**HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation**

基础

1. 超网络：为其他网络生成权重的网络
2. 分组卷积和DW卷积
3. PW卷积

摘要

我们提出了一个新颖的、实时的、语义分割网络，其中编码器既编码又生成解码器的参数（权重）。此外，为了允许最大的适应性，每个解码器块的权重在空间上都有变化。为此，我们设计了一种新型的超网络，由一个嵌套的U型网组成，用于绘制更高层次的上下文特征，一个多头的权重生成模块，在解码器中的每个块的权重被消耗之前立即生成，以有效地利用内存，还有一个主网络，由新颖的动态补丁式卷积组成。尽管使用了不太传统的块，我们的架构仍获得了实时性能。就运行时间与准确性的权衡而言，我们在流行的语义分割基准上超过了最新技术水平（SotA）的结果。PASCAL VOC 2012（估值集）以及Cityscapes和CamVid的实时语义分割。代码可供查阅：<https://nirkin.com/hyperseg>。

github链接：<https://github.com/YuvalNirkin/hyperseg>

1.引言

语义分割在场景理解中起着至关重要的作用，无论场景是微观的、望远镜的、由移动车辆拍摄的，还是通过AR设备观看的。新的移动应用不仅寻求准确的语义分割，而且还需要实时处理，这刺激了对实时语义分割的研究。此后，这一领域已成为新架构和训练方法的领先测试平台，其目标是提高准确性和速度。最近的工作增加了容量[5，6]和注意力机制[20，45，49]来提高性能。当运行时间不是问题时，图像往往被模型多次处理，结果被累积。在本文中，我们试图以一种不同的方式来提高性能：通过为网络提供额外的适应性。我们使用元学习技术来增加这种适应性，通常被称为动态网络或超网络[13]。这些网络被用于从文本分析[13, 50]到三维建模[26, 42]的任务，但很少用于生成类似图像的地图。原因是，正如以前的方法所建议的那样，超网络不能完全捕捉高分辨率图像的信号。语义分割图是一个特别有趣的例子。它们是由一个从粗到细的金字塔产生的，其中每一级的过程都可以从适应中受益，因为这些效果从一个区块积累到下一个区块。此外，由于图像的每一部分都可能包含不同的对象，这种适应最好是在本地完成。因此，我们提供了一种新的编码器-解码器方法，其中编码器的主干是基于该领域的最新进展。编码信号通过内部U-Net映射到动态网络权重，而解码器则由具有空间变化权重的动态块组成。所提出的架构在该任务最广泛使用的基准上实现了SotA精度与运行时间的权衡。PASCAL VOC 2012 [11], CityScapes [8], 和CamVid [2]。对于CityScapes和CamVid来说，SotA的精度结果是在实时条件下获得的。尽管我们使用了一个非常规的架构，即采用具有动态权重的局部连接层，但我们的方法非常有效（相对于其他方法，我们的运行时间/精度权衡见图1）。 总之，我们的贡献是  
- 一种新的超网络结构，在U-Net中采用UNet。  
- 新的动态补丁式卷积，其权重在每个输入和每个空间位置上都有变化。  
- 在该领域的主要基准上对SotA准确性与运行时间进行了权衡。

2.相关工作

超网络

超网络[13]，是指为其他网络（通常被称为主网络）生成权重值的网络。超网络作为一种建模工具是非常有用的，例如，作为图像到图像翻译的隐含函数[9，24]，三维场景表示[26，42]，也可以在神经结构搜索（NAS）[55]和持续学习[48]过程中避免计算和数据的沉重训练周期。然而，据我们所知，超网络从未被提议用于语义分割，正如我们在这里提议的那样。局部连接层。

局部连接层的连接性遵循一种空间模式，类似于传统的卷积层，但没有权重共享。这种层在早期的深度学习中发挥了重要作用，主要是由于计算方面的原因[10, 36, 47]。在人脸识别的背景下，局部连接层被引入作为提高准确率的组件，这些组件的动机是需要以不同的方式对人脸的每个部分进行建模[43]。然而，随后的人脸识别方法倾向于使用传统的卷积，例如，[41]。部分共享权重，即在图像斑块内共享卷积，被提议用于分析面部动作[58]。就我们所知，我们是第一个提出将局部连接层与超网络结合起来，或在语义分割的背景下，或更广泛地在图像到图像的映射中提出的。

语义分割

早期的语义分割方法使用了特征工程，并经常依赖于数据驱动的方法[15, 16, 17, 46]。据我们所知，Long等人[28]是第一个展示用于语义分割的卷积神经网络（CNN）的端到端训练。他们的完全卷积网络（FCN）基于分类网络主干，输出密集的、每像素的可变分辨率预测。他们在早期和最后一层之间加入了跳过连接，以结合粗略和精细信息。随后的方法增加了一个基于条件随机场（CRF）的后处理步骤，以进一步细化分割掩码[3, 4, 59]。Nirkin等人通过利用视频中的运动克服了有限的、稀缺的分割标签[31]。U-Nets[38]使用编码器-解码器对，将编码器的最后一个特征图与解码器的相应的上采样特征图串联起来，在每个stride中。一些人提议用扩张卷积（又称曲折卷积）来代替串联卷积[4, 54]。这种方法通过扩大对数的接受域产生了更详细的分割，但也极大地增加了计算成本。另一种扩大感受野的方法被称为空间金字塔集合（SPP）[18, 57]，其中来自不同步幅的特征被平均集合并串联在一起，之后信息被后续卷积层所融合。随后的工作将无序卷积与SPP（ASPP）相结合，实现了更高的精度，但计算成本更高[4, 5, 6]。为了进一步提高准确性，一些人提出了在输入图像的多尺度和水平翻转的版本上多次应用网络的推理策略，并使用平均池将结果结合起来[5，6]。最近，Tao等人[45]利用注意力来更好地结合推理策略的预测，把尺度考虑在内。最后，其他人提出了轴向注意，沿着高度和宽度轴分别执行注意，以更好地模拟长距离依赖 [20, 49] 。

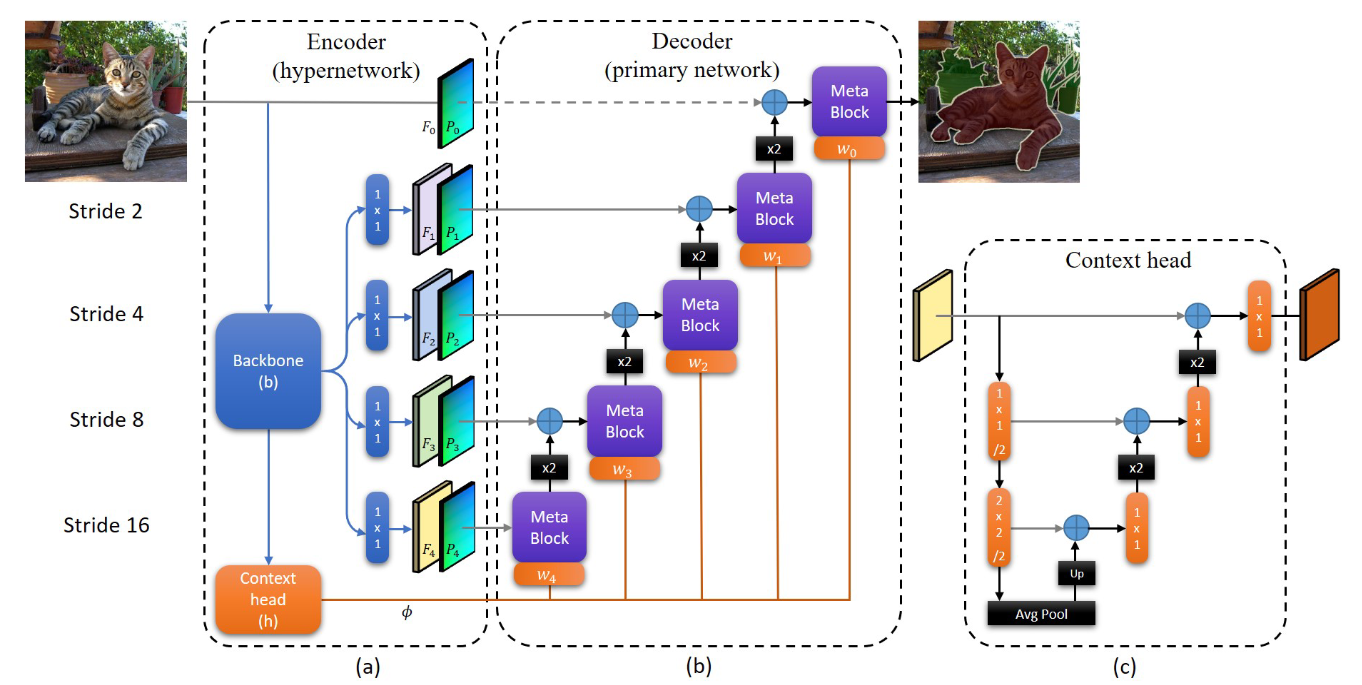
实时分割。

一些方法的目标是在准确性和计算量之间实现最佳权衡，重点是保持实时性能。实时方法通常采用由基于高效主干的编码器和相对较小的解码器组成的架构。廉价语义分割的一个早期例子是SegNet[1]，它使用一个带有跳过连接和转置卷积的编码器-解码器架构来进行上采样。ENet[34]提出了一个基于ResNet的瓶颈块[19]的架构，实现了高帧/秒（FPS）率，但牺牲了相当大的精度。ICNet[56]利用从图像金字塔中提取的融合特征，报告了比以前的方法更好的准确性。GUNNet[29]在编码器的融合特征图的指导下，对从多尺度输入图像中提取的特征进行了上采样。SwiftNet[32]提出了一个带有SPP和1×1卷积的编码器-解码器，用于减少每个跳过连接前的维数。随后的方法受益于高效网络结构的进展[25, 51]，如深度可分离的卷积[7, 21]和倒置的残差块[40]，我们在工作中也使用了这些方法。BiSeNet[52]提出了一个额外的、较粗的下采样路径，在上采样之前与较细的分辨率主网络融合。BiSeNetV2[51]对BiSeNet进行了扩展，对两个分支进行了更精细的融合，并从中间层增加了预测头以提高训练效果。最后，TDNet[22]提出了一个用于视频语义分割的网络，通过在连续的帧上循环分布子网络，利用时间的连续性。

3.

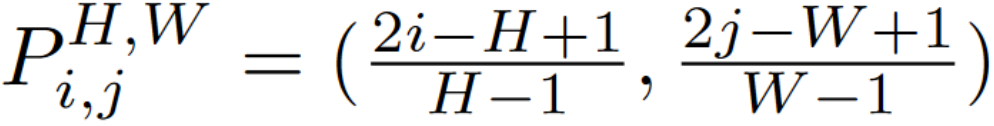
概述。

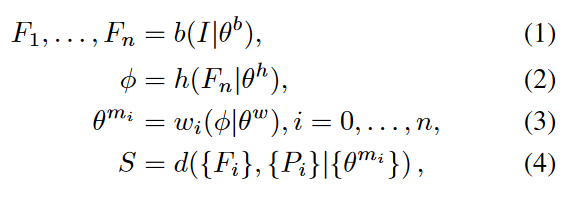
我们提出的超网络编码器-解码器方法如图2和图3所示。与基于U-Net的方法[38]类似，我们在编码器和解码器的相应层之间采用跳过连接。然而，我们的网络使用编码器和后续块，我们称之为上下文头和权重映射器，符合超网络设计的精神。因此，跳过的连接将不同的编码器级别与作为我们的解码器的分层主网络的级别连接起来。此外，我们的解码器的权重在每个阶层的块之间是不同的。我们提出的模型包括三个子网络：骨干网络b（在图2(a)中以蓝色显示），上下文头h（图2(a)中的橙色方框，在图2(c)中也有详细说明），以及作为解码器的主网络d（图2(b)）。此外，解码器由多个元块组成，详见图3（a）。每个元块，i=0 ... n，包括一个额外的权重映射网络组件wi，在图2（b）中表示为橙色方框。



信息流。

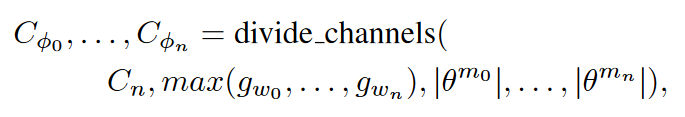
三个网络的权重：θb、θh和θw，在推理过程中是固定的，并在训练过程中学习，而θmi，即解码器元块的权重，mi，在推理时是动态预测的。编码器的主干b将输入图像I∈R3×H×W映射到一组不同分辨率的特征图Fi∈RCi×H/2i×W/2i，i∈[1，5]，其中H和W是沿图像高度和宽度的相应像素数。上下文头h（大小不变）：RCn×H/2n×W/2n→RCn×H/2n×W/2n，将b的最后一个特征图映射为一个信号φ，然后将这个信号输入w。RCn×H/2n×W/2n→R(∑i |θmi |)×H/2n×W/2n，生成主网络元块的权重d。注意，这些权重在不同的空间位置是不同的。我们定义一个固定的位置编码，P H,W∈R2×H×W，这样在每个位置（i，j），



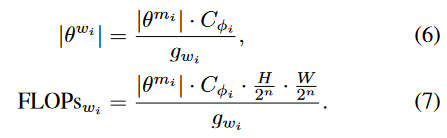
，i∈[0，H]，j∈[0，W）。最后，给定输入图像和特征图，F1, . . . ，Fn，它们对应的相同分辨率的位置编码，P0， . . Pn，以及权重θd，解码器d输出分割预测，S∈RC×H×W，其中C是语义分割任务中的类别数量。因此，我们的整个网络是由以下一组方程定义的。  


其中每个网络的权重在分隔符后明确指定。

3.1. 编码器和超网络  
超网络的第一个组成部分是主干网络，b（图2(a)中的蓝框）。它是基于EfficientNet[44]系列的模型。第4节提供了该网络的细节。在我们的工作中，骨干网络架构的头部被我们的上下文头部h所取代。骨干网络在每个步长中输出特征图Fi。为了减少解码器的大小，我们用额外的1×1卷积来降维，使每个Fi的通道数减少一个系数ri。ri的具体数值详见补充材料中的B节。b的最后一张特征图，即骨干网络，大小为H/2n × W/2n。这个特征图中的每个像素都编码了输入图像中的一个斑块。这些斑块几乎没有重叠，如果是跨越多个斑块的大型物体，有限的感受野会导致结果不佳。因此，上下文头h，结合了来自多个斑块的信息。我们在图2（c）中详细介绍了h的结构。网络h使用Xuebin等人[35]介绍的嵌套U-Net结构。在我们的实现中，我们采用了2×2的卷积，跨度为2，输出的通道数是输入的一半。这样的卷积在计算上要比3×3卷积便宜，因为3×3卷积需要对h处理的低分辨率特征图进行填充，而这种填充可以显著提高空间分辨率。最底层的特征图被平均汇集以提取最高级别的上下文，然后使用近邻插值上采样到之前的分辨率。最后，在h的上采样路径中，在每一级，我们将特征图与相应的上采样特征图连接起来，然后是一个全连接层。虽然权重映射网络w=[w0, ..., wn]是我们超网络的关键部分，但在我们的分层网络中，将w分成若干部分并将这些部分连接到主网络块上是比较有效的（图2(b)）。因此，权重映射网络的各层w0, ..., wn不是直接跟随h，而是嵌入到d的每个元块中。其原理是，从上下文到权重的映射会在内存中产生大量的扩展，这可能成为性能瓶颈。相反，权重是在它们被消耗之前产生的，最大限度地减少了内存的消耗，更好地利用了内存缓存。每个wi是一个1×1的与通道组的卷积，gwi，接下来将详细介绍。  
3.2. 解码器（主网络）

解码器d，如图2(b)所示，由n+1个元块组成，m0, . . , mn，如图3(a)所示。块，m0，对应于输入图像，而每个块，mi，i=1...n，对应于编码器的特征图，Fi，。每个区块之后都要进行双线性上采样，并与下一个更精细分辨率的特征图相连接。与传统方案不同，通过采用超网络，解码器的权重d取决于输入图像。此外，d的权重不仅以输入图像为条件，而且在图像的不同区域之间也有变化。通过这种方法，我们可以有效地将来自网络干层的低层次信息与来自底层的高层次信息相结合。这使我们的方法能够使用更小的解码器达到更高的精度，从而实现实时性能。超网络可以被看作是一种注意力，与一些基于注意力的方法[20, 33]相似，D从知道像素的位置信息中获益。为此，我们对输入图像和编码器的特征图进行了额外的位置编码。m0, ... . m0, ..., mn的设计是基于MobileNetV2[40]的倒置残差块：一个点式卷积，pw1，然后是深度式卷积，dw，以及另一个点式卷积，pw2，没有激活函数。我们的网络没有采用常规的卷积，而是采用了动态的、补丁式的卷积，这在下一节中描述。对于非常小的斑块--在我们的大模型中小于4×4；在我们的小模型中小于8×8--元块只包括pw1。每个mi所需要的总元参数，θmi∈R(|θpw1|+|θdw|+|θpw1|)×H 2n×W 2n，是mi中所有动态卷积的综合元参数：θmi=θpw1∪θdw∪θpw2。考虑到信号φi∈RCφi ×H 2n ×W 2n，权重θmi由嵌入mi的wi层生成。在推理时，mi的批量规范化层与wi融合；更多细节在补充材料的E部分提供。在每个mi中采用完整的信号在计算上和可训练参数的数量上都是低效的，因为φ被直接映射到大量的权重中。因此，我们把φ的通道分成几个部分，Cφ0, .... Cφn，其大小与每个元块所需的权重数量相对应。通道的划分是通过以下程序定义的：  


其中divide channels(-)详见补充材料中的A节。这个程序确保每个部分与它所分配的信号通道成正比，对于w中的分组卷积来说，可以被max(gw0, ..., gwn)所分割，并且被分配的通道数量最少。组的数量，gwi，是一个重要的超参数，因为它控制了在产生mi的权重时投入的计算量和可训练参数。增加gwi可以直接减少计算量和可训练参数，这可以从以下公式中看出：



在补充材料中，我们研究了不同gwi值的影响（C节），并报告了我们测试中使用的gwi的确切值（B节）。

3.3. 动态补丁式卷积

简单理解：当训练结束后，卷积核参数不再是一个定值，而是一个由输入决定的变量。

我们在图3(b)中说明了动态补丁式卷积（DPWConv）的操作，即mi的层，pw1，dw和pw2。给定一个输入特征图，X∈RCin×H×W，和一个权重网格，θ∈RCout×Cin G×Kh×Kw×Nh×Nw，其中Cin和Cout是输入和输出的通道号，G是通道组的数量，H和W是输入的高度和宽度，Kh和Kw是内核的高度和宽度，Nh和Nw是沿高度和宽度轴的斑块数量，我们定义输出斑块如下。



其中∗是卷积运算，i∈[0，Nh]和j∈[0，Nw]是补丁索引，Xi,j是X在网格位置（i，j）的一个补丁，θi,j是来自权重网格的相应权值。我们首先对整个输入特征图X进行填充，然后在每个补丁Xi,j处，我们对相邻补丁的相邻像素进行包裹。

5. 结论

我们建议将自动编码器与超网络结合起来，以完成语义分割的任务。在我们的方案中，超网络是由三个网络组成的：语义分割编码器的主干b、内部U-Net形式的上下文头h、以及多个权重映射头wi。解码器是一个多块解码器，其中每个块，di，实现局部连接的层。其结果是一种新型的U-Net，能够动态地、局部地适应输入，从而有可能使分割过程更好地适应输入图像。正如我们的实验所显示的，在这个竞争激烈的领域，我们的方法在多个基准中的表现优于SotA方法。

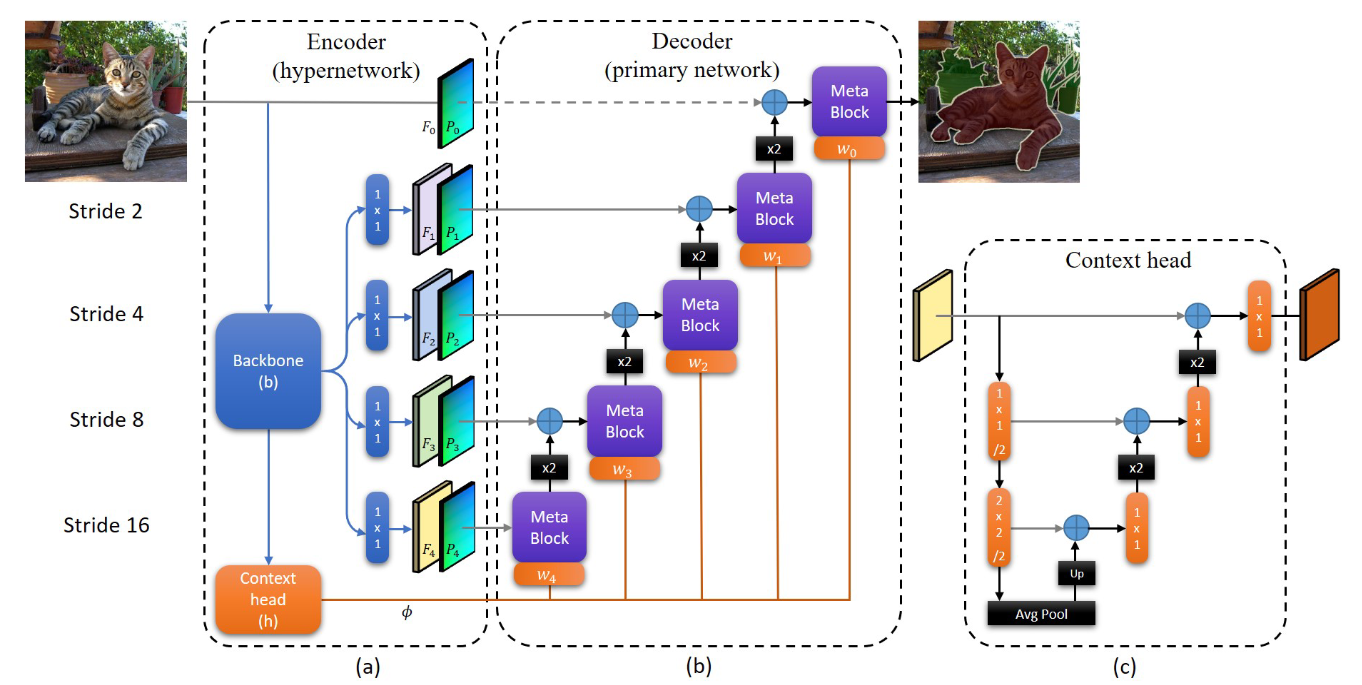


图2. 方法概述。(a) 基于EfficientNet[44]骨干网的超网络编码器，b，其最后一层由上下文头h代替。 (b) 主网络解码器，d，以及嵌入在每个元块中的权重映射网络的层wi。解码器的输入，d，是输入图像和特征，Fi，与位置嵌入，Pi相连接。它的权重是为图像中的每个斑块动态确定的。灰色箭头代表跳过连接，×2块是双线性升采样，蓝色的'+'号是连接。(c) 情境头被设计成一个嵌套的U-Net。更多细节请见第3章。

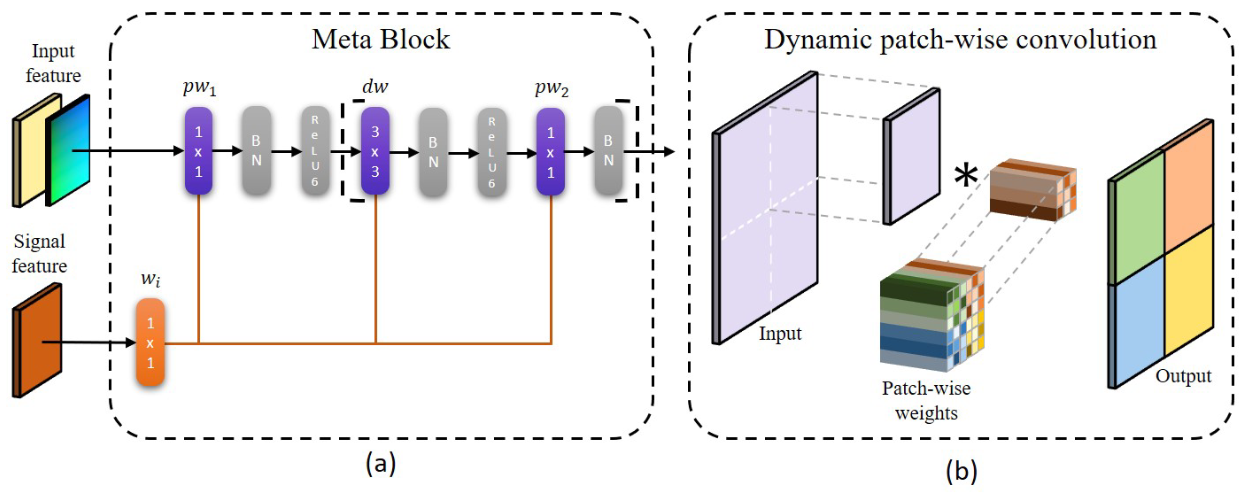


图3. (a) 基于倒置残差块的元块[40]。(b) 动态斑块卷积操作的可视化。每种颜色代表对应于特定补丁的权重，"\*"是卷积操作。更多细节请见第3.3节。

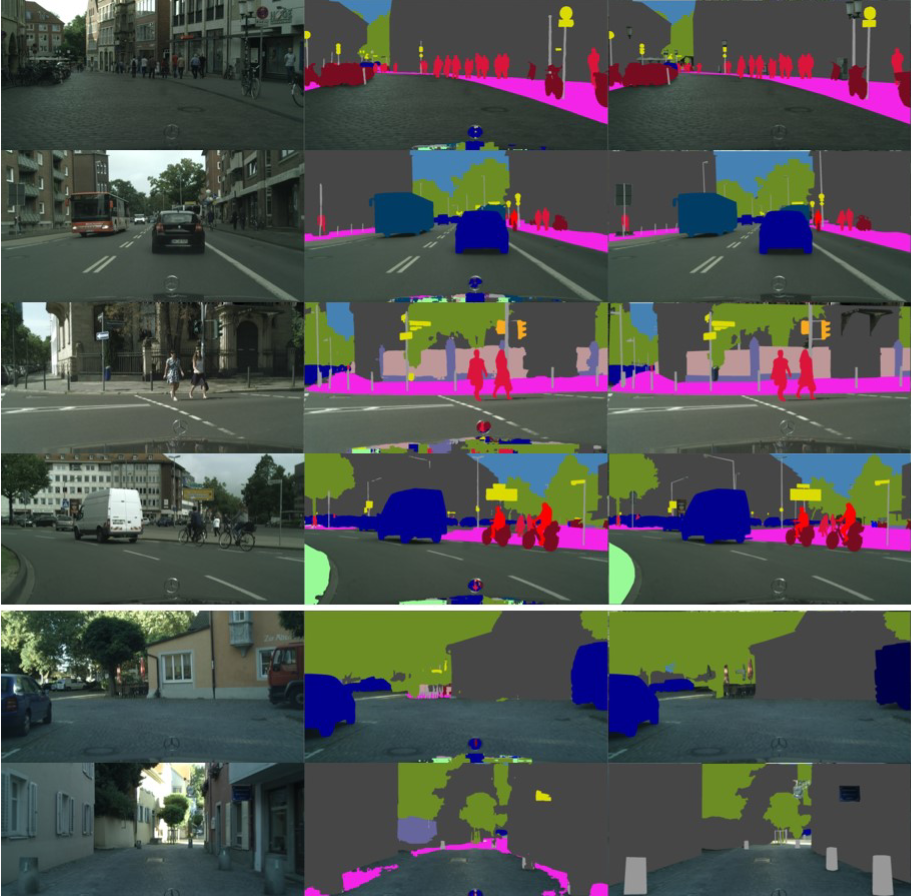


图5. 对Cityscapes[8]验证集图像的定性结果。从左到右：输入，我们的结果，和地面真实。前四行展示了我们的模型在不同场景中的表现。最后两行提供了失败的例子。请注意，在评估中忽略了反射的汽车引擎盖区域。