

硕 士 研 究 生 读 书 报 告



题目 《ACV-Net》读书报告

作者姓名 周杰

作者学号 22251177

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 二○二二年十二月

摘要

双目深度估计一直是计算机视觉领域的一大经典问题，在辅助驾驶、三维重建等方面都有广泛的应用场景。传统的基于深度学习的深度估计模型通过卷积层的堆叠，在增加模型参数的同时并没有获得很好的精度提升。论文提出的ACV-Net模型，利用注意力机制，对连接体的冗余信息进行过滤，得到轻量级的网络模型的同时获得了明显的精度提升，在各大深度估计数据集上都有很好的表现。

关键词：深度估计 注意力机制 深度学习

1引言

近些年，增强现实、自动驾驶等领域的研究火热，深度估计是其中不可或缺的一个研究分支。深度估计通过在左右图像中寻找匹配的像素点，计算目标位置的视差，完成立体匹配的任务。尽管这一领域研究广泛，但少有模型能够同时达到实际应用需要的精度和计算效率。

目前，各类基于CNN的模型实现深度估计的流程可以分为四步：特征提取、代价立方体构建、代价聚合和视差回归。其中，代价立方体的构建是决定模型最终效能的关键。在代价立方体的构建方面，目前的主流算法分为三类：基于连接体(Concatenation Volume)、基于相关体(Correlation Volume)以及两者的结合。

基于连接体的模型代表是GC-Net[1]，将左右图像的特征图进行连接以构建4D代价立方体。这类方法忽略了左右图像之间的相似性度量，需要大量的3D卷积对代价立方体进行处理。

基于相关体的模型代表是DispNetC[2]，通过计算左右特征映射之间的单通道全相关量构建代价立方体。这类方法为左右图像之间提供了有效的相似性度量，却忽略了图像的内容信息。

将二者结合的模型代表是GwcNet[3]，将相关体进行分组并连接构建连接体。然而，不加考虑的直接连接并没有带来效率的优化，模型仍然需要大量的3D卷积层进行代价聚合。

基于当前的研究现状，作者提出了两个假设：连接体包含丰富但冗余的内容信息，而相关体可以衡量左右图像特征相似性，隐含地反映图像中相邻像素之间的关系。

基于上述结论，作者提出了ACV-Net[4]，利用相关体编码像素关系帮助连接体抑制冗余信息，巧妙利用连接体的信息改进模型精确度的同时，还利用了相关体的信息降低了模型复杂度，提高了运算效率。

2算法介绍

2.1基于注意力机制的连接体构建

2.1.1初始连接体构建

给定一个大小为H×W×3的输入立体图像对，对于每张图像，分别从CNN特征提取中获得左右图像的一元特征图和。特征图大小为 Nc×H/4×W/4 (Nc=32)。然后通过连接每个视差级别的和来形成初始拼接体，如公式1所示。

|  |  |
| --- | --- |
|  | (1) |

连接体的尺寸为2Nc×D/4×H/4×W/4，其中D为最大视差值。

2.1.2注意力权重生成

注意力权重旨在过滤初始拼接体，以强调有用信息并抑制不相关信息。传统的相关体是通过计算像素之间的相似性获得的，但是这种方法对于无纹理区域并不适用。为此，文章提出了鲁棒性更强的多级自适应块匹配(Multi-level Adaptive Patch Matching)(MAPM)方法。

首先，从特征提取模块获得三个层级的特征图、和，通道数分别为64，128和128。对于不同层级的特征图，使用不同感受野和扩张率的卷积核对其特征进行提取，并进行相关性度量，如图1所示。

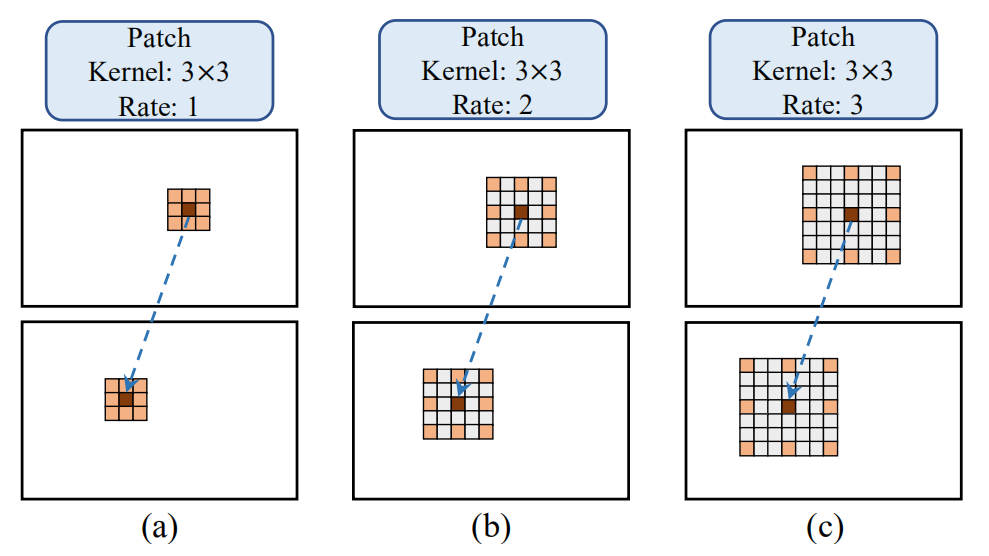


图1 MAPM机制示例

图1中，(a)、(b)、(c)分别代表、和的特征图，卷积核的感受野和特征图层级相关，特征图层级越高，感受野越大。由于使用空洞卷积，每个像素提取的特征只与本身和周围的八个像素点相关。

其次，将、和三个级别的特征图连接起来形成通道一元特征图(=320)，并平均分为组(=40)，其中前8组来自，中间16组来自，最后16组来自，不同层次的特征图不会相互干扰。

将第g个特征组表示为、，并且多级匹配代价计算如公式2和公式3所示。

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |

其中，表示不同特征层的匹配代价；表示级特征图上卷积核的权重，并在训练过程中自适应地学习；代表内积运算；代表像素点坐标；为候选视差值；为卷积核中被加权采样的像素点集合。

然后，通过拼接三个级别的匹配代价值，得到多级的匹配代价体，如公式4所示。

|  |  |
| --- | --- |
|  | (4) |

最后，使用2个3D卷积和1个3D沙漏网络来正则化，然后用另一个卷积层把通道数压缩为1得到注意力权重。为了获得不同视差的准确注意力权重来过滤初始拼接体，我们使用视差真值来监督。用soft argmin方法从中得到基于权重的视差值，计算和真实值的平滑L1损失值。

2.1.3注意力权重过滤

获得注意力权重后，用于消除初始拼接体中的冗余信息，进而增强其表示能力。通道处的注意力级联体的计算如公式5所示。

|  |  |
| --- | --- |
|  | (5) |

其中，代表点乘操作，注意力权重将应用于初始拼接体的所有通道。

2.2整体网络构建

基于上述注意力连接体机制，构造ACV-Net模型，结构如图2所示，一共分为特征提取、注意力连接体构造、代价聚合和视差预测四个部分，接下来对四个部分进行展开介绍。

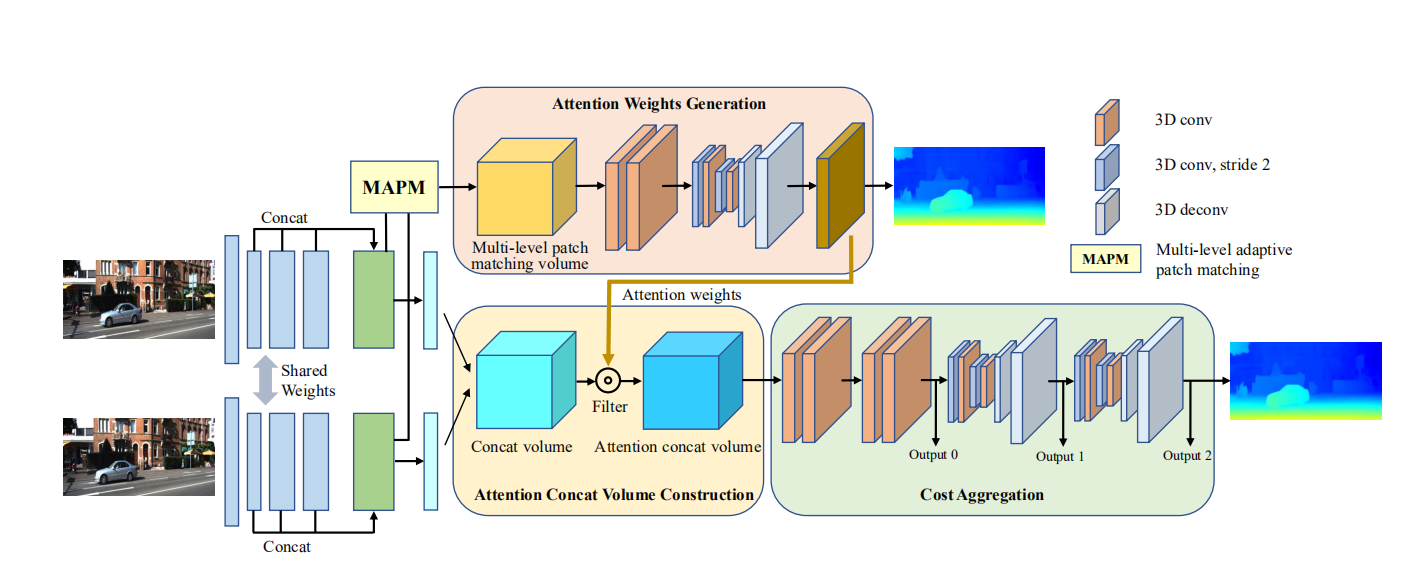


图2 ACV-Net整体架构

2.2.1特征提取

在网络的前3层，使用3个卷积层对图片进行下采样，得到不同层级的特征图、和，将三者拼接形成320通道的特征图，用于生成注意力权重。然后，再通过两次卷积压缩到32通道，得到和，用于构建初始连接体。

2.2.2注意力连接体构造

这一部分的工作主要是生成注意力权重，对初始连接体进行过滤，并最终生成一个4D的代价立方体。

2.2.3代价聚合

这一部分包括4个3D卷积层，两个Encoder-Decoder结构的沙漏网络，用于处理上个模块生成的代价立方体。

2.2.4视差预测

代价聚合模块获得三个输出，通过3D卷积层得到单通道代价立方体，并通过上采样和softmax输出置信值。最终视差值的计算如公式6所示，其中代表候选视差值，代表置信值。

|  |  |
| --- | --- |
|  | (6) |

2.2.5损失函数

使用L1损失函数，计算如公式7所示，其中代表注意力生成模块的输出，代表视差预测层的三个输出结果，为视差真值。

|  |  |
| --- | --- |
|  | (7) |

3实验结果

3.1精确度消融实验

通过设置消融实验，作者对ACV-Net的两个重要模块进行了评估，结果如表1所示。

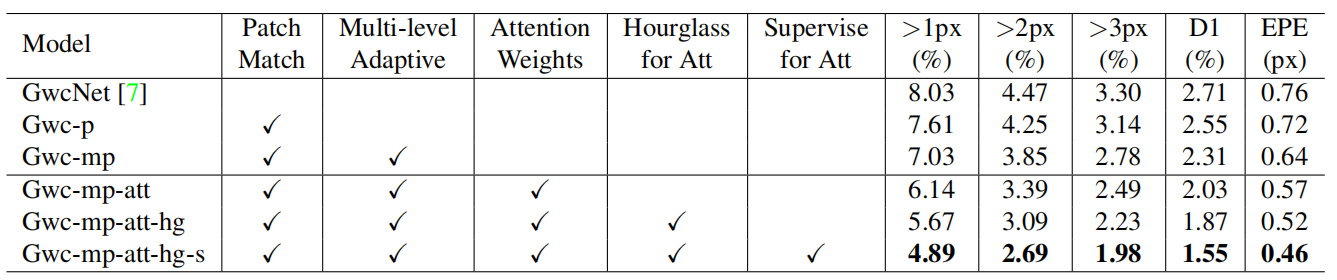


表1 消融实验结果

作者以GwcNet为基准，评估了MAPM模块的效能。结果表明，加入了MAPM模块的模型能够显著降低损失值。

为了评估ACV模块的效果，作者设计了三组ACV模块，如图3所示。

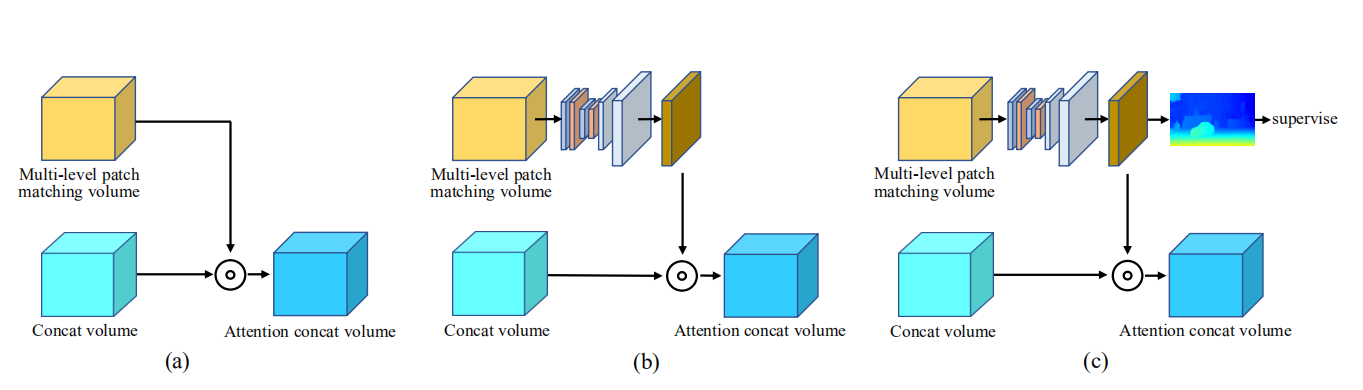


图3 ACV模块的消融实验

同样以GwcNet为基准，作者评估了三类ACV模块的最终效能。从表1中可以得出，加入ACV模块以后，模型的损失值得到了明显降低，在加入沙漏模块和真值监督环节之后，损失值还有进一步的降低，最终模型在D1和EPE上分别比GwcNet提高了42.8%和39.5%，证明了ACV的有效性。

3.2模型复杂度分析

如前文所述，作者在解决模型精确度问题的同时，还对模型的复杂性进行了分析，结果如表2所示。

同样以GwcNet为基准，作者评估了聚合模块的沙漏网络层数对输出结果的影响。结果表明，即使加入两个沙漏模块，模型的总参数量依然低于GwcNet，同时在精确度上仍然有所提高，最终作者选择2个沙漏模块作为最终模型。

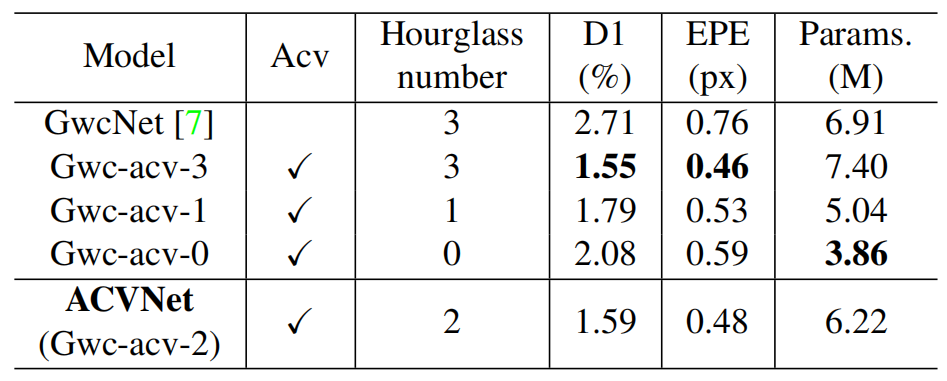


表2 模型复杂度分析

3.3定量分析

ACV-Net在KITTI2012和KITTI2015两个数据集上的表现如表3所示。

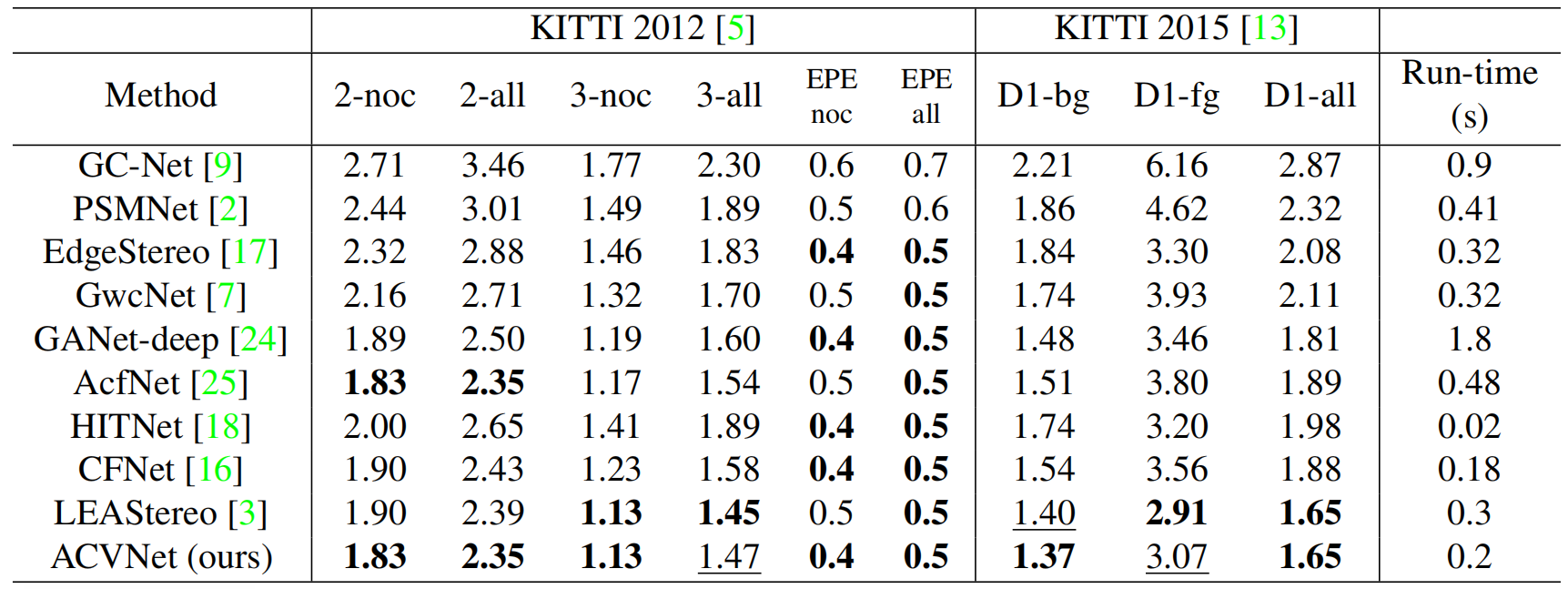


表3 模型定量分析

结果表明，ACV-Net在两个数据集上比起当前最先进的算法提升了38.4%的精确度，同时拥有更快的计算速度。

模型在两个数据集上的输出结果如图4所示。

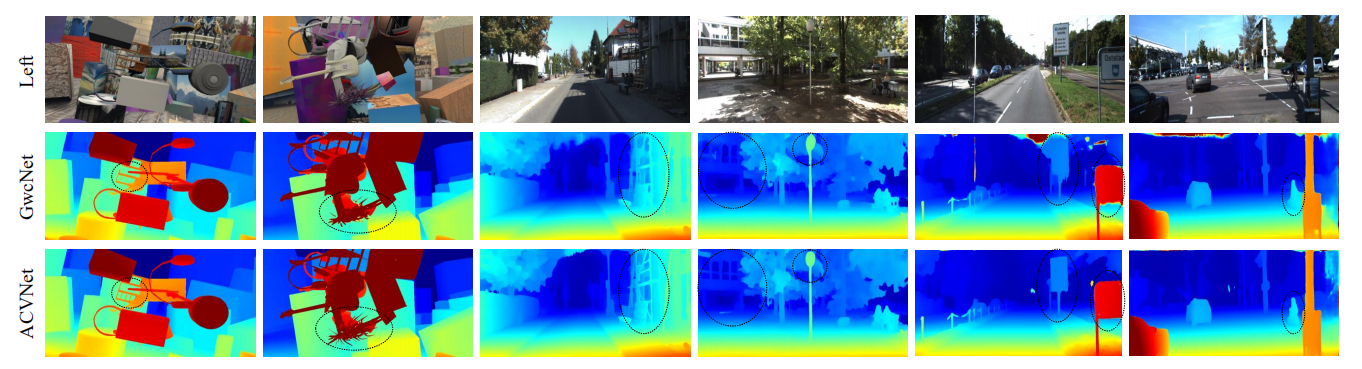


图4 模型最终输出效果

4总结

论文提出了一种新的代价立方体，作者称之为注意力机制的连接体，即ACV，同时提出了多级自适应块匹配机制，并基于这些机制设计了全新的模型架构ACV-Net。最终，模型在KITTI数据集上取得了优秀的成绩，在一定程度上解决了目前模型普遍存在的精度较低、速度较慢的问题，为今后的研究做出了重要贡献。

参考文献

[1] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-End Learning of Geometry and Context for Deep Stereo Regression: IEEE, 10.1109/ICCV.2017.17[P]. 2017.

[2] Mayer N, Ilg E, Hausser P, et al. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[3] Guo X, Yang K, Yang W, et al. Group-Wise Correlation Stereo Network[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[4] Xu G, Cheng J, Guo P, et al. ACVNet: Attention Concatenation Volume for Accurate and Efficient Stereo Matching[J]. 2022.