Modeling and Simulation 建模与仿真, 2024, 13(3), 2868-2880
Published Online May 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.133260

Hans汉斯

混合CNN和ViT的自监督知识蒸馏单目深度估计方法

郑千惠[1]，孔玲君[2]

[1]上海理工大学出版印刷与艺术设计学院，上海
[2]上海出版印刷高等专科学校，上海

收稿日期：2024年4月22日；录用日期：2024年5月21日；发布日期：2024年5月29日

摘　要

单目深度估计是一项具有挑战性的任务，现有的方法无法高效利用特征的长程相关性和局部信息。针对该问题，本文提出一种混合CNN和ViT (Vision Transformer)的自监督知识蒸馏单目深度估计方法HCVNet。HCVNet对CNN和Vision Transformer的有效组合进行研究，设计了CNN-ViT混合特征编码器，来建模局部和全局上下文信息，提取更具场景表达性的细节特征。采用通道特征聚合模块来捕获长距离依赖，通过在通道维度上聚合区分度高的特征，来增强场景结构的感知能力。引入自监督知识蒸馏，利用结构相同的教师模型为学生模型的训练提供更多监督信号，进一步提高网络性能。在KITTI和Make3D数据集上的实验结果表明，本方法的深度估计性能优于目前的主流方法，且具有较强的泛化能力，能够更好地估计出结构完整细节清晰的深度图。

关键词

单目深度估计，自监督学习，知识蒸馏，Vision Transformer

Hybrid CNN and ViT for Self-Supervised Knowledge Distillation Monocular Depth Estimation Method

Qianhui Zheng[1], Lingjun Kong[2]

[1]College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai
[2]Shanghai Publishing and Printing College, Shanghai

Received: Apr. 22nd, 2024; accepted: May. 21st, 2024; published: May. 29th, 2024

文章引用: 郑千惠, 孔玲君. 混合 CNN 和 ViT 的自监督知识蒸馏单目深度估计方法[J]. 建模与仿真, 2024, 13(3): 2868-2880. DOI: 10.12677/mos.2024.133260

Abstract

Monocular depth estimation is a challenging task, and existing methods cannot efficiently utilize feature long-range correlation and local information. To address this problem, this paper proposes HCVNet, a hybrid CNN and ViT (Vision Transformer) method for self-supervised knowledge distillation monocular depth estimation. HCVNet investigates the effective combination of CNN and Vision Transformer, and designs a hybrid CNN-ViT feature encoder to model local and global contextual information and extract more scene-expressive detailed features. Channel feature aggregation module is employed to capture long-range dependencies and enhance the perception of scene structure by aggregating discriminative features in the channel dimension. Self-supervised knowledge distillation is introduced to provide more supervised signals for the training of student models using structurally identical teacher models to further improve network performance. Experimental results on KITTI and Make3D datasets confirm that the depth estimation performance of this method is better than the current mainstream methods and has strong generalization ability, which can better estimate the depth map with complete structure and clear details.

Keywords

Monocular Depth Estimation, Self-Supervised Learning, Knowledge Distillation, Vision Transformer

1. 引言

单目深度估计[1]是计算机视觉的一项基础研究任务。高质量的深度信息可以为各个领域提供有用的信息，包括 3D 场景重建[2]和自动驾驶[3]等领域。虽然有监督的单目深度估计方法[4]已经可以取得不错的效果，但是它们受到地面实况数据收集成本高昂的限制。自监督方法[5]作为一种替代方案，可以在没有深度标签的情况下使用立体图像对或单目图像序列训练深度估计网络，该方法大大减轻了对深度标签的需求。其中，在利用立体图像对[6]时，摄像机的运动是已知的，因此可采用单个深度估计网络来预测深度。而仅使用单目图像序列[7]的训练需要联合估计深度图和相机的自我运动来计算光度损失的变换矩阵，这需要额外的姿态估计网络。尽管如此，使用单目图像序列的方法仍然更具有吸引力，因为它不需要额外的传感器，且无需复杂的立体数据处理操作。因此，本文将使用单目图像序列训练深度估计网络。

然而，仅使用单目图像序列的训练方法高度依赖于静态场景和朗伯表面假设，在存在一些动态对象的场景下(例如行人或汽车)，其模型性能会受到严重影响。为了缓解这些问题，Godard 等人[7]通过引入最小重投影损失和自动掩蔽来防止遮挡并隐藏场景中的移动物体。但是，该方法未能在边界模糊、形状复杂、高反射和颜色饱和的区域预测出精确的深度结果。Lyu 等人[8]在 Monodepth2 的基础上，研究用于高分辨率自监督单目深度估计的网络结构 HR-Depth，该网络重新设计了 DepthNet 的跳层来帮助预测更清晰的边缘。但是该网络对像素遮挡区域的深度信息估计不准确。为缓解该问题，Wang 等人[9]设计一种高质量深度解码器 HQDec，利用自适应细化模块建模像素依赖关系，以最大程度恢复深度图精细度。但该方法需要在已知相机内在参数的情况下训练，无法输出清晰度稳定的深度图。Ren 等人[10]利用单目图像序列训练自适应知识蒸馏框架 MUSTNet，并提出协同教学损失来实现更准确的深度估计。但是，该

方法在蒸馏过程中需要分别完成教师模型和学生模型的训练，整个网络训练资源占用率高、效率低。虽然这些方法取得了可观的效果，但很少有研究来改进深度估计网络本身的架构。使用卷积神经网络(CNN)[11]进行特征提取在现有的有监督和无监督训练中占主导地位。然而 CNN 缺乏对全局信息的感知力，卷积操作会使得输入数据的位置信息丢失，降低了对远处物体深度估计的准确性。Vision Transformer(ViT) [12] [13]最近在目标检测和语义分割等任务上表现出出色的结果。ViT 能够对像素之间的远程关系进行建模，从而获得全局感受野。但是传统的 ViT 架构[14]中的多头自注意力(MHSA)模块具有昂贵的计算成本，对硬件要求高。

为了获得整体场景更高精度深度估计结果的同时降低资源占用率，本文提出了一种混合 CNN 和 ViT 的自监督知识蒸馏单目深度估计方法 HCVNet (Hybrid CNN and ViT for Self-Supervised Knowledge Distillation Monocular Depth Estimation Method)。设计 CNN-ViT 混合的特征编码器，捕捉细粒度高的局部和全局信息，通过重新排列卷积操作缓解内存访问成本高的问题；设计通道特征聚合模块，从通道层面上增强场景结构的感知，解决深度不连续的远处区域估计不清晰的问题；引入单阶段同步知识蒸馏[15]知识，将同一个模型的前一轮训练作为教师模型来蒸馏当前训练的学生模型，避免了同时训练两个模型效率低下的问题。实验结果表明 HCVNet 在真实户外场景数据集 KITTI [16]和 Make3D [17]上展现了优秀的性能与泛化能力。

2. 研究方法

2.1. HCVNet 网络的基本原理

本文提出混合 CNN 和 ViT 的自监督知识蒸馏单目深度估计方法 HCVNet，框架结构如图 1 所示，主要由深度网络、姿态网络和损失函数三个部分组成。深度网络包含教师和学生两个分支，教师和学生模型都为具有跳跃连接的编码器－解码器架构。在该架构中，特征编码器以 CNN-ViT 混合网络为主干从输入图像 I_{t-1}, I_t, I_{t+1} 中提取具有表达力的空间信息特征。此后特征图进入通道特征聚合模块，该模块增强场景结构感知，从而生成新的特征。新特征经过深度解码器的四次上采样和卷积层处理，最终输出深度图 D_t，由 D_t 可计算得到平滑度损失。教师模型利用多视图检查过滤器来过滤掉视图中的异常值 M，得到过滤后的深度图 D_{tch}。由 M 和 D_t 可得到学生分支的输出深度图 D_{stu}，D_{tch} 和 D_{stu} 可被用来计算自蒸馏损失。姿态网络以 ResNet18 为主干，使用三个 1 × 1 卷积层来估计相邻帧之间的相机运动。网络通过执行投影和重投影操作得到合成目标图像 $\hat{I}_t$，根据 $\hat{I}_t$ 和 I_t 可计算光度损失。平滑度损失、自蒸馏损失和光度损失为网络训练提供更多监督信号，最终得到更高精度的深度图。

2.2. CNN-ViT 混合特征编码器

现有深度估计任务中的 CNN 模型[7] [8]只是简单的将当前层的输入特征和输出特征相加并送入下一层，来提取局部特征，难以捕获高精度全局特征。Vision Transformer 模型[18]有较强的长程依赖提取能力，但具有巨大的计算开销。为了高效提取到细粒度高的局部和全局信息，本文结合 CNN 和 Vision Transformer 的优点，提出了 CNN-ViT 混合特征编码器。CNN-ViT 混合特征编码器的结构如图 2 所示，它包含一个 Conv Stem 和四个 Stage，分别作用在不同的尺度上。

为了捕获更多低层局部细节信息，CNN-ViT 特征编码器从一开始就设置具有多个卷积层的 Conv Stem 来提取图像特征。如图 2(a)所示，Conv Stem 接收大小为 $3\times H\times W$ 的输入图像。该图像经过第一个步长为 2 的 3 × 3 常规卷积层和 3 × 3 深度可分离卷积层处理后，获得低级特征。1 × 1 Conv 层和 3 × 3 DWConv 层被用来对低级特征进行二次提取。此外，Conv Stem 在 3 × 3 Conv 层和 3 × 3 DWConv 层中间设置跳跃连接来避免信息丢失。最终得到大小为 $C\times\frac{H}{2}\times\frac{W}{2}$ 的特征 X_0。

为了改善计算延迟，该编码器重新排列卷积操作并删除部分非线性激活函数，在 Conv Stem 后设计

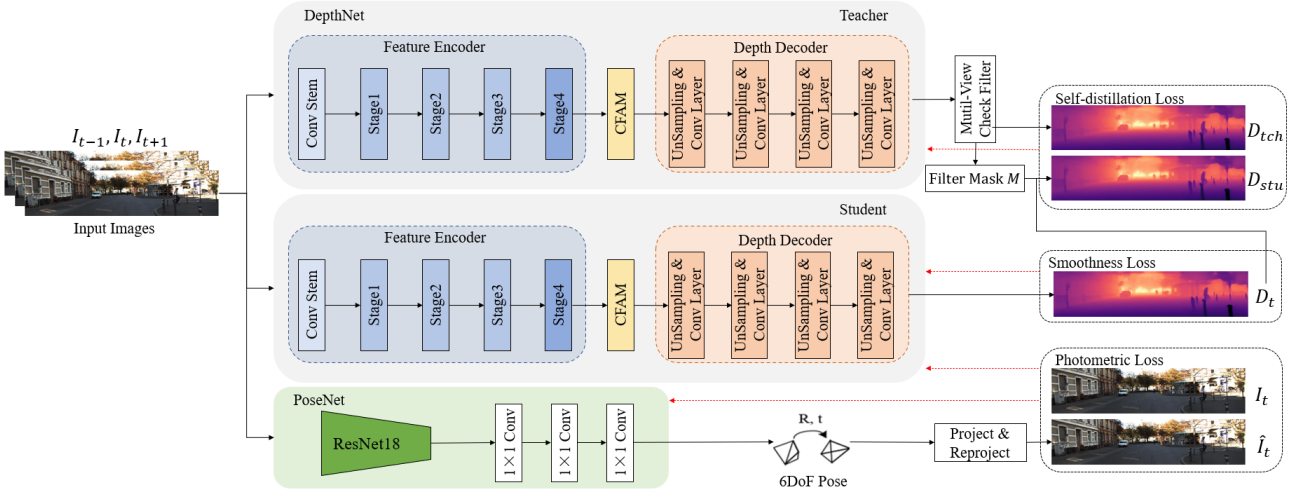

Figure 1. The overview of HCVNet
图 1. HCVNet 结构总览

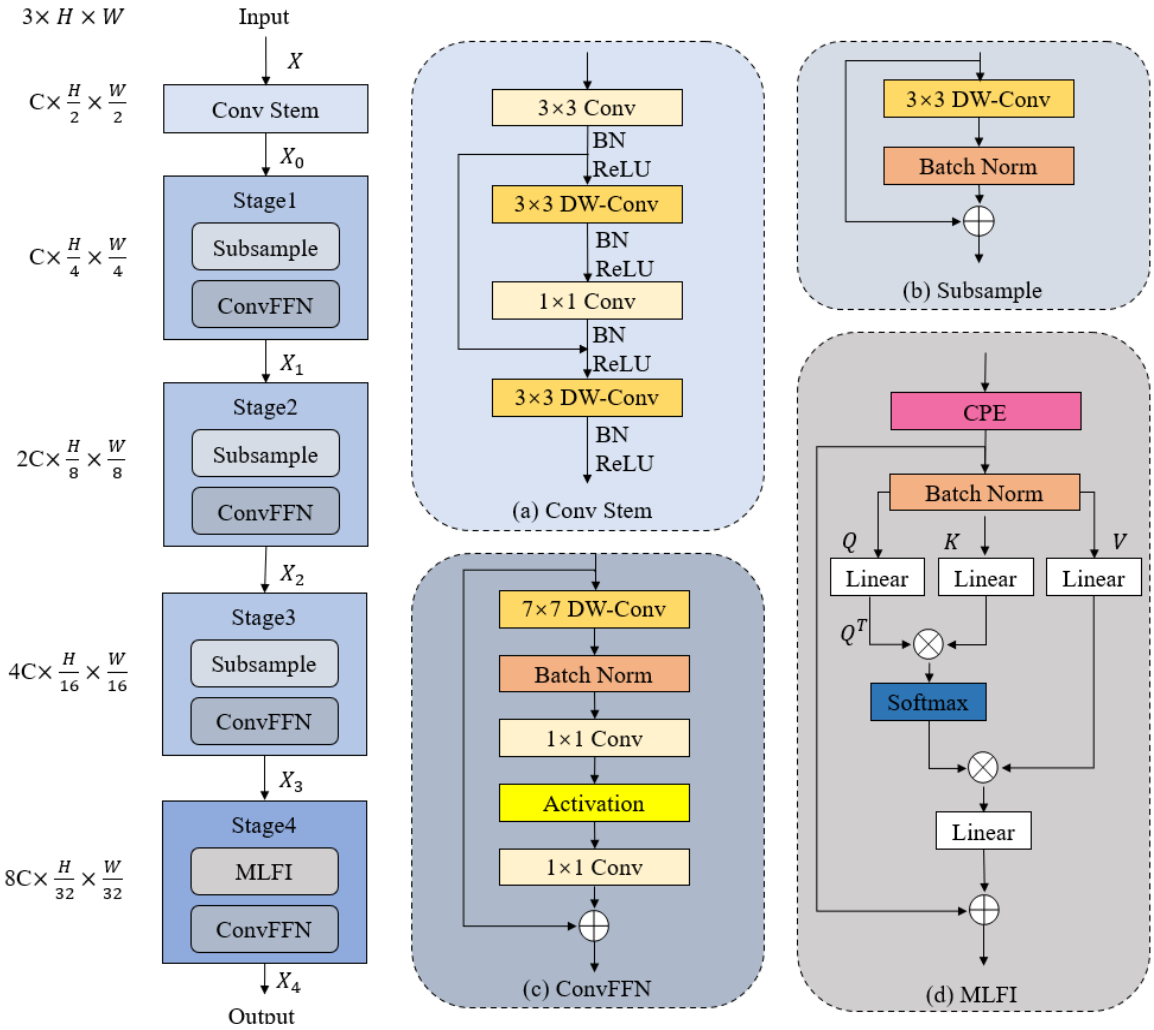

Figure 2. Hybrid CNN-ViT feature encoder
图 2. CNN-ViT 混合特征编码器

了 4 个 stage。其中，前 3 个 stage 拥有相同的内部卷积架构，进一步提取局部细节信息。第四个 Stage 利用注意力机制捕获全局上下文信息。

Stage 1~3 主要包括下采样层和 ConvFFN。如图 2(b)所示，下采样层使用 3 × 3 DWConv 将空间分辨率减半，通道数量加倍。随后使用批量归一化层，提高训练稳定性。以 Stage1 为例，该训练过程可表示为公式(1)：

$$X_1 = BN\left(DWConv_{3\times3}\left(X_0\right)\right) + X_0 \tag{1}$$

为了缓解计算时跳跃连接操作对内存带来的负担，推理时采用结构重构参数化技术来简化推理计算量。Stage1 下采样层的推理过程可表示为公式(2)：

$$X_1 = DWConv_{3\times3}\left(X_0\right) \tag{2}$$

在下采样层后，本文使用 CNN 风格的 ConvFFN 来融合局部信息，其结构如图 2(c)所示。ConvFFN 包含短连接、7 × 7 DWConv、批量归一化、1 × 1 Conv、激活层和 1 × 1 Conv。研究发现，大核卷积有助于提高模型的鲁棒性和性能，因此选用 7 × 7 卷积核。由于 DWConv 具有较低的运算参数，因此选用 7 × 7 DWConv 来缓解使用大核卷积对计算带来的不利影响。批量归一化的使用有助于加强推理时特征与前一层的融合。

Stage4 由多级特征交互 MLFI(Multi Level Feature Interaction)和 ConvFFN 组成。MLFI 的结构如图 2(d)所示。输入大小为 $4C\times\frac{H}{16}\times\frac{W}{16}$ 的特征 X_3，X_3 经过条件位置编码(CPE)处理，被线性地投影到相同维度的查询 Q、键 K 和值 V。MLFI 使用互协方差注意力来增强 X_3：

$$Attention\left(Q,K,V\right) = V \otimes Soft\max\left(Q^T \otimes K\right) \tag{3}$$

$$\tilde{X}_3 = Attention\left(Q,K,V\right) \tag{4}$$

MLFI 接着增强特征的非线性，该推理过程可表示为公式(5)：

$$\hat{X}_3 = CPE\left(X_3\right) + Linear\left(\tilde{X}_3\right) \tag{5}$$

MLFI 编码长程全局上下文，弥补了 CNN 只能提取局部特征的缺点。经过 MLFI 处理后的特征进入 ConvFFN，Stage4 最终输出大小为 $8C\times\frac{H}{32}\times\frac{W}{32}$ 的特征 X_4。

2.3. 通道特征聚合模块

为了从远处区域获得更多相对深度信息，并显著增强场景的结构感知，本文提出通道特征聚合模块 CFAM (Channel Feature Aggregation Module)。如图 3 所示，CFAM 模拟通道之间的相互依赖关系，通过加权求和来聚合来自所有通道映射的特征，同时融合来自非连续区域的不同局部深度响应。

通道特征聚合模块的输入为编码器生成的特征映射 $F \in R^{C\times H\times W}$。$F$ 经过重塑得到 $F' \in R^{C\times N}$，其中 N 为像素数($N = H\times W$)。CFAM 将 F 与 F 的转置 F^T 进行矩阵乘法，计算出特征相似度 $S \in R^{C\times C}$。第 i 个通道特征和第 j 个通道特征之间的相似度 S_{ij} 可表示为公式(6)：

$$S_{ij} = F_i \otimes F_j^T \tag{6}$$

通道之间的相似度 S 反映了区域响应的空间关系，即任何两个特征图的相似度越高，它们对相同区域的响应越强。为了融合来自不同区域的更多响应，CFAM 最大化某个通道的特征得到 $\max\left(S\right)$，通过执行逐元素相减法将相似度 S 转换为区分度 $D \in R^{C\times C}$。D_{ij} 测量第 j 个通道对第 i 个通道的影响，计算过程可表示为公式(7)：

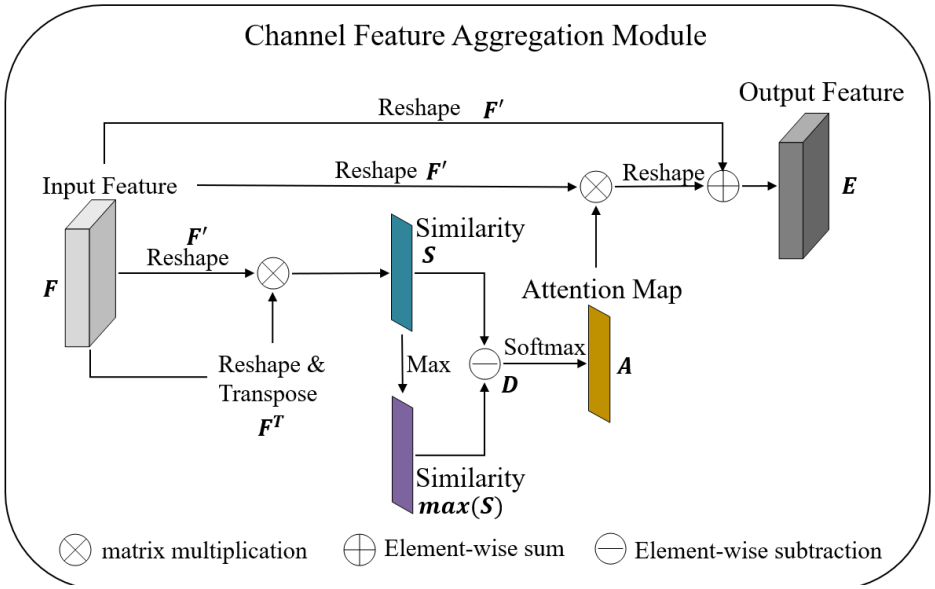

Figure 3. Channel feature aggregation module
图 3. 通道特征聚合模块

$$D_{ij} = \max_i \left(S\right) \ominus S_{ij} \tag{7}$$

对于每一个通道映射，具有不同区分特征的其他通道在特征聚合时将获得更高的区分度 D_{ij} 值。CFAM 将 D_{ij} 应用到 Softmax 层得到注意力映射 $A_{ij} \in R^{C\times C}$，见公式(8)：

$$A_{ij} = \frac{\exp\left(D_{ij}\right)}{\sum_{j=1}^{C}\exp\left(D_{ij}\right)} \tag{8}$$

此外，CFAM 在 A 和 F' 之间进行矩阵乘法，并将结果重塑为 $AF' \in R^{C\times H\times W}$。$F$ 和重塑结果进行按元素求和运算，得到最终输出 $E \in R^{C\times H\times W}$。第 i 个通道的输出特征 E_i 可表示为公式(9)：

$$E_i = \sum_{j=1}^{C} A_{ij}F'_j \oplus F_i \tag{9}$$

通过捕获特征映射之间的长程依赖关系，HCVNet 获得了具有丰富上下文信息的场景结构聚合特征。深度解码器利用该聚合特征与特征解码器提取的多尺度特征来估计输出深度图。

2.4. 知识蒸馏的原理

知识蒸馏的目的是利用结构复杂的教师模型来引导轻量化的学生模型，获得更好的性能。传统的知识蒸馏需要分别训练两个模型，存在资源占用率过高的问题。为此，HCVNet 在深度估计中引入了自蒸馏[15]的知识，即使用两个权重独立的相同模型来相互学习。

HCVNet 需要训练一个性能良好且复杂的教师模型，然后使用冻结权重的教师模型来蒸馏学生模型。在网络方面，教师模型与学生模型具有相同的编解码器架构，唯一的区别是模型参数。如图 1 所示，HCVNet 的学生模型首先进行自监督学习，该模型的权重将在训练结束后保存。来自前一个训练时期的权重被加载并应用于第二个训练时期，第二个训练时期对应于图 1 中的教师分支。教师模型输出的深度伪标签在通过多视图检查过滤器过滤出深度信息离群值 M 之后，计算自蒸馏损失来向学生模型提供更多的监督信息，引导后续训练的学生模型获得更好的深度估计性能。自蒸馏损失 L_d 可以用公式表示为：

$$L_d = \frac{1}{4}\sum_{i=1}^{4}\left(M\left\|D_{teacher} - \hat{D}_{student}\right\|_1\right) \tag{10}$$

2.5. 自监督损失函数

光度损失：光度损失 $pe(\cdot)$ [7]由 L1 和结构相似度损失 SSIM 组成，计算过程可表示为公式(11)。为了处理被遮挡的对象，计算最小光度损失 L_{pe}，如公式(12)所示：

$$pe\left(\hat{I}_t, I_t\right) = \frac{\alpha}{2}\left(1 - SSIM\left(I_t, \hat{I}_t\right)\right) + \left(1-\alpha\right)\left\|\hat{I}_t - I_t\right\| \tag{11}$$

$$L_{pe} = \min pe\left(\hat{I}_t, I_t\right) \tag{12}$$

平滑度损失：当图像出现低纹理的区域时，光度损失的影响变弱。为了生成平滑的逆深度图，本文使用逆深度图在 x 和 y 上的梯度，以及第 0 帧在 x 和 y 上的梯度计算平滑度损失 L_s。计算过程可表示为公式(13)：

$$L_s = \left|\partial_x D_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y D_t^*\right| e^{-\left|\partial_y I_t\right|} \tag{13}$$

式中 $D_t^* = D_t / \hat{D}_t$ 表示经过归一化的逆深度图。

综上，HCVNet 训练中的总损失函数 L_{total} 由自蒸馏损失 L_d、光度损失 L_{pe} 和平滑度损失 L_s 组成，

$$L_{total} = \mu L_d + \lambda L_{pe} + \gamma L_s \tag{14}$$

其中，μ 和 γ 为自蒸馏损失和平滑度损失的训练权重。λ 用来判断重投影的光度误差是否小于原光度误差，若小于则设为 1，反之设为 0。

3. 实验结果与分析

3.1. 数据集及实施细节

实验在 KITTI 和 Make3D 两个公开数据集上进行。KITTI 数据集[16]提供了 61 个用于自动驾驶和机器人研究的立体道路场景。为了训练和评估所提出的方法，本实验遵循 Eigen 等人[19]提出的 KITTI 数据分割方法，其中 39,810 个图像用于训练，4424 个图像用于评估，697 个图像用于测试。本文在训练期间对所有图像使用相同的内在参数。Make3D 数据集[17]是一个户外单目深度估计的数据集，它包含 534 个 RGB-D 室外场景图像对，即 RBG 图像及其对应的深度图像。其中，400 个图像对用于训练，134 个图像对用于测试。为了评估所提出方法的泛化能力，在 KITTI 数据集上训练的模型直接在 Make3D 的测试图像上进行加载和推断。

实验运行环境为 Ubuntu16.04，python3.7，pytorch1.8，cuda11.3，处理器为 Intel Core，内存 16 GB，图形处理卡为一张 Nvidia GTX1070 (8GB)。该方法使用 AdamW 优化器以端到端的方式在 KITTI 数据集上训练了 20 个 epoch，输入图像分辨率为 640 × 192，批大小设置为 4，初始学习率设置为 10^{-4}。该实验采用数据扩充[7]作为预处理步骤，来提高训练的鲁棒性。

3.2. 评估指标

为了评价不同单目深度估计方法的效果，本文使用七个常用度量[19]作为深度评估指标：绝对相对误差(Absolute Relative Error, Abs Rel)、平方相对误差(Square Relative Error, Sq Rel)、均方根误差(Root Mean Square Error, RMSE)、均方根对数误差(Root Mean Square Logarithmic Error, RMSElog)、$\delta < 1.25$、$\delta < 1.25^2$、$\delta < 1.25^3$。各指标具体定义见公式(15)-(19)：

$$Abs\ Rel = \frac{1}{n}\sum\left|\frac{D_{pred} - D_{gt}}{D_{gt}}\right| \tag{15}$$

$$Sq\ Rel = \frac{1}{n}\sum\left(\frac{D_{pred} - D_{gt}}{D_{gt}}\right)^2 \tag{16}$$

$$RMSE = \sqrt{\frac{1}{n}\sum\left(D_{pred} - D_{gt}\right)^2} \tag{17}$$

$$RMSE\log = \sqrt{\frac{1}{n}\sum\left(\log\left(D_{pred}\right) - \log\left(D_{gt}\right)\right)^2} \tag{18}$$

$$\delta = \max\left(\frac{D_{gt}}{D_{pred}}, \frac{D_{pred}}{D_{gt}}\right) \tag{19}$$

其中，D_{gt} 表示地面真实深度，D_{pred} 表示网络预测的相对尺度深度，n 表示有效像素。误差指标 Abs Rel、Sq Rel、RMSE、RMSE log 的数值越低越好，而精度指标 $\delta < 1.25^i, i = 1,2,3$ 的数值越高越好。

3.3. 对比与分析

Table 1. Comparison results of performance metrics on the KITTI dataset
表 1. KITTI 数据集上的性能指标比较结果

方法	误差↓				精度↑		
	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [7]	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Sun [20]	0.117	0.863	4.813	0.192	0.871	0.959	0.982
SGDepth [21]	0.113	0.835	4.693	0.191	0.879	0.961	0.981
SAFENet [22]	0.112	0.788	4.582	0.187	0.878	0.963	0.983
VC-Depth [23]	0.112	0.816	4.715	0.190	0.880	0.960	0.982
PackNet-SfM [24]	0.108	0.727	4.426	0.184	0.885	0.963	0.983
Mono-Uncertainty [25]	0.111	0.863	4.756	0.188	0.881	0.961	0.982
HR-Depth [8]	0.109	0.792	4.632	0.185	0.884	0.962	0.983
Johnston [26]	0.106	0.861	4.699	0.185	0.889	0.962	0.982
CADepth [11]	0.105	0.769	4.535	0.187	0.892	0.964	0.983
DIFFNet [27]	0.102	0.749	4.445	0.179	0.897	0.965	0.983
MonoFormer [28]	0.108	0.806	4.627	0.184	0.889	0.962	0.983
Lite-Mono [29]	0.107	0.765	4.561	0.183	0.886	0.963	0.983
HCVNet(Ours)	0.101	0.730	4.417	0.177	0.899	0.966	0.984

为了验证 HCVNet 的先进性，将 HCVNet 与当前先进的 13 种自监督单目深度估计方法在 KITTI 上进行性能比较。表 1 展示了 HCVNet 在 KITTI 数据集上的深度估计性能指标比较结果。在 KITTI 数据集上，HCVNet 的 Abs Rel、Sq Rel、RMSE 和 RMSElog 指标值分别达到了 0.101、0.750、4.417、0.177，$\delta < 1.25^i, i = 1,2,3$ 指标值分别达到了 0.899、0.966、0.984。与最佳对比方法 DIFFNet [27]相比，HCVNet

的误差指标 Abs Rel、Sq Rel、RMSE、RMSElog 分别降低了 0.98%、1.87%、6.29%、1.12%，精度指标 $\delta < 1.25^i, i = 1,2,3$ 分别提高了 0.22%、0.10%、0.10%。HCVNet 在所有指标设置上都达到最佳性能。这是因为以往的方法仅使用单一的 U-Net 架构，而 HCVNet 设置教师模型和学生模型来进行自监督同步知识蒸馏。教师模型可为学生模型提供更多增益信息，从而有效提高了学生模型的精度与性能。

Table 2. Comparison results of performance metrics on the Make3D dataset
表 2. Make3D 数据集上的性能指标比较结果

方法	误差 ↓			
	Abs Rel	Sq Rel	RMSE	RMSElog
Monodepth2 [5]	0.321	3.378	7.252	0.163
HR-Depth [6]	0.305	2.944	6.857	0.157
CADepth [11]	0.319	3.564	7.152	0.158
DIFFNet [27]	0.298	2.901	6.753	0.153
Lite-Mono [29]	0.305	3.060	6.981	0.158
HCVNet (Ours)	0.279	2.781	6.537	0.146

为了评估 HCVNet 在不同户外现实场景中的泛化能力，在 Make3D 数据集上运行实验评估模型性能。表 2 展示了 HCVNet 与其他 5 种方法的性能指标比较结果。在 Make3D 数据集上，HCVNet 的 Abs Rel、Sq Rel、RMSE 和 RMSElog 指标值分别达到了 0.279、2.781、6.537、0.146。与最具竞争力的方法 DIFFNet [27]相比，HCVNet 的 Abs Rel、Sq Rel、RMSE 和 RMSElog 误差指标值分别降低了 6.38%、4.14%、4.58%。由此可见，HCVNet 的深度估计性能优于其他方法，且拥有先进的泛化效果。主要原因为 CNN-ViT 混合特征编码器具有强大的特征提取能力与鲁棒性。该编码器发挥 CNN 网络在局部特征提取方面的优势以及 Transformer 在全局信息建模方面的优势，使得 HCVNet 估计出的深度图具有更好的特征表示。

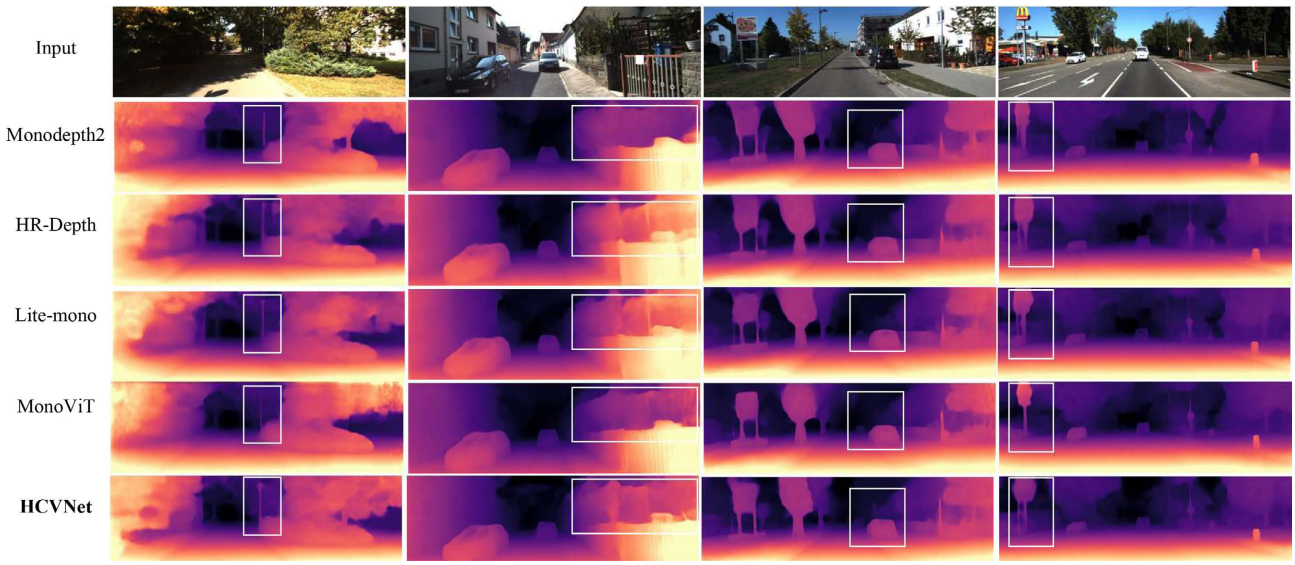

Figure 4. Comparison of visualization results on the KITTI dataset
图 4. KITTI 数据集上的可视化比较结果

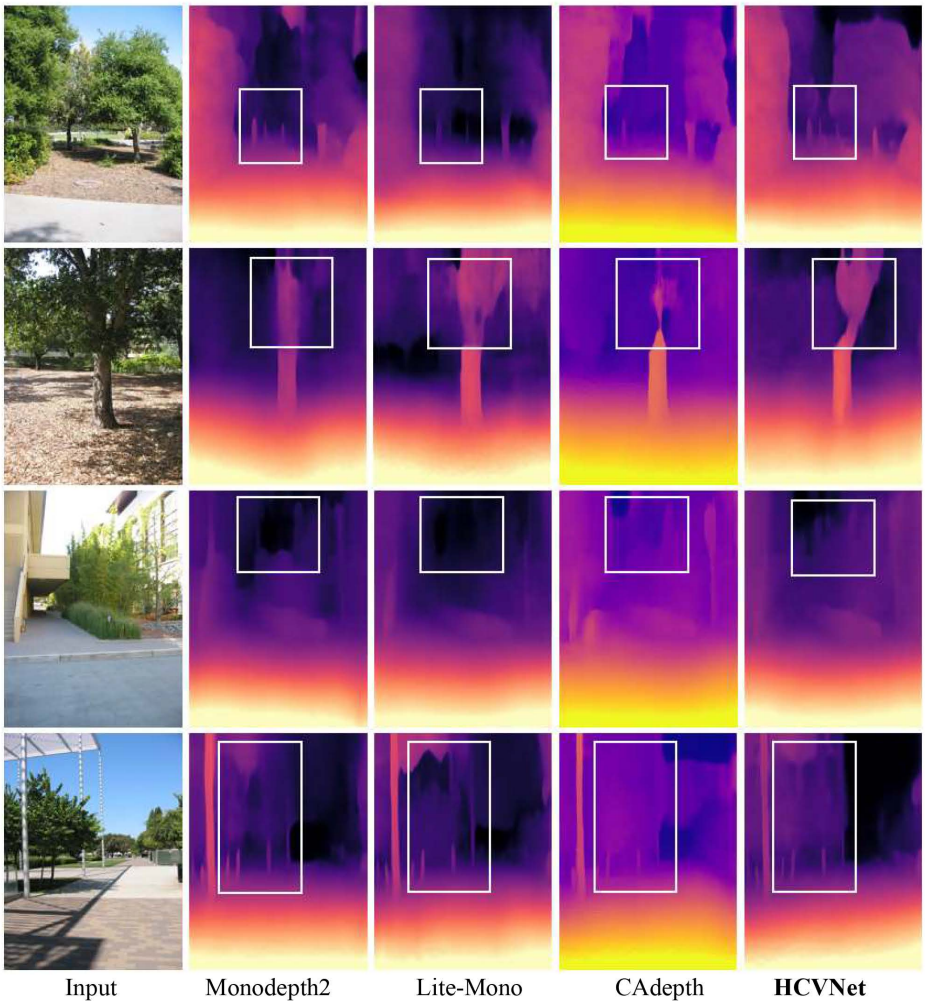

Figure 5. Comparison of visualization results on the Make3D dataset
图 5. Make3D 数据集上的可视化比较结果

为了更加直观地表现本文方法的有效性，本文比较了 HCVNet 与其他先进方法在 KITTI 和 Make3D 数据集上的可视化效果图。图 4 展示了在 KITTI 数据集上的可视化比较结果。图 5 展示了在 Make3D 数据集上的可视化比较结果。相比于目前的主流方法，HCVNet 预测的深度图中物体更贴合其本身形状，且物体边缘和轮廓细节更加清晰。如远处路灯、建筑、广告牌和树木，见图中白色线框标记处。在图 4 第一列中，其他方法仅估计出灯珠的轮廓，HCVNet 准确感知到圆灯的深度信息。由此可见，HCVNet 对深度不连续区域的深度估计性能优于其他方法。在图 5 第四行中，其他方法存在树木与背景模糊的问题，HCVNet 的树木轮廓最为接近现实。由此可见，HCVNet 能够比其他方法更准确地建模对象的结构。这得益于 CNN-ViT 混合特征编码器具有全局感受野，通道特征聚合模块丰富了上下文信息的整体场景几何感知。HCVNet 可以有效地对前景和背景进行建模，最终获得细粒度更高的深度估计结果。

3.4. 消融实验

为了进一步证明所提出各结构的有效性，本实验在 KITTI 数据集上对 CNN-ViT 特征编码器(CNN-ViT)、知识蒸馏(KD)和特征聚合模块(CFAM)三个组件进行消融实验，实验结果如表 3 所示。本实验使用 Monodepth2 [7]为 Baseline。从表 3 的结果可发现，本文所提出的三个组件均能有效提高单目深度

估计网络的性能。与 Baseline 相比，HCVNet 的误差 Abs Rel、Sq Rel、RMSE、RMSElog 指标值分别降低了 12.17%、16.94%、9.17%、8.29%，精度 $\delta<1.25^{i}, i=1,2,3$ 指标值分别提高了 2.51%、0.73%、0.31%。与 Baseline + CNN-ViT 相比，HCVNet 的误差 Abs Rel、Sq Rel、RMSE、RMSElog 指标值分别降低了 7.82%、14.40%、8.40%、4.15%。与 Baseline + CNN-ViT + KD (full)相比，HCVNet 的误差 Abs Rel、Sq Rel、RMSE、RMSElog 指标值分别降低了 9.57%、15.28%、9.10%、6.22%。由此可见，三个模块组合使用具有最佳增益效果，完整的 HCVNet 网络具有最好的深度估计性能。

Table 3. Ablation experiment of modules
表 3. 模块消融实验

方法	误差 ↓				精度 ↑		
	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^{2}$	$\delta<1.25^{3}$
Baseline	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Baseline + CNN-ViT	0.106	0.773	4.456	0.185	0.889	0.962	0.981
Baseline + CNN-ViT + KD	0.104	0.765	4.421	0.181	0.895	0.965	0.983
Baseline + CNN-ViT + KD + CFAM (full)	0.101	0.750	4.417	0.177	0.899	0.966	0.984

4. 结论

为了解决场景结构感知不足，深度估计精度与效率低下的问题，本文提出一种混合 CNN 和 ViT 的自监督知识蒸馏单目深度估计方法 HCVNet。HCVNet 使用 CNN-ViT 混合特征编码器，其中 CNN 关注局部细节，ViT 关注全局结构，有效避免了局部-全局特征丢失的问题。此外，编码器重新排列卷积操作符，能够以更高的效率提取整体结构特征。本文设计通道特征聚合模块，在通道维度上捕获场景的长程依赖关系。该设计丰富了上下文特征表示，缓解了复杂场景整体布局感知不完整的问题。与传统的知识蒸馏框架相比，本文设置相同结构的教师模型和学生模型来进行自监督蒸馏，能够在提高模型性能的同时并减少资源占用率。在 KITTI 和 Make3D 数据集上的实验结果表明，HCVNet 拥有最先进的性能和较强的泛化能力，可获得更清晰的深度估计结果。在未来工作中，我们将从去噪的角度来进一步优化图像特征，尝试引入扩散模型作为网络主干，以获得更高分辨率的深度图。

参考文献

[1] Xu, Q., Tan, C., Xue, T., *et al*. (2021) Overview of Monocular Depth Estimation Based on Deep Learning. 5*th International Conference on Cognitive Systems and Signal Processing* (*ICCSIP* 2020), Zhuhai, 25-27 December 2020, 499-506. https://doi.org/10.1007/978-981-16-2336-3_47

[2] Sun, J., Xie, Y., Chen, L., *et al*. (2021) Neuralrecon: Real-Time Coherent 3D Reconstruction from Monocular Video. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR* 2021), Nashville, 20-25 June 2021, 15593-15602. https://doi.org/10.1109/CVPR46437.2021.01534

[3] Luo, Y., Ren, J.S.J., Lin, M., *et al*. (2018) Single View Stereo Matching. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 155-163. https://doi.org/10.1109/CVPR.2018.00024

[4] Zhang, Z., Xu, C., Yang, J., *et al*. (2018) Progressive Hard-Mining Network for Monocular Depth Estimation. *IEEE Transactions on Image Processing*, **27**, 3691-3702. https://doi.org/10.1109/TIP.2018.2821979

[5] Wang, Z. (2022) Self-Supervised Learning in Computer Vision: A Review. 12*th International Conference on Computer Engineering and Networks* (*CENet* 2022), Haikou, 4-7 November 2022, 1112-1121. https://doi.org/10.1007/978-981-19-6901-0_116

[6] Wang, R., Yu, Z. and Gao, S. (2023) PlaneDepth: Self-Supervised Depth Estimation via Orthogonal Planes. 2023

IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 18-22 June 2023, 21425-21434. https://doi.org/10.1109/CVPR52729.2023.02052

[7] Godard, C., Aodha, O.M., Firman, M., *et al*. (2019) Digging into Self-Supervised Monocular Depth Estimation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 3827-3837. https://doi.org/10.1109/ICCV.2019.00393

[8] Lyu, X., Liu, L., Wang, M., *et al*. (2021) HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 2294-2301. https://doi.org/10.1609/aaai.v35i3.16329

[9] Wang, F. and Cheng, J. (2023) HQDec: Self-Supervised Monocular Depth Estimation Based on a High-Quality Decoder. arXiv: 2305.18706.

[10] Ren, W., Wang, L., Piao, Y., *et al*. (2022) Adaptive Co-Teaching for Unsupervised Monocular Depth Estimation. 17*th European Conference on Computer Vision*, Tel Aviv, 23-27 October 2022, 89-105. https://doi.org/10.1007/978-3-031-19769-7_6

[11] Yan, J., Zhao, H., Bu, P., *et al*. (2021) Channel-Wise Attention-Based Network for Self-Supervised Monocular Depth Estimation. 9*th International Conference on* 3*D Vision*, London, 1-3 December 2021, 464-473. https://doi.org/10.1109/3DV53792.2021.00056

[12] Zhao, C., Zhang, Y., Poggi, M., *et al*. (2022) MonoViT: Self-Supervised Monocular Depth Estimation with a Vision Transformer. 10*th International Conference on* 3*D Vision*, Prague, 12-16 September 2022, 668-678. https://doi.org/10.1109/3DV57658.2022.00077

[13] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 18*th IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[14] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Annual Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5999-6009.

[15] Kim, K., Ji, B., Yoon, D., *et al*. (2021) Self-Knowledge Distillation with Progressive Refinement of Targets. 18*th IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 6547-6556. https://doi.org/10.1109/ICCV48922.2021.00650

[16] Geiger, A., Lenz, P., Stiller, C., *et al*. (2013) Vision Meets Robotics: The KITTI Dataset. *International Journal of Robotics Research*, **32**, 1231-1237. https://doi.org/10.1177/0278364913491297

[17] Saxena, A., Sun, M. and Ng, A.Y. (2009) Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 824-840. https://doi.org/10.1109/TPAMI.2008.132

[18] Shim, D. and Kim, H.J. (2023) SwinDepth: Unsupervised Depth Estimation Using Monocular Sequences *via* Swin Transformer and Densely Cascaded Network. 2023 *IEEE International Conference on Robotics and Automation*, London, 29 May 2023-2 June 2023, 4983-4990. https://doi.org/10.1109/ICRA48891.2023.10160657

[19] Eigen, D., Puhrsch, C. and Fergus, R. (2014) Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. 28*th Annual Conference on Neural Information Processing Systems* 2014, Montreal, 8-13 December 2014, 2366-2374.

[20] Sun, Q., Tang, Y., Zhang, C., *et al*. (2022) Unsupervised Estimation of Monocular Depth and VO in Dynamic Environments *via* Hybrid Masks. *IEEE Transactions on Neural Networks and Learning Systems*, **33**, 2023-2033. https://doi.org/10.1109/TNNLS.2021.3100895

[21] Klingner, M., Termohlen, J.A., Mikolajczyk, J., *et al*. (2020) Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. 16*th European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 582-600. https://doi.org/10.1007/978-3-030-58565-5_35

[22] Choi, J., Jung, D., Lee, D.H., *et al*. (2020) Self-Supervised Monocular Depth Estimation with Semantic-Aware Depth Features. arXiv: 2010.02893.

[23] Zhou, H., Greenwood, D., Taylor, S., *et al*. (2020) Constant Velocity Constraints for Self-Supervised Monocular Depth Estimation. 17*th ACM SIGGRAPH European Conference on Visual Media Production*, 7-8 December 2020, 1-8. https://doi.org/10.1145/3429341.3429355

[24] Rares, V.G., Ambrus Pillai, S., *et al*. (2020) 3D Packing for Self-Supervised Monocular Depth Estimation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14-19 June 2020, 2482-2491.

[25] Poggi, M., Aleotti, F., Tosi, F., *et al*. (2020) On the Uncertainty of Self-Supervised Monocular Depth Estimation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 3224-3234. https://doi.org/10.1109/CVPR42600.2020.00329

[26] Johnston, A. and Carneiro, G. (2020) Self-Supervised Monocular Trained Depth Estimation Using Self-Attention and

Discrete Disparity Volume. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 4755-4764. https://doi.org/10.1109/CVPR42600.2020.00481

[27] Zhou, H., Greenwood, D. and Taylor, S. (2021) Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. 32*nd British Machine Vision Conference*, 22-25 November 2021, 730-734.

[28] Bae, J.H., Moon, S. and Im, S. (2022) Deep Digging into the Generalization of Self-Supervised Monocular Depth Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 187-196. https://doi.org/10.1609/aaai.v37i1.25090

[29] Zhang, N., Nex, F., Vosselman, G., *et al*. (2023) Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 18537-18546. https://doi.org/10.1109/CVPR52729.2023.01778