基于内容注意力机制的生成图像修复

Jiahui Yu[1] Zhe Lin[2] Jimei Yang[2] Xiaohui Shen[2] Xin Lu[2] Thomas S. Huang[1]

[1]University of Illinois at Urbana-Champaign [2]Adobe Research

图 1：我们的方法在自然场景，面部和纹理图像上的修复结果示例。 缺失区域以白色显示。 在每对图像中，左边是输入图像，右边是我们训练的生成神经网络的直接输出，没有任何后处理。

摘 要

最近基于深度学习的方法已经显示出在图像中修复大的缺失区域这一挑战性任务的有希望的结果。这些方法可以生成视觉上合理的图像结构和纹理，但通常会产生扭曲的结构或模糊纹理，与周围区域保持一致。这主要是由于卷积神经网络在明确地借用或复制来自遥远的空间位置的信息时的无效性。另一方面，传统的纹理和补丁合成方法在需要从周围区域借用纹理时特别适合。在这些观察的推动下，我们提出了一种新的基于深度生成模型的方法，该方法不仅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的图像特征作为参考来进行更好的预测。该模型是一个前馈的完全卷积神经网络，可以在任意位置处理具有多个孔的图像，并且在测试时间内具有可变大小。在包括面部（CelebA，CelebA-HQ），纹理（DTD）和自然图像（ImageNet，Places2）在内的多个数据集上的实验表明，我们提出的方法比现有方法产生更高质量的修复结果。代码，演示和模型可在以下网址获得：

https://github.com/JiahuiYu/generative_inpainting

1. 介绍

填充图像的缺失像素（通常称为图像修复或完成）是计算机视觉中的重要任务。它在照片编辑，基于图像的渲染和计算摄影中有许多应用[3,25,30,31,36,41]。图像修复的核心挑战在于为缺失区域合成视觉逼真和语义合理的像素，这些像素与现有区域共存。

早期作品[3,14]试图使用类似于纹理合成的思想来解决问题[10,11]，即通过将背景补片匹配和复制到从低分辨率到高分辨率或从孔边界传播的孔中。这些方法尤其适用于后台修复任务，并在实际应用中得到广泛应用[3]。然而，由于他们假设可以在背景区域的某处找到错过的补丁，因此他们不能为新的图像内容产生幻觉，因为其中的修复区域涉及复杂的，非重复的结构（例如，面部对象）。而且，这些方法无法捕获高级语义。

深度卷积神经网络（CNN）和生成对抗网络（GAN）[12]的快速进展启发了最近的作品[17,27,32,41]，将绘画中的条件图像生成问题表达为高级识别和低级级别像素合成被制定成卷积编码器 - 解码器网络，

与对抗性网络共同训练，以鼓励生成和现有像素之间的一致性。这些作品被证明可以在高度结构化的图像中生成合理的新内容，例如面部，物体和场景。

不幸的是，这些基于 CNN 的方法经常产生边界伪影，扭曲的结构和与周围区域不一致的模糊纹理。我们发现这可能是由于卷积神经网络在模拟远距离上下文信息和空洞区域之间的长期相关性方面的无效性。例如，为了允许像素受到 64 个像素的内容影响，它需要至少 6 层 3×3 的卷积，具有扩张因子 2 或等效物[17,22]。然而，扩张的卷积样本具有来自规则和对称网格的特征，因此可能无法权衡其他感兴趣的特征。请注意，最近的一项工作[40]试图通过优化生成的补丁与已知区域中的匹配补丁之间的纹理相似性来解决外观差异。虽然提高了视觉质量，但是这种方法被数百个梯度下降迭代所拖累，并且花费几分钟来处理 GPU 上分辨率为 512 × 512 的图像。

我们提出了一个统一的前馈生成网络，它具有一个新颖的上下文关注层，用于图像修复。我们提议的网络包括两个阶段。第一阶段是一个简单的扩张卷积网络，训练有重建损失以粗略地丢失缺失的内容。背景注意力集中在第二阶段。上下文关注的核心思想是使用已知补丁的特征作为卷积过滤器来处理生成的补丁。它是通过卷积设计和实现的，用于将生成的补丁与已知的上下文补丁匹配，通道 softmax 用于权衡相关补丁和解卷积以使用上下文补丁重建生成的补丁。上下文注意力模块还具有空间传播层以鼓励注意力的空间一致性。为了让网络产生想象的新内容，我们还有另一个与情境关注通道并行的卷积通道。这两个通道被聚合并馈入单个解码器以获得最终输出。整个网络经过端到端的重建损失和两次 Wasserstein GAN 损失[1,13]，其中一位评论家关注全局图像，而另一位评论家则查看失踪区域的局部补丁。

在包括面部，纹理和自然图像在内的多个数据集上进行的实验表明，所提出的方法可以产生比现有数据更高质量的修复结果。示例结果如图 1 所示。

我们的贡献总结如下：

- 我们提出了一种新颖的语境关注层，以便在远处的空间位置明确地参与相关的特征片。
- 我们介绍了几种技术，包括修复网络增强，全局和局部 WGAN [13]以及空间折扣重建损失，以提高基于当前最先进的生成绘图网络的训练稳定性和速度[17]。 因此，我们能够在一周而不是两个月内训练网络。
- 我们统一的前馈生成网络在各种挑战性数据集上实现了高质量的修复效果，包括 CelebA 面部[28]，CelebA-HQ 面部[22]，DTD 纹理[6]，ImageNet [34]和 Places2 [43]。

2. 相关工作

2.1. 图像修复

用于图像修复的现有工作可以主要分为两组。第一组代表传统的基于扩散或基于补丁的方法，具有低级别的特征。第二组试图通过基于学习的方法来解决修复问题，例如，训练深度卷积神经网络以预测缺失区域的像素。

传统的扩散或基于补丁的方法，例如[2,4,10,11]，通常使用变分算法或补丁相似性来将信息从背景区域传播到孔。这些方法适用于静止纹理，但仅限于非固定数据，如自然图像。 Simakov 等人[36]提出了一种基于双向补丁相似性的方案，以更好地模拟非静态视觉数据，以重新定位和修复应用程序。然而，补丁相似性的密集计算[36]是一种非常昂贵的操作，它禁止这种方法的实际应用。为了应对这一挑战，我们提出了一种名为 PatchMatch [3]的快速最近邻域算法，它已经为包括修复在内的图像编辑应用提供了显着的实用价值。

最近，深度学习和基于 GAN 的方法已成为图像修复的有前途的范例。最初的努力[23,39]训练卷积神经网络用于小区域的去噪和修复。 Con-word Encoders [32]首先训练深度神经网络，用于绘制大孔。它被训练成在 128×128 图像中完成 64×64 的中心区域，具有 2 像素重建损失和生成对抗性损失作为目标函数。最近，Iizuka 等人[17]通过引入全局和局部鉴别器作为经验损失来改进它。全局鉴别器评估完成的图像是否作为一个整体是连贯的，而局部的鉴别者则关注以生成的区域为中心的小区域以强制局部一致性。另外，Iizuka 等人 [17]使用扩张卷曲修复网络，

替换在 Con-word 编码器中采用的通道式完全连接层，两种技术都被提出用于增加输出神经元的感受域。同时，有几项研究侧重于生成性面部修复。 Yeh 等人 [41]在受损图像的潜在空间中搜索最接近的编码并解码以获得完成的图像。李等人 [27]为面部完成引入额外的面部解析损失。然而，这些方法通常需要后处理步骤，例如图像混合操作，以增强孔边界附近的颜色一致性。

一些作品[37,40]遵循图像样式[5,26]的想法，将修复制定为优化问题。例如，杨等人 [40]提出了一种基于图像内容和纹理约束的联合优化的多尺度神经补片合成方法，它不仅可以保留上下文结构，还可以通过匹配和调整补丁与最相似的中间层来产生高频细节。深度分类网络的特征相关性。这种方法显示了有希望的视觉效果，但由于优化过程而非常缓慢。

2.2. 注意力模型

在深度卷积神经网络中已经有许多关于学习空间注意力的研究。在这里，我们选择回顾一些与提议的情境关注模型相关的代表性问题。 Jaderberg 等人 [19]首先提出一种称为空间变换网络（STN）的参数空间注意模块，用于对象分类任务。该模型具有一个定位模块，用于预测全局仿射变换到扭曲特征的参数。但是，这个模型假定了一个全局变换，因此不适合建模补丁方面的注意力。周等人 [44]引入外观流程以预测偏移矢量，该偏移矢量指定应移动输入视图中的哪些像素以重建用于新颖视图合成的目标视图。根据我们的实验，该方法被证明对于匹配相同物体的相关视图是有效的，但是在预测从背景区域到孔的流场方面是无效的。最近，戴等人[8]和 Jeon 等人[20]建议学习空间注意力或主动卷积核。这些方法可以更好地利用信息来在训练期间使卷积核形状变形，但是当我们需要从背景中借用精确特征时，这些方法可能仍然有限。

3. 改进的生成修复网络

我们首先通过对最近最先进的修复模型[17]进行复制和改进来构建我们的绘图网络中的基线生成图像，该模型已经显示出有希望的视觉效果，用于修复面部图像，建筑物外墙和自然的图像。

粗到精网络架构 我们改进模型的网络架构如图 2 所示。我们遵循与[17]中相同的输入和输出配置进行训练和推理，即生成器网络采用白色像素的图像填充孔和二进制掩码，指示孔区域作为输入对，并输出最终完成的图像。我们将输入与相应的二元掩模配对，以处理具有可变大小，形状和位置的孔。网络的输入是 256×256 图像，在训练期间随机采样矩形缺失区域，并且训练的模型可以拍摄具有多个孔的不同尺寸的图像。

在图像修复任务中，感受野的大小应该足够大，Iizuka 等人 [17]为此目的采用相反的卷积。为了进一步扩大感知领域并稳定培训，我们引入了两阶段粗到细网络架构，其中第一个网络进行初始粗略预测，第二个网络将粗略预测作为输入并预测重新定义的结果。对粗网络进行明确的重建损失训练，同时通过重建和 GAN 损失训练细化网络。直观地，细化网络看到比具有缺失区域的原始图像更完整的场景，因此其编码器可以比粗网络学习更好的特征表示。这种两阶段网络架构在 spir 中类似于剩余学习[15]或深度监督[24]。

此外，我们的修复网络设计为一个薄而深的方案，以提高效率，并且与[17]中的参数相比具有更少的参数。在层实现方面，我们对所有卷积层使用镜像填充并删除批量标准化层[18]（我们发现它会降低颜色一致性）。此外，我们在[17]中使用 ELU [7]作为激活函数而不是 ReLU，并剪切输出滤波器值而不是使用 tanh 或 sigmoid 函数。此外，我们发现，在[17]中，GAN 训练的全局和局部特征表示优于特征连接。更多细节可以在补充材料中找到。

全局和局部的 Wasserstein GANs 与先前依赖 DCGAN [33]进行对抗性监督的生成网络[17,27,32]不同，我们提出使用 WGAN-GP 的修改版本[1,13]。我们将 WGAN-GP 损失附加到第二阶段细化网络的全局和局部输出，以实现全局和局部的一致性，受到[17]的启发。众所周知，WGAN-GP 损失在图像生成任务方面优于现有的 GAN 损失，并且当它们都使用 1 个距离度量时，它与“1 重建损失”结合使用时效果很好。

具体而言，WGAN 使用 EM 距离（a.k.a.Wasserstein-1）距离 W（Pr; Pg）来比较生成的和实际的数据分布。它的目标函数是通过应用 Kantorovich-Rubinstein 构造的

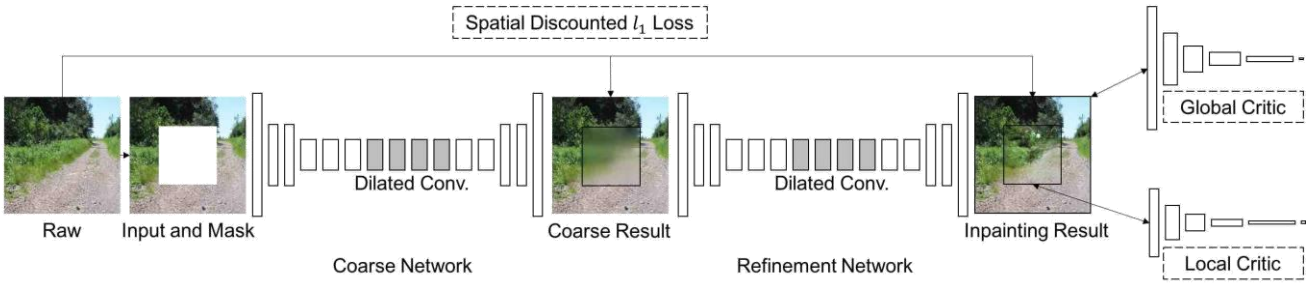

图 2:我们改进的生成性修复框架概述。 粗略网络明确地训练有重建损失，而精炼网络训练有重建损失，全局和局部 WGAN-GP 对抗性损失。

二元性:

$$\min_G \max_{D\in\mathcal{D}} E_{\mathbf{x}\sim\mathbb{P}_r}[D(\mathbf{x})] - E_{\tilde{\mathbf{x}}\sim\mathbb{P}_g}[D(\tilde{\mathbf{x}})],$$

其中 D 是 1-Lipschitz 函数的集合，Pg 是由 x~ = G(z) 隐式定义的模型分布，z 是生成器的输入。Gulrajani 等 [13] 提出了一个带有梯度惩罚项的 WGAN 的改进版本

$$\lambda E_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}(\|\nabla_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}})\|_2 - 1)^2,$$

其中 x^是从分布 Pg 和 Pr 采样的点之间的直线采样。原因是 D 的梯度在所有点 x^=(1-t)+tx^上直线应直接指向当前样本 x^，意思是 $\nabla_{\hat{\mathbf{x}}}D^*(\hat{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}-\hat{\mathbf{x}}}{\|\tilde{\mathbf{x}}-\hat{\mathbf{x}}\|}$

对于图像修复，我们仅尝试预测孔区域，因此梯度损失应仅应用于孔内的像素。 这可以通过梯度和输入掩码 m 的乘法来实现，如下所示:

$$\lambda E_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}(\|\nabla_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}}) \odot (\mathbf{1}-\mathbf{m})\|_2 - 1)^2,$$

其中掩码值为 0 表示缺失像素，1 表示 else-where。 在所有实验中都设置为 10。

我们使用像素方式的 1 损失（而不是[17]中的均方误差）和 WGAN 对抗性损失的加权和。 请注意，在原始空间中，WGAN 中的 Wasserstein-1 距离基于 1-EM 距离:

$$W(\mathbb{P}_r,\mathbb{P}_g) = \inf_{\gamma\in\prod(\mathbb{P}_r,\mathbb{P}_g)} E_{(\mathbf{x},\mathbf{y})\sim\gamma}[\|\mathbf{x}-\mathbf{y}\|],$$

其中（Pr; Pg）表示所有联合分布的集合，（x; y）其边缘分别为 Pr 和 Pg。 直观地，像素方式的重建损失直接将空洞图像回归到当前真实图像，而 WGAN 明确地学习匹配潜在的正确图像并用对抗性梯度训练发生器。 由于两种损耗均以像素为单位测量 1 个距离，因此组合损失更容易训练并使优化过程更稳定。

空间代价的重建损失 修复问题涉及像素的想象，因此它可以为任何给定的环境提供许多合理的解决方案。在困难的情况下，合理的完整图像可能具有与原始图像中的图像非常不同的斑块或像素。由于我们使用原始图像作为计算重建损失的唯一基础事实，因此强制执行这些像素中的重建损失可能会误导卷积网络的训练过程。

直观地，孔边界附近的缺失像素比靠近孔中心的像素具有更少的模糊度。这类似于加强学习中观察到的问题。当长期奖励在采样过程中有很大的变化时，人们会使用时间贴现的重新数据而不是采样轨迹[38]。受此启发，我们使用权重掩模 M 引入空间代价的重建损失。掩模中每个像素的权重计算为γ，其中γ是像素与最近的已知像素的距离。在所有实验中设定为 0.99。

在[32,41]中也探讨了类似的加权思想。在[41]中提出的重要性加权上下文损失由固定窗口内的未损坏像素的比率（例如 7×7）进行加权。 Pathak 等人[32]预测在边界地区有一个稍大的补丁，损失权重更高（×10）。对于修复大孔，建议的代价损失对于改善视觉质量更有效。我们在实施中使用贴现的 1 重建损失。

通过以上所有改进，我们的基线生成修复模型收敛速度比[17]快得多，从而获得更准确的修复结果。对于 Places2 [43]，我们将训练时间从[17]报告的 11,520 GPU 小时（K80）减少到 120 GPU 小时（GTX 1080），这几乎是 100 加速。而且，不再需要后处理步骤（图像混合）[17]。

4. 基于上下文注意力模型的图像修复

卷积神经网络逐层处理具有局部卷积核的图像特征，因此不是

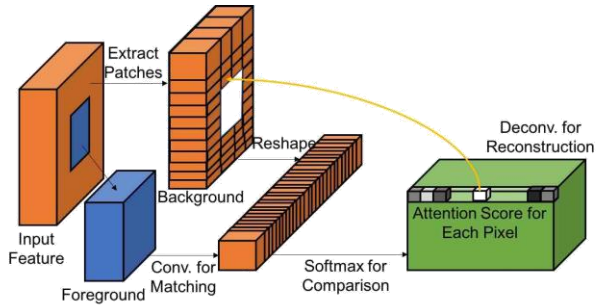

图 3：上下文关注层的图示。 首先，我们使用卷积来计算前景补丁与背景补丁的匹配分数（作为卷积滤波器）。 然后我们应用 softmax 来比较并获得每个像素的注意力得分。 最后，我们通过对注意力得分进行去卷积来重建具有背景斑块的前景补丁。 上下文关注层是可区分的并且是完全卷积的。

有效地借用遥远的空间位置的特征。 为克服这一局限，我们考虑了注意力机制，并在深层生成网络中引入了一种新的上下文关注层。 在本节中，我们首先讨论上下文关注层的细节，然后讨论如何将其集成到我们的统一修复网络中。

4.1. 上下文注意力模型

上下文关注层学习从已知背景补丁借用或复制特征信息的位置以生成缺失补丁。 它是可微分的，因此可以在深度模型中进行训练，并且可以完全卷积，这允许在任意分辨率下进行测试。

匹配和参加 我们考虑我们想要将缺失像素（前景）的特征与周围环境（背景）匹配的问题。 如图 3 所示，我们首先在背景中提取补丁（3×3）并将它们重新整形为卷积滤波器。 为了匹配前景补丁{fx,y }与背景{bx',y'}，我们用正常化的内积（余弦相似度）进行测量

$$s_{x,y,x',y'} = \langle \frac{f_{x,y}}{||f_{x,y}||}, \frac{b_{x',y'}}{||b_{x',y'}||} \rangle,$$

其中 Sx,y,x',y'表示以背景（x',y'）和前景（x,y）为中心的补丁的相似性。 然后我们用 x'y'的维度与缩放的 softmax 来衡量相似度,得到每个像素的注意力得分:

S*x,y,x',y'=SOFTMAXx',y'(λSx,y,x',y')，其中 λ 是一个常数值,这个有效地实现为卷积和通道方式 softmax。 最后，我们重新使用提取的补丁{bx',y'}作为反卷积滤波器来重建前景。重叠像素的值被平均。

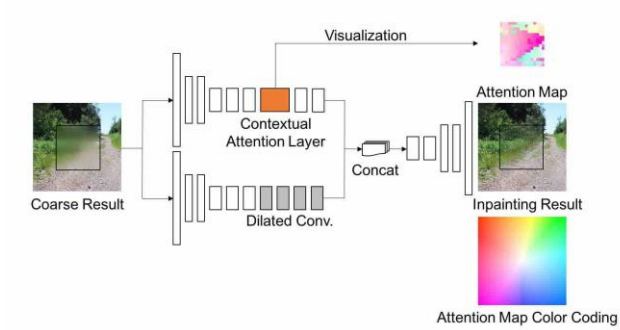

图 4：基于第一编码器 - 解码器网络的粗略结果，引入两个并行编码器，然后合并到单个解码器以获得修复结果。 对于注意力图的可视化，颜色表示前景中每个像素的最感兴趣的背景色块的相对位置。 例如，白色（彩色编码图的中心）表示像素在其自身上，粉红色在左下方，绿色表示在右上方。

注意力传播 我们通过传播（融合）进一步鼓励注意力的一致性。一致性的想法是前景补丁的移位可能对应于背景补丁中的相同移位以引起注意。 对于样例, S*x,y,x',y'通常具有接近，S*x+1,y,x'+1,y' 的值。 为了模拟和鼓励注意力图的一致性，我们进行左右传播，然后是内核大小为 k 的自上而下的传播。以左右传播为例，我们得到了新的注意力得分：

$$\hat{s}_{x,y,x',y'} = \sum_{i \in \{-k,\ldots,k\}} s^*_{x+i,y,x'+i,y'}.$$

传播有效地实现为具有单位矩阵作为核的卷积。 注意力传播显著改善了测试中的绘画效果并丰富了训练中的渐变。

内存使用效率 假设 128×128 特征映射中缺少 64×64 区域，则从背景中提取的卷积滤波器数量为 12,288。 这可能会导致 GPU 的内存开销。 为了克服这个问题，我们引入了两个选项：1）提取具有跨步的背景贴片以减少滤波器的数量;以及 2）在卷积之前缩小前景输入的分辨率并且在传播之后放大注意力图。

4.2.统一修复网络

为了集成注意力模块，我们引入了两个并行编码器，如图 4 所示。基于图 2. bottom 编码器专门用于逐层（扩散）卷积的想象内容，而顶层编码器则尝试参加背景感兴趣的功能。 来自两个编码器的输出特征被聚合并馈入一个

单个解码器获得最终输出。 为了解释对文本的关注，我们以图 4 所示的方式对其进行可视化。我们使用颜色来指示每个前景像素的最感兴趣的背景补丁的相对位置。 例如，白色（颜色编码图的中心）表示像素在其自身上，左下方为粉红色，右上方为绿色。 对于不同的图像，偏移值的缩放方式不同，以便最佳地显示最有趣的范围。

对于训练，给定原始图像 x，我们在随机位置采样二进制图像掩模 m,输入图像 z 从原始图像中被破坏为 z=x⊙m。修补网络 G 将 z 和 m 的连接作为输入，并输出具有与输入相同大小的预测图像 x'=G(z,m)。 将 x0 的掩蔽区域粘贴到输入图像，我们得到修复输出 x^=z+x'⊙(1-m)。 输入和输出的图像值线性缩放为[-1，1]在所有实验中。 训练程序如算法 1 所示。

Algorithm 1 Training of our proposed framework.

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

5. 实验

我们在四个数据集上评估所提出的修复模型，包括 Places2 [43]，CelebA 面[28]，CelebA-HQ 面[22]，DTD 纹理[6]和 ImageNet [34]。

首先定性比较，我们在图 5 中显示，我们的基线模型通过比较我们的输出结果和从主要论文复制的结果，与先前的最新技术[17]产生可比较的修复结果。 请注意，我们的基线模型没有执行后处理步骤，而图像混合应用于[17]的结果。

接下来，我们使用最具挑战性的 Places2 数据集，通过比较我们的基线两阶段模型来评估我们的完整模型，并将其与先前的最新技术进行比较[17]。 对于训练，我们使用第 4.2 节中描述的分辨率为 256×256 且最大孔尺寸 128×128 的图像。 这两种方法都基于完全卷积神经网络，因此可以填充多个在不同分辨率的图像上的孔。验证集中各种复杂场景的可视化比较如图 6 所示。为了测试的一致性，这些测试图像的大小均为 512×680。报告的所有结果都是来自训练模型的直接输出，而不使用任何后处理。对于每个例子，我们还在最后一栏中对我们模型的潜在注意力图进行了可视化（颜色编码在 4.2 节中进行了解释）。

如图所示，我们具有上下文关注的完整模型可以利用周围的纹理和结构，从而生成更逼真的结果，并且比基线模型更少的伪像。注意力图的可视化表明我们的方法知道上下文图像结构，并且可以自适应地借用周围区域的信息来帮助合成和生成。

在图 7 中，我们还展示了我们在 CelebA，DTD 和 ImageNet 上训练的完整模型的一些示例结果和注意图。由于篇幅限制，我们在补充材料中包含了这些数据集的更多结果。

定量比较 与其他图像生成任务一样，图像修复缺乏良好的定量评估指标。为评估 GAN 模型而引入的初始评分[35]不是用于评估图像修复方法的良好度量，因为修复主要集中在背景填充（例如，对象去除情况），而不是其生成各种类对象的能力。

由于存在许多可能的解决方案与原始图像内容不同，因此在重建误差方面的评估度量也不是完美的。然而，我们在 Places2 上的验证集上的平均“1 误差，平均值”2 误差，峰值信噪比（PSNR）和总变差（TV）损失方面报告我们的评估，仅供参考表 1 所示。如图所示。在表中，基于学习的方法在 l1，l2 错误和 PSNR 方面表现更好，而直接复制原始图像补丁的方法具有较低的总变异损失。

Method	ℓ_1 loss	ℓ_2 loss	PSNR	TV loss
PatchMatch [3]	16.1%	3.9%	16.62	**25.0%**
Baseline model	9.4%	2.4%	18.15	25.7%
Our method	**8.6%**	**2.1%**	**18.91**	25.3%

表 1：在 Places2 上的验证集上的平均 l1 个错误，平均值 l2 错误，PSNR 和 TV loss 的结果以供参考。

我们的完整模型总共有 2.9M 参数，大约是[17]中提出的模型的一半。 模型在 TensorFlow v1.3，CUDNN v6.0，CUDA v8.0 上实现，并在具有 CPU Intel（R）Xeon（R）CPU E5-2697 v3（2.60GHz）和 GPU GTX 1080 Ti 的硬件上运行。 我们的完整型号在 GPU 上每帧 0.2 秒，在 CPU 上每帧 1.5 秒，对于平均分辨率为 512×512 的图像。

图 5：我们的基线模型与 Iizuka 等人的比较[17]。 从左到右，我们显示输入图像，从主要工作文件[17]复制的结果，以及我们的基线模型的结果。 请注意，我们的基线模型没有执行后处理步骤，而图像混合适用于[17]的结果。 放大效果最佳。

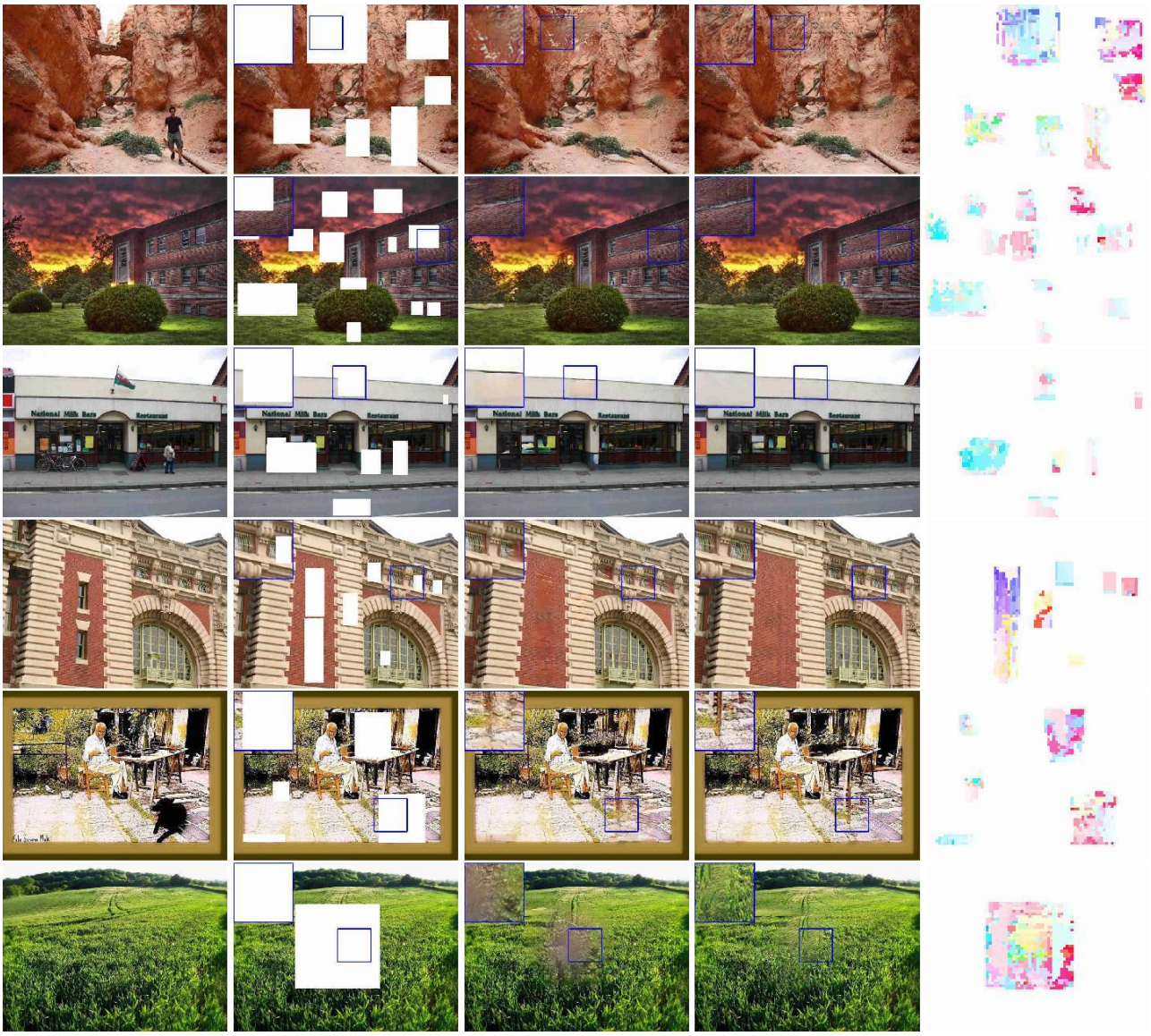

图 6：定性结果和与基线模型的比较。 我们从左到右显示我们的完整模型的原始图像，输入图像，基线模型的结果和注意力图（放大 4 倍）放大效果最佳。

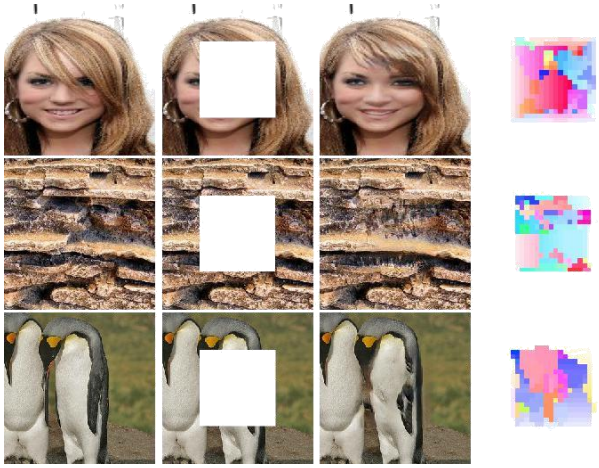

图 7：我们的 CelebA 面部模型，DTD 纹理和 ImageNet 从上到下的样本结果。 从左到右的每一行分别显示原始图像，输入图像，结果和注意力图（放大 4 倍）。

5.1. 消融研究

图 8：我们使用三个不同的注意模块显示输入图像，结果和注意力图：空间变换器网络（左），外观流（中间），我们的上下文注意（右）。

语境关注与空间变换网络和外观流程我们研究了语境关注与其他空间关注模块相比的有效性，包括用于图像修复的外观流[44]和空间变换网络[19]。对于外观流[44]，我们在相同的框架上训练，除了用卷积层替换上下文关注层以直接预测 2-D 像素偏移作为注意。如图 8 所示，对于非常不同的测试图像对，外观流返回非常相似的注意力图，这意味着网络可能陷入不良的局部最小值。为了改善外观流动的结果，我们还研究了多个注意力聚集和基于补丁的注意力的想法。这些想法都不能很好地改善绘画结果。此外，我们在图 8 的框架中将空间变换器网络[19]的结果显示为注意力。如图所示，基于 STN 的注意力不适用于修复，因为其全局仿射变换太粗糙。

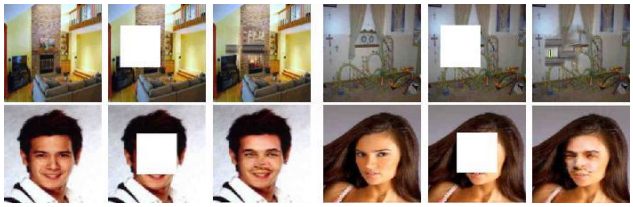

图 9：当模式崩溃时，在 Places2（顶部）和 CelebA（底部）上使用 DC-GAN 训练的模型的修复结果。

图像修复的 GAN 损失的选择我们的绘画框架从其学习曲线和更快/更稳定的收敛行为验证的 WGAN-GP 损失中获益很大。使用 DC-GAN 训练的相同模型有时会折叠到有限模式进行修复任务，如图 9 所示。我们还尝试了 LSGAN [29]，结果更糟。

基本重建损失我们还进行了测试，如果我们能够摒弃 1 重建损失并纯粹依靠对抗性损失（即改进的 WGAN）来产生良好的结果。为了得出结论，我们在精炼网络中训练我们的绘画模型而没有“1 重建损失”。我们的结论是，逐像素重构的损失虽然趋于使结果模糊，但却是图像修复的重要组成部分。重建损失有助于捕获内容结构，并作为训练 GAN 的强大正则化术语。

感知损失，风格损失和总变差损失我们还没有发现感知损失（VGG 特征的重建损失），风格损失（根据 VGG 特征计算的 Gram ma-trix 的 Frobenius 范数）[21]和总变差（TV）损失在我们的框架中为图像修改带来了显着的改进，因此没有使用。

6. 结论

我们在绘图框架中提出了粗到细的生成图像，并引入了我们的基线模型以及具有新颖的上下文关注模块的完整模型。 我们通过学习用于明确匹配和参与相关背景补丁的特征表示，显示了上下文关注模块显着改善了图像修复结果。作为未来的工作，我们计划使用类似于 GAN 渐进式增长的想法将方法扩展到高分辨率的修复应用[22]。 所提出的修复框架和上下文关注模块还可以应用于条件图像生成，图像编辑和计算摄影任务，包括基于图像的渲染，图像超分辨率，引导编辑等等。

参考文献

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. IEEE transactions on image processing, 10(8):1200–1211, 2001.

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH 2009), 2009.

[4] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 417– 424. ACM Press/Addison-Wesley Publishing Co., 2000.

[5] T. Q. Chen and M. Schmidt. Fast patch-based style transfer of arbitrary style. arXiv preprint arXiv:1612.04337, 2016.

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3606–3613, 2014.

[7] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289, 2015.

[8] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. arXiv preprint arXiv:1703.06211, 2017.

[9] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen. Image melding: Combining inconsistent images using patch-based synthesis. ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH 2012), 2012.

[10] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In Proceedings of the 28th an-nual conference on Computer graphics and interactive tech-niques, pages 341–346. ACM, 2001.

[11] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 2, pages 1033–1038. IEEE, 1999.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028, 2017.

[14] J. Hays and A. A. Efros. Scene completion using millions of photographs. In ACM Transactions on Graphics (TOG). ACM, 2007.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learn-ing for image recognition. In Proceedings of the IEEE con-ference on computer vision and pattern recognition, pages 770–778, 2016.

[16] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf. Image com-pletion using planar structure guidance. ACM Transactions on Graphics (TOG), 33(4):129, 2014.

[17] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics (TOG), 36(4):107, 2017.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pages 448– 456, 2015.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems, pages 2017–2025, 2015.

[20] Y. Jeon and J. Kim. Active convolution: Learning the shape of convolution for image classification. arXiv preprint arXiv:1703.09076, 2017.

[21] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, pages 694–711. Springer, 2016.

[22] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[23] R. Kohler,¨ C. Schuler, B. Scholkopf,¨ and S. Harmeling. Mask-specific inpainting with deep neural networks. In Ger-man Conference on Pattern Recognition, pages 523–534. Springer, 2014.

[24] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In Artificial Intelligence and Statistics, pages 562–570, 2015.

[25] A. Levin, A. Zomet, S. Peleg, and Y. Weiss. Seamless image stitching in the gradient domain. Computer Vision-ECCV 2004, pages 377–389, 2004.

[26] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2479–2486, 2016.

[27] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. arXiv preprint arXiv:1704.05838, 2017.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of International Con-ference on Computer Vision (ICCV), December 2015.

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smol-ley. Least squares generative adversarial networks. arXiv preprint ArXiv:1611.04076, 2016.

[30] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Perez´. Video inpainting of complex scenes. SIAM Journal on Imaging Sciences, 7(4):1993–2019, 2014.

[31] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. arXiv preprint arXiv:1703.02921, 2017.

[32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2536–2544, 2016.

[33] A. Radford, L. Metz, and S. Chintala. Unsupervised repre-sentation learning with deep convolutional generative adver-sarial networks. arXiv preprint arXiv:1511.06434, 2015.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015.

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Rad-ford, and X. Chen. Improved techniques for training gans. In Advances in Neural Information Processing Systems, pages 2234–2242, 2016.

[36] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summa-rizing visual data using bidirectional similarity. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

[37] X. Snelgrove. High-resolution multi-scale neural texture synthesis. In SIGGRAPH ASIA 2017 Technical Briefs. ACM, 2017.

[38] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. MIT press Cambridge, 1998.

[39] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In Advances in Neural Information Processing Systems, pages 1790–1798, 2014.

[40] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. arXiv preprint arXiv:1611.09969, 2016.

[41] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. arXiv preprint arXiv:1607.07539, 2016.

[42] F. Yu and V. Koltun. Multi-scale context aggregation by di-lated convolutions. arXiv preprint arXiv:1511.07122, 2015.

[43] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelli-gence, 2017.

[44] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In European Conference on Computer Vision, pages 286–301. Springer, 2016.

A. More Results on CelebA, CelebA-HQ, DTD and ImageNet

CelebA-HQ [22] We show results from our full model trained on CelebA-HQ dataset in Figure 10. Note that the original image resolution of CelebA-HQ dataset is 1024 1024. We resize image to 256 256 for both training and evaluation.

CelebA [28] We show more results from our full model trained on CelebA dataset in Figure 11. Note that the orig-inal image resolution of CelebA dataset is 218 178. We resize image to 315 256 and do a random crop of size 256 256 to make face landmarks roughly unaligned for both training and evaluation.

ImageNet [34] We show more results from our full model trained on ImageNet dataset in Figure 12.

DTD textures [6] We show more results from our full model trained on DTD dataset in Figure 13.

B. Comparisons with More Methods

We show more results for qualitative comparisons with more methods including Photoshop Content-Aware Fill [3], Image Melding [9] and StructCompletion [16] in Figure 14 and 15. For all these methods, we use default hyper-parameter settings.

C. More Visualization with Case Study

In addition to attention map visualization, we visualize which parts in the input image are being attended for pix-els in holes. To do so, we highlight the regions that have the maximum attention score and overlay them to input im-age. As shown in Figure 16, the visualization results given holes in different locations demonstrate the effectiveness of our proposed contextual attention to borrow information at distant spatial locations.

D. Network Architectures

In addition to Section 3, we report more details of our network architectures. For simplicity, we denote them with K (kernel size), D (dilation), S (stride size) and C (channel number).

Inpainting network Inpainting network has two encoder-decoder architecture stacked together, with each encoder-decoder of network architecture:

K5S1C32 - K3S2C64 - K3S1C64 - K3S2C128 - K3S1C128 - K3S1C128 - K3D2S1C128 - K3D4S1C128 - K3D8S1C128 - K3D16S1C128 - K3S1C128 - K3S1C128 - resize (2) - K3S1C64 - K3S1C64 - resize (2) - K3S1C32 - K3S1C16 - K3S1C3 - clip.

Local WGAN-GP critic We use Leaky ReLU with = 0:2 as activation function for WGAN-GP critics.

K5S2C64 - K5S2C128 - K5S2C256 - K5S2C512 - fully-connected to 1.

Global WGAN-GP critic K5S2C64 - K5S2C128 - K5S2C256 - K5S2C256 - fully-connected to 1.

Contextual attention branch K5S1C32 - K3S2C64 - K3S1C64 - K3S2C128 - K3S1C128 - K3S1C128 - contex-tual attention layer - K3S1C128 - K3S1C128 - concat.

图 10：我们完整模型的更多修复结果，以及对 CelebA-HQ 面孔的背景关注。 每个三元组从左到右显示原始图像，输入蒙版图像和结果图像。 所有输入图像都从验证集中屏蔽（训练和验证拆分在已发布的代码中提供）。 所有结果都是来自同一训练模型的直接输出，无需后处理。

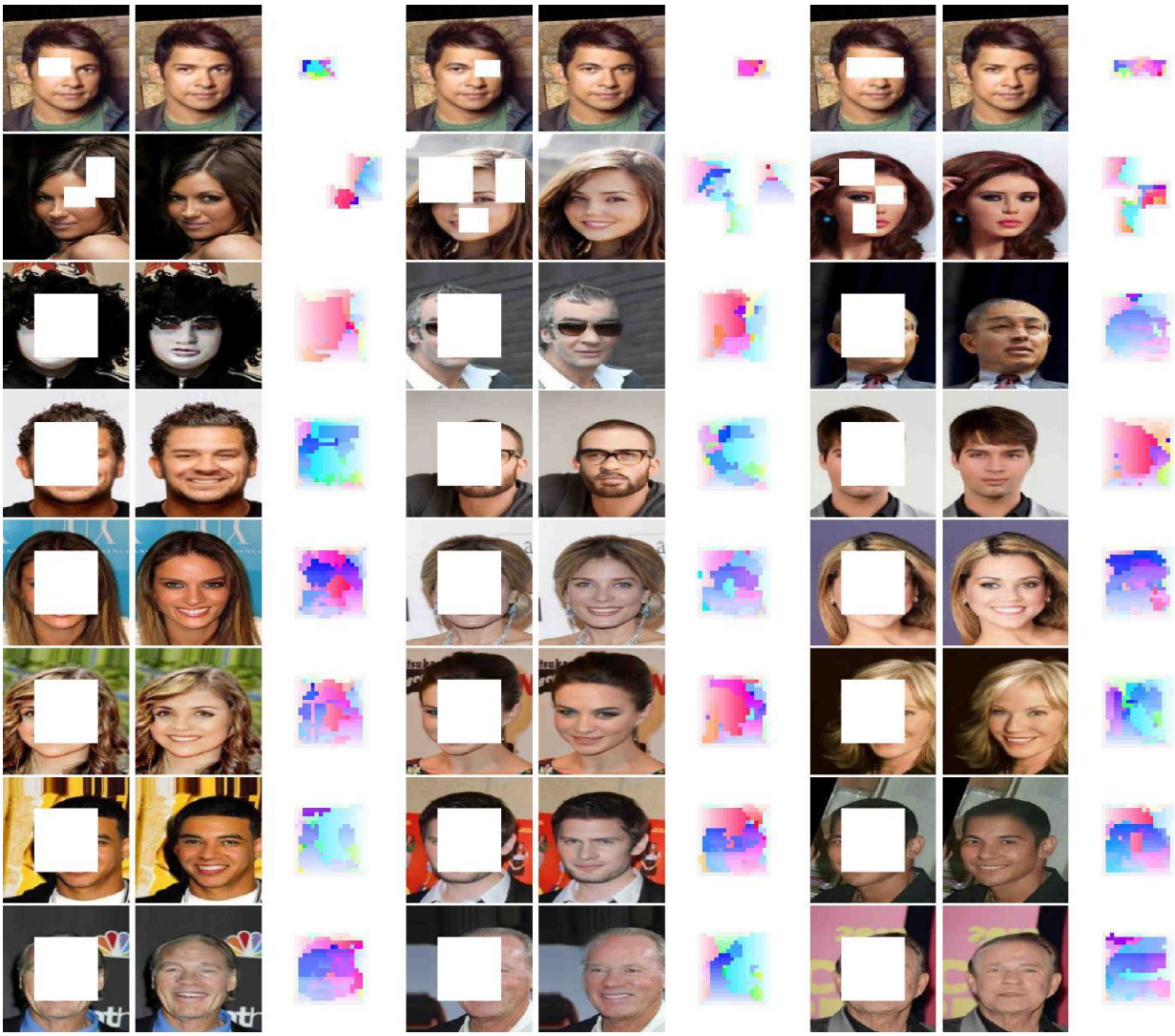

图 11：我们完整模型的更多修复结果，以及对 CelebA 面孔的背景关注。 每个三元组从左到右显示输入图像，结果和注意力图（放大 4 倍）。 所有输入图像都从验证集中屏蔽（面部标识在训练集和验证集之间不重叠）。 所有结果都是来自同一训练模型的直接输出，无需后处理。

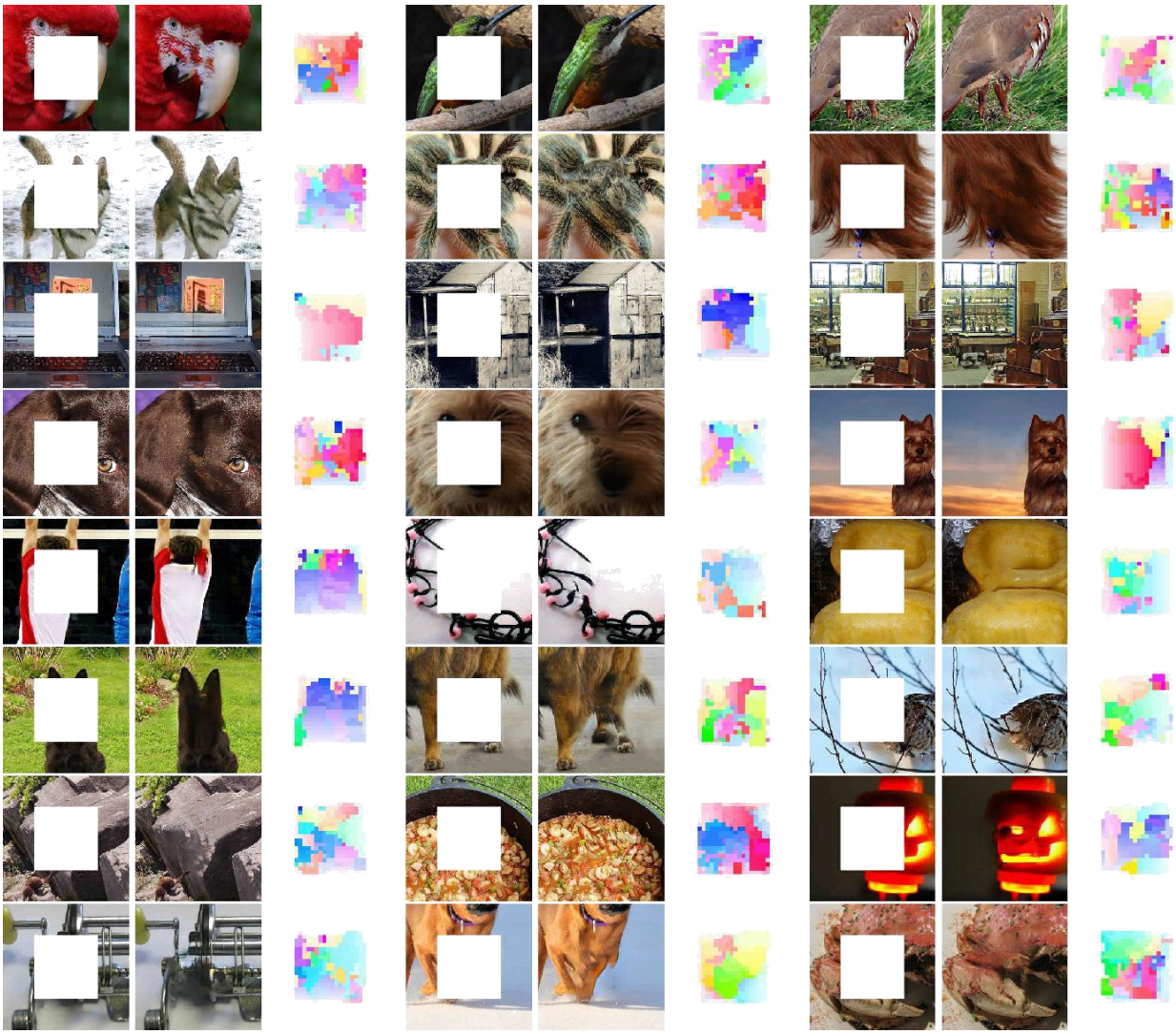

图 12：我们的完整模型的更多修复结果，并在 ImageNet 上进行了上下文关注。 每个三元组从左到右显示输入图像，结果和注意力图（放大 4 倍）。 所有输入图像都从验证集中屏蔽。 所有结果都是来自同一训练模型的直接输出，无需后处理。

图 13：我们的完整模型的更多修复结果以及对 DTD 纹理的上下文关注。 每个三元组从左到右显示输入图像，结果和注意力图（放大 4 倍）。 所有输入图像都从验证集中屏蔽。 所有结果都是来自同一训练模型的直接输出，无需后处理。

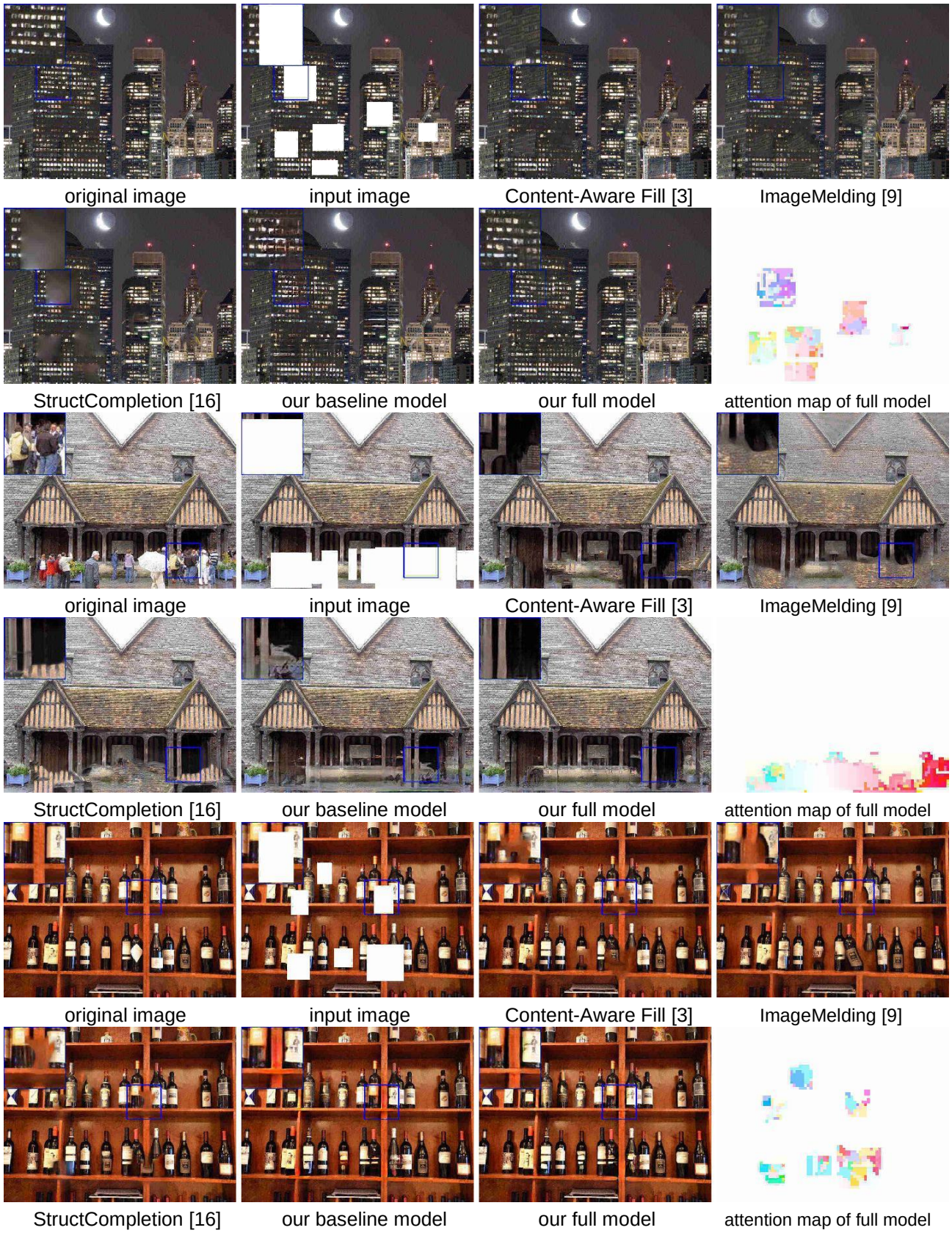

图 14：更多定性结果和比较。 所有输入图像都从验证集中屏蔽。 我们所有的结果都是来自同一训练模型的直接输出而没有后处理。 放大效果最佳。

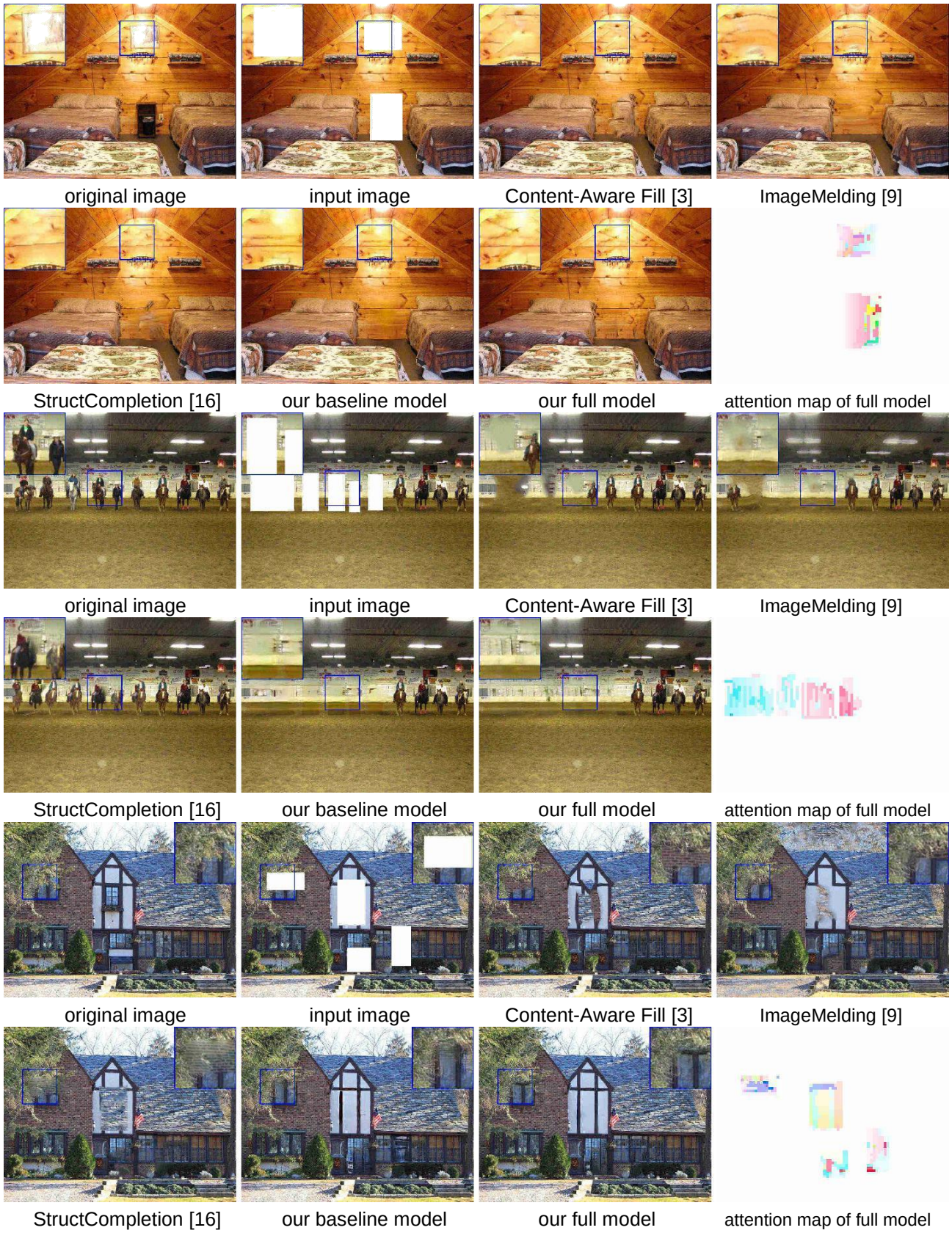

图 15：更多定性结果和比较。 所有输入图像都从验证集中屏蔽。 我们所有的结果都是来自同一训练模型的直接输出而没有后处理。 放大效果最佳。

图 16：可视化（突出显示的区域），输入图像中的部分参与其中。 每个三元组从左到右显示输入图像，结果和注意力可视化。