Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting

Yanhong Zeng1,2*, Jianlong Fu3, Hongyang Chao1,2, Baining Guo3

1School of Data and Computer Science, Sun Yat-sen University, Guangzhou, P.R. China

2The Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, Guangzhou, P.R. China

3Microsoft Research, Beijing, P.R. China

zengyh7@mail2.sysu.edu.cn, {jianf,bainguo}@microsoft.com, isschhy@mail.sysu.edu.cn

June 27, 2019

Outline

- Brief Introduction
  - Motivation
  - Existing Methods and Limitations
- Proposed Network
  - PEN-Net
  - Pyramid-context encoder
  - Attention Transfer Network
  - Multi-scale decoder
- Contribution

Brief Introduction

Motivation

过去的工作只在 feature-level 或者 image-level 上进行修复，各有缺点，本工作是第一个将两者同时用到一起的工作。

Existing Methods and Limitations

1. The first group inspired by texture synthesis techniques attempts to fill regions at image-level. Specifically, such approaches usually sample and paste full image resolution patches from source images into missing regions, which allows synthesizing results with details.
2. As the lack of high-level understanding of an image, such approaches often fail in generating semantically-reasonable results.
3. The second group of approaches proposes to encode the semantic context of an image into a latent feature space by deep neural networks and then generate semantic-coherent patches by generative models.
4. It remains challenging to generate visually-realistic results from a compact latent feature, as full image resolution details can be usually smoothed by stacked convolutions and poolings.

Proposed Network

PEN-Net

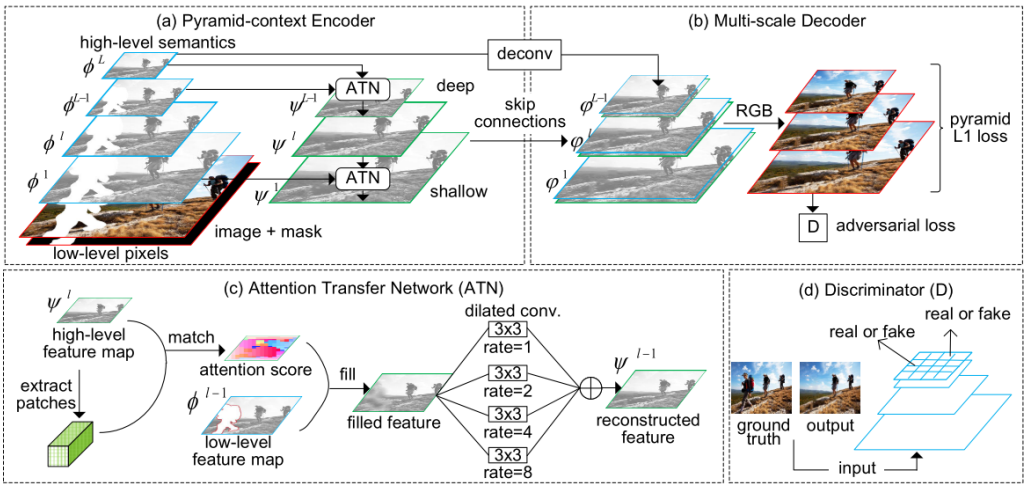

Figure 2: The **P**yramid-context **En**coder **Net**work (PEN-Net) is proposed to boost the capability of U-Net in image inpainting with three tailored components, *i.e.*, a pyramid-context encoder (a), a multi-scale decoder (b), and an adversarial training loss (d). First, once the compact latent feature has been encoded, the pyramid-context encoder further improves the encoding effectiveness by filling regions from high-level feature maps to low-level feature maps (with richer details) through the proposed **A**ttention **T**ransfer **N**etwork (ATN) (c). Second, the multi-scale decoder takes as input the reconstructed features from ATNs through skip connections and the latent features for decoding. Finally, the decoder decodes the features back into an image. The whole network is optimized by minimizing pyramid L1 losses and an adversarial loss. [Best viewed in color.]

Pyramid-context encoder

$$\begin{aligned} \psi^{L-1} &= f(\phi^{L-1}, \phi^{L}), \\ \psi^{L-2} &= f(\phi^{L-2}, \psi^{L-1}), \\ &\cdots, \\ \psi^{1} &= f(\phi^{1}, \psi^{2}) = f(\phi^{1}, f(\phi^{2}, \ldots f(\phi^{L-1}, \phi^{L}))), \end{aligned} \tag{1}$$

Attention Transfer Network

$$s_{i,j}^{l} = \langle \frac{p_i^l}{\|p_i^l\|_2}, \frac{p_j^l}{\|p_j^l\|_2} \rangle, \tag{2}$$

where p_i^l is the i-th patch extracted from ψ^l outside mask, p_j^l is the j-th patch extracted from ψ^l inside the mask. Then softmax is applied on the similarities to obtain the attention score for each patch:

$$\alpha_{j,i}^{l} = \frac{\exp(s_{i,j}^{l})}{\sum_{i=1}^{N} \exp(s_{i,j}^{l})}. \tag{3}$$

$$p_j^{l-1} = \sum_{i=1}^{N} \alpha_{j,i}^{l} \; p_i^{l-1}, \tag{4}$$

Multi-scale decoder

$$\begin{aligned} \varphi^{L-1} &= g(\psi^{L-1} \oplus g(\phi^{L})), \\ \varphi^{L-2} &= g(\psi^{L-2} \oplus \varphi^{L-1}), \\ &\cdots, \\ \varphi^{1} &= g(\psi^{1} \oplus \varphi^{2}), \end{aligned} \tag{5}$$

where g denotes transposed convolution operation, $\oplus$ denotes feature concatenation, and ψ^l is the reconstructed feature from an ATN in the l-th layer of the encoder.

Pyramid L1 losses

$$L_{pd} = \sum_{l=1}^{L-1} \left\| x^l - h(\varphi^l) \right\|_1, \tag{6}$$

Contribution

1. We propose a novel network, ATN, to learn region affinity from high-level feature maps (e.g., the compact latent features in the encoder).

2. Our model can fill holes multiple times (depends on the depth of the encoder) by repeating using ATNs from deep to shallow, which can re-store an image with more fine-grained details.

Generative Image Inpainting with Contextual Attention Jiahui

Jiahui Yu1 Zhe Lin2 Jimei Yang2 Xiaohui Shen2 Xin Lu2 Thomas S. Huang1

1University of Illinois at Urbana-Champaign

2Adobe Research Figure

June 27,2019

Outline

- Brief Introduction
  - Motivation
  - Existing Methods and Limitations
- Proposed Network
  - Proposed Network
  - Contextual Attention
  - Unified Inpainting Network
- Contribution

Brief Introduction

Motivation

1. Recent deep learning based approaches have shown promising results for the challenging task of inpainting large missing regions in an image.
2. These methods can generate visually plausible image structures and textures, but often create distorted structures or blurry textures inconsistent with surrounding areas.
3. This is mainly due to ineffectiveness of convolutional neural networks in explicitly borrowing or copying information from distant spatial locations.

Existing Methods and Limitations

1. Early works attempted to solve the problem using ideas similar to texture synthesis , i.e. by matching and copying background patches into holes starting from low-resolution to high-resolution or propagating from hole boundaries.
2. However, as they assume missing patches can be found somewhere in background regions, they cannot hallucinate novel image contents for challenging cases where inpainting regions involve complex, non- repetitive structures (e.g. faces, objects).
3. CNN+GAN
4. Unfortunately, these CNN-based methods often create boundary artifacts, distorted structures and blurry textures inconsistent with surrounding areas.
5. We found that this is likely due to ineffectiveness of convolutional neural networks in modeling long-term correlations between distant contextual information and the hole regions.

Proposed Network

Proposed Network

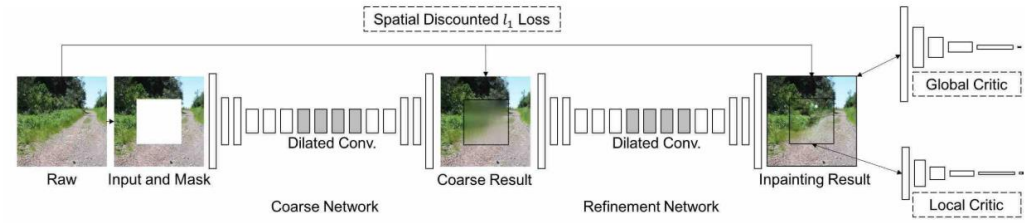

Figure 2: Overview of our improved generative inpainting framework. The coarse network is trained with reconstruction loss explicitly, while the refinement network is trained with reconstruction loss, global and local WGAN-GP adversarial loss.

Spatially discounted reconstruction loss

对于图像修复问题，一个缺失区域可能会有多种可行的修复结果。一个可行的修复结果可能会和原始图像差距很大，如果使用原始图像作为唯一的参照标准（ground truth），计算重构损失（reconstruction loss）时就会误导卷积网络的训练过程。

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（边界区域的取值范围要比中心区域小）。这与强化学习中的问题类似。当长期奖励（long-term rewards）有着很大的取值范围时，人们在采样轨迹（sampled trajectories）上使用随着时间衰减的奖励（随着时间的流逝，网络得到的奖励会越来越小）。

具体的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。

Global and Local WGAN

Contextual Attention

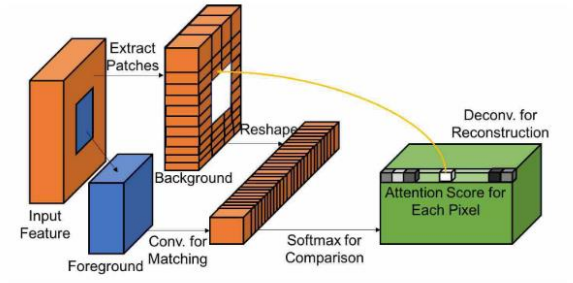

Figure 3: Illustration of the contextual attention layer. Firstly we use convolution to compute matching score of foreground patches with background patches (as convolutional filters). Then we apply softmax to compare and get attention score for each pixel. Finally we reconstruct foreground patches with background patches by performing deconvolution on attention score. The contextual attention layer is differentiable and fully-convolutional.

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

dimension to get attention score for each pixel $s^*_{x,y,x',y'} = softmax_{x',y'}(\lambda s_{x,y,x',y'})$, where λ is a constant value. This

Unified Inpainting Network

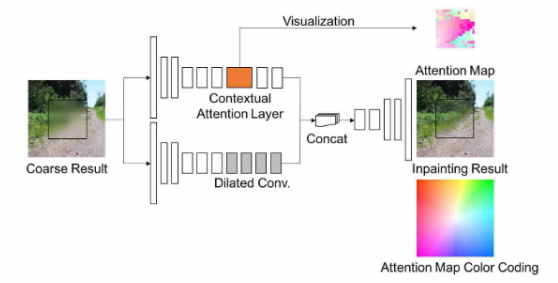

Figure 4: Based on coarse result from the first encoder-decoder network, two parallel encoders are introduced and then merged to single decoder to get inpainting result. For visualization of attention map, color indicates relative location of the most interested background patch for each pixel in foreground. For examples, white (center of color coding map) means the pixel attends on itself, pink on bottom-left, green means on top-right.

Contribution

1. We propose a novel contextual attention layer to explicitly attend on related feature patches at distant spatial locations.
2. We introduce several techniques including inpainting network enhancements, global and local WGANs and spatially discounted reconstruction loss to improve the training stability and speed based on the current the state-of-the-art generative image inpainting network.
3. Our unified feed-forward generative network achieves high-quality inpainting results on a variety of challenging datasets including CelebA faces , CelebA-HQ faces , DTD textures, ImageNet and Places2 .

Foreground-aware Image Inpainting

Wei Xiong1* Jiahui Yu2 Zhe Lin3 Jimei Yang3 Xin Lu3 Connelly Barnes3

Jiebo Luo1

1University of Rochester2University of Illinois at Urbana-Champaign 3Adobe Research

1{wxiong5,jluo}@cs.rochester.edu 2jyu79@illinois.edu

3{zlin, jimyang, xinl, cobarnes}@adobe.com

June 27, 2019

Outline

- Brief Introduction
  - Motivation
  - Existing Methods and Limitations
- Proposed Network
  - Proposed Network
  - Loss
- Contribution

Brief Introduction

Motivation

在实际中一些 mask 会与前景的实际物体重叠或解除，这样，围绕背景图像的像素进行修复的效果就不好，因此提出了一种前景图像感知修复方法，先对前景物体进行轮廓的预测再根据物体的轮廓进行图像修复。

Existing Methods and Limitations

1. Patches-based
2. 但是，这些方法仍然依赖于现有的补丁和低级特征，因此无法处理漏洞与前景对象重叠或接近的挑战性情况。
3. Learning-based
4. 但是，默认情况下，所有这些方法都假定生成网络可以学习隐式地预测或理解图像中的结构，而无需在学习过程中对结构或前景/背景层进行显式建模。

Proposed Network

Proposed Network

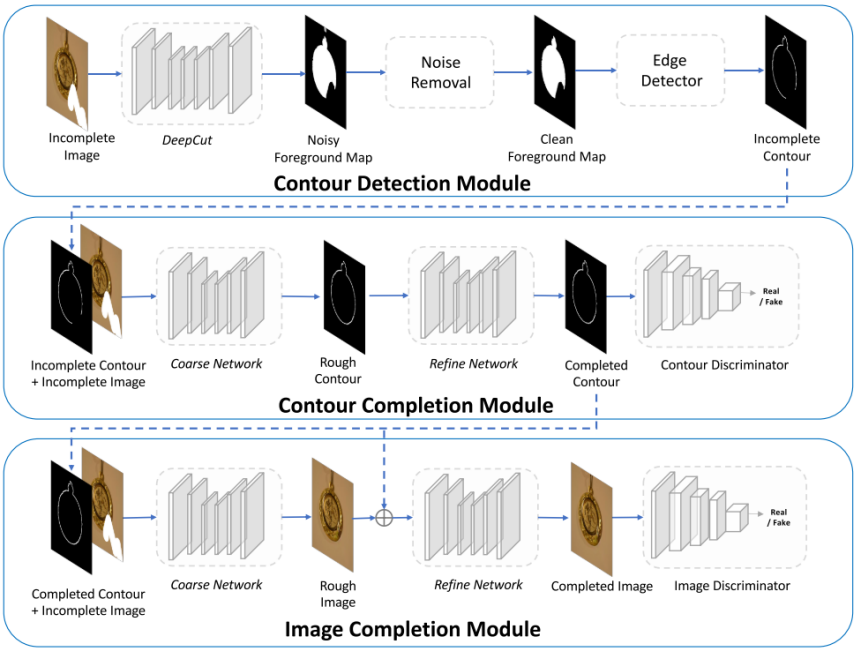

Figure 2. The overall architecture of our inpainting model.

Content Loss:

Contour Completion Module:

因为 mask 中的轮廓稀疏，导致数据不平衡问题。

我们建议利用轮廓 mask 的固有性质，即掩模中的每个像素可以解释为像素是原始图像中的边界像素的概率。

Therefore we can take the contour map as samples of a distribution, and calculate the distance with the ground-truth contour by calculating their binary cross-entropy between each pixel

$$
\begin{aligned}
&\mathcal{L}_{con}^{C}\left(C_c^{cos}, C_{gt}\right) \\
&= \frac{\lambda}{N}\sum_{p}\left(H[p](C_c^{cos}[p]-C_{gt}[p])^2\mathcal{L}_e(C_c^{cos}[p], C_{gt}[p])\right) \\
&+ \frac{1}{N}\sum_{p}\left((1-H[p])(C_c^{cos}[p]-C_{gt}[p])^2\mathcal{L}_e(C_c^{cos}[p], C_{gt}[p])\right) .
\end{aligned}
\tag{1}
$$

$$
\mathcal{L}_{con}^{C} = \mathcal{L}_{con}^{C}(C_c^{cos}, C_{gt}) + \mathcal{L}_{con}^{C}(C_c^{ref}, C_{gt}) . \tag{2}
$$

Image Completion Module:

$$
\mathcal{L}_{con}^{I} = \frac{1}{N}\sum_{p}\left(|I_c^{cos}[p]-I_{gt}[p]| + |I_c^{ref}[p]-I_{gt}[p]|\right) .
\tag{5}
$$

Contribution

1. We propose to explicitly disentangle structure inference and image completion to address challenging scenarios in image inpainting where holes overlap with or touch foreground objects.
2. To infer the structure of images, we propose a contour completion module trained explicitly to guide image completion.
3. Our experiments demonstrate that the system produces higher-quality inpainting results compared to existing methods.

Meta-SR: A Magnification-Arbitrary Network for Super-Resolution

Xuecai Hu*1,2 , Haoyuan Mu* 4, Xiangyu Zhang3, Zilei Wang1, Tieniu Tan1,2, Jian Sun3

1 University of Science and Technology of China

2 Center for Research on Intelligent Perception and Computing, NLPR, CASIA

3 Megvii Inc (Face++) 4 Tsinghua University

huxc@mail.ustc.edu.cn, muhy17@mails.tsinghua.edu.cn

{zhangxiangyu, sunjian}@megvii.com, zlwang@ustc.edu.cn, tnt@nlpr.ia.ac.cn

June 27, 2019

Outline

- Brief Introduction

  Motivation

- Proposed Method

  Meta-SR

Brief Introduction

Motivation

用一个模型实现任意 scale factor 下的 SR image.

Proposed Method

Assumption: 认为 SR image 的一个像素点（i，j）是由 LR feature 中的相应的像素点（i',j'）和对应权重的卷积 W(i,j)得到的。

$$\mathbf{I}^{SR}(i,j) = \Phi(\mathbf{F}^{LR}(i',j'), \mathbf{W}(i,j)) \qquad (1)$$

三个主要步骤：先根据 resnet 提取 LR image 的 feature，再根据 SR image 的像素点(i,j)和 scale factor r 映射的到 LR feature 的像素点(i',j')，再根据两个像素点和 r 预测不同的 W(i,j)，再根据 W 和 LR feature 得到 SR image.(HW 是 LR 与 SR 中对应 pixel 之间的关系向量组成的矩阵)

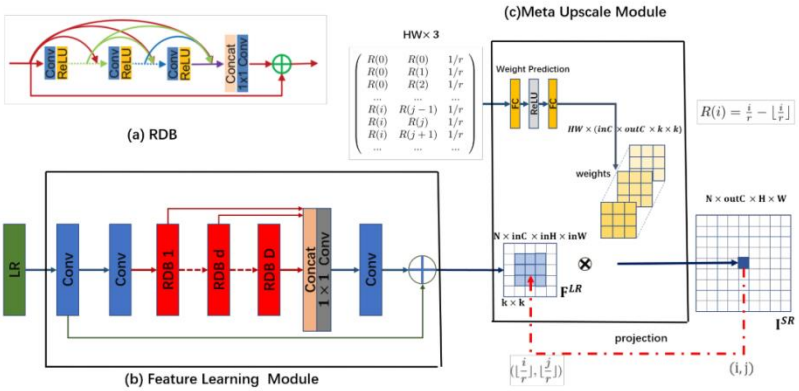

Figure 1. An instance of our Meta-SR based on RDN [36]. We also call the network Meta-RDN. (a) The Residual Dense Block proposed by RDN [36]. (b) The Feature Learning Module which generates the shared feature maps for arbitrary scale factor. (c) For each pixel on the SR image, we project it onto the LR image. The proposed Meta-Upscale Module takes a sequence of coordinate-related and scale-related vectors as input to predict the weights for convolution filters. By doing the convolution operation, our Meta-Upscale finally generate the HR image.

Location Projection

$$(i',j') = T(i,j) = \left(\left\lfloor \frac{i}{r} \right\rfloor, \left\lfloor \frac{j}{r} \right\rfloor\right) \qquad (2)$$

Weight Prediction

$$\mathbf{W}(i,j) = \varphi(\mathbf{v}_{ij};\theta) \tag{3}$$

$$\mathbf{v}_{ij} = \left(\frac{i}{r} - \left\lfloor \frac{i}{r} \right\rfloor, \frac{j}{r} - \left\lfloor \frac{j}{r} \right\rfloor, \frac{1}{r}\right) \tag{5}$$