大能够达到 1/2 的人脸区域，对于验证原模型及改进模型的有效性非常有效，因此本次实验固定遮挡区域为眼睛以下的鼻子及口部区域。

2.3.3 修改 mask 属性

论文中选用的是纯白(255,255,255)的颜色作为遮挡部分的颜色。但是人脸不可能是纯白色。为了提高生成效果，应选取 mask 颜色较为接近人脸颜色。为此，遍历 CelebHQ 高清人脸数据集中所有人脸，计算出各种人脸中 R、G、B 三个颜色通道的均值，作为修改后的 mask 颜色。

2.3.4 增加关键点约束

人脸是具有高度结构性信息的物体，日常生活中能很轻易的分辨不同的人，有很大一部分都是由于人脸的结构性信息导致的。原始论文中并没有把应用聚

焦于人脸，故没有办法去增加这种与实际应用高度相关的上层信息约束。所以，为了能够使得生成效果更好，在模型中加入 landmark 约束，用于提取人脸的结构化信息。

对于提取关键点，使用 Insightface 项目。InsightFace 是一个开源的 2D&3D 深度人脸分析工具箱，基于各流行 AI 框架，高效地实现了丰富多样的人脸识别、人脸检测和人脸对齐的最先进算法，针对训练和部署进行了优化。选取身边 50 个人，每个人拍 10 张不同的照片，再选取 CelebA 的 950 个人，共 1000 人 5000 张人脸图片的数据集，使用 InsightFace 进行识别，识别效果高达 99.6%。因此，使用 Insightface 提取人脸 2D 图片的 106 个关键点。在原始图片和 Inpainting 结果的 2D 图片中分别提取 106 个关键点，利用关键点计算l_1损失。使用这个损失进一步优化 Refinement Network。这样，整个模型的 Refinement Network 优化总共有四个部分（如下图 4-1 所示）：第一部分是关于像素级的损失；第二、三部分使用了双判别器，进行 feature map 级约束；第三部分就是本次实验新增的 landmark 结构化信息的约束。由于增加了一个约束，最终得到的效果较原论文更好，有了更强的人脸结构信息。

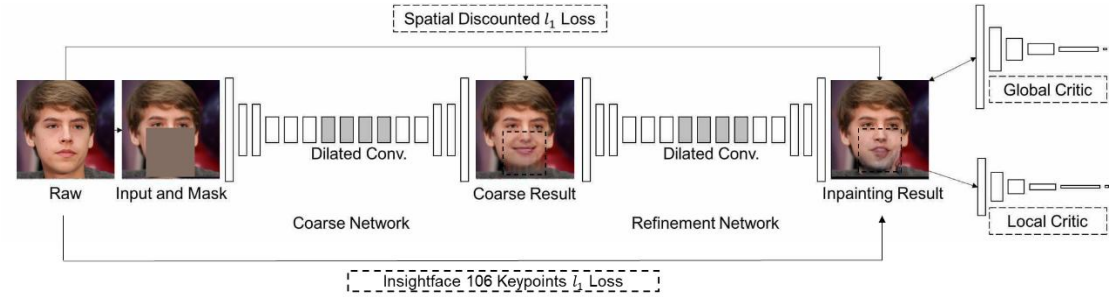

图 4-1 Refinement Network 优化

三、拓展实验结果

在对真实照片完成遮挡后分别在使用 landmark 约束和不使用 landmark 约束方法下对照片进行修复，通过结果可以看出：未使用 landmark 约束修复出来的照片，由一些严重偏离了人脸的基本结构，例如不使用 landmark 组的第二张和第五张照片，人脸的嘴巴和鼻子几乎被挤压到了一起，而使用 landmark 约束修复出的照片则能较好的维持人脸的基本结构，有着更好的修复效果，具体效果见下图 3-1：

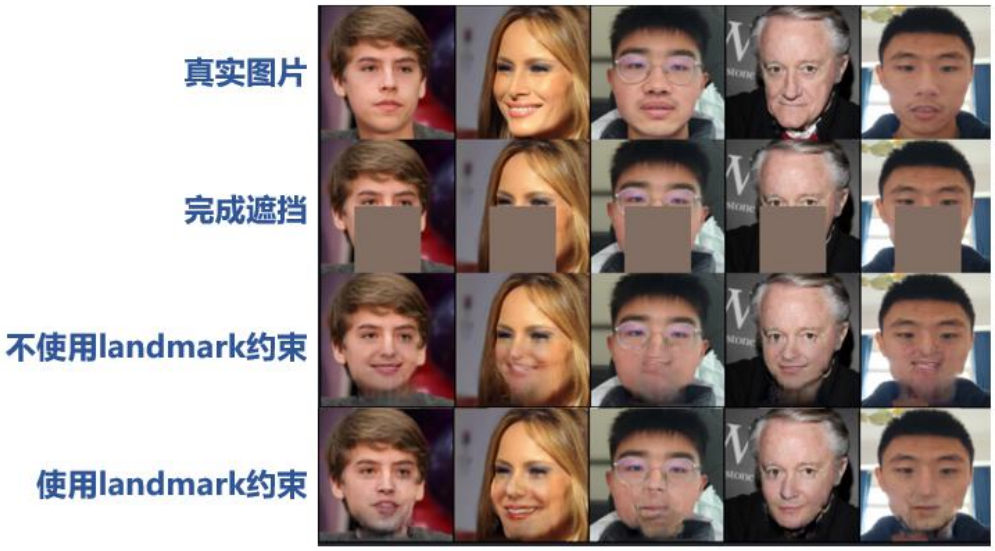

图 3-1 landmark 约束人脸修复效果

四、结论

参考论文的作者提出了粗细网络图像生成修复的框架并介绍了带有内容感知的模型，这一工作在提高图像修复结果上有着很大的意义。我们在作者工作的基础上，聚焦于有遮挡的人脸修复重建，在现有的模型基础上增加了 landmark 约束，经过实验验证，该约束的添加相较于未添加该约束，有着更好的人脸修复重建效果。

参考文献

[1] Yu J， Lin Z， Yang J , et al. Generative Image Inpainting with Contextual Attention[J]. IEEE, 2018.

[2] Iizuka S， Simo-Serra E， Ishikawa H . Globally and locally consistent image completion[J]. ACM Transactions on Graphics (TOG), 2017, 36(4CD):107.1-107.14.

[3] Deng Y， Yang J， Xu S , et al. Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2020.

成员签名