Rezolution: 一种基于图像特征锐化的

高效上采样方法

赵轩磊1 严茹丹1 苏秦1 董思琪1

**摘要** 特征上采样是当前许多视觉网络架构中的关键操作，例如特征金字塔。它的设计对于密集预测任务至关重要，例如语义/实例分割。在这项工作中，我们提出了Rezolution，一种轻量且高效的上采样算法。不同于常用的插值或生成式方法，Rezolution将特征上采样定义为图像特征锐化，对插值后的图片重新赋予重点。因此，Rezolution拥有几个优越的特点：(1)动态更新重点。Rezolution分别提取图片插值前与插值后的特征，根据两者的差异动态更新图像特征，使其符合原有分布。(2)兼顾全局与局部感受野。与之前仅利用子像素邻域的工作的插值方法或拥有领域感受野的生成式方法不同，Rezolution可以兼顾全局与局部，聚合上下文信息。(3)高效计算。无论放大倍率的大小，Rezolution引入的计算开销都很小，可以很容易地集成到大多数网络架构中。我们在语义分割任务上进行了一系列评估，Rezolution相比于当前的SOTA方法，能够在VOC12 val提高了0.9% mIoU，同时减少96%的FLOPS与80%的参数量。

**关键词** 上采样，注意力, 计算机视觉，深度学习

**Rezolution: An effective upsample method based on**

**image feature sharpening**

Xuanlei Zhao1 Rudan Yan1 Qin Su1 Siqi Dong1

**Abstract**  Feature upsampling is a key operation in a number of modern convolutional network architectures, e.g. feature pyramids. Its design is critical for tasks such as semantic/instance segmentation. In this work, we propose Rezolution, a lightweight and highly effective way to deal with it. Different from regular interpolation or generative methods, Rezolution regard upsampling as image feature sharping, in which we redistribute its focus after upsample. Thus, Rezolution has several appealing properties: (1) Dynamically feature updating. Rezolution extracts features from images both before and after upsample. Then it reorganizes the features according to their difference to fit the original attention distribution. (2) Comprehensive receptive field. Unlike previous interpolation or generative works that only exploit a limited neighborhood, Rezolution can aggregate contextual information within a globle receptive field as well as a neighbor one. (3) Lightweight and fast to compute. Rezolution introduces little computational overhead and can be readily integrated into most network architectures. We conduct comprehensive evaluations on standard benchmarks in semantic segmentation. Rezolution shows a gain by 0.9% mIoU over SOTA works and reduces FLOPs and parameters by 96% and 80% respectively.

**Key words**  Upsample, Attention, Computer Vision, Deep Learning

# 

特征上采样是深度神经网络中最基本的操作之一。一方面，对于密集预测任务中的解码器（例如[1, 2]，图像修复[3, 4]和语义分割[5, 6]），特征图通常会被被上采样以匹配高分辨率监督。在另一方面，特征上采样还涉及将高、低分辨率的特征图相互融合，这在许多最先进的架构中被广泛采用，例如，

1.华中科技大学电子信息与通信学院 湖北 430074

1. School of Electronic Information and Communications, Huazhong University of Science and Technology Hubei 430074

特征金字塔网络[7]、U-Net [8]、堆积沙漏[9]和HRNet[10]。因此，如何设计有效的特征上采样算子成为一个关键问题。

最广泛使用的特征上采样算子是最近邻和双线性插值，它们采用像素之间的空间距离来指导上采样过程。然而，最近邻和双线性插值只考虑子像素邻域，未能捕捉到密

集预测任务所需的丰富语义信息。自适应上采样的另一条途径是反卷积[11]。 反卷积层作为卷积层的逆算子，它学习一组与实例无关的上采样内核。但是，它有两个主要缺点。

首先，反卷积算子在图像所有位置上应用相同的卷积核，忽视了图像中重要的内容，这限制了它处理局部变化的能力。其次，使用大尺寸卷积核时，参数数量与计算量会大量增加。这使得其难以覆盖小邻域外的更大区域，从而限制了其表达能力和性能。

最近，一些工作提出了生成式的上采样算法，例如Pixel Shuffle[12]和CARAFE[13]。虽然这样的算法在性能上取得了一定的提升，但是也受到了一些缺点的限制。第一，在生成图片时，自生成算法将原图生成的特征重新排列为空间块来完成特征上采样，而每个空间块内忽略了局部像素的位置关系，这会造成上采样图片的空间关系不匹配。第二，此类算法的计算量与感受野成平方关系，这导致感受野依旧被限制在一小块领域范围内，对性能提升造成了阻碍，在处理高分辨率图片时影响尤为明显。第三，在进行高比例上采样时，算法中需要生成的参数也会激增，导致对计算能力的要求攀升，影响了实际的应用。

在这项工作中，为了克服这些限制，需要寻找一种全新的特征上采样算子，能够满足以下的要求：1)兼顾全局与局部信息，2)根据当前的数据自动生成，3)建立通道信息相关性，4)保持计算效率。为此，我们提出了一种轻量但高效的全新方法，称为Rezolution。不同与上述的生成式算法，该方法考虑的是如何对双线性插值上采样的图片进行后处理精调，从而修正需要关注的重点。

在我们看来，我们进行的后处理本质上是对上采样后的图片进行特征锐化操作。对于传统的双线性插值方法，它被广泛使用的原因在于能够高效完成上采样计算，并且保留图像的空间关系。但是就像上述提到的，双线性插值只考虑子像素邻域，未能捕捉到密集预测任务所需的丰富语义信息，对画面中所有信息进行等比例的放大，失去了关注的重点。因此，要解决的问题就是如何对双线性插值放大后的图片进行图像特征锐化，突出对于画面正真重要的内容，是该方法关注的重点。

具体来说，Rezolution对原图与上采样后的新图片，分别提取沿着两个不同方向聚合的特征，再通过原图特征修正新图特征，最后将新特征映射回新图，完成上采样图片修正的过程，得到最终的上采样图片。

为了证明Rezolution的普遍有效性，我们对语义分割任务进行了综合评估。Rezolution改进了HRNet的特征图融合部分，相对于SOTA方法，在语义分割的数据集VOC12 val上提高了0.9% mIoU。在OCRNet上对具有256个通道的特征图进行2-8倍上采样时，相比于基准算法CARAFE[13]，Rezolution将计算量缩减了96%，参数量缩减了80%。所有任务的实质性收益表明，Rezolution是一种有效且高效的特征上采样算子，具有作为未来研究的强大基石的巨大潜力。

# 相关算法概述

## 上采样运算符

最常用的上采样方法是最近邻和双线性插值。这些插值利用距离来测量像素之间的相关性，并在其中使用手工制作的上采样内核。在深度学习时代，提出了几种使用可学习算子对特征图进行上采样的方法。例如，反卷积[11]是卷积的逆算子，是那些可学习的上采样器中最著名的。Pixel Shuffle [12] 提出了一种不同的上采样器，它将通道空间上的深度重塑为空间空间上的宽度和高度。最近，[14]提出了引导上采样（GUM），它通过对具有可学习偏移量的像素进行采样来执行插值。然而，这些方法要么利用小邻域中的上下文信息，要么需要昂贵的计算来执行自适应插值。在超分辨率和去噪领域，其他一些作品 [15, 16, 17] 也探索在低级视觉中空间使用可学习内核。本着类似的设计精神，我们在这里展示了内容感知特征重组在多个视觉感知任务中进行上采样的有效性和工作机制，并提供了一个轻量级的解决方案。

## 语义分割任务

语义分割[18, 19]需要对给定图像进行像素

图1: **Rezolution的整体框架**。Rezolution由三个步骤组成。第一步是对原图与上采样后的图片分别提取特征向量，第二步是用原图的特征向量更新上采样后的特征向量,再将新的特征向量映射回图片。尺寸为 的特征图以上采样倍率k(=2)被放大

级语义预测。PSPNet[20]在多个网格尺度上引入了空间池化。UperNet[21]设计了一个基于PSPNet的更通用的框架。U-net[8]在最近的作品[22, 3]中很受欢迎，并且采用了多个上采样算子。[23]引入部分卷积层来减轻缺失区域对卷积层的影响。HRNet[10]通过反复交换跨分辨率子网络信息来丰富语义信息和细节信息。OCRNet[24]以HRNet为骨干，在构建上下文信息时显式地增强了来自于同一类物体的像素的贡献。我们的Rezolution在语义分割任务中，对于不同模型展示了普遍的有效性。

# Rezolution

在许多现代卷积网络架构中，特征上采样是一个关键的算子，这些架构是为目标检测、实例分割和场景解析等任务而开发的。

在这项工作中，我们提出了Rezolution来对特征图进行上采样。，并获得比主流上采样算子更好的性能，例如插值或生成式算法。

## 公式

Rezolution由两个步骤组成。第一步是对原图与上采样后的图片分别提取特征向量，第二步是用原图的特征向量更新上采样后的特征向量,再将新的特征向量映射回图片。给定大小为的特征图和上采样率(假设是整数),首先会通过双线性插值生成大小为的特征图，之后Rezolution将调整特征图，最终生成大小为的新特征图。

在第一步中，特征提取模块将通过池化分别对上采样前与上采样后特征图的特征图提取关键特征，如公式(1)所示。在第二步中，利用模块将原图的特征向量拓展，并使用更新上采样图的特征向量，最后将特征向量映射回原图，如公式(2)(3)所示。我们将在本节后续部分中说明、、的详细信息。

（1）

（2）

（3）

## 特征提取

在特征提取部分，我们参考CANet[25]的方法。全局池化通常用于通道注意力以全局编码空间信息，但它将全局空间信息压缩

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 算 法 | 骨干网络 | 上采样方法 | mIoU | mAcc | aAcc |
| PSPNet | ResNet50 | Bilinear | 75.4 | 86.1 | 94.4 |
| CARAFE | 75.6 | 86.3 | 94.3 |
| Rezolution | **76.1** | **87.1** | **94.4** |
| UperNet | ResNet50 | Bilinear | 74.0 | 84.6 | 93.9 |
| CARAFE | 74.3 | 84.9 | 94.1 |
| Rezolution | **74.7** | **85.4** | **94.2** |
| OCRNet | HRNet18 | Bilinear | 73.4 | 85.0 | 93.8 |
| CARAFE | 73.9 | 85.4 | 94.0 |
| Rezolution | **74.8** | **85.5** | **94.1** |

表1 不同算法的性能比较

到通道描述符中，因此难以保留位置信息，这对于在视觉任务中捕获空间结构至关重要。为了鼓励注意力块以精确的位置信息在空间上捕获远程交互，我们将全局池化分解为为一对一维特征编码操作，如公式(4)(5)。具体来说，给定输入 X，我们使用池化内核的两个空间范围或，分别沿水平坐标和垂直坐标对每个通道进行编码。因此，高度为的第个通道的输出可以表示为：

（4）

类似地，宽度为的第个通道的输出可以写为:

（5）

上述两种变换分别沿两个空间方向聚合特征，产生一对方向感知特征图。这两种转换还允许我们的注意力块沿一个空间方向捕获远程依赖关系，并沿另一个空间方向保留精确的位置信息，这有助于网络更准确地定位感兴趣的对象。

## 特征更新

如上所述，方程(4)和方程(5)启用全局感受野并编码精确的位置信息。为了更新上采样后的特征向量，我们必须先对原图的特征向量进行扩展，使得原图特征向量的长度与上采样后的特征长度保持一致。对于，特征向量的扩张，我们使用了经典方法PixelShuffle，即先用卷积增加通道深度，再利用通道信息将每个位置的向量空间上扩大相应的倍数，如公式(6)所示，其中代表卷积函数。

（6）

获得了扩展后的原图特征后，我们希望使用原图特征去指导上采样后的特征，通过特征向量中捕获的空间和通道信息，将被双线性插值稀释的重点信息重新标记出来。为了完成这个目标，我们提出了一个简单但是有效的方法来更新上采样特征——门控机制。具体来说，对于扩展后的原图特征与上采样后的特征，我们通过计算与的内积，获取这两个向量在高维空间中的几何关系，将内积后的结果作为门控，如公式(7)所示，再利用门控机制对上采样后的特征进行更新，如公式(8)所示。

（7）

. （8）

# 实现

语义分割要求模型在整个图像上输出每像素级别的预测，因此通常使用高分辨率的特征图输出结果。上采样被广泛用于放大特征图，并在自上而下的路径中以一定比例将特征图融合在一起。我们选择了3个目标检测中具有代表性的模型：PSPNet[20]，UperNet[21]，OCRNet[24]，实现了我们的上采样方法。

## PSPNet

PSPNet中PPM将输入特征图分层下采样到多个尺度，然后使用双线性将它们上采样回原始大小。这些特征最终通过串联与原始特征融合。由于上采样率非常大，我们采用了两步策略作为

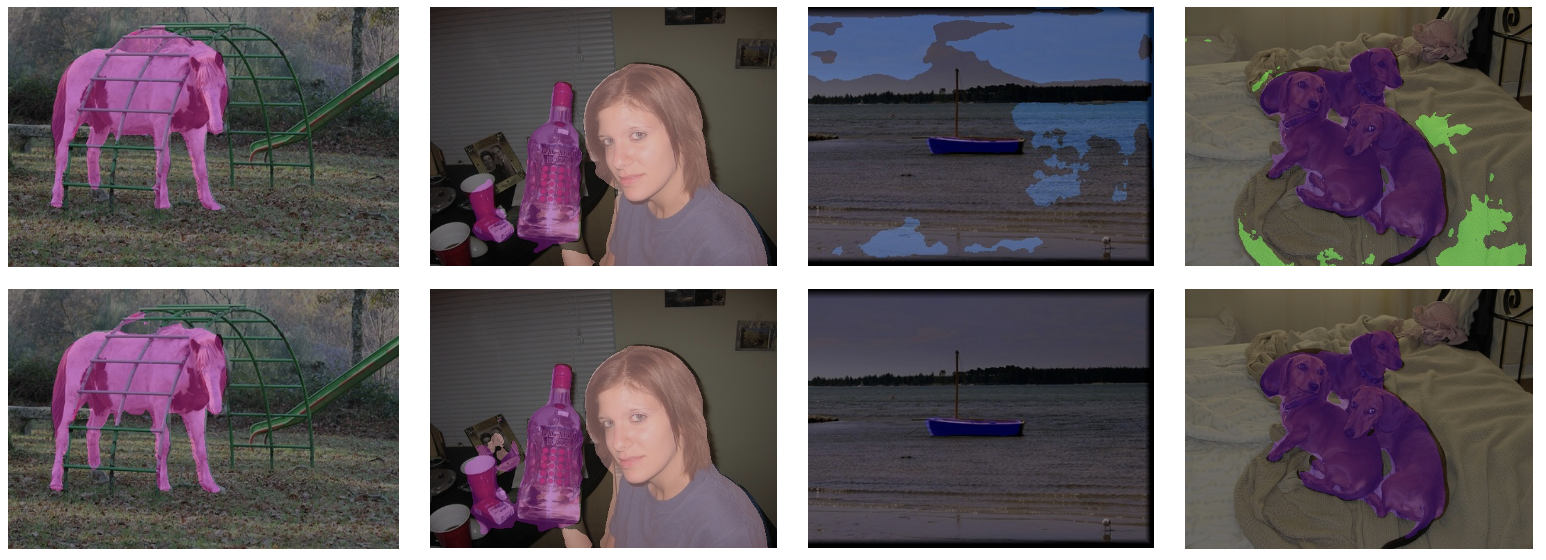


图2：**Rezolution与CARAFE效果对比。**第一行为CARAFE的结果，第二行为Rezolution的结果。左侧两列图片中Rezolution对于局部信息理解的更好，右侧两侧图片中Rezolution优化了全局的效果

性能和效率之间的权衡。首先我们使用双线性插值将特征上采样到原始特征图大小的一半，然后使用 Rezolution将它们进一步上采样2倍。

## UperNet

在UperNet中，我们修改了PPM和FPN两个模块中的上采样。PPM的修改方法与3.1节中提到的相同。对于FPN，UperNet采用它来丰富语义特征。它只有四个不同的特征级别，步幅为。我们依次从高级特征到低级特征上采样两倍。

## OCRNet

OCRNet使用HRNet作为backbone，在backbone的最后分别将4个不同尺度的特征图分别放大倍，最后将他们拼接在一起获得最终的特征图，在此处我们直接用Rezolution上采样对应的倍数。

## CARAFE

我们将目前SOTA的方法CARAFE作为对比之一，直接使用mmdetection中提供的代码。CARAFE使用生成式的方法，对图片中每一个位置按照上采样比例生成代码块，最终将代码块合成新的图片。在上采样时，我们直接用CARAFE替代双线性插值。

# 实验

在本节中，我们首先描述我们的实验设置，然后进行一系列对比实验证明我们算法的有效性，我们报告了与其他上采样方法在语义分割上的结果的对比。之后，我们对不同上采样算法的参数、计算量与mIoU进行了对比。接着我们对比了不同算法对于不同比例缩放的性能，证明我们的算法在进行大比例上采样时明显优于以往方法。最后我们进行了一系列消融实验以证明我们的门控方法对性能的贡献。

## 实验设置

我们使用基于PyTorch的mmsegmenta-tion工具来实现我们所有的实验。在训练期间，我们使用衰减和动量为0.9的标准SGD优化器来训练所有模型。权重衰减始终设置为。我们使用两NVIDIA 2080Ti GPU进行训练，批量大小设置为8。在没有额外声明的情况下，我们训练所有模型 200 个 epoch。 对于数据增强，我们使用与mm-segmentation相同的方法。我们报告了VOC12-val中语义分割的结果。

## 基线结果

在基线结果对比中，我们挑选了3个语义分割的主流算法：PSPNet，UperNet与OCRNet，前两者的骨干网络为ResNet50，OCRNet的骨干网络为HRNet。对于每种算法，我们分别测试了双线性插值，CARAFE和Rezolution三种算法的效果，结果如表1所示，可以从表中看出，对于三种不同结构的算法，Rezolution方法都取得了明显的提

升。同时，我们将分割结果可视化出来，如图2所示，可以看到Rezolution无论是对局部细节还是全局理解都更胜一筹。

## 性能分析

在性能分析中，我们对比了在不同放大倍率下，不同采样方法的区别。

首先，在大比例上采样的情况下，我们对比了CARAFE与Rezolution的性能提升。如表2所示，我们能够发现在上采样倍率为2时，Rezolution相比于CARAFE能够提高0.4%的mIoU。而当上采样倍率变大时，Rezolution的提升变得更为明显，能够达到0.9%。证明我们的方法无论是何种放大比例，都能够取得更优的效果。造成的稿件如果第一作者非文章通信作者或通信作者地址变动, 务请及时通知编辑部.

表2 不同上采样倍率下的提升

|  |  |  |  |
| --- | --- | --- | --- |
| 算 法 | 倍率 | 采样方法 | mIoU |
| PSPNet | 2x | Bilinear | 75.4 |
| CARAFE | 75.6 |
| Rezolution | **76.1** |
| UperNet | 2x | Bilinear | 74.0 |
| CARAFE | 74.3 |
| Rezolution | **74.7** |
| OCRNet | 2-8x | Bilinear | 73.4 |
| CARAFE | 73.9 |
| Rezolution | **74.8** |

最终指标以外，我们也对三种不同上采样方法的计算性能进行了评估，主要考虑计算量与参数量。我们使用了OCRNet作为模型，HRNet作为骨干模型，评估三种上采样算法的计算量与参数量，如表3所示。可以发现，相比于轻量型算法CARAFE，在取得相当的性能提升以外，我们将计算量缩减了96%，参数量缩减了80%。

表3 不同算法的计算量与参数量

|  |  |  |  |
| --- | --- | --- | --- |
| 上采样方法 | mIoU | Flops | Params |
| Bilinear | 73.4 | 1M | 0 |
| CARAFE | 73.9 | 981M | 1.25M |
| Rezolution | **74.8** | 32M | 0.26M |

## 消融实验

在消融实验中，我们评估了特征更新算法的性能，我们使用三种方法进行对比：特征拼接，GRU更新，门控更新。如表4所示，门控方法虽然很简单，但是相比于其他两种方法取得了显著的提升，证明了我们特征融合方法的高效性。

表4 不同上采样倍率下的提升

|  |  |  |  |
| --- | --- | --- | --- |
| 融合方法 | mIoU | mAcc | aAcc |
| Concat | 74.0 | 85.1 | 94.0 |
| GRU | 73.5 | 84.9 | 93.8 |
| Gate | **74.8** | **85.4** | **94.1** |

# 结论

本文中，我们提出了一种全新的上采样方法Rezolution。不同于当前流行的生成式算法它将上采样过程定义为一种图片特征锐化的过程，对双线性插值后的结果重新赋予重点，将语义分割的标准基准性能提高了0.9% mIoU。更重要的是，Rezolution引入的计算开销非常小，并且可以很容易地集成到各种网络架构中。

# References

1. Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2016
2. Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2017
3. Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics, 36(4):107, 2017
4. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In IEEE Conference on Computer Vision and Pattern Recognition, 2016
5. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In IEEE Conference on Computer Vision and Pattern Recognition, 2017
6. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision, 2018
7. Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In IEEE Conference on Computer Vision and Pattern Recognition, July 2017.
8. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, 2015.
9. Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision, 2016.
10. Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. In Computer Vision and Pattern Recognition, 2019.
11. Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. IEEE International Conference on Computer Vision, Dec 2015.
12. Wenzhe Shi, Jose Caballero, Ferenc Husz´ar, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In IEEE Conference on Computer Vision and Pattern Recognition, 2016.
13. J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, “CARAFE: Content-aware reassembly of features,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-Octob, pp. 3007–3016, 2019, doi: 10.1109/ICCV.2019.00310.
14. ﻿Davide Mazzini. Guided upsampling network for real-time semantic segmentation. arXiv preprint arXiv:1807.07466, 2018.
15. Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.
16. Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.
17. Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-SR: A magnification-arbitrary network for super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.
18. Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In IEEE International Conference on Computer Vision, 2015
19. Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade. In IEEE Conference on Computer Vision and Pattern Recognition, 2017
20. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.
21. Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In European Conference on Computer Vision, 2018.
22. Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.
23. Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In European Conference on Computer Vision, 2018.
24. ﻿Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object- contextual representations for semantic segmentation. In 16th European Conference Computer Vision (ECCV 2020), August 2020. 7
25. Q. Hou, D. Zhou, and J. Feng, “Coordinate Attention for Efficient Mobile Network Design,” 2021, [Online]. Available: http://arxiv.org/abs/2103.02907.