基于解耦区域校准的高分辨率超像素生成算法

王亚雄[1)] 魏云超[2)] 钱学明[3)] 朱 利[3)]

[1)](合肥工业大学计算机与信息学院 合肥 230009)

[2)](北京交通大学计算机与信息技术学院 北京 100044)

[3)](西安交通大学电信学部 西安 710049)

摘 要 超像素分割是计算机视觉领域的一项重要任务,该任务将具有相似属性的像素分组到称为超像素的簇中.图像超像素不仅可以增益图像注释,而且还是各种下游应用的基础,如分割、光流估计和深度估计.尽管超像素分割技术取得了显著进展,特别是随着深度学习方法的出现,但现有解决方案由于 GPU 内存和计算能力的限制,一直无法有效处理高分辨率图像.针对这个问题,作者提出了一种名为区域解耦校准的高分辨率超像素网络(Patch Calibration Network,PCNet)的新型深度学习框架,通过采用解耦的一致性学习策略,解决了现有方法的局限性.这种方法允许通过从低分辨率输入预测高分辨率输出来高效生成高分辨率超像素结果,从而绕过了 GPU 内存限制.PCNet 的一个关键贡献是解耦的区域块校准(DPC)分支,它将高分辨率图像块作为额外输入,以保留细节并增强边界像素分配.为了改善边界像素的识别,作者利用二进制掩模设计了一种动态引导训练机制.这种机制鼓励网络专注于区域内的主要边界,将任务从多类分类简化为二分类问题.这一创新策略不仅减少了网络优化的复杂性,而且显著提高了边界检测的精度.本文通过在包括 Mapillary Vistas、BIG 和新创建的 Face-Human 数据集在内的多样化数据集上进行广泛的实验,证明了 PCNet 的有效性.结果表明,PCNet 能够成功处理 5K 分辨率图像,并与现有的最先进的 SCN 方法相比,实现了更优越的性能,后者在处理高分辨率输入时存在困难.作者的贡献包括开发了 PCNet,一种针对高分辨率超像素分割的深度学习解决方案,引入了解耦的区域校准架构,并构建了一个超高分辨率基准测试集,用于评估高分辨率场景中超像素分割算法的性能.本文首先回顾了超像素分割领域的相关工作,然后详细介绍了 PCNet 框架,接着展示了实验结果并与最先进的方法进行了比较.结论部分总结了研究结果并概述了未来研究的潜在方向.代码、预训练模型和新的基准数据集的可用性无疑将促进高分辨率超像素分割领域的进一步发展.总之,本文在超像素分割领域提供了一个重要的进步,提供了一种能够高效、准确处理高分辨率图像的解决方案.所提出的 PCNet 框架,凭借其创新的 DPC 分支和动态引导训练机制,为未来在计算机视觉领域的研究和应用提供了一个有前景的方向.本文的代码、预训练模型以及新构建的评估基准数据集可在 https://github.com/wangyxxjtu/PCNet 上获取.

关键词 超像素分割;图像分割;高分辨率视觉;深度学习;人工智能

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.02664

Generating Superpixels for High-Resolution Images with Decoupled Patch Calibration

WANG Ya-Xiong[1)] WEI Yun-Chao[2)] QIAN Xue-Ming[3)] ZHU Li[3)]

[1)](*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei* 230009)

[2)](*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)

[3)](*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049)

Abstract Superpixel segmentation is a significant task in the field of computer vision that involves grouping pixels with similar attributes into coherent clusters known as superpixels. These superpixels are not only useful for image annotation but also serve as a foundation for various downstream

收稿日期:2023-06-07;在线发布日期:2024-07-16.本课题得到国家自然科学基金(62302140)、中央高校基本科研业务费专项资金(合肥工业大学学术新人提升计划(JZ2024HGTB0261)资助.**王亚雄**,博士,副教授,主要研究方向为图像理解与跨模态计算.E-mail:wangyx15@stu.xjtu.edu.cn.**魏云超**(通信作者),博士,教授,主要研究领域为弱监督学习与图像语义理解.E-mail:wychao1987@gmail.com.**钱学明**(通信作者),博士,教授,主要研究领域为跨模态内容理解.E-mail:qianxm@mail.xjtu.edu.cn.**朱 利**,博士,教授,主要研究领域为图像处理与计算机网络.

applications such as segmentation, optical flow estimation, and depth estimation. Despite the substantial progress in superpixel segmentation techniques, particularly with the advent of deep learning methods, existing solutions have been unable to effectively handle high-resolution images due to constraints in GPU memory and computational power. The authors propose the Patch Calibration Network (PCNet), a novel deep learning framework that addresses the limitations of current methods by employing a decoupled consistency learning strategy. This approach allows for the efficient generation of high-resolution superpixels by predicting high-resolution outputs from low-resolution inputs, thereby bypassing the GPU memory limitations. A key aspect of PCNet is the Decoupled Patch Calibration (DPC) branch, which incorporates high-resolution image patches as additional inputs to preserve fine details and enhance boundary pixel allocation. To improve the identification of boundary pixels, the authors introduce a dynamic guidance training mechanism that utilizes a binary mask. This mechanism encourages the network to focus on the primary boundaries within a region, simplifying the task from multi-class classification to a binary classification problem. This innovative strategy not only reduces the complexity of network optimization but also significantly enhances the precision of boundary detection. The paper demonstrates the effectiveness of PCNet through extensive experiments on diverse datasets, including Mapillary Vistas, BIG, and a newly created Face-Human dataset. The results indicate that PCNet can successfully process 5K resolution images and achieve superior performance compared to the state-of-the-art SCN method, which struggles with high-resolution inputs. The authors' contributions include the development of PCNet, a deep learning solution for high-resolution superpixel segmentation, the introduction of a decoupled regional calibration architecture, and the construction of an ultra-high-resolution benchmark dataset for evaluating the performance of superpixel segmentation algorithms in high-resolution scenarios. The paper is structured to first review the related work in the field of superpixel segmentation, then present the PCNet framework in detail, followed by experimental results and comparisons with state-of-the-art methods. The conclusion summarizes the findings and outlines potential directions for future research. The availability of code, pre-trained models, and the new benchmark dataset will undoubtedly facilitate further advancements in the field of high-resolution superpixel segmentation. In summary, this paper presents a significant advancement in the domain of superpixel segmentation, providing a solution that can handle high-resolution images efficiently and accurately. The proposed PCNet framework, with its innovative DPC branch and dynamic guidance training mechanism, offers a promising direction for future research and applications in computer vision. Our code, pre-trained models, and the newly constructed evaluation benchmark dataset are available at https://github.com/wangyxxjtu/PCNet.

Keywords　superpixel segmentation; image segmentation; high-resolution vision; deep learning; artificial intelligence

1　引　言

超像素分割的目标是将具有相似颜色或其他低层属性的像素分配到相同的簇中，这个过程可以视为图像上的聚类. 生成的像素簇称为超像素，超像素可以直接用于图像标注[1]或增益下游任务[2-7]. 受深度卷积神经网络的推动，诸多基于深度学习的方法已被提出来提升超像素分割的性能，并取得了良好的结果[8-10]. 基于深度网络的超像素分割的一般做法是首先将图像分割成网格，接着利用卷积网络为每个像素预测一个 9 维向量，该向量表示每个像素被分

配到其周围 9 个网格的概率[9-10]. 然而,由于 GPU 显存限制,现有方法无法处理较高分辨率图像.

受益于硬件以及移动设备的发展,高分辨率图像的获取变的相对容易,另一方面,某些领域的数据本身具有高分辨率的特征,例如医学图像和 SAR 图像等. 此外,高分辨率的超像素分割也对高分辨的图像标注具有重要意义. 因此,在计算机视觉方面,尤其是像素级任务(如分割[11-15]、光流估计[16-19]、深度估计[20-22]等),高分辨率问题受到了较大的关注. 对于超像素分割,虽然已经提出了一些基于深度学习的方法,但高分辨率场景尚未得到很好的探索. 例如,目前最先进的方法 SCN[9] 只能为低分辨率图像生成超像素,而如果遇到高分辨率图像,其推理速度将会变很慢. 在英伟达 1080Ti GPU 上,SCN[9] 无法在输入图像分辨率超过 3.5K 的情况下正常工作,如图 1 所示,我们的方法能够在保持边界精度的同时,高效地获取高分辨率图像(5K)的超像素(结果在单个 NVIDIA1080Ti GPU 和 Face-Human 数据集上获得). 为了扩大可处理的分辨率上限,一个直观的解决方案是从低分辨率图像预测高分辨率输出. 通过在解码器阶段引入额外的上采样层,我们可以使得网络从较低分辨率的 $H/4\times W/4$ 图像中学习到 $H\times W$ 的关联映射,如图 2(b)所示. 采用这种设计,网络可以成功处理更大尺寸的图像.

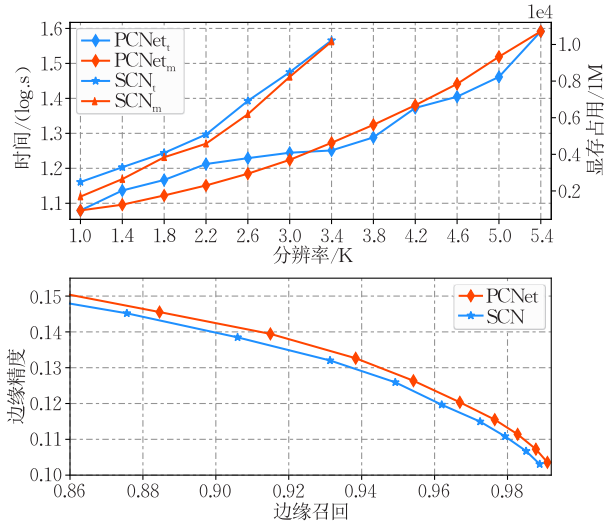


图 1 PCNet 与最先进的 SCN[9] 在时间、内存和性能的比较(其中下标"t"和"m"分别表示时间耗费和内存占用)

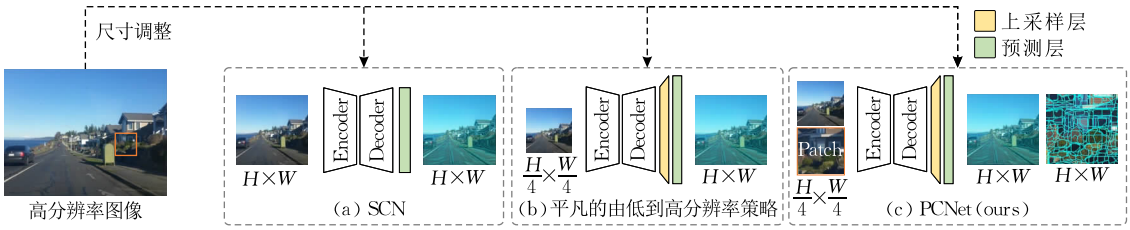


图 2 不同训练架构的比较(SCN[9] 在输入输出具有相同的分辨率. 直接从低分辨率图像预测高分辨率图像的朴素解决方案无法取得满意的效果. 通过引入校准分支,我们的 PCNet 可以处理更高分辨率的图像并同时实现令人满意的性能)

然而,低分辨率的输入会损失大量信息,这会阻碍网络对图像的纹理背景进行感知,特别是语义边界细节,从而导致性能的显著下降. 为了感知更多边界细节,我们提出通过直接从原始高分辨率图像中裁剪补丁作为附加输入来校准边界像素分配,从而提出了区域块校准网(Patch Calibration Network, PCNet). 我们 PCNet 的核心思想如图 2(c)所示,我们直接从原始高分辨率图像中裁剪出一个局部图像块,以保留所有细节. 然后,我们将其送入解耦的区域块校准(DPC)分支,从而帮助对粗略全局预测的边界细节进行校准. 这里"解耦"指的是将局部区域块从原始图像中裁剪下来,并与原始图像协同作为输入的训练方式. 通过与主分支共享权重,从 DPC 学习到的知识能够得到很好的转移,使主超像素分支能够准确感知更多边界像素.

不同于通过将像素分类为其语义类别来感知整体边界布局的全局输入,我们裁剪的区域块的目的仅仅在于帮助网络准确分配边界周围的像素,而不关注其所包含的对象的语义信息. 为了实现这一点,我们设计了动态引导训练机制. 具体地,我们设计一个二值掩码来鼓励网络仅关注当前区域的主要边界,而不是贪心地识别区域块中所有像素的类别,这样可以将多类语义标签降级为二值掩码作为指导. 通过我们的动态引导训练,每次迭代,网络只需区分前景和背景以突显主要边界,而无需识别所有像素的多个类别. 这种策略不仅可以减轻网络优化的难度,还可以使网络准确识别更多边界像素.

从图 1 可以看出,我们的 PCNet 可以在单个 NVIDIA 1080Ti GPU 上可以成功处理 5K 分辨率图像(5000×5000 至 6000×6000 的分辨率范围). 与

最先进的 SCN[9] 相比，虽然输入分辨率缩小了 4 倍，但 PCNet 的性能仍然可以略微超过它. 对于 Mapillary Vistas[23]、BIG[11] 和我们创建的 Face-Human 数据集的定量和定性结果表明，所提出的方法可以有效地处理高分辨率 5K 图像，并获得更好的性能. 在这项工作中，我们做出的贡献可以总结如下：

(1) 一个基于深度学习的高分辨率超像素分割解决方案. 我们针对高分辨率的超像素分割进行了早期探索，并提出了一个解决方案. 所提出的 PCNet 可以有效地处理 5K 图像，并在五个基准数据集上取得了令人满意的性能.

(2) 解耦区域块校准架构. 我们提出了一种补丁校准训练范式. 采用这种架构，全局图像和局部图像块可以实现信息互补. 该架构采用动态引导训练策略以提高性能，并设计了局部判别损失函数以帮助实现更好的边界精度.

(3) 超高分辨率基准测试集. 我们构建了一个超高分辨率基准测试集，它的图像大小范围为 1810×2066 到 10 000×6675，用于评估超高分辨率超像素分割性能.

接下来，我们将首先在第 2 节介绍与本文相关的现有工作；并在第 3 节详细阐述我们的 PCNet；第 4 节将呈现实验结果，并与最先进的方法进行比较；最后，在第 5 节中将给出结论.

2　相关工作

超像素分割可以被视为对图像进行像素级聚类的过程，该问题的关键在于估计每个像素与其潜在聚类中心的隶属关系[8-9,24-36]. 传统方法使用聚类[25-27,37] 或关联图[24,38] 的技术来进行归属关系估计. 一般来说，基于聚类的方法通常使用聚类策略来计算锚点像素与其邻居之间的连通性，这种做法也较为直观. 著名的超像素方法 SLIC[25] 将 k-均值算法应用于超像素分割，SLIC 是超像素领域中的一个经典算法. Liu 等人[26] 以计算内容敏感度扩展了 SLIC 算法，设计的模型可以在内容密集区域生成小的超像素，在内容稀疏区域生成大的超像素. Li 等人[37] 通过引入一个精心设计的高维空间，显式地利用了加权 k-均值和归一化分割来优化目标之间的联系. Wang 等人[39] 提出了一种新颖的关联植入方法来捕捉像素网格上下文. 而基于图的方法将超像素分割视为图划分问题，并通过估计像素之间的连接强度来执行超像素分割. Felzenszwalb 等人[38] 利用图像的图的表示，并定义了用于衡量两个区域边界相关性的连通度准则，作者基于该连通度设计了一种高效的超像素分割算法 FH. 在文献[24]中，Liu 等人提出了一个新的超像素分割目标函数，并通过互异性约束下最大化目标函数的图拓扑来给出分割结果. 受到深度卷积网络在许多视觉任务上的成功的鼓舞，研究人员最近尝试利用深度卷积网络来提升超像素分割. Tu 等人[8] 提出通过来自深度网络的像素级深度特征来改进超像素分割的效果. Jampani 等人[10] 提出一种软聚类机制，并将其与深度网络结构相结合，开发了第一个基于深度神经网络的端到端的超像素分割解决方案. 在文献[9]中，作者进一步简化了在文献[10]中的框架，并提出了一种更快、更简单的超像素分割算法. 与基于深度学习的超像素分割相比，传统方法受制于推理速度慢和性能较差的问题. 在为高分辨图像生成超像素时，传统方法的低效率和效果不佳的缺点更为明显. 对于基于深度学习的方法，SSN 旨在设计一个端到端可训练的超像素网络. SCN 更注重效率，并通过简化 SSN 来提供更有效的网络. 尽管 SSN 和 SCN 相对于非学习的方法具有出色的性能，但它们无法执行高分辨率超像素分割. 因此，本文旨在基于深度网络开发一种高效的高分辨率超像素方法. 像 SLIC[25]、ERS[24] 等传统方法也可以执行高分辨率超像素分割. 然而，它们的推理速度通常较慢，并且性能较差. SSN[10] 开发了一种软 k 均值方法，并提出了一个超像素适配的损失函数来训练网络，这是第一个可训练的深度超像素模型. SCN[9] 旨在开发一种更高效的超像素网络，作者进一步简化了 SSN 模型并提出了一种更高效的网络. 本文提出的方法采用了与 SSN 和 SCN 类似的训练范式，但本文的重点在于高分辨率超像素分割，这项任务尚待探索.

3　方法论

图 3 展示了我们提出的 PCNet 的框架示意图. 如图所示，在训练阶段，分别将全局图像 $G \in R^{3\times H\times W}$ 和局部块 $L \in R^{3\times H\times W}$ 作为输入送入到主分支和解耦的区域校准分支中. 输入图像在压缩阶段下采样 4 次，然后在解码阶段上采样 6 次. 因此，两个分支都输出大小为 $9\times 4H\times 4W$ 的关联映射矩阵，分别由语义标签和我们设计的二值引导掩膜进行监督.

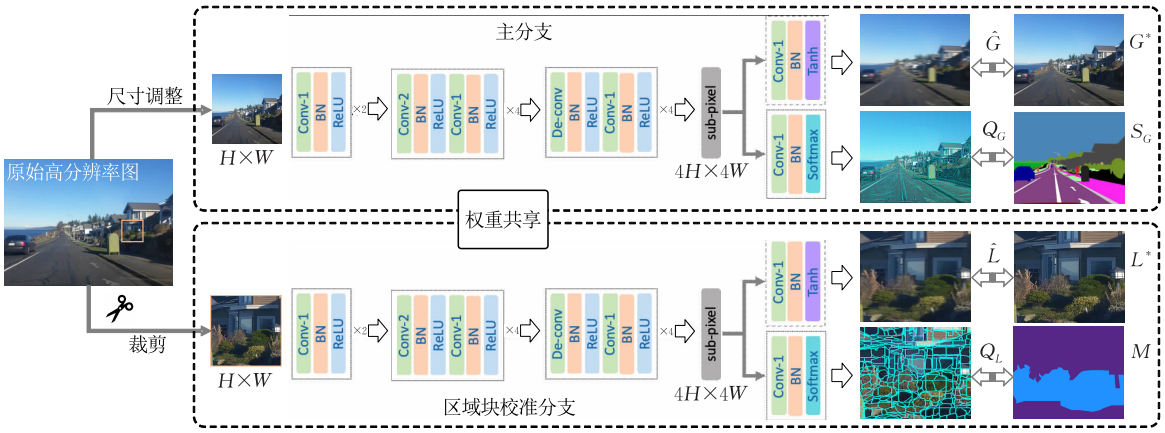


图 3 PCNet 的网络架构图($Q_G \in R^{9\times H\times W}$ 是全局输入的预测关联图,每个像素 $Q_{ij} \in R^9$ 表示像素 $p(i,j)$ 分配给其 9 个周围网格的概率."SR"和"SP"分别表示超分辨率和超像素头.在每次迭代中,低分辨率的全局和局部输入被送入到网络中,并预测 4 倍分辨率的关联概率矩阵,这些关联矩阵预测由真实语义标签和我们的二值引导掩膜进行监督.超分辨率分支作为辅助模块,在训练过程中帮助恢复更多的纹理细节,该分支在推理时被丢弃."Conv-#"表示步长为 # 的卷积,"sub-pixel"表示尺度因子为 4 的子像素卷积操作[40]."高分辨率图像"指数据集中的原始图像.训练过程中,每次迭代中会随机裁剪局部图像块)

超分辨率分支作为辅助模块,在训练时帮助网络从低分辨率输入中恢复更多的细节,其在推理时被丢弃.我们在两个分支之间共享权重,以实现所学习的知识的传播和互通.

3.1 区域校准结构

如图 3 所示,为了提高内存和计算效率,我们的 PCNet 采用编码器-解码器的框架结构,使用低分辨率 $H\times W$ 的输入预测高分辨率 $4H\times 4W$ 的关联映射.我们不使用对称的编码器和解码器层,而是在预测层之前引入一个缩放因子为 4 的子像素层[40],以帮助生成形状为 $9\times 4H\times 4W$ 的关联映射 Q,其中每个 9 维向量表示该像素分配给邻近的九个网格的概率.在训练期间,主分支负责从 $G\in R^{3\times H\times W}$ 中捕捉全局边界布局,该分支的输入是通过调整原始高分辨率图像大小得到的.而区域校准分支则专注于通过从局部区域 L 中获取更精细的边界来校准全局分支的结果.通过共享区域校准分支和主分支的学习权重,主分支可以同时感知全景布局和边界细节,从而有效地防止低分辨率输入的信息损失所导致的性能下降.

两个分支的输出都以类似的方式进行监督.给定全局输入的关联预测 Q_G 和相应的监督标签 S_G,网络训练第一步是利用超像素 S 的周围像素来计算其属性:

$$h(s)=\frac{\sum_{p;s\in N_p} S_G(p)\cdot Q_G(p,s)}{\sum_{p;s\in N_p} Q_G(p,s)} \tag{1}$$

接着在第二步,根据预测的关联矩阵以及第一步中的超像素属性重构每个像素的属性:

$$S'_G=\sum_{s\in N_p} S_G(p)\cdot Q_G(p,s) \tag{2}$$

其中,N_p 是 p 相邻的网格集合,$Q_G(p,s)$ 表示将像素 p 分配给超像素 s 的概率.图 4 中给出了预测关联图的解释.因此,网络的优化目标是最小化重建标签与真实标签之间的差异,其中 S_G 是 one-hot 语义标签的编码向量.与 Yang 等人[9]类似,也考虑了二维空间坐标 p,因此完整的超像素损失函数为

$$L_{SP}(Q_G,S_G)=\sum_p CE(S'_G(p),S_G(p))+\|p'-p\|_2 \tag{3}$$

其中 $CE(\cdot,\cdot)$ 表示交叉熵损失,p' 是根据式(1)~(2)从 p' 重建的向量.

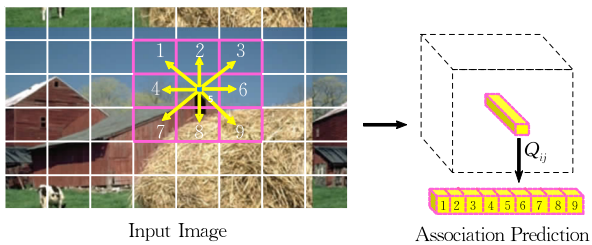


图 4 超像素网络的输出:关联图的直观示意图

为了从低分辨率输入 G 中恢复更多的细节,我们进一步引入了超分辨率分支,给出高分辨率图像 G^* 的重建 $\hat{G}$,通过一个掩膜 L_1 重建损失进行优化,具体细节将在第 3.3 节中详细说明.全局分支的完整损失为 $L_G=L_{SP}(Q_G,S_G)+L_{SR}(\hat{G},G^*)$.

3.2 解耦的区域校准分支

我们的 PCNet 试图从低分辨率输入中生成高

分辨率输出，这种模式允许我们训练一个能够处理更高分辨率图像的网络. 然而，在下采样输入中丢失的纹理细节和模糊的边界上下文使得网络难以很好地识别边界像素. 为了弥补这个缺点，我们设计了解耦的区域校准(Decoupled Patch Calibration，DPC)分支. 具体来说，我们首先从原始高分辨率数据中以边界像素为锚点裁剪局部区域块 $L^* \in R^{3\times 4H\times 4W}$，然后将其调整大小以获得低分辨率输入 $L \in R^{3\times H\times W}$，将其送入 DPC 分支中以捕捉更精细的边界细节. 最后，通过共享 DPC 的学习权重，我们可以赋予主分支较好的边界细节识别能力.

给定 DPC 的输出 Q_L，利用其对应的真实语义标签 S_L 作为监督是最直观的选择，它可以通过将像素分类到它们的对象类别中来提升超像素分割. 这种策略在全局图像上可以很好地工作[9-10]，但在局部输入上表现不佳，因为在我们的实践中裁剪的区域块通常只涵盖对象的一部分，如图 3 所示. 缺乏全局上下文使得多类别分类过程变得非常困难.

实际上，与需要捕捉全局上下文以识别完整对象的语义分割任务不同，超像素分割主要关注是否可以准确地识别边界[9-10,41]. 换句话说，超像素网络只需要区分边界周围的相邻区域以感知边界，而不是像素的所有对象类别. 考虑到这一特点，我们提出了二值引导的掩模来监督局部关联预测. 具体而言，我们首先对应于局部补丁采样语义标签，并找到具有最长边界的类 c，二值引导掩模定义如下：

$$M(p)=\begin{cases}1, & S_L(p)=c\\ 0, & \text{其他}\end{cases} \tag{4}$$

通过我们的二值引导掩模，多类物体识别退化为显著区域检测，这使得网络只需要区分类 c 和其他类别，从而简化了网络优化. 如果前景类别 c 能够被很好地感知，根据类 c 的选择，当前区域块的大多数边界可以被识别出来. 对于当前的掩模生成过程中所忽略掉的边界，他们有几率在其他的采样迭代中进行强调，这是由于局部块是随机裁剪的. 在我们的实践中，与多类标签 S_L 相比，我们的二值引导掩模可以带来更多的性能提升，如图 5 所示. 值得注意的是，二值动态引导训练策略不适用于主分支，因为全局输入的边界足够丰富，在执行式(4)时太多的边界被忽略. 此外，G 的随机性要少得多，这意味着在不同的迭代中，最长类 c 有很高的概率是相同的. 因此，在训练过程中无法很好地捕捉忽略的边界.

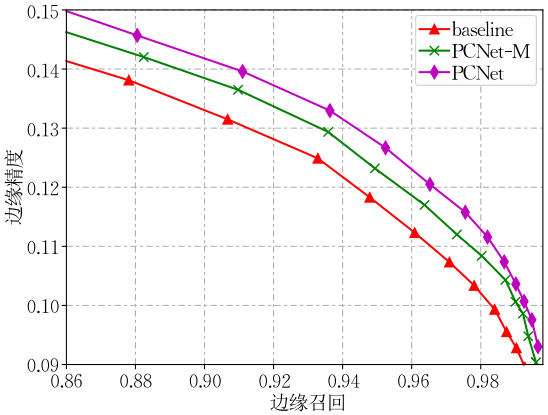


图 5　PCNet 在使用多类标签(用后缀“M”表示)和二值掩模对 DPC 分支的输出进行监督时的性能对比

在训练阶段，二值掩模 G 将替换 S_L 来监督式(3)中的局部输出 Q_L. 超分辨率也被用作 DPC 的辅助分支，以帮助还原更多的纹理. 因此，DPC 分支的完整损失为 $L=L_{SP}(Q_L,M)+L_{SR}(\hat{L},L^*)$，其中 $\hat{L}\in R^{3\times 4H\times 4W}$ 是局部图像块的重建，$L^*\in R^{3\times 4H\times 4W}$ 是 L 对应的高分辨率图像.

3.3　网络训练

我们的网络通过超像素损失、超分辨率损失和我们提出的局部判别损失进行训练. 超分辨率损失是一个遮挡的 L_1 重建损失：

$$L_{SR}(\hat{G},G^*)=\frac{1}{|B|}\left\|B\odot(G^*-\hat{G})\right\|_1 \tag{5}$$

其中 B 是用于指示边界像素的二值掩模. 为了获得 B，我们首先从真实的训练语义标签中提取边界，并通过 16×16 的核进行膨胀操作一次，以包含更多边界上下文. 超分辨率损失专注于恢复边界周围的像素，使模型能更好地感知边界上下文.

除了超像素损失和超分辨率损失，我们还设计了局部判别损失，该损失在隐藏特征层面来强调边界像素. 具体而言，令 $E\in R^{H\times W\times D}$ 表示由子像素层生成的像素嵌入特征，其中 D 是像素嵌入的特征维度. 由于在训练期间真实标签是可用的，因此我们可以从 D 中采样一个包围边界像素的小局部块 $B\in R^{K\times K\times D}$. 为了简化起见，我们只采样覆盖两个不同语义区域的图像块，即 B 是来自两个类别的特征组：$\{f_1,\cdots,f_m,g_1,\cdots,g_n\}$，其中 $f,g\in R^D$，$m+n=K^2$. 直观上，我们试图使同类别的特征更接近，而来自不同类别的像素嵌入应该相互远离. 为此，我们均匀地将同类别的特征分成两组，$\boldsymbol{f}^1,\boldsymbol{f}^2,\boldsymbol{g}^1,\boldsymbol{g}^2$，并最小化同组类内离散度，同时最大化组间离散度：

$$L_B=\frac{\left\|\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right\|_2}{\boldsymbol{S}_f^2+\boldsymbol{S}_g^2} \tag{6}$$

其中 $\boldsymbol{\mu}_f$ 和 $\boldsymbol{S}_f$ 分别是 $\{f_i\}_{i=1}^m$ 的均值特征和离散度：

$$\boldsymbol{\mu}_f=\frac{1}{|\boldsymbol{f}|}\sum_{f\in\boldsymbol{f}}f,\ \boldsymbol{S}_f=\sum_{f\in\boldsymbol{f}}|f-\boldsymbol{\mu}_f|_2^2 \tag{7}$$

考虑所有采样的图像块$\mathcal{B}$,局部判别损失被定义为

$$L_D=\frac{1}{|\mathcal{B}|}\sum_{B\in\mathcal{B}}L_B \tag{8}$$

使用此损失,边界像素可以在网络中间的隐藏特征级别进行预区分,这可以简化后续的超像素头的语义边界识别.综上,我们的完整的训练损失可以表示为

$$L=L_G+\alpha L_L+\beta L_D \tag{9}$$

其中$\alpha,\beta\in R$是权重各损失的超参数.

我们的PCNet的训练过程总结在算法1中,其中$Tensor.resize(H,W)$表示将张量调整为$H\times W$大小,而$RandPick(S)$代表根据语义标签S随机选择一个边界像素.推断时,超分辨率分支被丢弃.本文所提出的局部判别损失并非框架依赖的,可以应用在任何基于深度学习的超像素分割框架下.

算法1. PCNet训练.

输入:训练集,PCNet参数Θ,优化器(Θ),超参数α,β

输出:优化过后的网络参数Θ

```
While 未收敛 Do
  For I,S∈训练集 Do
    #输入准备
      4H×4W←原始分辨率
      G*,SG=I.resize(4H,4W), S.resize(4H,4W)
    #根据S随机挑选一个边界像素点
    #之后分别从I和S上裁剪4H×4W区域块
      b=RandPick(S);
      L*,SL=Crop(I,b),Crop(S,b)
      H×W←4H×4W
      G,L=G*.resize(H,W), L*.resize(H,W);
    #网络前向
      4H×4W←H×W
      Ĝ,QG,E=PCNet(G) #全部输入
      L̂,QL=PCNet(L) #局部输入
    #损失计算
      依据方程(3)计算 LSP(QG,SG),LSP(QL,M);
      依据方程(5)计算 LSR(Ĝ,G*),LSR(L̂,L*);
      依据方程(6)~(9)计算 LD(E);
    #整合上述损失,计算全部的损失
      L=LG+αLL+βLD
    #反向传播并且更新网络参数Θ
      L.backward()
      optimizer.step()
  End For
End While
```

和DSRL以及GLNet的比较. DSRL[42]从低分辨率输入生成高分辨率分割图,并引入超分辨率分支以帮助还原更多的结构信息.与DSRL不同,我们的超分辨率仅关注恢复边界周围的内容(式(5)).GLNet采用了调整大小的全局输入和裁剪的图像块以促进高分辨率语义分割[43],而我们的PCNet仅裁剪单个图像块进行校准.尽管存在一些细微差别,我们的机制,比如超分辨率分支和裁剪区域块作为辅助输入,确实已经在现有的工作中探索过.但是本文的贡献并不在于上述两种设计,而是动态引导掩码策略和局部判别损失.在我们的实验中,将这些现有的高分辨率分割方法直接应用于超像素分割甚至不能超过我们的基线方法,如图6所示.而我们的二值掩膜机制和局部判别损失都是根据超像素分割的特点专门设计的,更加适配超像素分割任务.

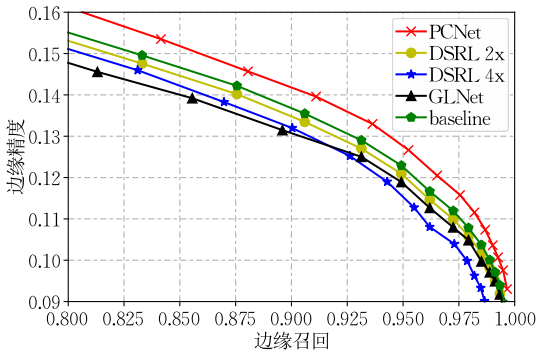


图6 在Face-Human数据集上,本文方法与高分辨率分割方法DSRL[42]和GLNet[43]的性能比较

4 实验验证与结果分析

实验部分,我们在包括三个高分辨率数据集和两个常规尺寸的常见基准数据集上进行了大量实验.我们系统地与最先进的和经典的方法进行比较,以评估我们的PCNet的性能.

4.1 数据集

为了彻底验证我们提出的方法的有效性,我们收集了一个高分辨率数据集:Face-Human.Face-Human数据集包含了共250张人脸和250张人体图像,尺寸范围从1810×2066到10 000×6675不等.所有样本都由专家仔细进行了像素级的标注,具体地来说,人脸图像手动标记为21个类别,而人体样本被分配了24个标签.图7展示了Face-Human数据集中的人脸和人体的数据实例.此外,我们也随机抽取了一部分Mapillary Vistas数据集[23]的样本,

该数据集包含 600 个样本，分辨率范围从 1024×768 到 5248×3936，用于测试我们的方法在公开基准上的性能. 其中，400 个图像用于训练，50 个和 150 个图像用作验证和测试数据.

图 7　Face-Human 数据集的两个示例样本（上面一行展示了一个人体样本，下面一行展示了一个人脸样本. 出于隐私安全考虑，我们遮挡了重要的身份识别区域）

除了以上两个数据集之外，我们还使用 BIG 数据集来评估性能. BIG 数据集包含 150 张图像，其分辨率范围从 2048×1600 到 5000×3600，图像注释遵循与 PASCAL VOC 2012[44] 相同的标注规则. 由于 BIG 数据集非常小，我们仅将其用于测试以评估训练模型的泛化性. 以上数据集包括人脸与人像（Face-Human）、街景图（Mapillary-Vistas），以及动物、物体、风景图（BIG）等，提供了足够多样的评测基准来对方法的有效性以及泛化性进行评估.

除了高分辨率数据集外，我们还在广泛使用超像素基准数据集 BSDS500[45] 和 NYUv2[46] 上进行实验，以进一步验证我们提出的方法的有效性. BSDS500 由 200 个训练、100 个验证和 200 个测试图像组成，每个图像由不同专家提供多个语义标签的注释. 为了公平比较，我们和之前的研究[8-10] 保持一致，将每个注释视为单独的样本. 因此，可以获得 1087 个训练样本、546 个验证样本和 1063 个测试样本. NYUv2 是一个室内场景理解数据集，包含 1449 张带有对象实例标签的图像. 为了评估超像素方法，Stutz 等人[41] 会移除边界附近未标记的区域，并收集一个大小为 608×448 的子集，用于超像素评估.

4.2　实现细节

在训练过程中，我们使用 128×128 的输入来预测 512×512 的关联矩阵和重构图像. 我们首先将原始高分辨率图像调整大小为 768×768，随机裁剪一个 512×512 的图像作为全局样本 G^*，而局部块 L^* 则是直接从原始图像中裁剪一个 512×512 的图像. 全局和局部样本都将下采样 4 倍以作为全局和局部输入. 编码器包含 5 个模块，除第一个模块外，每个模块都使用步幅为 2 的卷积将分辨率降低为 1/2，而解码器首先通过反卷积操作将分辨率恢复为 128×128，然后使用子像素卷积来输出一个 512×512 的特征图，这个特征图进一步输入到超分辨率和超像素预测头，用于重构图像和给出关联矩阵预测. 需要说明的是，在推断过程中超分辨率分支被丢弃. 我们在主分支的像素嵌入 E 上执行局部判别损失以强调边界像素，我们将块大小设置为 5，即 $K=5$. 超参数 α 和 β 分别设置为 0.1 和 0.5. 我们使用 Adam 优化器[47] 进行网络训练，批大小为 8，并进行 $4k$ 次迭代，验证集上最优的模型用于性能评估. 初始学习率设置为 5e-5，每 $2k$ 次迭代降低为 1/10. 进行推断时，对于测试图像，我们首先将其下采样 4 倍，然后将其馈送到网中，以产生与原始分辨率相同大小的关联矩阵. 我们遵循先前的做法[16,39]，网格大小是固定的，通过变动输入图像的分辨率来产生不同数量的超像素. 在推断过程中，像素分配给具有最高关联概率的相邻网格，因此，属于同一个网格的像素点形成了一个超像素. 我们对比下列方法，包括经典方法：SLIC[1]、LSC[21]、ERS[26]、SEEDS[9]、SNIC[2] 以及最先进的深度模型 SCN[9]. 我们使用 OpenCV 简单实现了 SLIC、LSC 和 SEEDS 方法. 对于其他方法，我们使用作者的官方实现. 至于另一个优秀的方法 SSN[16]，由于在我们的实践中只能处理 1K 分辨率的图像，因此在高分辨率比较中，我们不包含该方法.

4.3　定量比较

对于超像素分割而言，准确识别边界非常重要，因此我们使用边界召回率（Boundary Recall，BR）和边界精度（Boundary Precision，BP）来评估性能[41]. 为了全面地评估模型，我们在 Mapillary-Vistas 或 Face-Human 数据集上训练网络，并在所有三个基准测试上评估模型的能力和泛化性，特别是深度模型 SCN[9] 和我们提出的 PCNet.

图 8 展示了三个基准测试上 BR-BP 曲线的性能比较，其中图 8(a)显示在 Face-Human 上训练并在三个测试集上评估的所有模型的性能比较，而图 8(b)类似地展示了在 Mapillary-Vistas 数据集上的比较. 借助于可微卷积神经网络的优势，深度学习方法 SCN 和 PCNet 可以取得比传统方法 SLIC[25]、SNIC[27]、SEEDS[48] 和 ERS[24] 更好的性能. 与最先进的方法 SCN 相比，我们的 PCNet 在 Face-Human 上性能略好，在 Mapillary-Vistas 数据集上相当. 关于泛化性，从图 8(a)(3)和图 8(b)(3)中可以看到，

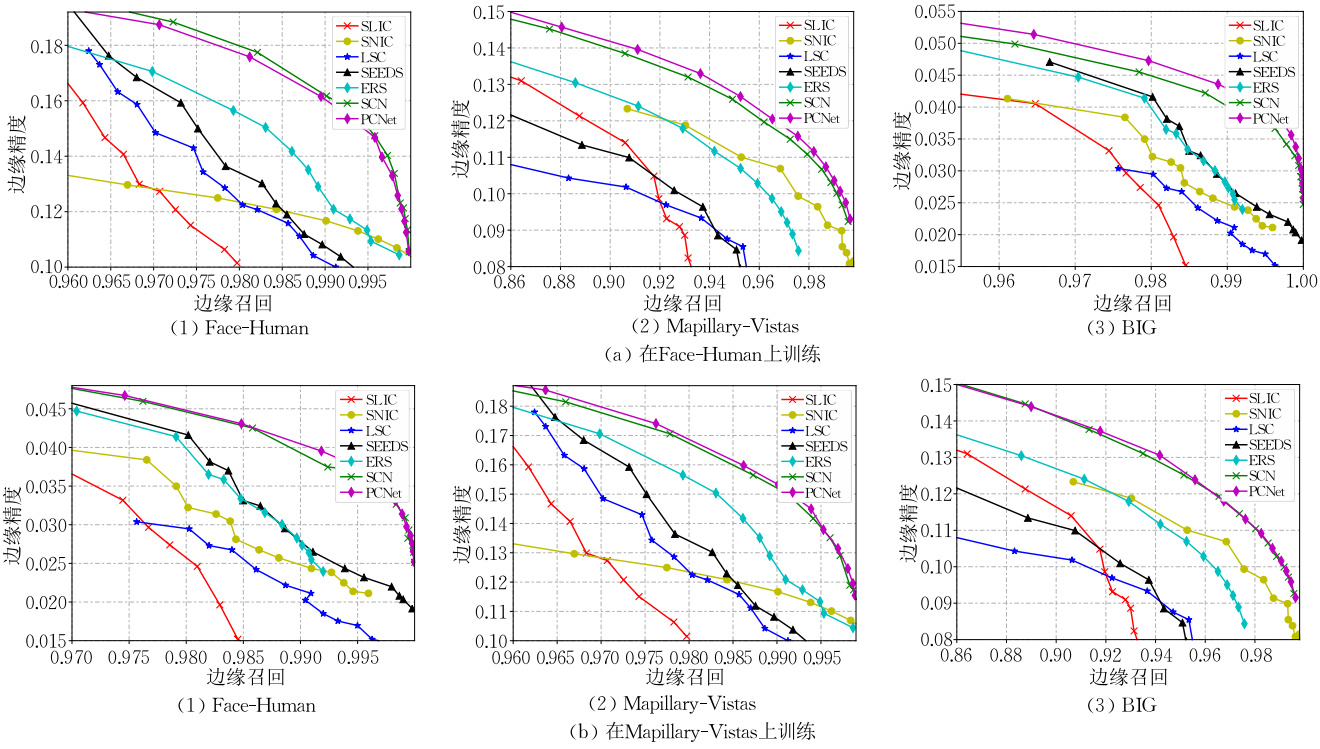


图 8 三个高分辨率基准上的性能比较(其中(a)从左到右展示了在 Face-Human、Mapillary-Vistas 和 BIG 数据集上的性能对比,模型只在 Face-Human 上训练;(b)则类似地展示了在 Mapillary-Vistas 数据集上训练的性能比较)

我们的方法在 BIG 数据集上的表现更好. 当推广到 Face-Human 或 Mapillary-Vistas 时,PCNet 和 SCN 相当. 除了高分辨率数据集外,我们还在一个常用的样本大小正常的数据集 BSDS500[45] 和 NYUv2[46] 上进行了实验. 我们沿用 Yang[9] 和 Jampani[10] 的做法,在 BSDS500 上训练模型,并在 BSDS500 和 NYUv2 数据集上进行测试,以评估性能和算法的泛化性. 结果报告在图 9 中,我们可以看到 PCNet 在 BSDS500 上的性能仍然与模型 SCN 相当,但比 SSN 差. 我们认为这是由于低分辨率图像块会包含更多的边界,这导致在生成二值引导掩码时会忽略更多边界. 因此,我们的块校准机制的优势不能充分体现. 当推广到 NYUv2 数据集时,我们的模型可以比 SCN 和 SSN 方法略好,这验证了我们的 PCNet 的有效性. 此外,由于输入分辨率较低,我们的 PCNet 还可以实现更好的推断效率. 从图 8、图 9 中可以看出,我们的 PCNet 可以使用测试图像的 1/4 分辨率即可实现略好或与 SCN 相媲美的性能. 较低的分辨率输入不仅允许网络处理非常高分辨率的 5K 图像,而且由于前向传播过程中较少的 FLOPs,还可以显著加速推断,正如图 1 所示. 图 10 可视化了五个数据集的超像素分割结果,包括三个高分辨率数

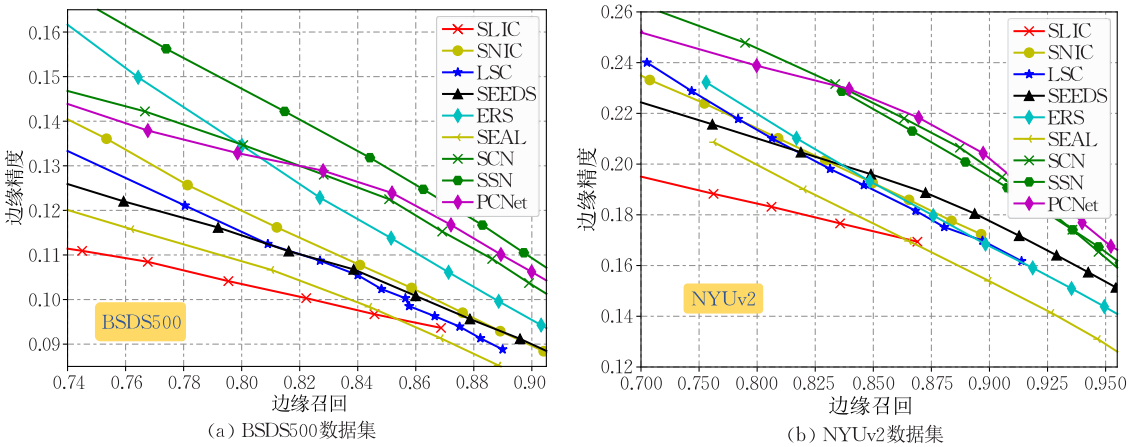
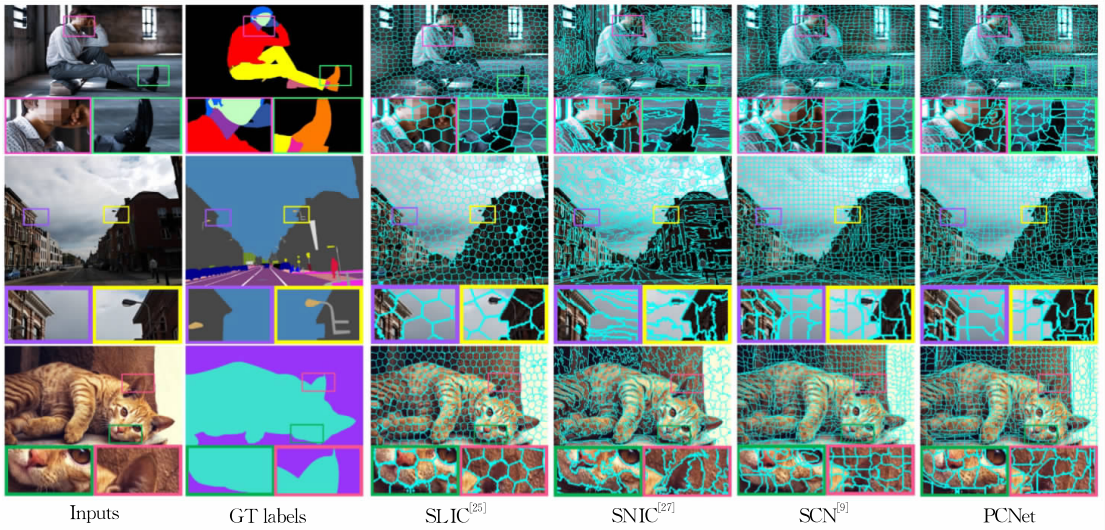
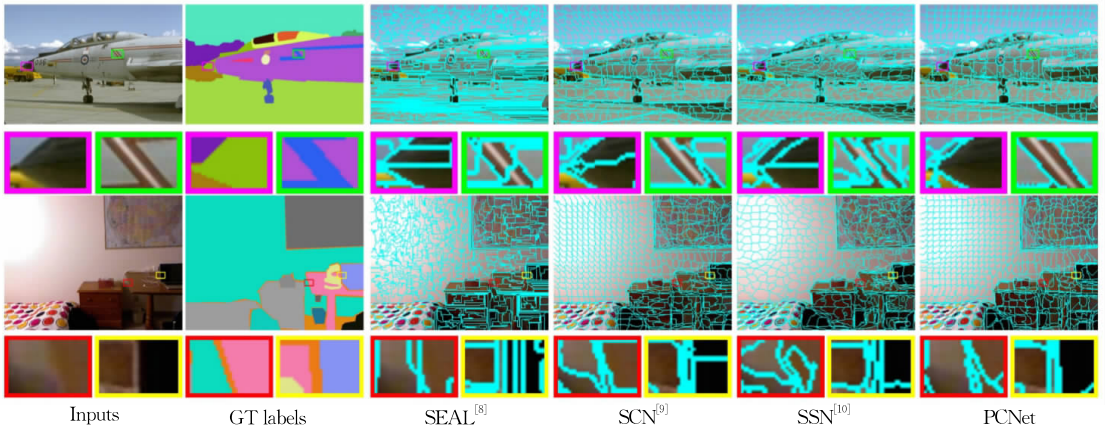


图 9 BSDS500 and NYUv2 数据集上的性能对比



(a) 在三个高分辨率数据集上的超像素分割结果(从上至下依次展示了Face-Human、Mapillary-Vistas和BIG数据集上的结果)



(b) 两个常规数据集上的超像素分割结果(第一行和第二行分别展示了BSDS500、NYUv2上的结果)

图 10　超像素分割结果展示(图中(a)展示了高分辨率的结果,包括 Face-Human、Mapillary-Vistas 和 BIG 数据集上的超像素分割结果;(b)展示了在常规数据集上的结果,第一行和第二行分别展示了 BSDS500 和 NYUv2 上的结果)

据集和两个常用的普通数据集. 图 10(a)从上到下依次显示了来自 Face-Human、Mapillary-Vistas 和 BIG 数据集的三个结果,而图 10(b)展示了来自两个普通数据集 BSDS500 和 NYUv2 的结果. 从这些图中,我们可以观察到 PCNet 可以更准确地识别语义边界,这直观地展示了我们提出的方法的优越性. 图 11对比了 PCNet、SCN 以及 SSN 在处理各个分辨率图片时的时间耗费情况. 从该图可以看到,我们的 PCNet 模型可以以更小的时间代价处理更高分辨率的图片.

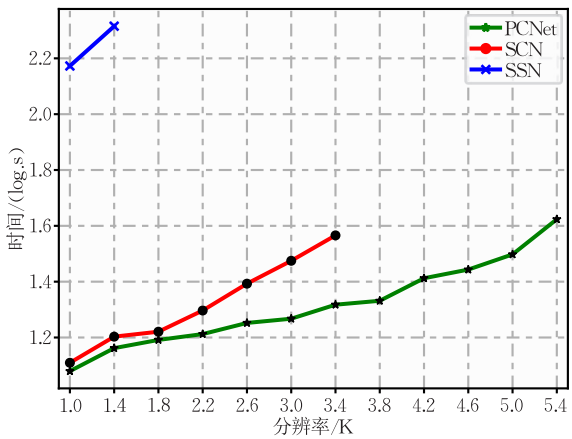


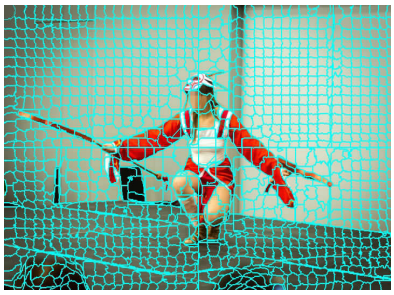
图 11　SSN、SCN 以及 PCNet 可处理的分辨率以及时间对比

4.4　消融实验

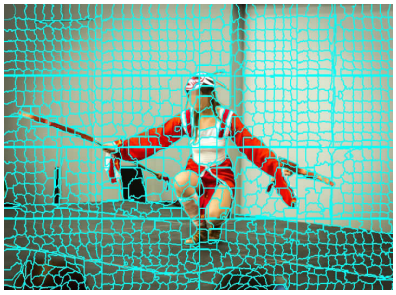
第 4.4 节中,我们首先讨论分而治之策略,以更清楚地阐明我们框架的动机,然后通过一系列实验证实 PCNet 中每个组件的贡献.

(1)分而治之策略. 由于超像素生成是一种图像过度分割的表示,因此我们可以通过分治策略为高分辨率图像生成超像素. 例如,我们可以将输入图像划分为四个不重叠的子图像,并为每个子图像运行超像素生成,最后拼接起来形成高分辨率的超像素结果. 直观地说,分治策略是高分辨率超像素分割的直接选

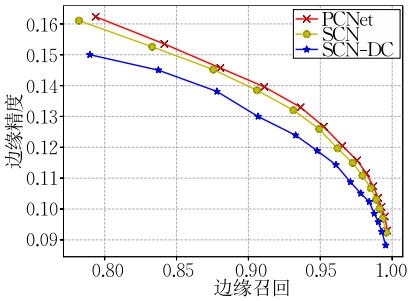
择.实际上,在我们的初始探索中,我们尝试基于 SCN 模型实验分治解决方案,但结果并不理想.这种策略的第一个弱点是在不同图像块之间会有明显的边界,由于块之间边界的存在,阻止了超像素的合并.如果在划分过程中拆分了一个语义紧凑的区域,那么就不可能为处于不同块中的像素形成一个超像素,如图 12(a)~(b)所示.此外,分治策略的性能也不理想.如图 12(c)所示,虽然边界召回率可接受,但精度太低.SCN-DC 的表现也不如 SCN,我们认为这是因为分割图像的块无法提供相似的宽视角的上下文.因此,我们放弃了这种朴素的策略,并提出了我们的 PCNet,以端到端的方式生成超像素,并取得了更好的性能.



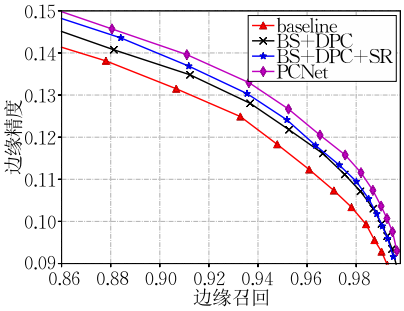
(a) 2×2划分



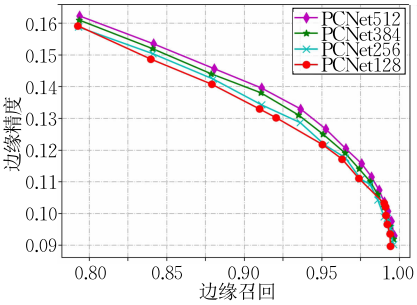
(b) 4×4划分



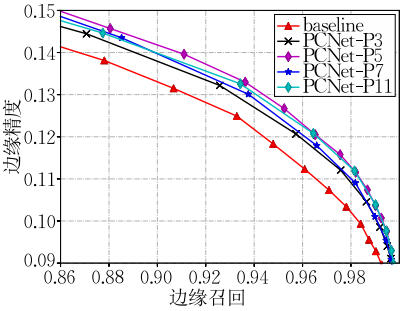
(c) 分而治之的性能



(d) 各部分的消融实验



(e) DPC分支中的区域大小讨论



(f) LD损失中的区域大小讨论

	Input	Output	Conv-3×3	BN	ReLU	In Total
SCN	$4H\times4WC_1$	$4H\times WC_2$	$16\times9\times HWC_1C_2$	$16HWC_2$	$16HWC_2$	$144HWC_1C_2+32HWC_2$
PCNet	HWC_1	HWC_2	$9HWC_1C_2$	HWC_2	HWC_2	$9HWC_1C_2+2HWC_2$

(g) 计算复杂度分析

图 12 消融和讨论实验(图中(a)和(b)展示了使用分而治之策略的 SCN 的超像素分割结果;(c)是对应的性能比较;(d)验证了我们方法中每个组件的贡献;(e)和(f)分别讨论 DPC 分支和 LD 损失中的图像块大小对性能的影响;(g)分析了基本计算模块的时间复杂度)

(2)各模块的贡献.为了验证 PCNet 中每个模块的贡献,我们在 Face-Human 数据集上进行了消融实验,以验证它们各自的贡献,包括 DPC 分支、超分辨率分支和局部判别损失.结果报告在图 12(d)中,基线(BS)方法是图 2(b)所示的基于 SCN 由低到高分辨率预测的朴素策略,仅使用全局图像,并使用损失函(式(3))训练网络.正如图 12(d)所示,基线方法表现非常差,因为非常低的分辨率输入牺牲了太多的纹理,特别是语义边界上下文.通过我们的 DPC 分支,可感知更详细的边界信息,进而提高了性能.超分辨率分支进一步使网络能够从低分辨率输入中恢复更多细节.从图 12(d)可以看到,当配备超分辨率分支时,性能进一步提升.图 5 还表明,我们的动态引导训练比朴素选择,即基于真实的多类标签表现更佳,这验证了我们动态引导训练机制的有效性.当进一步应用局部判别损失时,我们的方法可以获得最佳结果.

(3)图像块尺度的影响.我们的 DPC 分支和 LD 损失的设计中都包含了图像块,因此,在本小节中,我们进行实验来研究它们对于最终性能的影响.在训练过程中,我们的 DPC 分支以 128×128 像素图像作为输入,输出 512×512 像素的预测结果.在这部分中,我们还尝试了其他的图像块尺寸的选择,如 384、256、128(输出尺寸),在 Face-Human 数据集上的结果如图 12(e)所示.我们发现,图像块大小为 512 表现的最好,这是因为更大的图像块可以保留更多的细节.因此,在训练中,我们使用 512 大小的图像块送入 DPC 分支以获得更好的性能.

在我们的 LD 损失中,我们在边界像素周围采样 5×5 大小的图像块.在这个小节中,我们将图像块尺寸从 3 变化到 11,以探究其带来的性能差异.结果报告在图 12(f)中,其中各 PCNet-P 表示不同图像块尺寸的 PCNet.从图 12(f)可以看出,较小的图像块尺寸对性能提升的较小.当图像块尺寸增加

到 5 时，性能得到了提高. 然而，进一步增大图像块尺寸到 7 或 11 并不能使性能进一步提升，反而比 PCNet-P5 的表现略差. 我们认为原因是过大的图像块尺寸会引入更多不够接近边界的像素，导致 LD 损失无法很好地集中注意力于边界上下文. 因此，在我们的 PCNet 中，我们将图像块尺寸设置为 5.

(4) 计算复杂度. 在我们的 PCNet 中，复杂度降低的主要原因是较低分辨率的输入. 如图 2 所示为了获得一个 $4H\times4W$ 的预测结果，我们的模型仅需 $H\times W$ 大小的输入，与需要 $4H\times4W$ 大小输入的 SCN 模型相比显著降低了计算成本. 我们的网络的基本块由卷积批量归一化-ReLU 组成，我们分析了基本块的计算复杂度并在图 12(g)中进行了报告，其中 H,W 是输入/输出图像的高度和宽度，$C1$ 和 $C2$ 分别表示输入和输出通道数. 对于一个基本块，我们的模型的计算复杂度为 $9HWC_1C_2+2HWC_2$，比 SCN 的 $144HWC_1C_2+32HWC_2$ 快了 $135HWC_1C_2+30HWC_2$.

5　结论与展望

这项工作提出了一个基于深度学习的框架来进行高分辨率超像素分割. 所提出的 PCNet 可以以更快的速度处理更高分辨率的图像. 为了弥补下采样输入中失去的边界细节，我们设计了一个解耦的区域块校准分支来校正全局预测的边界像素. 我们提出了一个二值引导掩码，强制 DPC 分支专注于感知图像中的语义边界. 为了准确地识别更多的边界像素，我们提出了一个局部判别损失，以区分边界周围的像素嵌入. 我们还构建了一个超高分辨率基准 Face-Human 来评估超高分辨率的超像素分割. 我们在四个公共基准和我们收集的 Face-Human 数据集上进行了大量实验，评估其性能. 我们提出的 PCNet 可以高效地处理高分辨率的 5K 图像，同时保持与最先进的 SCN 相当的性能.

未来，我们将沿着三个方向继续探索超像素任务. 第一个值得研究的主题是如何有效地将超像素框架与分割和立体匹配等下游任务相结合. 作为基本的计算机视觉任务，如何利用超像素使更多的视觉任务受益是一个有前途的方向. 其次，我们将致力于面向语义区域感知的超像素分割. 虽然网格分割策略可以实现端到端的深度超像素分割网络的训练，但这种机制也阻碍了超像素合并从而形成更大的有意义的区域. 我们相信克服这个问题将推动超像素分割的研究. 第三，探索高分辨率超像素分割如何赋能高分辨率图像分割. 通过合理的训练集合选择或者使用 one-shot 的训练模式等策略，进一步挖掘和利用超像素分割，以更好地提升高分辨率图像的分割精度，也是一个有前景的研究方向.

参　考　文　献

[1] SuperAnnotate. https://www.superannotate.com/

[2] Chen Zixuan, Zhou Huajun, Lai Jianhuang, et al. Contour-aware loss: Boundary-aware learning for salient object segmentation. IEEE Transactions on Image Process, 2021, 30: 431-443

[3] Dong Xiaoyi, Han Jiangfan, Chen Dongdong, et al. Robust superpixel-guided attentional adversarial attack//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12892-12901

[4] Wang Hui, Shen Jianbing, Yin Junbo, et al. Adaptive nonlocal random walks for image superpixel segmentation. IEEE Transactions on Circuits System Video Technology, 2020, 30(3): 822-834

[5] Sun Wen, Liao Qingmin, Xue Jing-Hao, Zhou Fei. SPSIM: A superpixel-based similarity index for full-reference image quality assessment. IEEE Transactions on Image Processing, 2018, 27(9): 4232-4244

[6] Gaur U, Manjunath B S. Superpixel embedding network. IEEE Transactions on Image Processing, 2020, 29: 3199-3212

[7] Zhu Li-Yan, Luo Xiang-Yang, Zhang Yi, et al. Asymmetric distortion steganography method based on superpixel filtering. Chinese Journal of Computers, 2023, 46(7): 1473-1493(in Chinese)
(朱利妍，罗向阳，张祎等. 基于超像素滤波的非对称失真隐写方法. 计算机学报，2023，46(7)：1473-1493)

[8] Tu Wei-Chih, Liu Ming-Yu, Jampani V, et al. Learning superpixels with segmentation-aware affinity loss//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 568-576

[9] Yang Fengting, Sun Qian, Jin Hailin, Zhou Zihan. Superpixel segmentation with fully convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13961-13970

[10] Jampani V, Sun Deqing, Liu Ming-Yu, et al. Superpixel sampling networks//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 363-380

[11] Cheng Ho Kei, Chung Jihoon, Tai Yu-Wing, Tang Chi-Keung. CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8887-8896

[12] Zhou Peng, Price B L, Cohen S, et al. Deepstrip: High-resolution boundary refinement//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10555-10564

[13] Lin Shanchuan, Ryabtsev A, Sengupta S, et al. Real-time high-resolution background matting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8762-8771

[14] Wang Jingdong, Sun Ke, Cheng Tianheng, et al. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,

2021, 43(10): 3349-3364
[15] Qi Jiyang, Gao Yan, Hu Yao, et al. Occluded video instance segmentation: A benchmark. International Journal of Computation Vision, 2022, 130(8): 2022-2039
[16] Bar-Haim A, Wolf L. ScopeFlow: Dynamic scene scoping for optical flow//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7995-8004
[17] Teed Z, Deng Jia. RAFT: Recurrent all-pairs field transforms for optical flow//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 402-419
[18] Li Ruoteng, Tan R T, Cheong L F, et al. RainFlow: Optical flow under rain streaks and rain veiling effect//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 7303-7312
[19] Liu Pengpeng, King I, Lyu M R, Xu Jia. Flow2Stereo: Effective self-supervised learning of optical flow and stereo matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6647-6656
[20] Badki A, Troccoli A J, Kim K, et al. Bi3D: Stereo depth estimation via binary classifications//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1597-1605
[21] Garg R, Wadhwa N, Ansari S, Barron J T. Learning single camera depth estimation using dual-pixels//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Seoul, Republic of Korea, 2019: 7627-7636
[22] Zhang Haokui, Li Ying, Cao Yuanzhouhan, et al. Exploiting temporal consistency for real-time video depth estimation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 1725-1734
[23] Neuhold G, Ollmann T, Buló S R, Kontschieder P. The Mapillary Vistas dataset for semantic understanding of street scenes//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Republic of Korea, 2017: 5000-5009
[24] Liu Ming-Yu, Tuzel O, Ramalingam S, Chellappa R. Entropy rate superpixel segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 2097-2104
[25] Achanta R, Shaji A, Smith K, et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282
[26] Liu Yong-Jin, Yu Cheng-Chi, Yu Minjing, He Ying. Manifold SLIC: A fast method to compute content-sensitive superpixels //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 651-659
[27] Achanta R, Süsstrunk S. Superpixels and polygons using simple non-iterative clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4895-4904
[28] Li Hua, Liang Junyan, Wu Ruiqi, et al. Stereo superpixel segmentation via decoupled dynamic spatial-embedding fusion network. IEEE Transactions on Multimedia, 2024, 26: 367-378
[29] Pan Xiao, Zhou Yuanfeng, Chen Zhonggui, Zhang Caiming. Texture relative superpixel generation with adaptive parameters. IEEE Transactions on Multimedia, 2019, 21(8): 1997-2011
[30] Wang Yufeng, Ding Wenrui, Zhang Baochang, et al. Superpixel labeling priors and MRF for aerial video segmentation. IEEE Transactions on Circuits System Video Technology, 2020, 30(8): 2590-2603
[31] Shi Cheng, Pun Chi-Man. Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders. IEEE Transactions on Multimedia, 2020, 22(2): 487-501
[32] Li Hua, Kwong S, Chen Chuanbo, et al. Superpixel segmentation based on square-wise asymmetric partition and structural approximation. IEEE Transactions on Multimedia, 2019, 21(10): 2625-2637
[33] Liu Jiaying, Yang Wenhan, Sun Xiaoyan, Zeng Wenjun. Photo stylistic brush: Robust style transfer via superpixel-based bipartite graph. IEEE Transactions on Multimedia, 2018, 20(7): 1724-1737
[34] Zhu Lei, She Qi, Zhang Bin, et al. Learning the superpixel in a non-iterative and lifelong manner//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1225-1234
[35] Peng Hankui, Avilés-Rivero A I, Schönlieb C B. HERS superpixels: Deep affinity learning for hierarchical entropy rate segmentation//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 72-81
[36] Zhang Zhi-Long, Li Ai-Hua, Li Chu-Wei. Algorithm of superpixel segmentation based on density peak search clustering. Chinese Journal of Computers, 2020, 43(1): 1-15(in Chinese)
(张志龙，李爱华，李楚为. 基于密度峰值搜索聚类的超像素分割算法. 计算机学报，2020, 43(1): 1-15)
[37] Li Zhengqin, Chen Jiansheng. Superpixel segmentation using linear spectral clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1356-1363
[38] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation. International Journal of Computer Vision, 2004, 59(2): 167-181
[39] Wang Yaxiong, Wei Yunchao, Qian Xueming, et al. AINet: Association implantation for superpixel segmentation//Proceedings of the IEEE/CVF International Conference of Computer Vision. Montreal, Canada, 2021: 7058-7067
[40] Shi Wenzhe, Caballero J, Huszar F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1874-1883
[41] Stutz D, Hermans A, Leibe B. Superpixels: An evaluation of the state-of-the-art. Computer Vision and Image Understanding, 2018, 166: 1-27
[42] Wang Li, Li Dong, Zhu Yousong, et al. Dual super-resolution learning for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3773-3782
[43] Chen Wuyang, Jiang Ziyu, Wang Zhangyang, et al. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Long Beach, USA, 2019: 8924-8933

[44] Everingham M, Ali Eslami S M, Van Gool L, et al. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1): 98-136

[45] Arbelaez P, Maire M, Fowlkes C C, Malik J. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 898-916

[46] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images//Proceedings of the European Conference on Computer Vision (ECCV). Florence, Italy, 2012: 746-760

[47] Kingma D P, Ba J. Adam: A method for stochastic optimization //Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015

[48] Van den Bergh M, Boix X, Roig G, Van Gool L. SEEDS: Superpixels extracted via energy-driven sampling. International Journal of Computer Vision, 2015, 111(3): 298-314

WANG Ya-Xiong, Ph. D., associate professor. His research interests include image understanding and cross-modal computation.

WEI Yun-Chao, Ph. D., professor. His research interests include semi-supervised learning and semantic image understanding.

QIAN Xue-Ming, Ph. D., professor. His research interests lie in the content understanding of cross-modal data.

ZHU Li, Ph. D., professor. His research interests include image processing and computer network.

Background

Superpixel segmentation is an image processing technique that segments an image into multiple homogeneous regions, known as superpixels. These regions are visually similar and have clear boundaries, making them larger and more meaningful than individual pixels. Each superpixel is a group of pixels that share similar color, texture, and brightness, which helps in reducing the complexity of the image while retaining essential visual information. This method is particularly useful for various computer vision tasks such as image segmentation, object recognition, image compression, and 3D reconstruction.

The key characteristics of superpixel segmentation include:

Homogeneity: Pixels within a superpixel are similar in color and texture.

Clear Boundaries: The boundaries of superpixels often align with significant edges in the image.

Regular Shape: Superpixels tend to form regular shapes, approximating circles or ellipses.

There are two main approaches to superpixel segmentation:

Traditional Methods: These are typically based on clustering or graph-cut algorithms and include methods like Simple Linear Iterative Clustering (SLIC) and Superpixels Extracted via Energy-Driven Sampling (SEEDS).

Deep Learning-Based Methods: These leverage Convolutional Neural Networks (CNNs) to learn representations of the image for segmentation. They can handle more complex image structures and learn from vast amounts of data.

When it comes to high-resolution image superpixel segmentation, the challenge lies in processing the vast amount of data present in high-resolution images. Traditional methods may become computationally expensive or slow when applied to such images. However, with the advent of deep learning, it has become feasible to segment high-resolution images efficiently.

Deep learning models can be designed to handle high-resolution images by using strategies such as:

Patch-Based Processing: Breaking down the high-resolution image into smaller patches that fit within the memory constraints of the system.

Pyramidal Approaches: Utilizing a multi-scale representation of the image, where the segmentation is performed at multiple levels of resolution.

Efficient Network Architectures: Designing networks that are computationally efficient and can process high-resolution images without significant loss of detail.

High-resolution superpixel segmentation is particularly important in applications that require fine-grained analysis, such as medical imaging, satellite imagery, and high-definition video processing. By effectively segmenting these images into superpixels, the subsequent analysis can be made more robust and computationally efficient.

In this work, we introduce a deep learning-based framework for high-resolution superpixel segmentation, termed PCNet, which is capable of processing higher resolution images at a faster pace. To address the loss of boundary details due to downsampling in the input, a decoupled Patch Calibration branch (DPC) was designed to correct the boundary pixels in global predictions. A binary guidance mask is introduced to enforce the DPC branch to focus on perceiving semantic boundaries within the image. To accurately identify a greater number of boundary pixels, a local discrimination loss is proposed to differentiate the pixel embeddings surrounding the boundaries. Additionally, an ultra-high-resolution benchmark dataset, Face-Human, has been constructed to evaluate the performance of superpixel segmentation at extremely high resolutions.