基于人像分割的实时虚拟背景替换系统

郦运琛，马俊程，屈泽凯，魏昕原

2021K8009929028,2021K8009929021,2021K8009916005,2021K8009915002

1 问题分析

由于我们组对语义分割领域较感兴趣，在本次大作业将自主选题，选择和分割息息相关，且具有广泛应用价值的“虚拟背景”作为题目，目标是实现一个实时的、对环境具有鲁棒性的虚拟背景替换系统。

虚拟背景技术是一种可以将计算机生成的背景图像替换真实背景的技术，它主要用于视频会议、在线教育等场景中，可以消除背景噪声、保护隐私或者提升视频质量，从而来增强用户体验。虚拟背景技术本质上是将人的身体和脸部从视频流中提取出来，然后将其放置在虚拟背景中，而虚拟背景一般由预先制作的图像或视频组成，也可以根据用户的需求和喜好动态生成。虚拟背景技术的发展十分快速，现已成为了大多数视频会议和在线教育平台的基本功能之一，得到了广泛的应用，而其面临的最大挑战便是如何将人物与原始背景准确而迅速地分离。为了解决这一问题，语义分割、实时分割、人像分割等技术是虚拟背景技术研究的重点。

语义分割是一种计算机视觉领域的一项基本任务，旨在将图像中的每个像素都分配给特定的语义类别，例如人、车、路、树等等，而在虚拟背景中的应用则是聚焦于将人从背景中分离出来。语义分割技术的产生始于传统的图像分割技术，如阈值分割和边缘检测等。但这些方法只能将图像分割成块状区域或者边缘线条，并不能对不同语义类别进行精确区分。相较而言，语义分割能够更精确地理解图像中的细节，从而为图像理解和计算机视觉应用提供更多的信息。随着深度学习技术的发展，特别是卷积神经网络（CNN）的出现，语义分割技术得到了快速发展。早期语义分割模型大多是基于全卷积网络的编码器解码器结构，例如 FCN [1]、SegNet [2]、UNet [3] 等。后来，出现了应用空洞卷积的 DeepLab 系列模型 [4]，利用了空洞卷积的特点，能够在保证分辨率的同时扩大感受野。最近几年，又涌现出一些新的模型，如 HRNet [5]、BiSeNet [6] 等，这些模型在保持较高精度的同时，能够实现更快的速度和更少的参数。

随着移动设备需求的增大，许多工作开始针对实时语义分割设计轻量级模型。实时语义分割是指能够在较短时间内（通常是每秒钟几帧或几十帧）完成对图像中每个像素的标注，从而实现对图像或视频的快速、实时分析和处理。相比于传统的离线分割方法，实时分割方法需要满足更高的速度和实时性要求，并且在尽可能少的时间内完成分割任务。实时分割技术的发展历程与计算机硬件和深度学习技术的进步密不可分。随着 GPU 计算能力的提升和深度学习算法的发展，越来越多的实时分割方法被提出。在 2015 年左右，一些基于全卷积网络（FCN）的实时分割方法被提出，如基于编码器-解码器结构的 ENet [7] 和双分支结构的 BiSeNet [6] 等。这些方法采用了特殊的卷积和池化操作，能够在保证精度的同时，快速地对图像进行分割。

人像分割是语义分割的一个子问题，目标类别只有作为前景的人和背景两种。和通用语义分割相比，人像分割具有额外的挑战，如实时性，重视边界的分割精度等。人像分割的发展历程也与深度学习技术的进步密切相关。最早的人像分割方法主要基于传统算法，如基于边缘检测和区域生长的方法。然而，这些方法的准确度和鲁棒性受到限制，尤其是在复杂场景中。近年来，许多基于深度学习的人像分割算法被提出，如 PortraitNet [71],SiNet [?] 等。

我们分析了人像分割领域公开的数据集并对其进行整理，得到下表。其中 EG1800,AISeg 以及 Maadaa 都很少被用于视频会议任务。Maadaa 包含许多类似的图像和软件界面的无关信息，如按钮、窗口等；AISeg 为经人脸检测和区域裁剪后的半身人像集；而 PP-HumanSeg14K 主要是视频肖像方面的数据集，弥补了之前两个数据集的不足。我们将在其中挑选合适的数据进行训练。

数据集名称	网络链接	数据集简介
Maadaa	https://maadaa.ai/dataset/live-streamer-portrait-segmentation/	照片张数：15.6K 张；分辨率：$540\times960 \sim 720\times1280$, 为人体和背景的分割。
AISeg	$https://github.com/aisegmentcn/matting\_human\_datasets$	包含 34427 张图像和对应的 matting 结果图，经过人脸检测和区域裁剪后生成了 600*800 的半身人像。
EG1800	https://aistudio.baidu.com/aistudio/datasetdetail/155370	包含了 1736 张图片和对应的 label, 分辨率：$512 \times 512 \sim 512 \times 1024$
PP-HumanSeg14K	https://github.com/PaddlePaddle/PaddleSeg	包含来自 291 个会议场景的 23 个视频，帧为 14K，分辨率：1280×720 包含各种电话会议场景、参与者动作、照明变化。

2 相关工作调研

我们组将使用人像分割的方法实现虚拟背景任务，而人像分割作为语义分割的一个子领域，在实际应用时对实时性有较高的要求。因此，我们在这一部分将分别讨论基于监督学习的语义分割、实时语义分割和人像分割三个方面的相关工作。

2.1 语义分割

语义分割作为计算机视觉领域的基础任务之一，具有长久的发展历史。在深度学习方法流行之前，基于传统机器学习分类器的语义分割方法使用较多。但传统方法需要手工设计特征，无法使用一些深层和隐藏的特征，这导致图像特征的使用受到限制，在深度学习流行起来后很快被超过。我们在这一部分将简要介绍基于深度学习的语义分割方法，并分为基于 CNN 和基于 Transformer 两部分分别讨论。

2.1.1 基于 CNN 的语义分割

由于深度学习的快速发展，一系列深度神经网络取得了巨大的成就，特别是卷积神经网络，如 VGG [56]、GoogLeNet [58]、ResNet [52] 等。这些网络的出现为语义分割带来了新的解决方案。基于卷积神经网络的方法具有天然的优势，它可以自动提取语义特征，而不是原来方法中有偏差的人工特征提取。并且是端到端的处理结构，预测图可以直接在输出层得到。

和分类任务不同，语义分割作为一种密集预测任务需要对每个像素进行预测，因此，高层次的语义信息和低层次的细粒度信息对精度都有重要影响。语义信息为预测提供上下文特征，可以考虑全局对象的长距依赖，提供非局部的视角。它一般通过 CNN 下采样（跨步卷积，池化等）提取，在这个过程中特征图的分辨率不断降低，感受野不断增大，高层次的语义信息被编码到特征中。但只有语义信息无法给出精确的分割结果，因为每个像素点的局部特征在下采样的过程中逐渐丢失，无法给出精细的预测。包含空间结构信息的细粒度特征对分割边界的精度很重要，较浅卷积层的高分辨率特征图往往被认为含丰富的细粒度特征。由于分割任务的特殊性，感

受野与分辨率、语义特征与细粒度特征之间往往存在矛盾，语义分割工作大多围绕细粒度特征和上下文特征的提取和融合展开。本部分将按照提出的方法，对相关工作分类进行介绍

编码器-解码器结构 **FCN** [1] 是第一个将 CNN 用于语义分割的工作，仅包含卷积层，使其能够拍摄任意大小的图像并生成相同大小的分割图。通过使用跳跃连 接，其中来自模型最后一层的特征图被上采样并与较早层的特征图融合，从而结合了语义信息和细粒度信息，得到了不错的精度，被视为图像分割的一个里程碑。

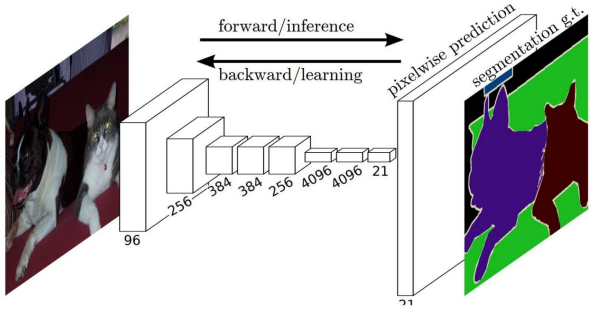


图 1: FCN [1]

U-Net [3] 最初为用于医学图像的分割模型，但后来广泛取得优秀结果，是经典的编码器-解码器结构。它由两部分组成，编码器部分采用了全卷积网络，通过下采样获得一系列特征图，将得到的低分辨率但富含语义信息的特征图输入解码器进行上采样，每次上采样都会将结果与下采样过程中对应的特征图裁剪后拼在一起，从而在这个过程中多次聚合语义特征和细粒度特征，实现很好的效果。由于上采样和下采样过程比较对称，形成一个 U 型结构，故起名为 U-Net。

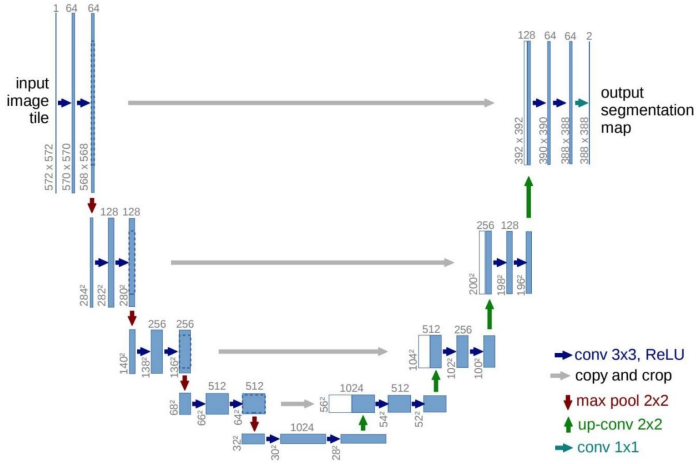


图 2: U-net [44]

Segnet [2] 提出最大反池化，创新了上采样的方式，即解码器使用相应编码器中最大池化时计算的池化索引来进行上采样，上采样后的图经过卷积层可产生密集的特征图。与反卷积相比，这种方式不需要学习，大大减小了计算量，并且只需要储存下采样时的最大池化索引而非特征图，节省了内存。

空洞卷积 扩张卷积（Dilated convolution，又称空洞卷积，atrous convolution）在卷积层中引入了另一个参数，即扩张率。例如，一个 3×3 的核，如果扩张率为 2，那么它将具有与 5×5 的核相同的感受野大小，但只使用了 9 个参数，从而在不增加计算成本的情况下扩大了感受野。

(a) (b)

图 3: (a) 空洞卷积原理；(b)Deeplab [4]

扩张卷积可以在保持分辨率的同时增大感受野，在语义分割中应用广泛。**DeepLab** [4,72–74] 系列网络用空洞卷积改造主干网络，使下采样后的特征图保持一个较高的分辨率，并提出一个提取多尺度上下文特征的 ASPP 模块，得到了很好的精度。

上下文模块 由于分割的目标有不同的尺寸，许多工作设计了提出多尺度上下文的模块。**PSPNet** [45] 提出金字塔池化模块，能更好地学习场景的全局上下文表示。这个网络使用残差网络（ResNet）作为特征提取器，其中应用了空洞卷积，从输入图像中提取不同模式的特征。这些特征图被送入金字塔池化模块，它们在四个不同的尺度上进行池化，获得不同尺度的上下文特征。

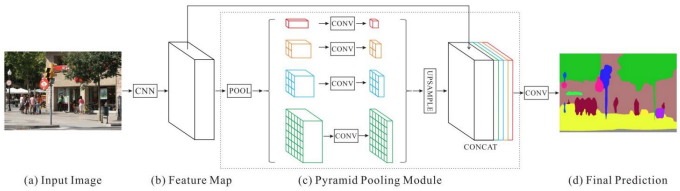


图 4: PSPNet [45]

注意力机制在捕获长程依赖性，提取全局特征上具有强大的优势，并且自注意力具有全局感受野，许多工作将注意力模块加入语义分割模型，获得了不错的效果。

EncNet [61] 提出了一个 CEM(Context Encoding Module) 模块，编码上下文信息，再对特征图的每个通道加权。**PSANet** [62] 引入逐点注意力，考虑全局相关性，每个点都自适应的通过一个可学习的注意力映射与其他所有点连接起来。**Non-local** [59] 将非局部的方法归纳成一个范式，并提出简化的自注意力操作。**DANet** [50] 为了更好的捕捉上下文信息（全局信息）和通道间的联系，提出了一种双注意力网络，分别提取空间注意力和通道注意力得到更好的特征表示。

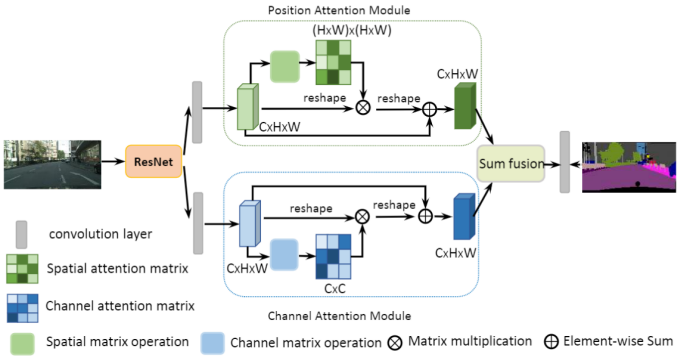


图 5: DANet [50]

由于自注意力的平方复杂度，出现了一系列简化计算的工作, 例如 **CCNet** [53] 仅在当前行和列计算注意力，**ANN** [64] 先对对 K 和 V 进行降维再计算注意力，**GCNet** [48] 使用共享注意力图。

2.1.2 基于 Transformer 的语义分割

Transformer 首先在自然语言处理领域取得巨大成功 [75]，ViT [49] 遵循 NLP 中的 Transformer 设计第一个 Vision Transformer ，可以在图像分类中实现 sota，之后又涌现了许多基于 Vision Transformer 的主干网络，如 Swin Transformer [54],BEiT [47],MAE [51] 等，证明了 Transformer 结构在视觉任务上的优势。因此，出现了一系列基于 Vision Transformer 的语义分割工作，取得了强大的效果。

SERT [63] 将 ViT 应用到语义分割领域，第一个证明了 Transformer 在此领域的优越性。**DPT** [55] 加入了更多卷积特性，对不同 Transformer 块的输出组装成不同分辨率的类似于图像的形式，其中较浅层的 Transformer 块会被组装成更大分辨率的表示，因为其中包含更多细粒度特征，最终融合得到多尺度的特征图。**Segmenter** [57] 提出一种新的解码器，学习了多个类嵌入，得到了不错的效果。**Segformer** [60] 提出了一种简单高效的基于 transformer 的语义分割模型，编码时通过合并补丁来获得不同分辨率的特征图，并使用了一个由全连接层组成的轻量解码器

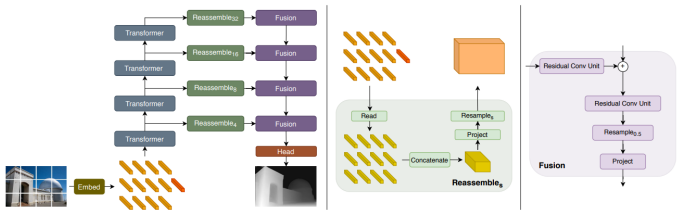


图 6: DPT [55]

2.2 实时语义分割

在上一部分中提到的语义分割方法使预测的准确度不断提升，但在准确性之外，计算复杂度也是一个重要的考虑因素. 现实世界的任务通常旨在在目标平台（硬件）和应用场景（如自动驾驶需要低延迟）给定的有限计算预算下获得最佳精度。作为机器人和自动驾驶汽车等复杂任务的重要一环，语义分割在许多实际情景中的应用要求它们在内存和计算能力有限的硬件上以非常低的延迟运行，这激发了一系列针对轻量级架构设计和更好的速度-准确性权衡的工作。

我们将在这一部分沿 *i) 高效编码器-解码器结构*；*ii) 多分支结构*两个主流方向介绍实时语义分割的相关工作。

2.2.1 高效编码器-解码器结构

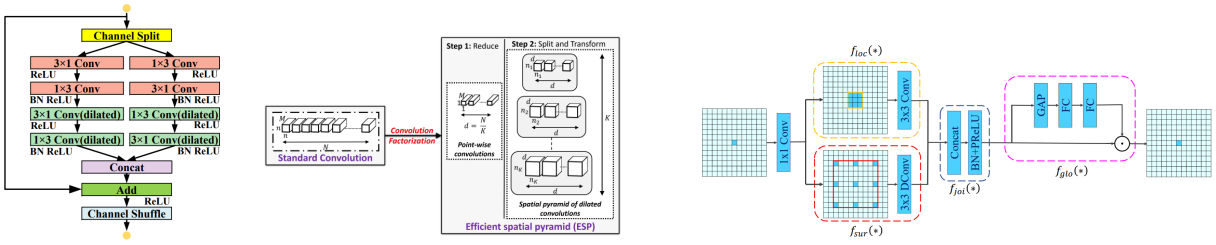
正如在上一部分中提到的，语义信息和细粒度特征对语义分割任务的精度都十分关键，而编码器-解码器结构是结合这两种特征的一种主要方法。编码器通常为一个特征提取网络，经过多次下采样，特征图的大小逐渐减小，感受野不断扩大，语义特征不断增加。解码器将富含语义信息的特征图经过多次上采样恢复分辨率，并在这个过程中通过横向连接融合细粒度特征，丰富细节信息。许多实时语义分割工作继续使用了编码器-解码器结构，在保证精度的同时，提出各种方法减小网络的计算量和参数量，使模型轻量化。

高效的主干网络 许多语义分割模型会选择优秀的主干网络作为编码器，这些主干网络往往针对分类任务设计，并经过了预训练，被认为具有较强的特征提取能力。其中，有一些为移动和资源受限环境设计的轻量级主干，显著减少了所需的操作和内存数量，同时保持较高的准确性。它们的组成模块和高效卷积的方法在实时语义分割任务中广为使用。

Mobilenet [11, 12, 69] 系列网络将深度可分离卷积应用到 CNN 中，把标准卷积分解为逐深度卷积和逐点卷积。**Mobilenet V1** [69] 用深度可分离卷积改进 VGG，参数量大大减小的同时，精度仅降低了 1%；**Mobilenet V2** [11] 改进了残差块，结合对兴趣流形的分析提出 Inverted Residual Block，提高了性能；**Mobilenet V3** [12] 利用 NAS 搜索最优架构，结合其他一些方法（如 SE 模块）进一步提高了性能。**shufflenet** [10] 使用分组卷积减小 1*1 卷积层的计算量，并通过通道混合实现组间特征交流；**shufflenet V2** [13] 取消了 V1 的分组卷积，使用通道的分割和混合提高效率。**GhostNet** [14] 减少特征图的固有维数，并通过线性变换生成更多的特征图，在保持相似识别性能的同时降低通用卷积层的计算成本。

设计高效卷积块 和分类任务相比，语义分割任务作为密集预测任务，还需要多尺度的上下文信息和高分辨率的细节信息。而主干网络通常是为分类任务设计的，其中的卷积块无法完全适应分割任务。因此，许多实时语义分割工作设计了新的轻量化模块，具有更强的提取多尺度特征和细粒度特征的能力。

早期的一些工作利用高效卷积设计新模块，显著降低计算量。**E-Net** [7] 是最早地针对实时语义分割的工作之一，它由一个（相对）大编码器和非常简单的解码器组成。整个网络由 bottleneck 残差块的几种变体构建而成，应用了深度可分离卷积和非对称卷积，上采样时应用了 Segnet 中的最大反池化方法，实现了一个非常紧凑和快速的架构，可以部署在嵌入式设备上。**ErfNet** [15] 利用不对称卷积改进了基础残差块，相比 E-Net 提高了精度。**LEDNet** [16] 提出 SS-nbt 块（如图所示），将输入的通道分割后，分别进行非对称卷积和非对称扩张卷积，将结果连接后经残差连接和通道混合得到输出。整体是一个非对称的编码器-解码器结构，在解码器使用特征金字塔，不断细化输出的语义特征。



(a) (b) (c)

图 7: (a)LEDNet [16] 提出的 SS-nbt 块；(b)ESPNet [17] 提出的 ESP 模块；(c)CGNet [19] 中提出的 CG 块

许多工作设计了提取上下文信息的高效模块。**ESPNet** [17] 提出的高效空间金字塔模块（如图所示），将标准卷积分解，使用 1×1 卷积降低输入维度，然后使用不同扩张率的并行卷积来增加感受野，提取到多感受野的特征。为了避免由不同的扩张率引起的网格伪影，输出被分层求和，结果被连接起来，最后通过残差连接添加到输入中。**Espnet V2** [18] 在原 ESP 模块的基础上，使用分组卷积计算 1*1 卷积，并将空洞卷积改为深度可分离的，大大减小了计算量。**CGNet** [19] 认为在特征提取时需要有上下文信息的指导，设计了 CG 块（如图所示），以其为主干设计了 CGNet，可以在每个阶段融合局部、上下文和全局特征。**MSFNet** [20] 针对高分辨率输入，设计了 SAP 模块，通过多尺度池化在每个感受野级别都有很好的空间信息恢复，并且将不同感受野层次相同分辨率的特征融合起来，在几乎不增加计算成本的情况下大大提高了性能，并且使用了针对边界的辅助任务，增强模型对边界的敏感度 **RegSeg** [21] 提出的 D 块使用通道分割、分组卷积、空洞卷积等方法，还加入了 SE 模块，融合全局信息。**DWRSeg** [22] 认为选择合适的感受野大小对提取特征的效率很重要，在浅层需要较小的感受野来捕捉局部特征，而高层需要更大的感受野捕捉语义特征，针对感受野设计了 DWR 块和 SIR 块。

一些工作改进了融合不同层次特征的方法。**SF-net** [34] 认为不同层次提取的特征在语义水平上存在差距，可能导致信息的无效传播。受视频中的“光流”启发，引入了语义流的概念，并提出了一种新颖的基于流的对齐模块（FAM）来学习相邻级别特征图之间的语义流场，用流场修正采样的标准网格，更有效地将高级语义特征融合到高分辨率特征。**SFNet-lite** [35] 在 SF-net 的基础上，把主干网络改为最新的 STDC，将门控加入 FAM，提出 GD-FAM，学习两个独立的语义流以同时细化高分辨率特征和低分辨率特征。由于 GD-FAM 出色的特征表示能力，整个网络进一步轻量化，相比 SF-Net 大大加速，并得到很好的表现。**AlignSeg** [36] 设计了对齐特征聚

合 (AlignFA) 模块和对齐上下文建模 (AlignCM) 模块，**FaPN** [37] 通过将变换偏移应用于可变形卷积来解决特征未对齐问题。

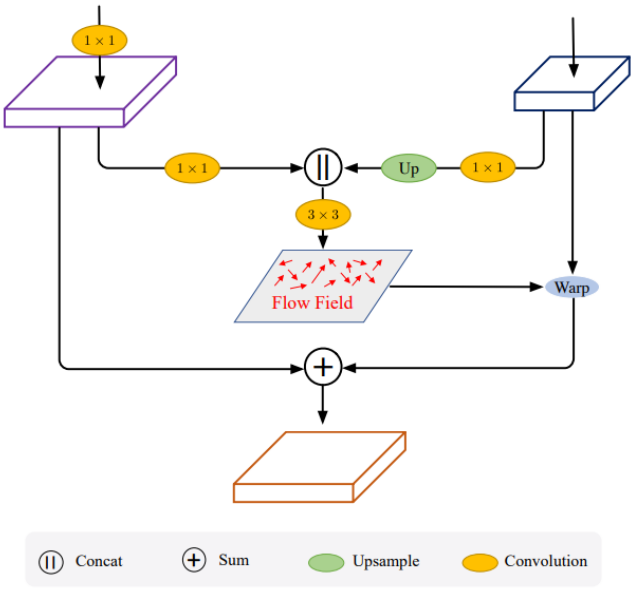


图 8: SF-net [34] 中的流对齐模块（FAM）

有些工作在改进模块的同时，对编码器-解码器结构进行了设计，如级联了多个编码器-解码器结构的 **shelfnet** [23]、将图片金字塔输入共享编码器并行计算的 **swiftnetRN** [24] 和回归简单编码器-解码器结构并使用多尺度头的 **FFNet** [25] 等。

引入注意力机制和 Transformer 块 注意力机制作为一种非局部方法，具有全局感受野，在建模全局相关性上很有效，已经在语义分割领域广泛使用。一些实时语义分割的工作创新了计算注意力的方法，降低其复杂度，提出了一些非局部模块。

DFANet [26] 编码器包括三个轻量级 Xception 分支，每个块输出与下一个分支的相应块的输出连接，并在每个分支的最后加入一个 FC 注意力模块，计算全局特征并对原特征图加权。**FANet** [27] 提出了快速注意力 (FA)，在编码器和解码器之间插入 FA 模块提取全局上下文。FA 将自注意力计算中的 softmax 替换为 L2 归一化，这本身更高效，并且通过更改矩阵乘法的顺序进一步提高效率。**PP-liteseg** [28] 提出了一个特征融合模块 UAFM(Unified Attention Fusion Module)，用于解码时融合下采样特征图。UAFM 通过空间注意力和通道注意力加权上下采样的特征图，得到输出。

Vision Transformers 已经在大量视觉任务中显示出相当强大的结果。但是，尽管取得了成功，基于自注意力机制的 Transformer 架构需要强大的计算资源，这超出了许多移动和嵌入式设备的能力。近两年，出现了许多简化 Transformer 块以适应实时语义分割的工作，取得了很好的效果。

TopFormer [29] 用 Transformer 块连接编码器和解码器，先使用级联的 mobilenet 块生成层次特征图，下采样到统一尺寸后连接作为 Transformer 块的输入。既融合了多尺度特征，又大大减少了输入的数量，减小了开销。**RTFormer** [30] 改进了自注意力的计算，借鉴了 EA(外部注意力) 提出 GFA(GPU 友好注意力)，实现线性复杂度；还提出一个跨分辨率注意力模块，实现了一个高效双分辨率网络。**AFFormer** [66] 利用聚类提取原型特征，用 Transformer 块对原型特征进行特征提取再恢复到原尺寸，因此可以保持一个较高分辨率的特征图，省去了解码器。并且从频率的角度解释自注意力，使用 AFF(Adaptive Frequency Filter) 计算自注意力，实现线性 Transformers，和聚类降维一起提高了效率，实验效果特别好。

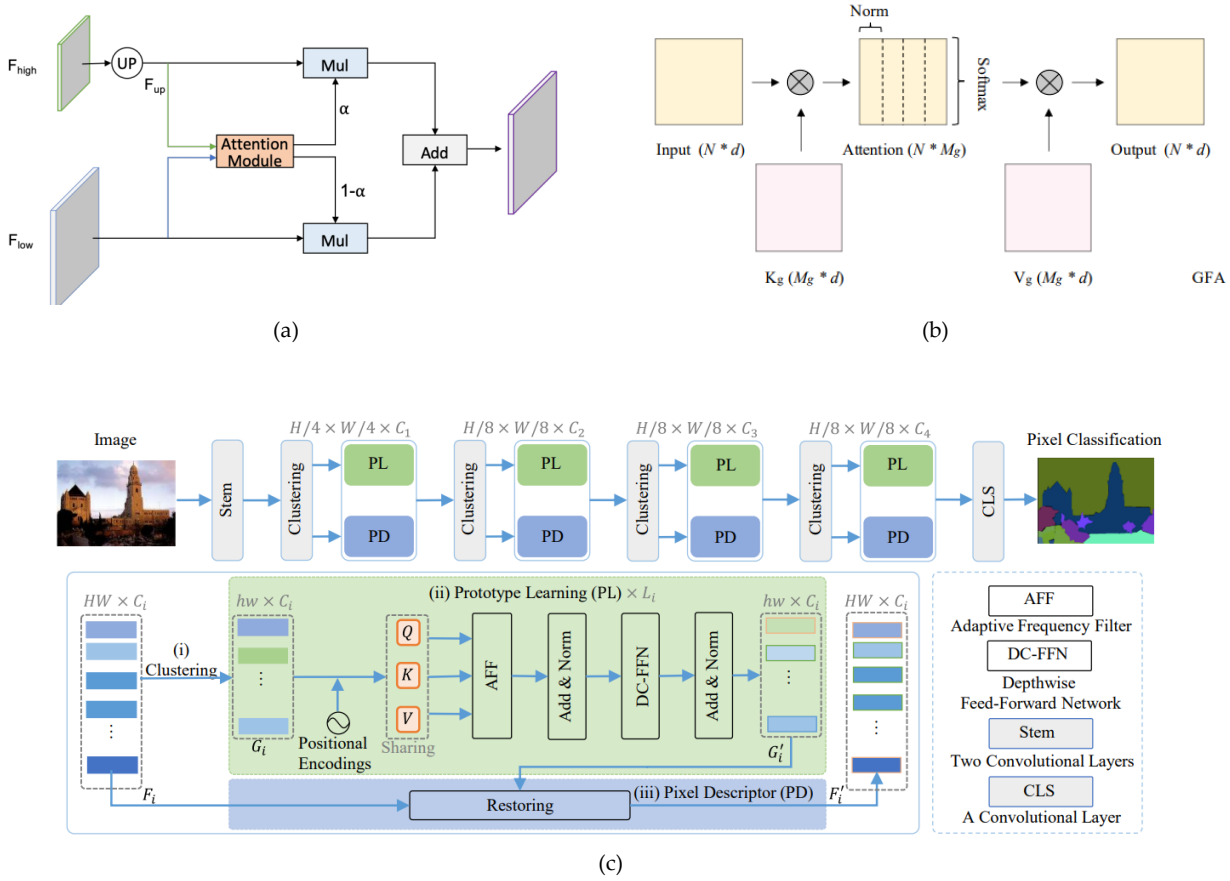


图 9: (a)PP-liteseg [28] 中的特征融合模块；(b)RTFormer [30] 中的 GPU 友好注意力 (GFA)；(c)AFFormer [66]

2.2.2 多分支结构

语义信息和细粒度信息的结合是追求语义分割精度的重要挑战，编码器-解码器结构的方法是通过在下采样提取语义信息，并在上采样过程中恢复细节信息。而实时分割领域还有另一种主流方法，即多分支结构，旨在通过多个分支独立提取不同尺度的特征来解决这个问题。

ICNet [31] 包括三个独立编码器分支，分别输入尺寸不同的图像。低分辨率分支使用较深的完整网络提取语义信息，中高分辨率分支利用简单的轻量级架构来提取更细粒度的信息以细化输出边界，大大减小参数量，最终用级联特征融合单元融合不同层次的特征。**ContextNet** [32] 利用两个分支有效地提取空间和上下文特征，上下文分支输入分辨率较小的图片，相对较深，空间分支输入高分辨率图片，保留丰富的空间特征。

BiSeNet [6] 提出了一种经典的双分支架构，由上下文路径和空间分支组成，如图 8 所示。空间路径非常简单，仅使用三个跨步卷积块，而上下文路径基于 Xception 主干快速下采样，扩大感受野，获得较低分辨率的含丰富语义特征的特征图。并设计了注意力细化模块 (ARM) 应用于上下文分支的最后两个阶段的输出，使用全局平均池化生成编码全局上下文的特征向量，然后重新加权特征图，两个 AFM 模块的输出通过特征融合模块 (FFM) 与空间分支的输出连接在一起。**BiSeNet V2** [33] 沿用了 Bisenet v1 的设计，本文改进了语上下文路径的模块，并用双边引导聚合层取代 FFM，将两个分支的特征按不同层次分别聚合，得到更好的表征。

Fast-SCNN [38] 让双分支结构共享前几层下采样，优化了冗余，是一种非常轻量级的架构，能够以极高的帧率提供良好的分割结果。**DDRnet** [39] 优化了双分支结构，除了共享前几层下采样，还精心设计了双边信息融合模块，两个分支进行了多次信息融合，并且提出了一个新的上下文模块 DASPP，可以捕捉多尺度且扩大有效感受野。

STDC [40] 设计了一个新的 STDC 块。它由级联的卷积块组成，在这个过程中获得不同尺度的感受野，并且维度不断降低（因为语义信息更集中），最终将不同卷积块的特征图连接起来。其中卷积块的数量对参数的影

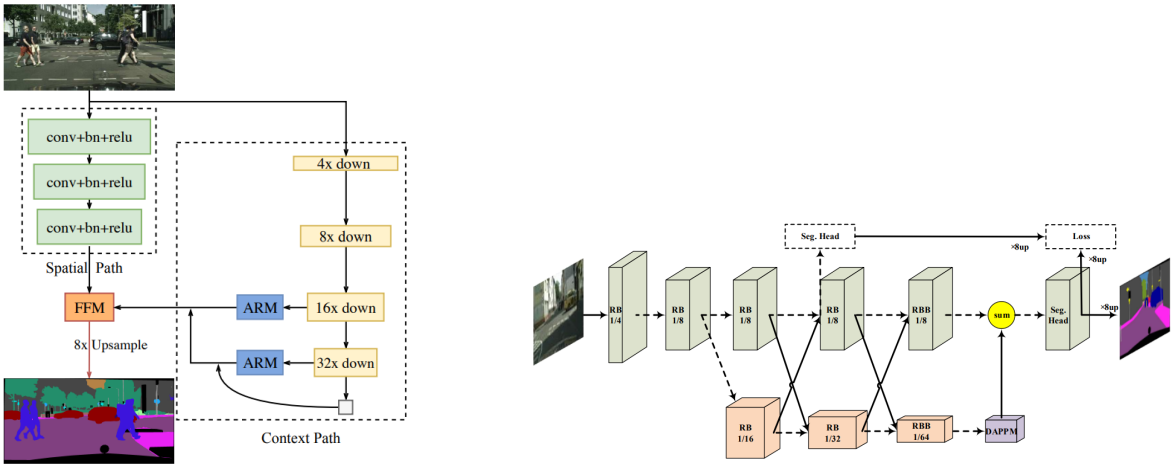


图 10: (a)BiSeNet [6]；(b)DDRNet [39]

响很小，通过 STDC 块，我们可以得到多尺度的特征，并可以通过改变 block 数量获得可扩展的感受野。通过级联 STDC 块作为双分支网络的主干，共享两个分支的前几层，提出 STDC 网络，在训练时还加入了细节分支作为辅助任务。这是一个在实时分割任务上表现很好的的主干网络，被很多后来的工作采用。

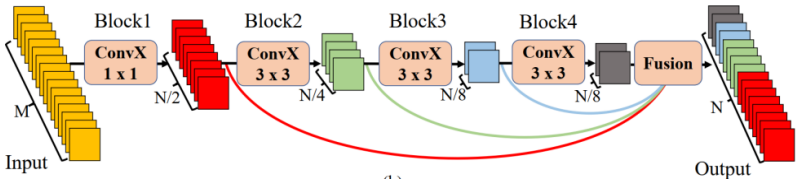


图 11: STDC [40] 中的 STDC 块

PIDNet [41] 从传统控制理论出发，认为整个语义分割网络相当于一个 PID 控制器，其中空间路径保持输入的大部分细节而相当于 P 分支，上下文路径不断聚合特征而相当于 I 分支。双分支结构在特征融合时可能因语义信息层次不同而损失表征能力，本文将其解释为 PI 控制器可能产生的“超调”现象，并引入边界路径作为 D 分支来抑制超调，用边界分支指导空间和上下文分支的特征融合，从而提出了一个三分支网络 PIDNet。PIDNet 认为 DDRNet 中的 DAPPM 模块太复杂而不能很好的并行，并且超过轻量级模型的表征能力，设计了简化的 PAPPM。借鉴 DDRNet 的思想，设计了单独的模块 Pag,Bag 进行三个分支之间的多次信息融合。为了增强模型对边界的敏感性，提高每个分支提取信息的能力，PIDNet 在训练时使用了多个辅助任务。

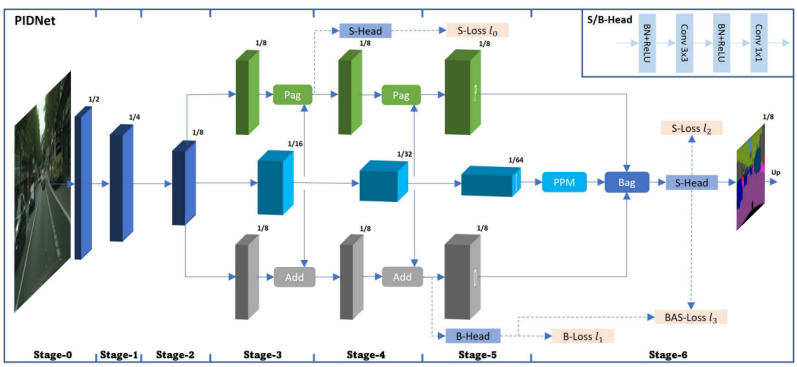
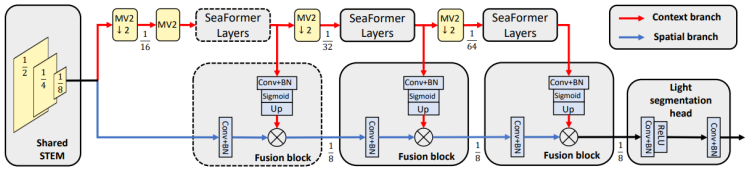
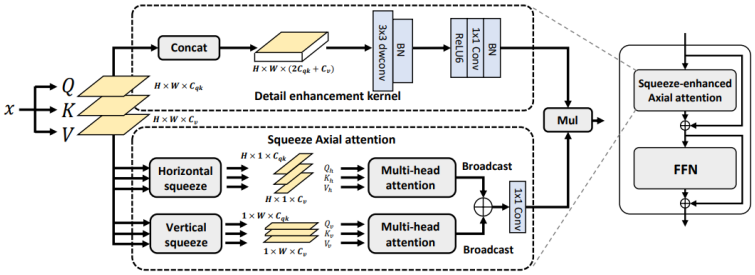


图 12: PIDNet [41]

SeaFormer [67] 使用提出的 Sea 注意力改进 Transformer 块，并将改进后的 Sea 块加入双分支结构中的上下文路径，大大提高了精度。Sea(squeeze-enhanced Axialattention) 注意力简化了自注意力的计算，将特征图按不同维度压缩，再计算压缩后的自注意力，将计算复杂度降到 O(HW)，与特征图大小呈线性。**StrideFormer** [65] 是百度在 2023 年 4 月刚发表的工作，使用了跨步 Sea 注意力，结合设计的聚合注意力模块 (AAM) 和有效插值模块 (VIM) 达到了多个数据集上的 sota。



(a)



(b)

图 13: (a)SeaFormer [67]；(b)Sea Attention

2.3 人像分割

人像分割一般被认为是语义分割的一个子问题，而它在下列三个方面不同于传统的分割。

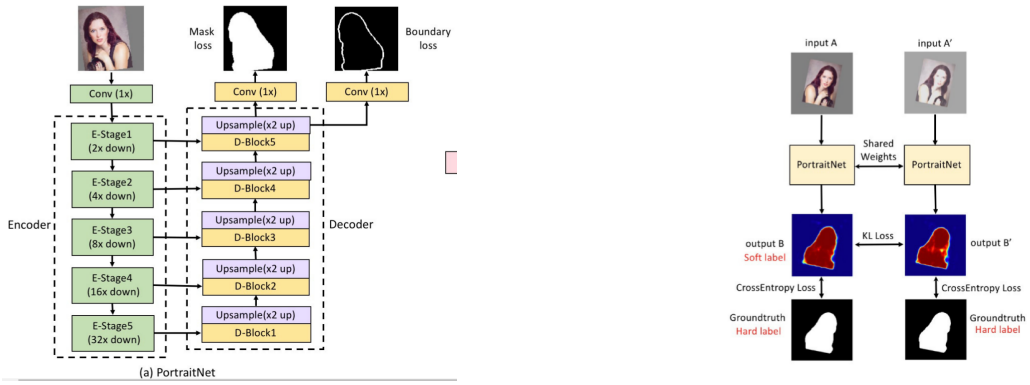
i) 人像分割是一个二元分割问题，前景对象仅为人，提供额外的先验信息；

ii) 人像分割对边界的精度要求更高，并且需要适应复杂的光照条件；

iii) 结合具体应用场景，人像分割通常仅作为实际应用中的几个步骤之一使用，由于其中许多应用程序在移动设备上运行，分割模型需要轻量级以确保实时速度

PortraitFCN+ [70] 构建了人像数据集，并提出了基于 FCN 的人像分割模型。

PortraitNet [71] 是基于 mobilenet v2 的 encoder-decoder 网络，并针对人像分割加入两个辅助损失，实现实时人像检测的精度和效率平衡。第一个辅助损失为边界损失，网络在解码器最后一层特征图后，加了一个边界的预测头，从而使分割对边界更敏感；第二个辅助损失为一致约束损失，将原图片 A 和经过纹理增强（改变亮度、对比度、锐度，加入随机噪声等）的 A′ 都输入网络，并分别预测。此时认为 A 的预测结果为更精细的分割，从而使用 KL 散度损失约束 A′ 向 A 靠拢，这可以增强网络对复杂光照环境的鲁棒性。



(a) (b)

图 14: (a)PortraitNet [71]；(b)PortraitNet 中的一致约束损失

BSN [68] 主干网络使用 Resnet 和 deeplab v2，引入了边界敏感内核来增强语义边界形状信息。**Sinet** [43] 创新了两个上下文模块，并加入了边界的辅助损失，相比 Portraitnet 大大降低了参数量，且精度损失较小。

PP-HumanSeg [42] 提出了一个超轻量级的人像分割模型 ConnectNet, 用极少的参数 (0.13M) 实现了很强的效果，关键是设计了一种新的连通性损失，实现自监督连接感知学习，让模型自我学习连通性。

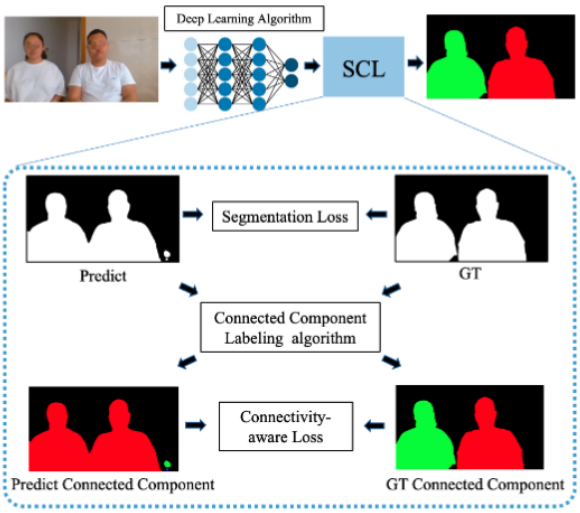


图 15: PP-HumanSeg [42] 中的自监督连接感知学习 (SCL)

3 我们的方案

为了实现虚拟背景替换，我们组的方案为

1. 设计一个高效的实时人像分割模型，并通过公开的数据集进行训练，后期可能还会采集一些国科大教室数据进行微调，在兼顾实时性的同时尽可能提高精度；

2. 将分割后的结果进行背景替换，并使用图像处理的方法使合成后的图片更加自然；

3. 搭建可现场演示的系统，进行实时虚拟背景替换

通过对相关工作的调研，我们将从下面几个方向出发设计并优化实时人像分割模型：

1. 测试目前最先进的实时语义分割方法在人像数据集上的效果，如基于编码器-解码器的 SFNet-lite, 多分支结构的 PIDNet，以 Transformer 为主干的 AFFormer 等，适应性的调整超参数后，选择效果最好的模型作为我们的 baseline

2. 向baseline中加入语义分割和实时语义分割工作中的高效特征提取模块，如Sea Attention,STDC块,PAPPM等，比较它们对效果的提升。尝试自主设计适应于人像分割任务的高效模块，并进行消融实验验证新设计模块的效果。

3. 测试不同特征融合方法对实验结果的影响，如加入 FAM 和 GD-FAM 模块, 加入边界分支等。

4. 进行多任务学习，由于人像分割的特殊性，我们将尝试在训练时加入辅助损失，如针对边界的损失、自我学习连通性的损失、针对鲁棒性的损失等，提高模型精度。

5. 收集并标注国科大教室的少量人像数据，对模型进行微调，以期在实时演示时达到更好的效果

参考文献

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

[2] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481-2495.

[3] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 234-241). Springer.

[4] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2018). DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[5] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y. (2019). HRNet: Deep high-resolution representation learning for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4854-4863).

[6] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (pp. 334-349).

[7] Paszke, A., Chaurasia, A., Kim, S., Culurciello, E. (2016). ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. arXiv preprint arXiv:1606.02147.

[8] Long, J., Shelhamer, E., Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3431-3440). IEEE.

[9] Chen, L. C., Papandreou, G., Schroff, F., Adam, H. (2018). Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587.

[10] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 6848–6856, doi: 10.1109/CVPR.2017.369.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, ”MobileNetV2: Inverted Residuals and Linear Bottlenecks,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[12] A. Howard et al., "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140.

[13] Ma, N., Zhang, X., Zheng, HT., Sun, J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision –ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11218. Springer, Cham.

[14] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, "GhostNet: More Features From Cheap Operations," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 1577-1586, doi: 10.1109/CVPR42600.2020.00165.

[15] E. Romera, J. M. Álvarez, L. M. Bergasa and R. Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, pp. 263-272, Jan. 2018, doi: 10.1109/TITS.2017.2750080.

[16] Wang, Y., Zhou, Q., Liu, J., Xiong, J., Gao, G., Wu, X., Latecki, L. J. (2019). Lednet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 3346-3350). IEEE.

[17] Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H. (2018). ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 552-568). Springer.

[18] Sachin Mehta, Mohammad Rastegari, Vikas Singh, and Hannaneh Hajishirzi. "ESPNetv2: A Light-weight, Power Efficient, and General Purpose Convolutional Neural Network." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9197-9206, 2019.

[19] Tianyi Wu, Sheng Tang, Rui Zhang, Yongdong Zhang. (2019). CGNet: A Light-weight Context Guided Network for Semantic Segmentation. IEEE Transactions on Image Processing, 29, 454-463. doi: 10.1109/TIP.2019.2934943.

[20] Haiyang Si, Zhiqiang Zhang, Feifan Lv, Gang Yu, and Feng Lu. Realtime semantic segmentation via multiply spatial fusion network. ArXiv,abs/1911.07217, 2019.

[21] Gao, R. (2021). Rethink Dilated Convolution for Real-time Semantic Segmentation. arXiv preprint arXiv:2111.09600.

[22] Wei, H., Liu, X., Xu, S., Dai, Z., Dai, Y., Xu, X. (2022). DWRSeg: Dilation-wise Residual Network for Real-time Semantic Segmentation. arXiv preprint arXiv:2212.00314.

[23] Zhuang, J., Yang, J., Gu, L., Dvornek, N. (2019). ShelfNet for Fast Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), (pp. 0-0).

[24] M. Oršic, I. Krešo, P. Bevandic and S. Šegvic, "In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 12599-12608, doi: 10.1109/CVPR.2019.01289.

[25] D. Mehta et al., "Simple and Efficient Architectures for Semantic Segmentation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 2627-2635, doi: 10.1109/CVPRW56347.2022.00296.

[26] H. Li, P. Xiong, H. Fan and J. Sun, "DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9514-9523, doi: 10.1109/CVPR.2019.00975.

[27] P. Hu et al., "Real-Time Semantic Segmentation With Fast Attention," in IEEE Robotics and Automation Letters, vol. 6, no. 1, pp. 263-270, Jan. 2021, doi: 10.1109/LRA.2020.3039744.

[28] Juncai Peng, Yi Liu, Shiyu Tang, Yuying Hao, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen,Zhiliang Yu, Yuning Du, Qingqing Dang,Baohua Lai, Qiwen Liu, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma.Pp-liteseg: A superior real-time semantic segmentation model. ArXiv,abs/2204.02681, 2022.

[29] Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G., Shen, C. (2022). TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12083-12093).

[30] J. Wang, C.-x. Gou, Q. Wu, H. Feng, J. Han, E. Ding, and J. Wang, "RTFormer: Efficient Design for Real-Time Semantic Segmentation with Transformer," arXiv preprint arXiv:2210.05861, Oct. 2022.

[31] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, Jiaya Jia. "ICNet for Real-Time Semantic Segmentation on High-Resolution Images." Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 405-420.

[32] Poudel, R. P. K., Bonde, U. D., Liwicki, S., Zach, C. (2018). ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-time. In Proceedings of the British Machine Vision Conference (pp. 139.1-139.14).

[33] Yu, C., Gao, C., Wang, J. et al. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. Int J Comput Vis 129, 3051–3068 (2021).

[34] Li, X. et al. (2020). Semantic Flow for Fast and Accurate Scene Parsing. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision –ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12346. Springer, Cham.

[35] Li, X., Zhang, J., Yang, Y., Cheng, G., Yang, K., Tong, Y., Tao, D. (2022). SFNet: Faster, Accurate, and Domain Agnostic Semantic Segmentation via Semantic Flow. arXiv preprint arXiv:2207.04097.

[36] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang and H. Shi, "AlignSeg: Feature-Aligned Segmentation Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 550-557, 1 Jan. 2022, doi: 10.1109/TPAMI.2021.3062772.

[37] S. Huang, Z. Lu, R. Cheng and C. He, "FaPN: Feature-aligned Pyramid Network for Dense Image Prediction," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 844-853, doi: 10.1109/ICCV48922.2021.00090.

[38] Poudel, R. P. K., Liwicki, S., Cipolla, R. (2019). Fast-SCNN: Fast Semantic Segmentation Network. In Proceedings of the British Machine Vision Conference (BMVC) (pp. 1-13).

[39] Hong, Y., Pan, H., Sun, W., Jia, Y. (2021). Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes. ArXiv.

[40] Rethinking BiSeNet for Real-Time Semantic Segmentation.Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, Xiaolin Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9716-9725

[41] Xu, J., Xiong, Z., Bhattacharyya, S. (2022). PIDNet: A Real-time Semantic Segmentation Network Inspired by PID Controllers. arXiv preprint arXiv:2206.01547.

[42] Chu, L., Liu, Y., Wu, Z., Tang, S., Chen, G., Hao, Y., Peng, J., Yu, Z., Chen, Z., Lai, B., Xiong, H. (2022). PP-HumanSeg: Connectivity-Aware Portrait Segmentation With a Large-Scale Teleconferencing Video Dataset. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops (pp. 202-209).

[43] Park, H., Sjosund, L., Yoo, Y., Monet, N., Bang, J., Kwak, N. (2020). SINet: Extreme Lightweight Portrait Segmentation Networks with Spatial Squeeze Module and Information Blocking Decoder. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2066-2074.

[44] W. Weng and X. Zhu, "Inet: Convolutional networks for biomedical image segmentation," IEEE Access, vol. PP, no. 99, pp. 1–1, 2021.

[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," IEEE Computer Society, 2016.

[46] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2016.

[47] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[48] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019, pages 1971–1980. IEEE, 2019.

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[50] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 3146–3154. Computer Vision Foundation / IEEE, 2019.

[51] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 15979–15988. IEEE, 2022.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society, 2016.

[53] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 603–612. IEEE, 2019.

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 9992–10002. IEEE, 2021.

[55] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 12159–12168. IEEE, 2021.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[57] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. CoRR, abs/2105.05633, 2021.

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 1–9. IEEE Computer Society, 2015.

[59] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 7794–7803. Computer Vision Foundation / IEEE Computer Society, 2018.

[60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 12077–12090, 2021.

[61] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 7151–7160. Computer Vision Foundation / IEEE Computer Society, 2018.

[62] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX, volume 11213 of Lecture Notes in Computer Science, pages 270–286. Springer, 2018.

[63] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 6881–6890. Computer Vision Foundation / IEEE, 2021.

[64] Zhen Zhu, Mengdu Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 593–602. IEEE, 2019.

[65] Shiyu Tang, Ting Sun, Juncai Peng, Guowei Chen, Yuying Hao, Manhui Lin, Zhihong Xiao, Jiangbin You, Yi Liu（2023）.PP-MobileSeg: Explore the Fast and Accurate Semantic Segmentation Model on Mobile Devices. https://doi.org/10.48550/arXiv.2304.05152

[66] Bo Dong, Pichao Wang, and Fan Wang. Head-free lightweight semantic segmentation with linear transformer. CoRR, abs/2301.04648, 2023.

[67] Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. CoRR, abs/2301.13156, 2023.

[68] Xianzhi Du, Xiaolong Wang, Dawei Li, Jingwen Zhu, Serafettin Tasci, Cameron Upright, Stephen Walsh, and Larry S. Davis. Boundary-sensitive network for portrait segmentation. In 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019, pages 1–8. IEEE, 2019.

[69] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.

[70] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian L. Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. Comput. Graph. Forum, 35(2):93–102, 2016.

[71] Song-Hai Zhang, Xin Dong, Hui Li, Ruilong Li, and Yong-Liang Yang. Portraitnet: Real-time portrait segmentation network for mobile device. Comput. Graph., 80:104–113, 2019.

[72] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[73] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. CoRR, abs/1706.05587, 2017.

[74] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, volume 11211 of Lecture Notes in Computer Science, pages 833–851. Springer, 2018.

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017.