基于 DeepLabV3plus 进行轻量化语义分割任务

陆志阳（31520211154071）

摘 要：随着深度学习的不断发展，对于图像的特征提取能力越来越强，逐渐超越了人工设计的特征提取手段。在计算机视觉领域，卷积神经网络的发展更是让许多视觉任务变得更有效，如图像分类，目标检测，语义分割等。因此越来越多工作对卷积操作进行改进：减小卷积参数使模型轻量化的深度可分离卷积，兼顾感受野和特征图分辨率的空洞卷积等等。在本次实验中，我将使用编码器解码器架构的 DeepLabV3plus 进行语义分割的任务，并采用轻量化网络 MobileNetV2 作为主干网络进行特征提取。

关键词：卷积;轻量化模型;语义分割;深度学习

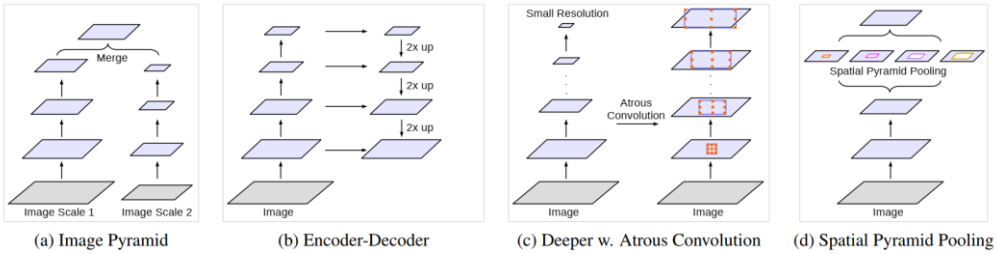
Lightweight Semantic Segmentation Based on DeepLabV3plus

Abstract: With the continuous development of deep learning, the ability to extract features from images is getting stronger and stronger, gradually surpassing the feature extraction methods designed by humans. In the field of computer vision, the development of convolutional neural networks has made many visual tasks more effective, such as image classification, object detection, semantic segmentation, etc. Therefore, more and more work has been done to improve the convolution operation: reducing the convolution parameters to make the model lightweight, depthwise separable convolution, taking into account the receptive field and feature map resolution, and so on. In this experiment, I will use the encoder-decoder architecture DeepLabV3plus for the task of semantic segmentation, and use the lightweight network MobileNetV2 as the backbone network for feature extraction.

Key words: Convolution; Lightweight model; Semantic segmentation; Deep learning

1 引言

语义分割是计算机视觉下游任务中必不可少的任务之一，和图片分类、目标检测不同，语义分割是一种像素级别的任务。但是语义分割任务也面临着许多的挑战，目前存在两个重要的挑战：将深度卷积网络应用于图片上时，为了提取图像中更加抽象的特征，会导致特征图的分辨率不断下降，但是语义分割任务是一种像素到像素的任务，需要对图像上的每个像素进行分类，而随着特征图分辨率的不断下降，详细的空间信息会缺失，这对于密集型的预测任务无疑是一种降维打击。为了解决这个问题，许多 Encoder-Decoder 网络模型不断涌现[1, 2]，利用编码解码的网络模型，在编码器中提取图像的抽象特征进行预测，再用解码器还原到原始像素级，并在编码解码层之间用拼接前向传递卷积前的空间信息。还有的方法采用空洞卷积（Atrous convolution）来进行特征分辨率的调整。这对于空间信息的保存和分辨率的要求得到了解决。第二个挑战来自于图像的多尺度特征。在一张图像中，，因为远景近景的原因，目标的尺度大小可能有很大的不同，所以对于多尺度特征的提取是一个重要的环节。应用于多尺度特征提取的方法大概分为四种[3]，如图一所示：（a）把图像进行不同尺度的缩放，然后用模型对于缩放后的图像进行特征提取；（b）采用编码器解码器架构，利用编码器进行不同尺度的特征取；（c）级联其他模块进行来捕获长范围的特征信息，常用的有条件随机场[4]；（d）还有一种方法是采用空间金字塔结构进行特征提取，通过扩大卷积核来不断增加感受野从而提取不同尺度的特征。



(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

图一：对于多尺度特征的提取

从以上来看，编码器解码器架构适合用于密集型的预测任务，既满足了分辨率恢复的要求，又可以对图像中多尺度的物体特征进行特征编码，无疑是一个很好的选择。DeepLabV3plus 利用编码器解码器结构进行语义分割[5]，DeepLabV3plus 改进于 DeepLabV3，同样采用不同比率的空洞卷积金字塔进行特征提取，并将这种方法于编码器解码器架构结合起来使用，DeepLabV3plus 的编码器部分输出两个特征图，一个是只经过利用空洞卷积进行深度特征提取的浅层特征图，另一个是经过特征提取空间金字塔提取的不同尺度的特征，然后进行全局平局池化的特征图。网络提取的浅层特征，对于图像空间信息的的损失很小，有利于进行图像中物体边界空间信息的保存，与深层全局抽象特征结合送入到解码器中，细化了物体边界的特征解码，提高了网络模型的精度。

在本次实验中，我采用 MobileNetV2 作为 DeepLab 网络的骨干网络[6]。MobileNet 系列模型是谷歌提出的一类轻量化模型，其模型参数更小但是精度损失不大。卷积神经网络推动计算机视觉向前迈进了一大步，卷积神经网络设计的历史始于 LeNet 风格的模型，这是用于特征提取的简单卷积堆栈和用于空间子采样的最大池运算，在最大池操作之间，卷积操作被重复多次，使网络能够在每个空间尺度上学习更丰富的功能。但是随着网络模型越做越大，越做越深，许多的模型只能在具有丰富的计算资源的特定场所才能训练的起来，对于边缘设备极不友好。比如无人驾驶汽车中，需要实时进行决策，而大规模的网络模型的计算量大，响应慢，有时甚至需要将数据上传云端再进行决策，耗时十分严重；在手机端我们也没有很多的计算资源可以使用，如果想在手机本地加载大规模的模型，可能需要云端服务器的帮助，这可能导致隐私的泄露的问题。所以对于边缘设备，一些小的模型十分的重要。为了降低卷积网络的参数，Inception 系列网络提出了一种可分离的卷积操作[7]，对于多通道的图像分别进行通道卷积和空间卷积，Inception 模块背后的想法是，通过明确地将其分解为一系列操作，使该过程更容易、更高效，这些操作将独立地查看跨通道相关性和空间相关性。具体来说，跨通道卷积首先通过一组 1x1 卷积来融合通道相关性，将输入数据映射到比原始输入空间小的独立空间中，然后通过常规 3x3 的单层卷积对于每个通道分别进行卷积操作。而 Xception 模型把这种方法做到了极致[8]。MoblieNetV2 网络结合了深度可分离卷积的的方法，还使用了反向残差连接的思想。

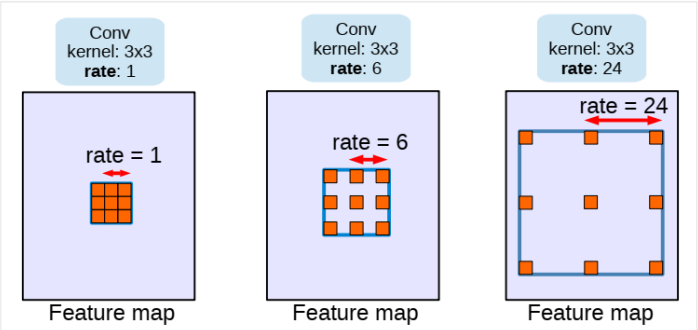
2 理论知识

2.1 深度可分离卷积

深度可分离卷积是许多高效神经网络结构的关键组成部分，如 Inception 系列网络就是在深度可分离卷积上做一系列改进。其基本思想是用通道卷积和空间卷积两个卷积过程来代替一个经典的常规卷积算子。首先对图像的每个通道分别用 3×3 或 5×5 的单层卷积核进行通道卷积提取每个通道上的抽象特征，然后使用与通道数量相同大小多个的 1×1 卷积核进行跨通道的空间卷积，融合不同通道的特征信息来构建新的特征。深度可分离卷积在产生相同效果的同时，减少了卷积网络的参数数量，为模型轻量化奠定了为基础。

2.2 空洞卷积（Atrous Convolution）

普通卷积神经网络随着网络层数不断加深，图像的特征也越来越抽象，特征图分辨率越来越低，这对于语义分割等密集型的任务会出现空间信息损失等问题。DeepLab 系列网络提出了空洞卷积的思想。如图所示，核大小为 3×3 且比率不同的空洞卷积。标准卷积对应于比率为 1 的空洞。使用较大的比率可扩大卷积野，从而实现多尺度的对象编码。实际上也就是在卷积参数不变的情况下向中间插入零值。

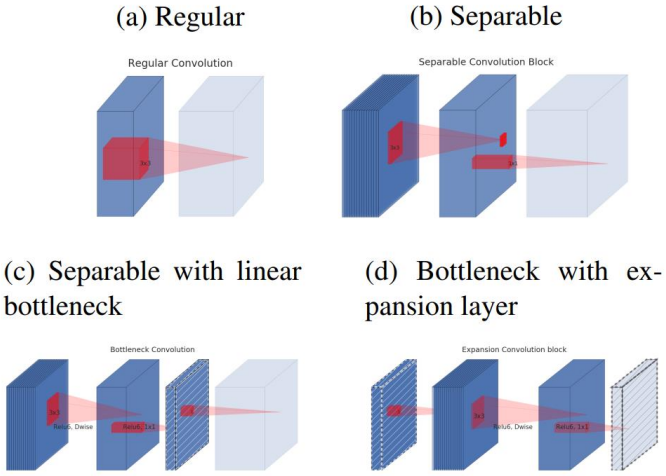


图二：不同比率的空洞卷积图例

2.3 MobileNetV2网络结构

MobileV2 是一个轻量型的网络架构，采用深度可分离卷积的方法进行多层叠加和残差连接。受到 Xception 对于卷积改进的启发，MobileNetV2 还对深度可分离网络进行了改进， MobileNetV2 先用 1×1 的卷积核进行空间卷积，再用 3×3 的卷积核进行通道

卷积。通过实验发现，高维特征信息的嵌入到低位流形时，如果直接在低维空间进行非线性（ReLU）操作再映射回原始空间时会产生信息丢失，MobileNetV2 为了避免特征信息的丢失，先通过可分离卷积方法中的空间卷积把特征维度映射到高维，让特征图的通道数增加，然后使用通道卷积对更高的特征图通道进行逐通道卷积，提取抽象的通道特征，最后再使用空间卷积 1×1 的卷积核把高维特征通道映射回原始的通道数。这样既可以避免信息的丢失，又可以减少卷积的参数数量。



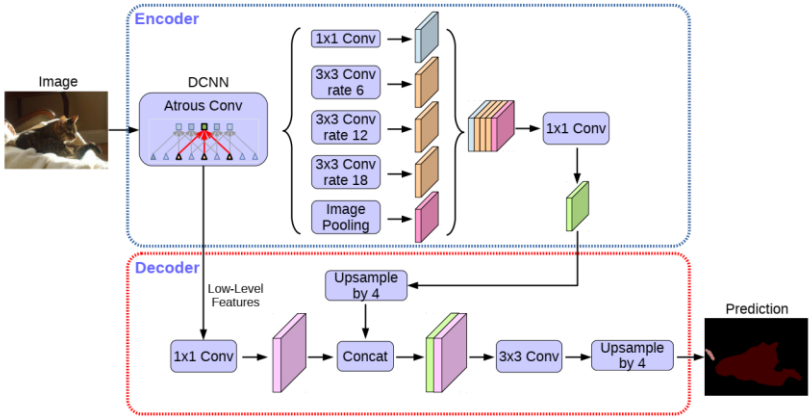
图三：MobileNetV2(d)相比于常规卷积(a)、可分离卷积(b)、降维可分离卷积(c)的改进

最后，MobileNetV2 利用残差连接将低维通道的特征连接起来，一般的残差连接是连接特征通道数量较多的特征块，但是 MobileNetV2 连接的时通道数较小的特征块，作者认为较小的瓶颈特征块包含更加抽象有效的特征，连接瓶颈特征块这样可以高效进行特征提取，故称为反向残差连接。

2.4 DeepLabV3plus网络结构

编码器解码器结构对于密集型任务十分友好，如这种结构可以用于解决语义分割的特征多尺度和分辨率调整等问题。DeepLabV3plus 采用编码器解码器架构，在之前的网络基础上进行改进。因为在家中设备不够，我在实验中使用轻量化网络 MobileNetV2 作为 DeepLabV3plus 特征提取的骨干网络，并且将依旧采用 DeepLab 系列的空洞卷积操作进行实验。

为了解决多尺度问题，DeepLabV3plus 采用了空洞卷积金字塔（ASPP）模块使用不同比率的空洞卷积对特征图进行特征提取。具有不同衰减率的 ASPP 有效地捕获多尺度信息。然而，随着采样率变大，有效滤波器权重的数量变小。极端情况下，3×3 滤波器不是捕获整个图像上下文，而是退化为简单的 1×1 滤波器，其外部的参数权重都变为零值，使得感受野较大的空洞卷积无法有效的提取相应感受野的抽象特征，模型将损失部分大尺度的特征信息导致精度下降。将全局上下文信息纳入模型是一个不错的方法，DeepLabV3plus 对图像的特征进行全局池化，相当于做一次全局感受野的池化操作，然后再用双线性插值将特征图上采样到对应的维度。再把全局平均池化得来的全局特征与一个 1×1 和三个 3×3 组成的空洞卷积金字塔提取的多尺度特征结合起来，这样解决了空洞卷积的比率过大产生卷积参数为零的问题，为网络模型增加了全局感受野的特征信息。



图四：DeepLabV3plus 网络结构图

网络的解码器接受一个从 MobileNetV2 提取出的浅层特征，在代码中使用 MobileNetV2 的前四层对浅层特征进行提取，这个浅层特征包含了图像中物体边界的信息，通过与深层抽象特征结合，让解码器对于物体边界的预测更加精确。

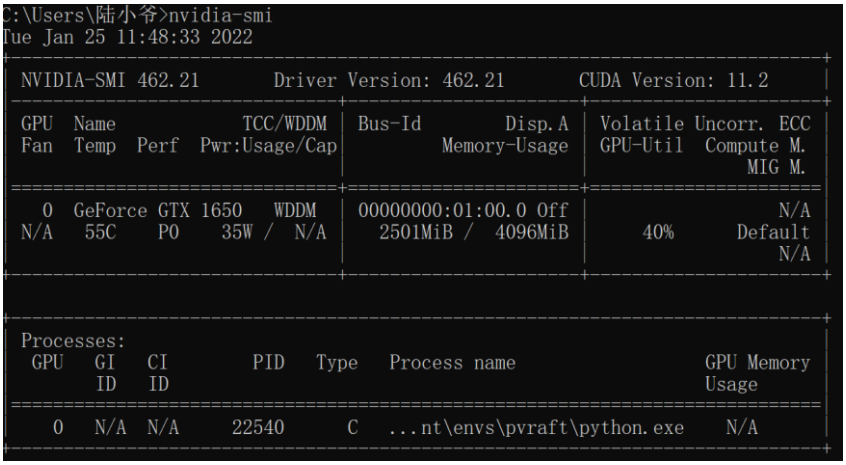
3 实验结果及分析

3.1 数据集

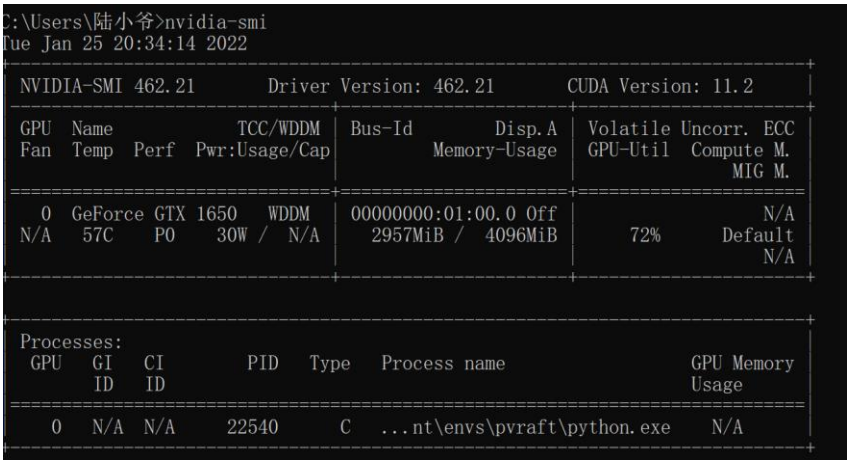
我使用 PASCAL VOC2012 数据集进行模型的训练[9]。PASCAL VOC2012 是语义分割任务中一个经典的数据集，包含了 20 个前景类和 1 个背景类。原始数据包含了 1464 张训练图片和 1449 张验证集图片。在训练中，我使用数据增强增加训练集数据，对训练数据及标签图像进行随机裁剪、缩放、反转、改变颜色。

3.2 训练过程

采用预训练模型加微调的方式进行模型的训练。先从网上下载 MobileNetV2 的预训练权重，载入到模型将特征提取骨干网络 MobileNetV2 冻住，训练其余的 ASPP 模块和解码器模块，训练 50 轮，因设备内存有限，BatchSize=4，LearningRate=0.0005。然后将骨干网络解冻，一起微调 50 轮，BatchSize=2，LearningRate=0.00005.学习率每过一轮缩减为原来的 0.94。 BatchSize 过小导致后面模型收敛性和泛化性能较低。损失函数采用标准的交叉熵损失函数。我使用笔记本电脑一个 GPU GTX1650 对模型训练了 12.3 个小时



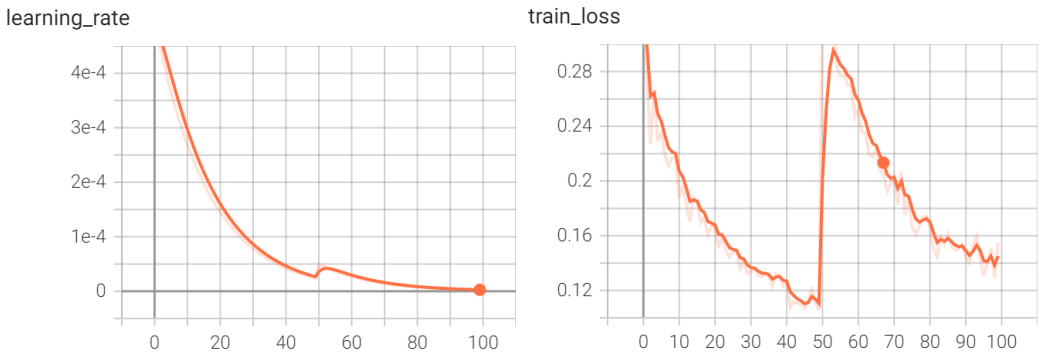
图五：采用预训练权重的骨干网络对模型进行训练时 GPU 情况



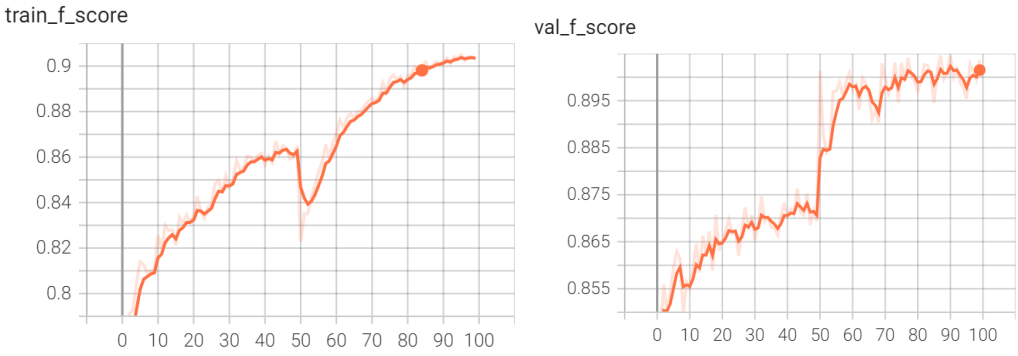
图六：对模型进行微调时 GPU 情况

3.3 实验结果

我使用 Tensorboard 库对网络模型的训练情况进行检测。使用标准的交叉熵损失函数，并使用 Dice 系数和 IoU 参数来计算逐像素的预测得分。前 50 轮训练除预训练权重的骨干网络的其余模型模块，后 50 轮微调模型所有参数。

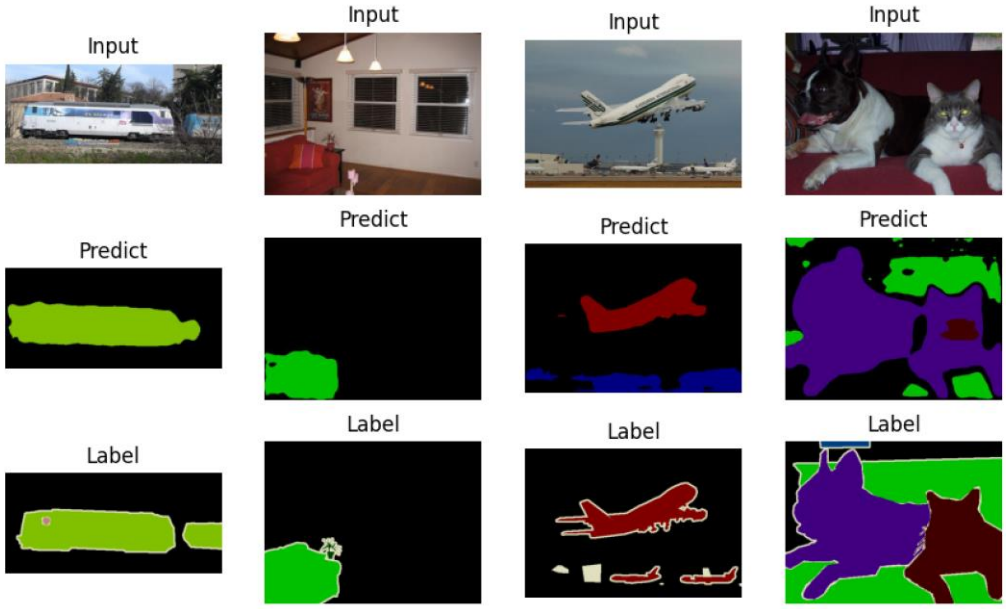


图七：100 轮学习率的变化和次训练集损失函数的变化曲线



图八：100 轮次训练集 mIoU 得分和测试集 mIoU 得分的变化曲线

因为网上测试集的数据大多数都没有标签，所以我从 PASCAL VOC2012 数据集的验证集中均匀随机挑选出 97 张图片和其标签作为测试集与模型的预测结果进行比对。



图九：使用未经训练的验证集进行测试的结果及标签

3.4 实验结果分析

在微调模型的时候，因为训练轮数过少，且每个批次大小只有 2，所以模型对于数据的分布的均值和方差学习的不准确，导致拟合能力减弱，在微调初期模型训练参数的忽然增大，可能学习率还是有点大，从而出现训练损失激增的现象。

4 结论

DeepLabV3plus 是编码器解码器架构的网络模型，其编码器有效的提取了图像的多尺度特征信息，并利用浅层特征保存物体边界的信息，从而在解码器中能有效的恢复图像物体的边界。网络使用改进的空洞卷积特征金字塔不仅在参数量不变时有效的提取了

图像多尺度特征，还可以通过空洞卷积比率来控制分辨率。受可分离卷积的启发，轻量级网络 MobileNetV2 可以在参数很小的情况下进行特征提取的任务，将 MobileNetV2 作为 DeepLabV3plus 的骨干网络用于语义分割任务可以有效地在边缘设备上进行训练，但是预测精度还有待提高

References:

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440.

[2] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520-1528.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[4] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," Advances in neural information processing systems, vol. 24, pp. 109-117, 2011.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801-818.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510-4520.

[7] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," International journal of computer vision, vol. 111, no. 1, pp. 98-136, 2015.