看的“深”，看的“清”

——深度学习在图像超清化上的应用

画家在知道人脸的知识之后，只要说几个关键特征，就可以画出完整的人脸。同样的，当拥有了低清图像上的图像信息后，也可以推断出其对应的高清细节，但这就需要算法能够“理解”图像内容。于是，传统的规则算法不堪重负，新兴的深度学习照耀着图像超清化的星空。

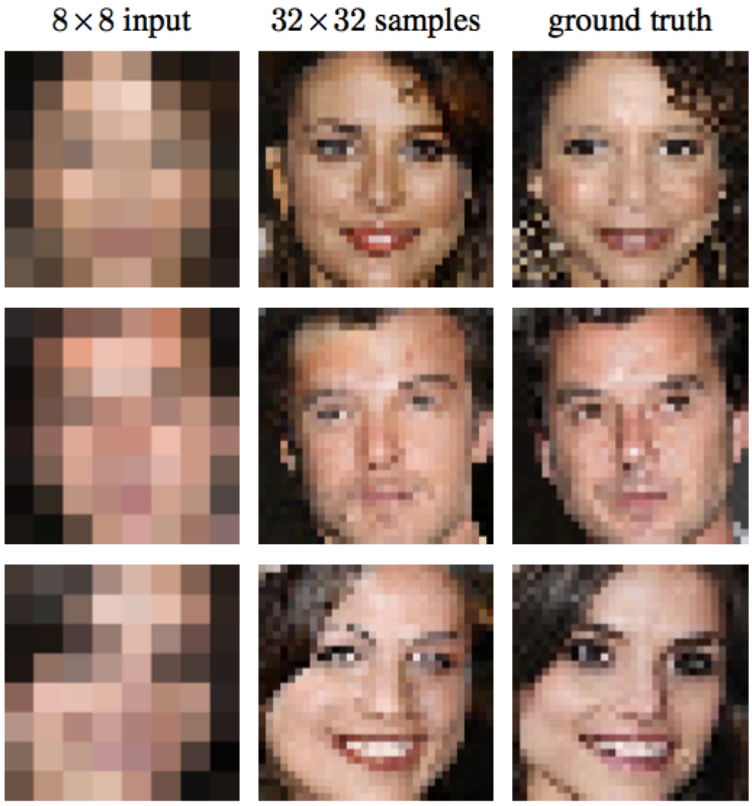

图 1. 最新的 Pixel 递归网络在图像超清化上的应用，左图为低清图像，右图为其对应的高清图像，中间为算法生成结果。这是 4 倍超清问题，即将边长扩大为原来的 4 倍。

得益于硬件的迅猛发展，短短数年间，手机已更新了数代，老手机拍下的照片在大分辨率的屏幕上变得模糊起来。同样的，分辨率的提升使得网络带宽的压力骤增。如此，图像超清化算法就有了用武之地。

存放多年的老照片，使用超清算法变得细节栩栩如生。网络传输，将图像压缩然后再用超清化算法复原便可大大减少传输量。

而传统的几何手段如三次插值，传统的匹配手段如碎片匹配，在应对这样的需求上皆有心无力。

深度学习的出现使的算法对图像的语义级别的操作成为可能。本文即是介绍深度学习技术在图像超清问题上的最新的研究进展。

深度学习最早的兴起是源于图像，其主要处理图像的技术是卷积神经网络，关于卷积神经网络的起源业界公认是 Alex 在 2012 年的 ImageNet 比赛中的煌煌表现。虽方五年之久，却已是老生常谈。因而卷积神经网络的基础细节本文不再赘述。在下文中，使用 CNN（Convolutional Neural Network）来指代卷积神经网络。

CNN 出现以来，产生了很多研究热点，但最令人印象深刻的五个热点是：

- 深广探索，VGG 网络的出现标志着 CNN 在深度和广度上有了初步的突破。
- 结构探索，Inception 及其变种的出现进一步增加了模型的深度。而 ResNet 的出现则使得深度学习的深度变得“名副其实”起来，可以达到上百层甚至上千层。
- 内容损失，图像风格转换是 CNN 在应用层面的一个小高峰，涌现了一批以 Prisma 为首的小型创业公司。但图像风格转换在技术上的真正贡献却是通过一个预训练好的模型上的特征图，可以在语义层面生成图像。
- 对抗神经网络（GAN），虽然这是针对机器学习领域的架构创新，但其最先应用却是在 CNN 上，通过对抗训练，使得生成模型能够借用监督学习的东风进行提升，将生成模型的质量提升了一个级别。
- Pixel CNN，将依赖关系引入到像素之间，是 CNN 模型结构方法的一次比较大的创新，用于生成图像，效果最佳，但失之于效率。

这五个热点，在图像超清这个问题上都有所体现。下面会一一为大家道来。

CNN 的第一次出手

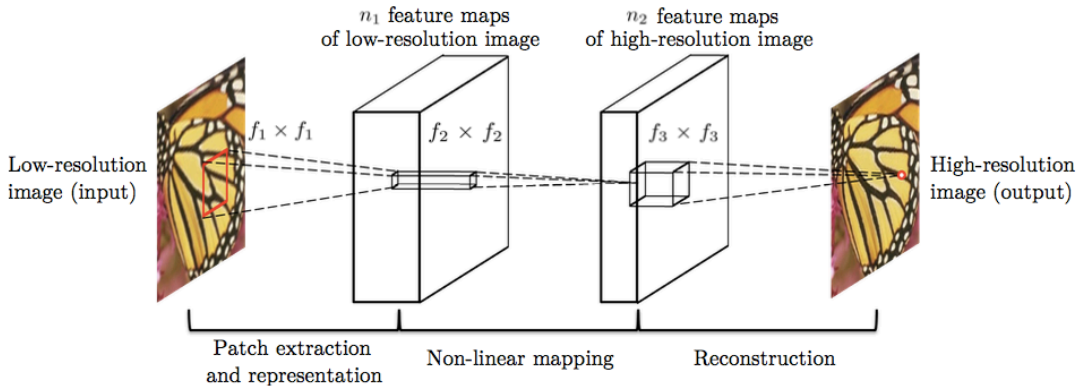

图 2. 首个应用于图像超清问题的 CNN 网络结构，输入为低清图像，输出为高清图像。该结构分为三个步骤，低清图像的特征抽取，低清特征到高清特征的映射，高清图像的重建。

图像超清问题的特点在于，低清图像和高清图像中很大部分的信息是共享的，基于这个观点，CNN 以前的解决方案是使用一些特定方法如 PCA、Sparse Coding 将低分辨率和高分辨率图像都变为特征表示，然后将特征表示做映射。

基于传统的方法结构，CNN 也将模型划分为三个部分，即特征抽取，非线性映射和特征重建。但由于 CNN 的特性，三个部分的操作都可以使用卷积完成。因而，虽然针对模型结构的解释与传统方法类似，但 CNN 却是可以同时联合训练的统一体，在数学上拥有更加简单的表达。

不仅在模型解释上可以看到传统方法的影子，在具体的操作上也可以看到。在这上述模型中，需要对数据进行预处理，抽取出很多 patch，这些 patch 可能互有重叠，但 Patch 取合集便是整张图像。而上述的 CNN 结构是应用到这些 Patch 而不是整张图像上，得到所有图像的 patch 后，将这些 patch 组合起来得到最后的高清图像，重叠部分则取均值。

更深更快更准的 CNN

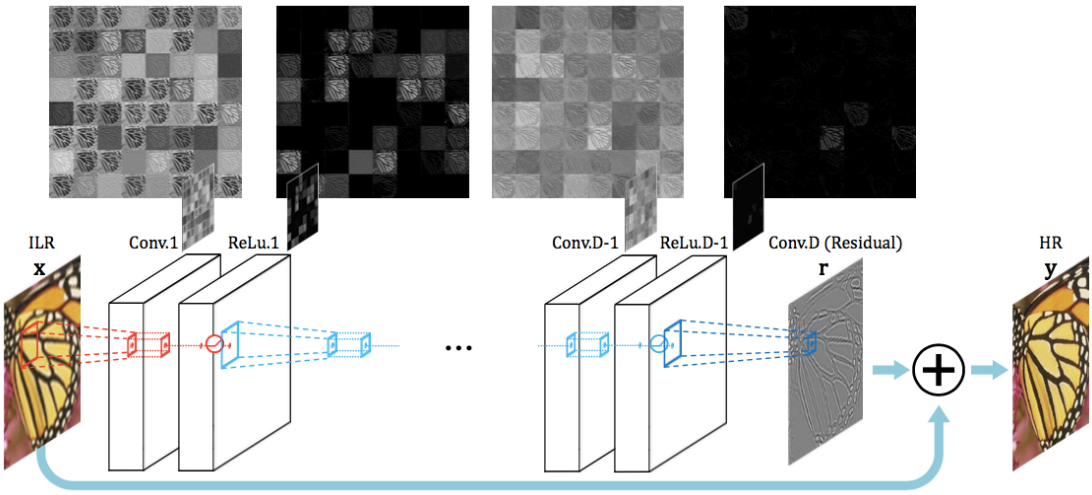

图 3. 基于残差的深度 CNN 结构。该结构使用残差连接将低清图像与 CNN 的输出相加得到高清图像，即仅用 CNN 结构学习低清图像中缺乏的高清细节部分

图 2 中的方法虽然效果远高于传统方法，但是却有若干问题：

- 训练层数少，没有足够的视野域
- 训练太慢，导致没有在深层网络上得到好的效果
- 不能支持多种倍数的高清化。

针对上述问题，图 3 算法提出了采用更深的网络模型，但为了克服之前的算法的问题，采用了三种技术进行解决。

第一种技术是残差学习，CNN 是端到端的学习，如果像上一个方法那样直接学习，那么 CNN 需要保存图像的所有信息，需要在恢复高清细节的同时记住所有的低分辨率图像的信息，如此，网络中的每一层都需要存储所有的图像信息，这就导致了信息过载，使得网络对梯度十分敏感，容易造成梯度消失或梯度爆炸等

现象。而图像超清问题中，CNN 的输入图像和输出图像中的信息很大一部分是共享的。残差学习正是只学习高清细节信息的算法。如上图所示，CNN 的输出加上原始低分辨率图像得到高分辨率图像，即 CNN 学习到的是高分辨率图像和低分辨率图像的差。如此，CNN 承载的信息量小，更容易收敛的同事还可以达到比非残差网络更好的效果。

高清图像之所以能够和低清图像做加减法，是因为，在数据预处理时，将低清图像使用插值法缩放到与高清图像同等大小。于是虽然图像被称之为低清，但其实图像大小高清图像是一致的。

第二种技术是高学习率，在 CNN 中设置高学习率通常会导致梯度爆炸，因而在使用高学习率的同时还使用了自适应梯度截断。截断区间为[- θ / γ， θ / γ]，其中 γ 为当前学习率， θ 是常数。

第三种技术是数据混合，最理想化的算法是为每一种倍数分别训练一个模型，但这样极为消耗资源。因而，同之前的算法不同，本技术将不同倍数的数据集混合在一起训练得到一个模型，从而支持多种倍数的高清化。

感知损失

在之前，使用 CNN 来解决高清问题是将 CNN 生成模型产生的图像和实际图像以像素为单位计算损失函数（一般为欧式距离），以此为损失函数得到的模型捕捉到的只是像素级别的规律，其泛化能力相对而言比较弱。

而感知损失，则是指将 CNN 生成模型和实际图像都输入到某个训练好的网络中，得到这两张图像在该训练好的网络上某几层的激活值。在激活值上计算损失函数。由于 CNN 能够提取高级特征，那么基于感知损失的模型能够学习到更鲁棒更令人信服的结果。

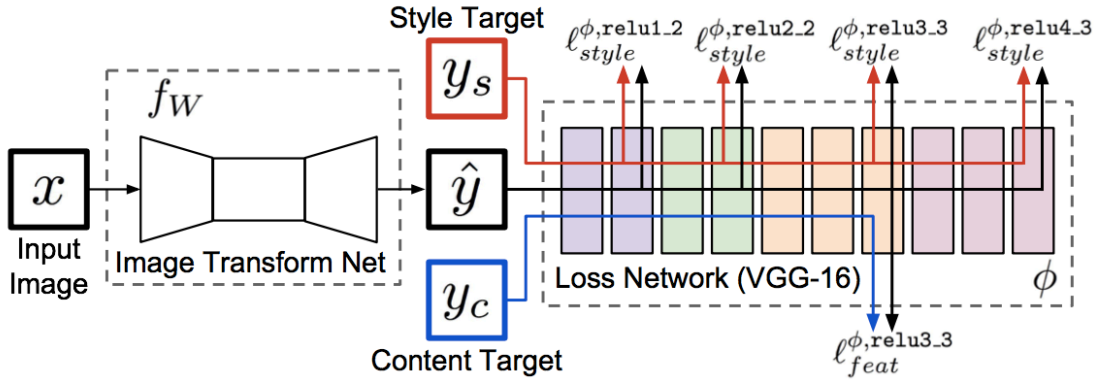

图 4. 基于感知损失的图像风格转换网络，该网络也可用于图像超清问题。左侧是一个待训练的转换网络，用于对图像进行操作；右侧是一个已训练好的网络，将使用其中的几层计算损失。

图 4 即为感知损失网络，该网络本是用于快速图像风格转换。在这个结构中，需

要训练左侧的 Transform 网络来生成图像，将生成的图像 Y 和内容图像与风格图像共同输入进右侧已经训练好的 VGG 网络中得到损失值。而如果去掉风格图像，将内容图像变为高清图像，将输入改为低清图像，那么这个网络就可以用于解决图像超清问题了。

对抗神经网络（GAN）

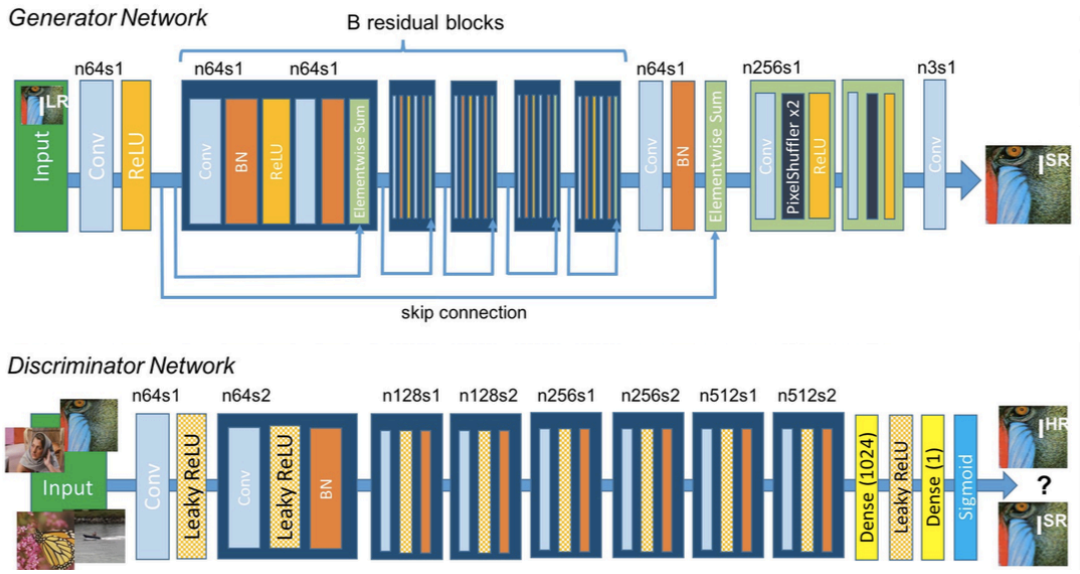

图 5. 对抗训练的生成网络 G 和判别网络结构 D，上-部分是生成网络 G，层次很深且使用了 residual block 和 skip-connection 结构；下一部分是判别网络 D。

对抗神经网络称得上是近期机器学习领域最大的变革成果。其主要思想是训练两个模型 G 和 D，G 是生成网络而 D 是分类网络，G 和 D 都用 D 的分类准确率来进行训练。G 用于某种生成任务，比如图像超清化或图像修复等，G 生成图像后，将生成图像和真实图像放到 D 中去进行分类。训练过程是保持 G 不变，训练 D 使分类准确率提升；保持 D 不变训练 G 使分类准确率下降，直到平衡。GAN 框架可以使得无监督的生成任务可以利用到监督学习的优势来进行提升。

基于 GAN 框架，只要定义好生成网络和分类网络，就可以完成某种生成任务。

而将 GAN 应用到图像高清问题的这篇论文，可以说是集大成之作。生成模型层次深且使用了 residual block 和 skip-connection；在 GAN 的损失函数的基础上同时添加了感知损失。

GAN 的生成网络和分类网络如图 5，其中，生成网络自己也可以是一个单独的图像超清算法。论文中分析了 GAN 和 non-GAN 的不同，发现 GAN 主要在细节方面起作用，但无法更加深入的解释。无法解释性也是 GAN 目前的缺点之一。

像素递归网络(Pixel CNN)

图 5 中的 GAN 虽然能够达到比较好的效果，但是由于可解释性差，难免有套用之嫌。

其实，对于图像超清这个问题来说，有一个关键性的问题，那就是一张低清图像可能对应着多张高清图像。那么问题来了。

假如我们把低分辨率图像中需要高清化的部分分成 A,B,C,D 等几个部分，那么 A 可能对应 A1,A2,A3,A4，B 对应 B1,B2,B3,B4，以此类推。假设 A1,B1,C1,D1 对应一张完美的高清图片。那么现有的算法可能生成的是 A1,B2,C3,D4 这样的混搭，从而导致生成的高清图像有了模糊。

为了验证上述问题的存在，设想一种极端情况。

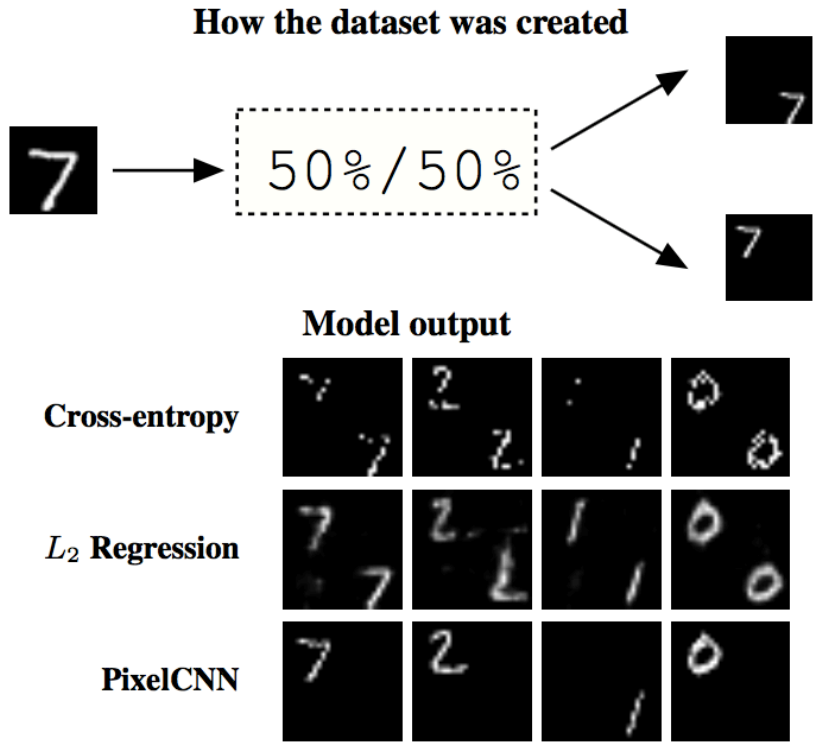

图 6. 图像超清模糊性问题分析图示。上半部分为分析问题所用数据集的构建。下半部分为现有的损失函数在这个问题上的效果，可以看到，PixelCNN 可以防止这种模糊的出现。

为了分析问题，在图 6 中，基于 MNist 数据集生成一个新的数据集，生成方法如下，将 MNIST 数据集中的图片 A 长宽各扩大两倍，每张图片可以生成两张图片 A1 和 A2，A1 中 A 处于右下角，A2 中 A 处于左上角。

把原图当做低清图片，生成的图当成高清图片。使用现在的方法进行训练，得到的模型，在生成图像的时候，会产生图 6 下半部分的结果。即每个像素点可能等

概率的投射到左上部分和右下部分，从而导致生成的图片是错误的。而引入 PixelCNN 后，像素之间产生了依赖关系，这种情况则不会发生。

为了解决上述问题，需要在生成图像的时候引入先验知识。画家在拥有了人脸的知识之后，就可以画出令人信服的高清细节。而这种先验知识在图像超清问题中就是告知算法该选择哪一种高清结果。

在图像超清中，这样的知识体现为让像素之间有相互依赖的关系，这样，就可以保证不同的部分，其高清版的选择是一致的。

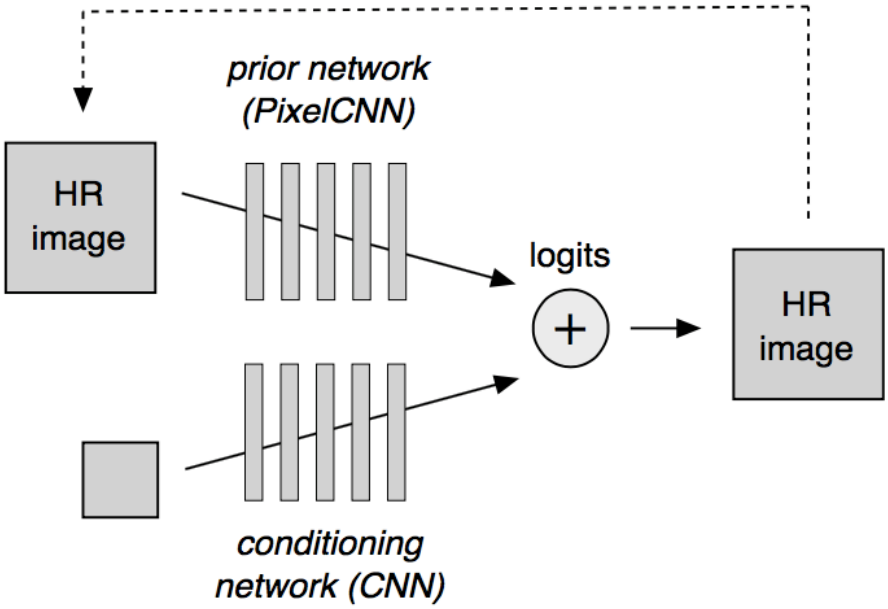

图 7. 基于 PixelCNN 的解决图像超清问题的 CNN 网络结构。其中先验网络(prior network)为 PixelCNN；条件网络（conditioning network）为图像生成网络，其结构与 GAN 中的生成网络，感知损失中的转换网络结构与作用均类似。

模型架构如图 7。其中条件网络是一个将低清图像生成高清图像的网络，它可以像素独立的生成高清图像，如同 GAN 中的 G 网络，感知损失中的转换网络。而先验网络则是一个 Pixel CNN 组件，它用来增加高清图像像素间的依赖，使像素选择一致的高清细节，从而看起来更加自然。

那么 Pixel CNN 是如何增加依赖的呢？在生成网络的时候，Pixel CNN 以像素为单位进行生成，从左上角到右下角，生成当前像素的时候，会考虑之前生成的像素。

而先验网络和条件网络的混合，则是在 PixelCNN 生成图像的时候，除了考虑前面生成的像素，还需要考虑条件网络的结果。

总结

上述算法只是图像超清问题中使用的较为典型的 CNN 结构，还有不少其他的结构也达到了比较好的效果。但随着 CNN 网络结构的日益加深，距离实用反而越来越远。譬如，基于 GAN 的网络结构的训练很难稳定，且结果具有不可解释性。而基于 PixelCNN 的网络在使用中由于要在 pixel 级别生成，无法并行，导致生成效率极为低下。

更进一步的，从实用出发，可以在数据方向上进行进一步的优化，譬如，现在的算法输入图像都是由低清图像三次插值而来，那么，是否可以先用一个小网络得到的结果来作为初始化的值呢？再如，多个小网络串联是否能得到比一个大网络更好的结果等等。

图像超清问题是一个相对来说比较简单的图像语义问题，相信这只是图像语义操作的一个开始，而越来越多的图像处理问题将会因为 CNN 的出现变得迎刃而解。

参考文献

1. Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(2): 295-307.
2. Kim J, Kwon Lee J, Mu Lee K. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1646-1654.
3. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]//European Conference on Computer Vision. Springer International Publishing, 2016: 694-711.
4. Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[J]. arXiv preprint arXiv:1609.04802, 2016.
5. Dahl R, Norouzi M, Shlens J. Pixel Recursive Super Resolution[J]. arXiv preprint arXiv:1702.00783, 2017.