文章编号　1004-924X(2024)24-3603-13

多尺度特征增强的多帧自监督单目深度估计

寇旗旗[1]，王伟臣[2]，韩成功[2]，吕　晨[2]，程德强[2]，姬玉成[3*]
(1. 中国矿业大学 计算机科学与技术学院，江苏 徐州 221116；
2. 中国矿业大学 信息与控制工程学院，江苏 徐州 221116；
3. 应急管理部 大数据中心，北京 100013)

摘要：针对目前的深度估计网络对室外场景下图像的空间特征提取不够充分的问题，导致输出深度图存在物体边缘失真、模糊和区域伪影等缺陷，本文提出了一种多尺度特征增强的多帧自监督单目深度估计模型。首先，该模型编码器引入大核注意力的激活模块，提高编码器对输入图像全局空间特征的提取能力，保留空间上下文信息；同时，提出了一种结构增强模块，使其能够在通道维度上判别重要特征，增强网络对图像结构特征的感知能力；最后，解码器中使用动态上采样方法代替近邻插值的上采样方法，恢复细节信息，优化了深度图的伪影现象。实验结果表明，本文提出的深度估计网络在KITTI和CityScapes室外数据集的测试结果优于目前的主流算法，尤其在KITTI数据集上的预测正确率达到90.3%。可视化结果也表明，本文网络模型生成的深度图边缘更加清晰准确，有效地提高了深度估计网络的预测精度。

关　键　词：单目深度估计；自监督；多帧；大核注意力；特征增强
中图分类号：TP391.41；　　**文献标识码**：A　　**doi**：10.37188/OPE.20243224.3603

Multi-frame self-supervised monocular depth estimation with multi-scale feature enhancement

KOU Qiqi[1], WANG Weichen[2], HAN Chenggong[2], LÜ Chen[2], CHENG Deqiang[2], JI Yucheng[3*]

(1. *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou* 221116, *China*;
2. *School of Information and Control Engineering, China University of Mining and Technology, Xuzhou* 221116, *China*;
3. *Department Big Data Center, Ministry of Emergency Management, Beijing* 100013, *China*)
* *Corresponding author, E-mail: j. yc@outlook. com*

Abstract: The current depth estimation networks do not sufficiently extract spatial features from images in outdoor scenes, leading to issues such as object edge distortion, blurriness, and regional pseudo-shadows in the output depth maps. To address these problems, this paper proposed a multi-frame self-supervised monocular depth estimation model with multi-scale feature enhancement. Firstly, the model′s encoder incorporated an activation module based on large kernel attention to enhance its ability to extract global spatial features from the input image, preserving more spatial context information. Simultaneously, a structur-

收稿日期：2024-06-08；**修订日期**：2024-08-14.
基金项目：国家自然科学基金项目(No. 52204177)；徐州市基础研究计划青年科技人才项目(No. KC23026)

al enhancement module was introduced that can discriminate important features across channel dimensions, enhancing the network's perception of the structural characteristics of the image. Finally, the decoder used a dynamic upsampling method instead of the traditional nearest interpolation upsampling method to restore detailed information, thereby optimizing the pseudo-shadow phenomenon in the depth map to some extent. Experimental results demonstrate that the depth estimation network proposed in this paper outperforms current mainstream algorithms in tests on the KITTI and CityScapes outdoor datasets, particularly achieving a prediction accuracy rate of 90.3% on the KITTI dataset. Visualization results also indicate that the depth maps generated by our network model have clearer and more precise edges, effectively improving the prediction accuracy of the depth estimation network.

Key words: monocular depth estimation; self-supervision; multi-frame; large kernel attention; feature enhancement

1 引　言

通过深度估计算法获取图像中的逐像素深度信息已经被证实具有广泛的应用场景,例如自动驾驶[1]、增强现实[2]、三维重建[3]等。目前大部分深度估计都是基于二维RGB图像到RGB-D图像的转化估计,虽然专业设备可以直接获取像素级的真实深度,但是这些深度获取成本昂贵。因此,仅需部署一个普通摄像头且不需要昂贵的硬件来获取深度训练数据的自监督单目深度估计方法引起了广泛关注。

最早Garg[4]等人第一次提出了自监督的单目深度估计方法,其将立体图像之间的颜色一致性作为损失训练单目深度估计模型。Zhou[5]等人利用深度网络和姿态网络来构建跨时间帧的光度损失。Godard[6]等人将左右一致性损失引入立体训练。Bian[7]等人认为相邻帧具有一致的深度预测而提出时间深度一致性损失。Monodepth2[8]显著提高了以往方法的性能,其通过使用处理遮挡的每像素最小光度损失、遮蔽静态像素的自动掩码方法以及缓解深度纹理复制问题的多尺度深度估计策略,在自监督单目深度估计框架上获得了良好的性能。

虽然这种基于单幅图像的方法看起来应用前景广阔,但其深度估计性能还不能与专业获取深度的硬件或多帧方法相提并论。在场景静态的前提下,多帧方法[9-10]通常可以依靠多视图几何线索获得更高的整体精度。然而,在具有非朗伯曲面、无纹理区域和移动物体的现实场景中,依靠多视图几何线索的多帧方法仍然受到不完美重建的挑战[11-12]。为了解决这些问题,Watson等[13]首次提出了一种新的自监督的多帧单目深度估计方法,不仅消除了对人工注释和深度监督的要求,还提出了一种师生训练架构。这种师生培训架构使用了额外的单目深度线索来引导多帧成本量中特征的利用,以鼓励网络忽略多帧成本量中的不可靠区域,从而实现了先进的自监督多帧单目深度估计精度,这种方法也被后续相关工作[14-15]沿用。此外,Wang等[16]人针对多视角立体视觉(Multiple View Stereo, MVS)的固有问题,通过融合单目信息和速度引导来改善多帧深度估计精度,这种方法通过将单目深度作为几何优先级的方式来构建MVS成本体,并在预测相机速度的指导下,逐渐调整成本体的深度候选项。Li等[17]人通过将静态区域的多视角线索学习得到的几何感知传递到动态区域的单目表示中,再利用单目线索来增强多视角成本体的表达。

这些工作在网络结构上的创新有效地提高了单目深度估计的精度,但是存在网络对空间特征的提取和细节处理不够充分,导致输出深度图存在物体边缘失真、模糊和区域伪影的问题。因此针对上述问题,本文提出了一种多尺度特征增强的多帧自监督深度估计网络,主要贡献如下:(1)在编码器的不同尺度上引入基于大核注意力[18]的激活模块捕获更广阔的空间特征关联性信息,提高网络对输入图像的理解和表征能力,减少空间维度的信息丢失;结构增强模块,其在通道维度上判别重要特征,从而增强网络对图像结构特征的感知;(2)在上采样阶段引入动态上采样[19]方法代替近邻插值的上采样方法,更准确

的恢复特征细节信息，优化局部特征。实验表明，本文提出的多帧自监督单目深度估计模型在KITTI和CityScapes数据集上都取得了先进的深度估计结果。

2　算法和网络模型

本节将首先介绍自监督深度估计算法、然后介绍多尺度特征增强的多帧自监督深度估计网络架构，最后介绍本文所使用的模块。

2.1　自监督深度估计算法

自监督深度估计的核心思想是利用无需人工标注的数据来训练深度神经网络进行深度估计，通过自监督学习可以生成无标注数据的伪标签，这些伪标签可以提供一种反映场景中深度信息的近似。自监督深度估计过程可以被视为一种视图合成问题的形式。在视图合成问题中，深度估计算法可以根据已知视角和相机参数，以及学习到的深度信息和位姿信息，生成新的合成视图。新合成的视图与原视图进行损失计算，来约束网络在合成过程中尽量减少与原视图的差异，从而生成更加逼真和准确的合成视图。

具体自监督深度估计算法的流程如图1所示，首先给定视频流中相邻的目标帧I_t和源帧$I_{t'}$，其次目标帧I_t通过深度估计网络得到深度图D_t，

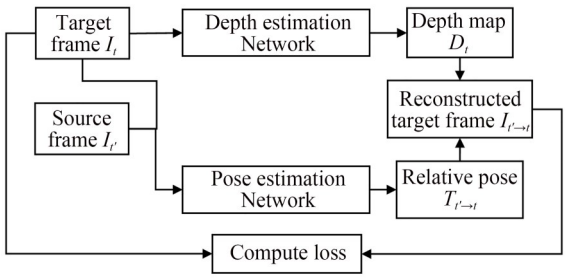


图1　自监督深度估计算法流程

Fig. 1　Process of self-supervised depth estimation algorithm

相邻的目标帧I_t与源帧$I_{t'}$通过位姿估计网络预测得到它们之间的相对位姿关系$T_{t\to t'}$，然后源帧$I_{t'}$根据深度图D_t、相对位姿$T_{t\to t'}$以及相机内参向目标帧I_t扭曲得到重构的目标帧$I_{t'\to t}$，最后根据损失函数Loss计算重构目标帧$I_{t'\to t}$与真实目标帧I_t图像之间的误差，从而实现对网络训练的约束。

2.2　本文深度估计网络模型

如2.1小节所述，重建帧$I_{t'\to t}$需要深度估计网络输出的深度图D_t作为依据，准确的深度图能够促使网络生成更加准确的重建帧，从而减少与真实帧的差异，因此本文主要对深度估计网络进行了改进，旨在生成更加准确的深度图D_t。在本小节，我们将介绍提出的多尺度特征增强的多帧自监督深度估计网络架构，该网络基于Watson等[13]提出的多帧自监督深度估计网络结构，网络架构如图2和图3所示。

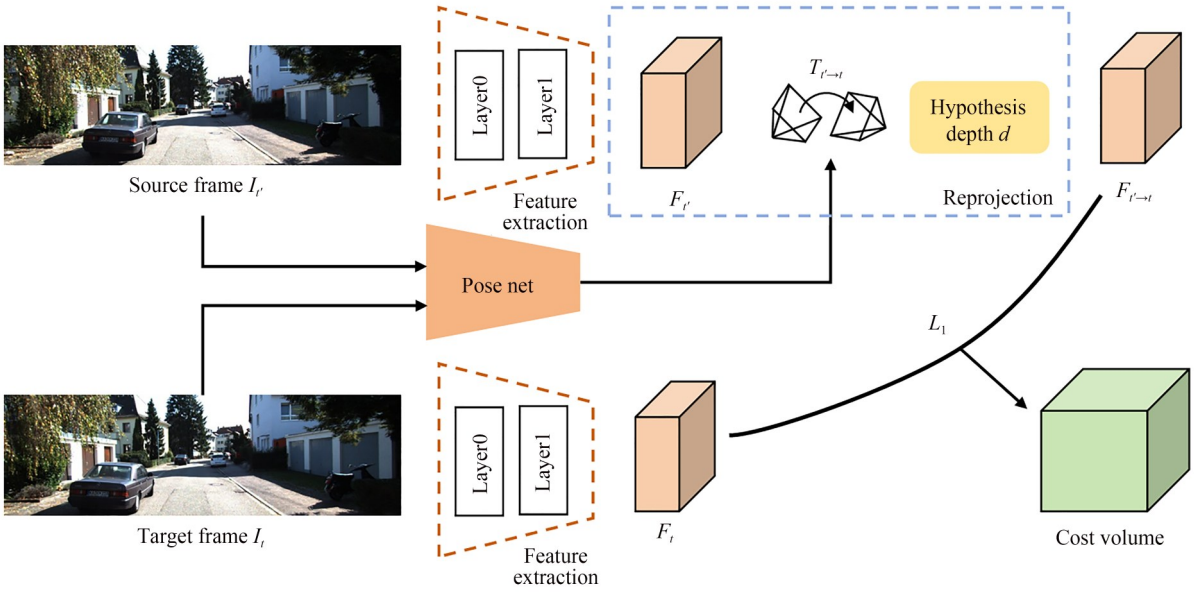


图2　多帧深度估计网络的成本量构建

Fig. 2　Cost volume construction of multi-frame depth estimation network

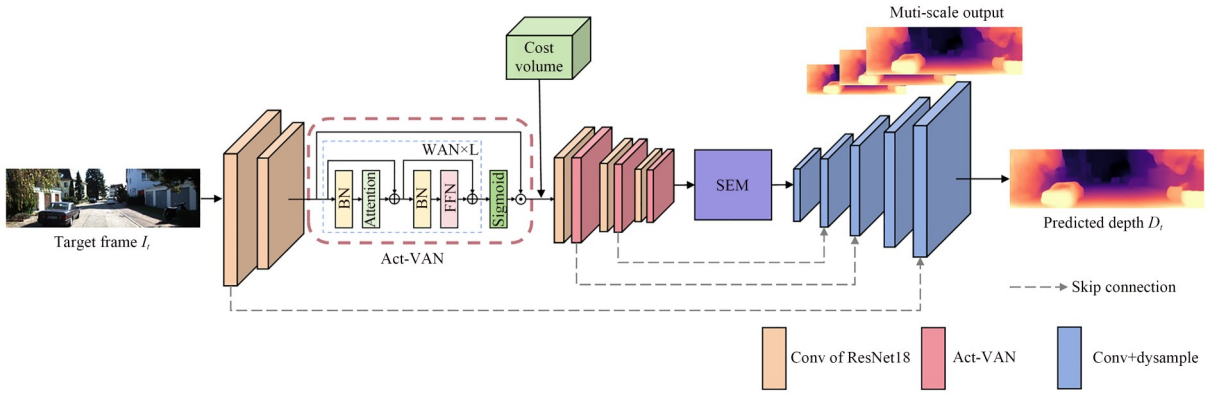


图 3　本文深度估计网络架构

Fig. 3　Our depth estimation network architecture

首先是多帧深度估计网络的成本量构建，如图 2 所示。与单帧深度估计网络不同的是，多帧深度估计网络的输入是前后相邻的两帧。为了利用两个输入帧，深度估计网络建立了一个成本量（Cost Volume，CV），用于测量输入视频中它和附近源帧的像素在不同深度值下的几何兼容性。即最终得到的成本量 $CV\in \boldsymbol{R}^{H\times W\times N_d}$ 中每个特征层可以表示为某一假设离散深度 d_n 下重建特征 $F_{t'\to t}$ 与目标帧特征 F_t 间的 L1 范数并在通道维度上的平均。其中 N_d 表示为假设的离散深度个数，取值为 96。

具体过程为，首先将源帧 $I_{t'}$ 与目标帧 I_t 分别送入相同的特征提取模块得到深度特征图 $F_{t'}$，F_t，其中特征提取模块由 Resnet18[20] 网络的前两层组成；同时也一起送入位姿估计网络，得到位姿关系 $T_{t\to t'}$。然后根据假设的候选深度 d 和位姿关系将 $F_{t'}$ 重投影到 I_t 的视角生成深度特征图 $F_{t'\to t}$。最终成本量是由重投影深度图 $F_{t'\to t}$ 和目标帧特征图 F_t 之间像素作差取绝对值得到。按照文献[21]的方法，将成本量与目标帧的特征图 F_t 进行通道维度上的融合，并作为解码器的输入，对深度进行 D_t 回归。

图 3 所示的是本文提出的多帧深度估计网络架构，其遵循基本的 U-net 结构，主要由编码器、跳跃连接、解码器三个部分组成。编码器是由 Resnet 与本文设计的基于大核注意力的激活模块组成，在多尺度特征上捕获更广阔的特征关联性信息，从而提高对图像空间特征的提取能力。为了细化编码器输出的最终特征空间信息，本文在编码器的最小尺度处设计了结构增强模块，该模块利用注意力机制增强了特征的结构语义表达。在解码器阶段通过跳跃连接来促进梯度和信息在整个模型中的流动并在多尺度上采用动态上采样的方法连续上采样深度特征图，逐步恢复空间分辨率，最终输出与输入尺度一致的深度图。实验表明，我们的网络与原有基础网络相比，提高了深度估计的精度，优化了深度图的细节。

2.3　基于大核注意力的激活模块

ResNet 作为一种深度学习中常用的编码器架构，它通过引入了残差连接来解决深层网络中的梯度消失和梯度爆炸问题。但它主要关注于低层和中层特征的传递和学习。这可能导致在高层特征表示方面略显有限，难以捕捉到数据中更复杂和抽象的特征。此外，ResNet 网络中的池化和卷积层会导致输入的空间信息的丢失，而在深度估计任务中，精确的空间信息对于准确的特征提取至关重要。因此，本文设计了一种基于大核注意力的激活模块，将其插入编码器下采样阶段的不同尺度上，希望在多尺度上捕捉更广阔的空间关联性信息，从而保留更多的空间上下文信息。此外，使用大核注意力模块可以使网络倾向于识别物体形状而非纹理，这在抑制区域伪影是有益的。

我们引入的模块如图 4 所示，它将输入特征图送入经过激活函数的视觉注意力网络（Visual Attention Network，VAN）并与原特征图相乘作为输出特征图。其中，视觉注意力块是基于 Guo 等[18]提出的大核卷积构成。如图 5 所示，大核卷积可分为三个部分：深度卷积、深度膨胀卷积和

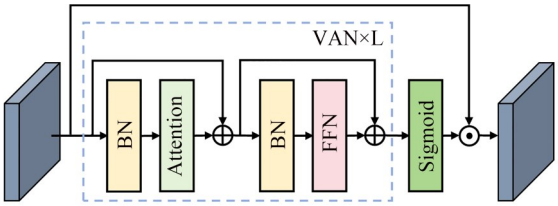


图 4　基于视觉注意力的激活模块

Fig. 4　Activation Module Based onVision Attention(Act-VAN)

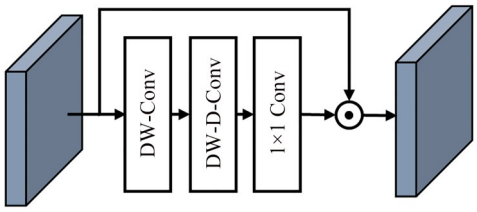


图 5　大核注意力

Fig. 5　Large kernel attention

信道卷积。具体来说，其将 $K\times K$ 卷积分解为一个 $\left\lceil\frac{K}{d}\right\rceil\times\left\lceil\frac{K}{d}\right\rceil$ 深度膨胀卷积（膨胀度为 d）、一个 $(2d-1)\times(2d-1)$ 深度卷积和一个 1×1 卷积。上述分解只需花费少量的计算成本和参数就能捕捉到长程关系，且同时实现了空间和通道维度的适应性。所以大核注意力（Large Kernel Attention，LKA）可写为：

$$Output=\mathrm{Conv}_{1\times1}(\boldsymbol{DW\text{-}D\text{-}}\mathrm{Conv}(\boldsymbol{DW}\text{-}\mathrm{Conv}(F)))\odot F, \tag{1}$$

其中：$F\in\boldsymbol{R}^{C\times H\times W}$ 表示输入特征，经过大核卷积获得注意力图，$\mathrm{Conv}_{1\times1}(\boldsymbol{DW\text{-}D}\text{-}\mathrm{Conv}(\boldsymbol{DW}\text{-}\mathrm{Conv}(F)))$ 的值表示每个特征的重要性，$\odot$ 表示逐元素乘积。

本文将提出的模块命名为 Act-VAN，其中视觉注意力块是由归一化层、1×1Conv、GELU 激活、大核注意力块和前馈网络（Feedforward Network，FFN）[22]组成。我们遵循文献[18]的方法，在训练时对 VAN 模块加载 VAN-B1 的预训练模型，提高编码器的特征提取能力和鲁棒性。对 VAN 输出的注意力加权特征利用 Sigmoid 激活函数捕捉特征的非线性关系并得到像素级特征权重然后与输入特征相乘作为模块的输出，实现网络对关键空间特征的关注。该模块保持了输入与输出特征通道数量的一致。其可表示为：

$$Output=F\odot sigmoid(\mathrm{VAN}_L(F)), \tag{2}$$

其中：$F\in\boldsymbol{R}^{C\times H\times W}$ 为输入特征，L 表示 VAN 模块的数量。经过实验验证，加入 Sigmoid 激活层的模块比直接引入 Guo 等[18]提出的 VAN 网络，在多帧单目深度估计任务上性能更优，在评价指标上有所提升。

2.4　结构增强模块

深度估计场景中的结构信息能够反映场景中物体的空间布局和相互关系。获取这些信息有助于网络模型更准确地理解和分析场景，为网络提供更多的语义信息和场景上下文。受 Yan 等人[23]与大核注意力的启发，我们设计了一个结构增强模块，如图 6 所示。该模块能够通过通道自注意力机制（Channel Self-Attention，CSA）模拟通道之间的相互依赖关系，将每个通道从其他通道捕获更多不同的区域响应，从而能够从较远区域获得更多相对深度信息。同时结合大核注意力机制，增强对场景结构的感知。经过实验证明，设计的结构增强模块能够有效提高多帧单目深度估计的性能。

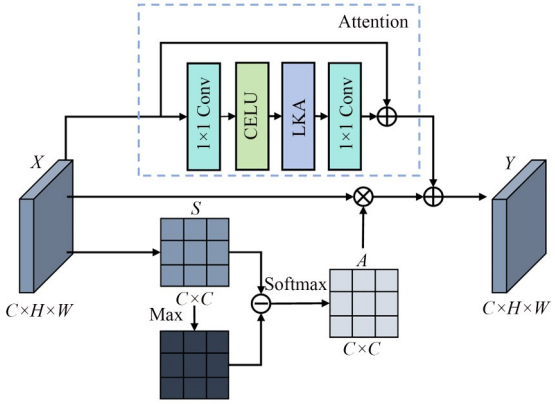


图 6　结构增强模块

Fig. 6　Structure enhancement module

首先将编码器最后一层输出特征 $\boldsymbol{X}\in\boldsymbol{R}^{C\times H\times W}$ 的二维尺寸合并成一维 $N=H\times W$ 得到变形后的 $\boldsymbol{X}'\in\boldsymbol{R}^{C\times N}$，然后将 $\boldsymbol{X}'$ 与它的转置进行矩阵相乘⊗得到通道特征相似度 $\boldsymbol{S}\in\boldsymbol{R}^{C\times C}$，可表示为：

$$\boldsymbol{S}_{ij}=\boldsymbol{X}_i'\boldsymbol{X}_j'^{\mathrm{T}}. \tag{3}$$

式(3)反映了区域响应的空间关系，$\boldsymbol{S}$ 值越高，两个通道特征图之间具有越高的相似性，意味着它们对同一区域也具有较强的响应，两个通

道之间的依赖性更强。然后通过逐元素减法⊖将通道相似度转换为区分度：

$$\boldsymbol{D}_{ij}=\max_i(\boldsymbol{S})-\boldsymbol{S}_{ij}, \tag{4}$$

其中，$\boldsymbol{D}_{ij}$表示第j个通道与第i个通道之间的差异性。通道间差异性越大，在特征聚合时会获得更高的分数$\boldsymbol{D}_{ij}$。接着将其通过softmax层生成注意力图$\boldsymbol{A}\in\boldsymbol{R}^{C\times C}$。注意力图与输入特征进行矩阵相乘获取所有通道特征。与此同时，将输入特征通过大核注意力模块获取空间结构特征。最终的对于每个通道输出特征为：

$$\boldsymbol{Y}_i=\sum_j^C(\boldsymbol{A}_{ij}\boldsymbol{X}_j)+Attention(\boldsymbol{X}_i). \tag{5}$$

式(5)表示每个通道的最终特征是所有通道特征与经过大核注意力模块的原始特征的加权和。

2.5 动态上采样

为了深度估计模型能够获取高质量的深度图。理想的上采样器应做到恢复细节的同时，保持平坦区域中深度值的一致性，并处理逐渐变化的深度值。

近邻插值法作为最常用的上采样方法，其遵循固定的规则，即根据最近的像素值直接增加维度。这可能导致图像中的目标边缘出现块状情况，减弱了边缘与背景之间的差异，不适用于深度估计这种密集预测任务。因此，为了针对深度估计的密集预测任务，增加上采样后深度图的精确度，本文引入了Dysample[19]来代替近邻插值的上采样方法。

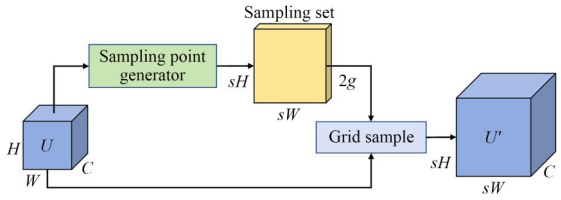


图7　动态上采样结构

Fig. 7　Structure of dynamic upsampling

动态上采样的结构如图7所示。给定一个输入特征$\boldsymbol{U}\in\boldsymbol{R}^{C\times H\times W}$和采样集$\boldsymbol{S}$，二者通过网格采样函数得到采样后的特征$\boldsymbol{U}'\in\boldsymbol{R}^{C\times sH\times sW}$。那么可将动态上采样过程定义为：

$$\boldsymbol{U}'=\text{grid\_sample}(\boldsymbol{U},\boldsymbol{S}). \tag{6}$$

Liu等人[19]将研究过程重点放在了对上述过程中采样点生成器的部分。他们将作用域因子的形式分为静态和动态，并根据偏移量生成方式共研究了四种Dysample变体。与一般的基于核函数的动态上采样不同，Dysample是从点采样的角度进行设计的，其具有快速，有效，通用的特点。在多种视觉任务中取得了优异的性能。

3 损失函数

本文沿用文献[13]的自监督损失函数。给定目标帧$\boldsymbol{I}_t$和源帧$\boldsymbol{I}_{t'}$，联合训练网络以预测目标图像的密集深度图$\boldsymbol{D}_t$和目标到原图像的相对相机姿态$\boldsymbol{T}_{t\to t'}$，构造光度重投影损失函数如式(7)：

$$L_p=\min_n\left[\frac{a}{2}(1-SSIM(\boldsymbol{I}_t,\boldsymbol{I}_{t'\to t}))+(1-a)\left\|\boldsymbol{I}_t,\boldsymbol{I}_{t'\to t}\right\|_1\right]. \tag{7}$$

它是光度损失L1和结构相似度损失$SSIM$的加权组合，这是因为在真实环境中光照条件不是固定不变的，加入结构相似度损失能够更好地处理复杂光照的变化，计算过程中一般取$a=0.85$，$SSIM$可以定量比较两张图像的相似性：

$$SSIM(\boldsymbol{I}_a,\boldsymbol{I}_b)=[l(\boldsymbol{I}_a,\boldsymbol{I}_b)]^{\alpha}[c(\boldsymbol{I}_a,\boldsymbol{I}_b)]^{\beta}[s(\boldsymbol{I}_a,\boldsymbol{I}_b)]^{\gamma}, \tag{8}$$

其中：α,β,γ为常数一般取1，$\boldsymbol{I}_{t'\to t}$是根据目标图像的深度(视差的倒数)和相机位姿扭曲到目标坐标系的源图像，可表示如式(9)：

$$\boldsymbol{I}_{t'\to t}=\boldsymbol{I}_{t'}\langle proj(\boldsymbol{D}_t,\boldsymbol{T}_{t\to t'},K)\rangle, \tag{9}$$

其中，$proj()$是转换函数，可将目标图像的像素p_t映射到源图像上$p_{t'}$：

$$p_{t'}\sim K\boldsymbol{T}_{t\to t'}\boldsymbol{D}_t(p_t)K^{-1}p_t. \tag{10}$$

而$\langle\cdot\rangle$是局部亚可微的双线性采样算子。K为相机内参，这里假设它固定不变。

当图像出现纹理较弱的区域时，边缘平滑损失能在上述光度重投影损失失效的情况下对深度预测值进行约束，具体损失函数为：

$$L_s=\left|\partial_x d_t^*\right|e^{-\left|\partial_x I_t\right|}+\left|\partial_y d_t^*\right|e^{-\left|\partial_y I_t\right|}, \tag{11}$$

其中，$d_t^*=d/\overline{d_t}$是平均归一化反深度，以阻止估计深度收缩。

此外，由于多帧深度估计网络中成本量对无纹理区域和移动物体上的预测效果要比单图像深度估计网络差，因此在训练时使用独立的单图像网络来纠正成本量有问题的区域。这个单图

像深度估计网络为每个训练图像生成深度图$\hat{\boldsymbol{D}}_t$，在训练结束后丢弃这些深度图。多帧深度估计输出中可能存在问题的像素由二进制掩码M识别，并在掩码区域对深度图$\boldsymbol{D}_t$采用L1损失来鼓励预测结果与深度图$\hat{\boldsymbol{D}}_t$相似：

$$L_c=\sum M\left|\boldsymbol{D}_t-\hat{\boldsymbol{D}}_t\right|, \tag{12}$$

其中：$\hat{\boldsymbol{D}}_t$的梯度被阻断，确保知识仅单向传播；二进制掩码M在不可靠区域为1，否则为0。在成本量可靠的区域$\hat{\boldsymbol{D}}_t$代表的深度应于成本量代表的深度$\boldsymbol{D}_{cv}$相似。因此将成本量代表的深度$\boldsymbol{D}_{cv}$与$\hat{\boldsymbol{D}}_t$进行比较，在两者显著不同的区域设为1，得到：

$$M=\max\left(\frac{\boldsymbol{D}_{cv}-\hat{\boldsymbol{D}}_t}{\hat{\boldsymbol{D}}_t},\frac{\hat{\boldsymbol{D}}_t-\boldsymbol{D}_{cv}}{\boldsymbol{D}_{cv}}\right)>1. \tag{13}$$

综上，本文总损失函数的计算公式如下：

$$L=(1-M)L_p+L_c+0.001L_s. \tag{14}$$

4　实验结果

4.1　实验设置

本文在两个具有挑战性的深度估计室外自动驾驶场景数据集上进行了测试评估，并与其他方法进行对比、分析。为确保公平对比，本文使用Eigen等人[24]对KITTI数据集的划分方法，测试集为来自29个场景中的697张图像，其他32个场景的22 600张图像与888张图像分别作为训练集和验证集。在测试集中，将处于序列起始位置的22张图像仅根据一帧进行预测并纳入评估。为了进一步证明本文方法的有效性，与KITTI一样，本文也在CityScapes数据集低于80 m的地面真实深度上进行评估。其中来自视频序列的69 731张图像作为训练集，1 525张图像作为测试集。深度估计网络与位姿网络均建立在ResNet18的基础上。

本文的网络使用Pytorch框架实现，Python版本为3.7，Pytorch版本为1.7，CUDA版本为11.3。实验设备为一张显存为24 G的NVIDIA RTX 3090显卡，操作系统为Ubuntu。本文使用Adam优化器进行训练，学习率为10^{-4}，训练迭代轮数（epoch）为20，批处理（batch size）大小为10。

4.2　评价指标

本文使用Eigen等人[25]提出的标准度量来作为深度估计模型性能的评估指标。分别为绝对相对误差（Abs Rel）、平方相对误差（Sq Rel）、均方根误差（RMSE）、均方根对数误差（RMSE log）以及不同阈值下的准确率（δ），它们分别具体表示如下：

$$\mathrm{AbsRel}=\frac{1}{|N|}\sum_{i\in N}\frac{\left\|d_i-d_i^*\right\|}{d_i^*}, \tag{15}$$

$$\mathrm{SqRel}=\frac{1}{|N|}\sum_{i\in N}\frac{\left\|d_i-d_i^*\right\|^2}{d_i^*}, \tag{16}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{|N|}\sum_{i\in N}\left\|d_i-d_i^*\right\|^2}, \tag{17}$$

$$\mathrm{RMSElog}=\sqrt{\frac{1}{|N|}\sum_{i\in N}\left\|\log(d_i)-\log(d_i^*)\right\|^2}, \tag{18}$$

$$\delta=\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)<thr,$$
$$thr=1.25,1.25^2,1.25^3, \tag{19}$$

其中：d_i表示像素点i的预测深度值，d_i^*表示像素点i的真实深度值，N表示可获取真实深度的像素点个数，准确率δ是取像素点预测深度值与真实深度值之间的比值的最大值，并统计小于thr阈值的像素点占总像素点的比值，δ越趋近于1说明深度估计的精度越准确。

4.3　结果对比

本文选取了近几年的主流先进算法与本文方法在KITTI和CityScapes数据集上进行了测试比较，如表1和表2所示。结果表明，本文的方法具有一定的先进性和有效性，在一些指标上取得了较之前的单目深度估计方法更优秀的测试结果。

本文采用连续视频的前后帧作为自监督单目深度估计的训练监督方法。本文方法与其他方法在KITTI数据集上进行结果对比时，除了使用大小为640×192的低分辨率输入图像，还将本文方法与其他方法在高分辨率输入图像的情况下进行比较。在CityScapes数据集上进行对比时，输入图像分辨率为416×128。将每项评价指标的最优结果加粗标注。

表1中可以看出，本文方法在视频帧监督方式下的评价指标结果几乎都优于基线方法[13]。

表 1　KITTI 数据集上的测试结果

Tab. 1　Test results on the KITTI dataset

Method (M)	Resolution	Indicator of Error (Lower is better)				Accuracy of Prediction (Higher is better)		
		AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Monodepth2[8]	640×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Johnston *et al.*[26]	640×192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
Packnet-SFM[27]	640×192	0.111	0.785	4.601	0.189	0.878	0.960	0.982
VA-Depth[28]	640×192	0.112	0.864	4.804	0.190	0.877	0.959	0.982
Zeeshan *et al.*[29]	640×192	0.113	0.903	4.863	0.193	0.877	0.959	0.981
STDepthFormer[30]	640×192	0.110	0.805	4.678	0.187	0.878	0.961	0.983
Patil *et al.*[31]	640×192	0.111	0.821	4.650	0.187	0.883	0.961	0.982
Suanders *et al.*[32]	640×192	0.100	0.747	4.455	0.177	0.895	**0.966**	**0.984**
Wang *et al.*[33]	640×192	0.106	0.799	4.662	0.187	0.889	0.961	0.983
Manydepth[13]	640×192	0.098	0.770	4.459	0.176	0.900	0.965	0.983
Ours	640×192	**0.095**	**0.743**	**4.374**	**0.175**	**0.903**	0.965	0.983
Monodepth2[8]	1 024×320	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Packnet-SFM[27]	1 280×384	0.107	0.802	4.538	0.186	0.889	0.962	0.981
Shu *et al.*[34]	1 024×320	0.104	0.729	4.481	0.179	0 893	0.965	**0.984**
Wang *et al.*[33]	1 024×320	0.106	0.773	4.491	0.185	0.890	0.962	0.982
Manydpeth[13]	1 024×320	0.093	0.715	4.245	0.172	0.909	**0.966**	0.983
Ours	1 024×320	**0.090**	**0.703**	**4.213**	**0.170**	**0.913**	**0.966**	0.983

注：**M** means that the training supervision method is video frames.

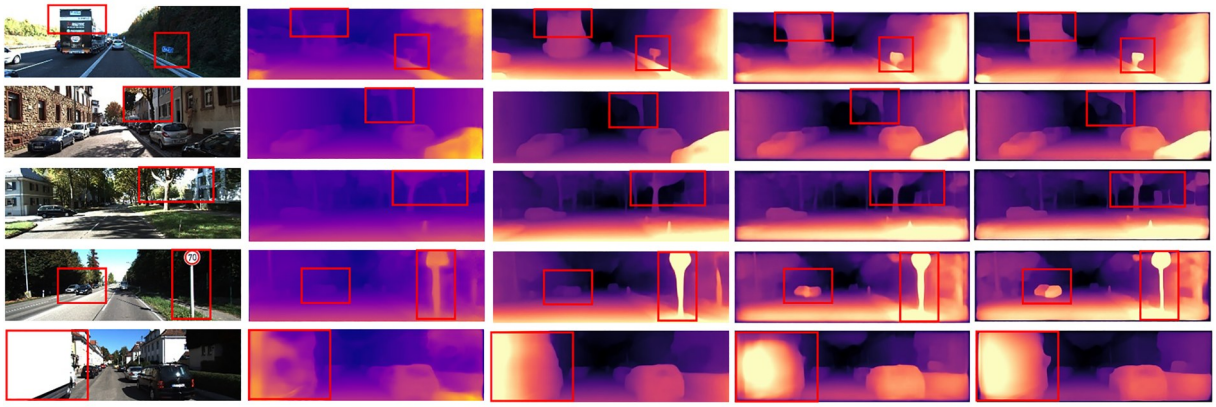
表 2　CityScapes 数据集上的测试结果

Tab. 2　Test results on the CityScapes dataset

Method	AbsRel	SqRel	RMSE	RMSElog
Monodepth2[8]	0.129	1.569	6.876	0.187
Li *et al.*[35]	0.119	1.290	6.980	0.190
Manydepth[13]	0.114	1.193	6.223	0.170
Ours	**0.111**	**1.165**	**6.196**	**0.168**

与基线方法相比，本文方法在基于视频帧(M)监督方式训练后测试的绝对相对误差、平方相对误差、均方根误差、均方根对数误差分别减少了3.1%，3.5%，1.9%，0.6%，在阈值为$\delta<1.25$的准确度指标上提升了0.33%。此外，对高分辨率输入图像的网络也进行了训练测试，并与基线算法进行对比，结果也表明本文方法的测试指标有显著的提高。与基线方法相比，本文方法在高分辨率输入图像训练后测试的绝对相对误差、平方相对误差、均方根误差以及均方根对数误差分别减少了3.2%，1.7%，0.8%，1.2%，在阈值$\delta<1.25$的准确度提升了0.4%。表2展示的是在CityScapes数据集上进行测试的评价结果，也表明了本文方法相比主流算法也具有更好的性能。

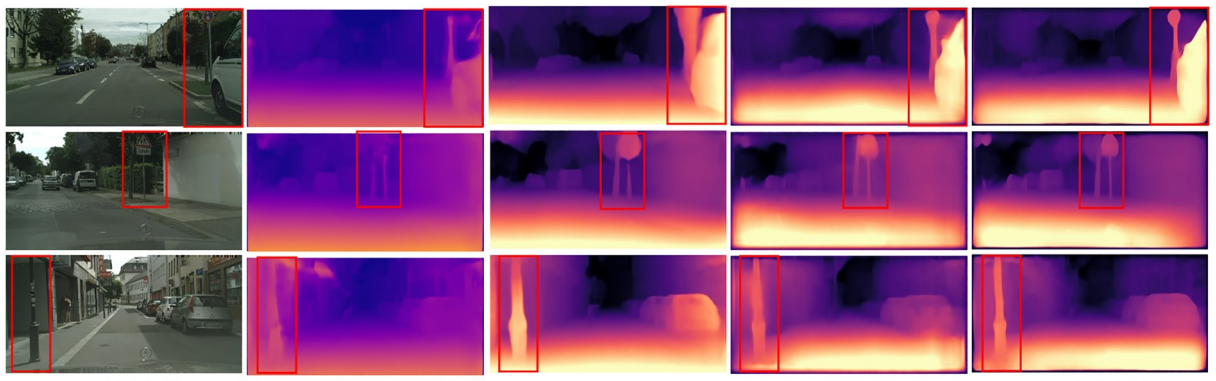
如图8和图9所示，为了更加直观地表现本文提出方法的有效性，分别选取了自监督单目深度估计的开山之作Monodepth、通过处理遮挡和纹理复制问题取得显著性能提高的经典之作Monodepth2和本文工作的基线方法Manydepth三个具有代表性的方法与本文方法在KITTI和CityScapes数据集上进行可视化对比。相较于之前的方法，可以看到本文方法对于深度图的预测效果有明显改善。首先，预测的深度图中物体的边缘更加完整，更好地贴合了物体本身的形状。而且不同深度区域之间的边缘更加清晰、锐利。此外，每个物体区域的预测效果保持了区域整体深度的一致性。从图8中第一行深度图中的货车车厢和第三行深度图中的道路限速牌可以明显看出本文方法对于物体轮廓的预测更加准确；此外，本文方法在对图8中第二行和第三行深度图中树杈以及第五行深度图中有着大片无纹理区域的货车的深度估计也都表现了本文方法相对于其他方法对同一物体预测出的深度区域更加



(a) 输入图像　(b) Monodepth[6]　(c) Monodepth2[8]　(d) Manydepth[13]　(e) 本文方法
(a) Input image　(b) Monodepth[6]　(c) Monodepth2[8]　(d) Manydepth[13]　(e) Ours

图 8　KITTI 数据集的可视化结果对比

Fig. 8　Comparison of visualization results on the KITTI dataset



(a) 输入图像　(b) Monodepth[6]　(c) Monodepth2[8]　(d) Manydepth[13]　(e) 本文方法
(a) Input image　(b) Monodepth[6]　(c) Monodepth2[8]　(d) Manydepth[13]　(e) Ours

图 9　CityScapes 数据集的可视化结果对比

Fig. 9　Comparison of visualization results on the CityScapes dataset

一致。另一方面，本文也选取了 CityScapes 数据集与其他方法进行定性对比。从图 9 中同样可以看出本文方法对物体边缘和轮廓细节的深度估计准确性明显优于其他方法（彩图见期刊电子版）。特别是在图 9 第一行深度图可以看出，红框圈出的道路指示牌以及汽车与周围环境之间的深度差都被准确地表现出来，而不存在类似方法（b）和方法（c）中存在的轮廓边缘模糊问题；且在第二行和第三行深度图中可以看出，本文方法预测的道路指示牌以及路灯杆的轮廓更加符合实际。上述的深度图的改善说明了本文模块针对场景整体空间结构以及物体边缘的感知是有效的。

本小节进一步测量了在输入图像尺寸为 640×192 时模型的参数量和计算复杂度，结果如表 3 所示。通过比较可知，本文提出的方法相较于其他方法具有相对较高的参数量和计算复杂度，这主要归因于所设计模块的复杂结构。该结构对整体方法的性能提升具有显著贡献，因此导致了整体模型参数量和计算复杂度的增多。

表 3　模型参数量和计算复杂度

Tab. 3　Model parametric and computational complexity

Model	Params/M	FLOPs/G
Monodepth2[8]	14.3	8.0
Manydepth[13]	14.4	10.2
Ours	22.0	16.3

4.4　消融实验

本文通过消融实验进一步验证了模块和设计的有效性，其中未使用的模块用×表示，使用

的模块用√表示，结果如表 4 所示。本节将文献[13]的方法作为基线方法，并在表 4 中以实验 1 表示。将只引入动态上采样的解码器方法在表 4 中以实验 2 表示。结果表明，本文设计的基于大核注意力的激活模块和 SEM 模块均能有效提高单目深度估计网络的性能。比较实验 2 和实验 3 指标可以发现，各指标有了显著的改善，这说明本文设计的基于大核注意力激活模块的编码器能够对图像进行更有效的特征提取。而比较实验 3 和实验 4 指标可以看出，本文提出的结构增强模块也能够提升深度估计网络的深度线索捕捉能力，从而提高深度估计精度降低误差。此外，通过对比实验 1 和实验 2 指标也能看出，动态上采样方法相比近邻插值上采样对于恢复特征的效果更好。

本小节也进一步验证了本章模块设计有效性，如表 5 所示。无激活层的 VAN 网络方法和不包含大核注意力模块的结构增强模块在表 5 中分别以 VAN 和 SEM(w/o)(LKA)表示。比较实验 2 和实验 3 指标可以看出，相比于无激活的模块，激活模块在一定程度上能够提升深度估计网络的精度。比较实验 4 和实验 5 指标可以看出，本文设计的结构增强模块有助于提高深度估计精度。为了更直观地展现 Act-VAN 和 SEM 的优势，绘制了在 SqRel 指标下模块设计的结果对比，如图 10 所示。

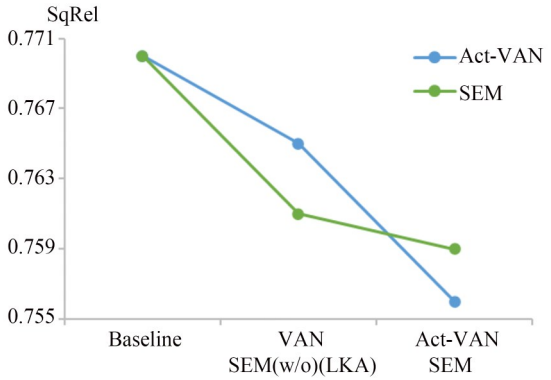


图 10　SqRel 指标下模块设计的结果对比

Fig. 10　Comparison of results for module design on SqRel

表 4　本文模块消融实验

Tab. 4　Ablation experiment of modules

Experiment	Act-VAN	SEM	DySample	Indicator of Error (Lower is better)				Accuracy of Prediction (Higher is better)		
				AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
1	×	×	×	0.098	0.770	4.459	0.176	0.900	0.965	0.983
2	×	×	√	0.098	0.765	4.441	0.176	0.900	0.965	0.983
3	×	×	√	0.096	0.753	4.379	0.175	0.902	0.965	0.983
4	×	√	√	0.096	0.755	4.436	0.176	0.901	0.965	0.983
5	√	√	√	**0.095**	**0.743**	**4.374**	**0.175**	**0.903**	0.965	0.983

表 5　本文模块设计消融实验

Tab. 5　Ablation experiment of modules design

Experiment	Method	Indicator of Error (Lower is better)				Accuracy of Prediction (Higher is better)		
		AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
1	Baseline	0.098	0.770	4.459	0.176	0.900	0.965	0.983
2	VAN	0.097	0.765	4.495	0.176	0.901	0.965	0.983
3	Act-VAN	0.096	0.756	4.392	0.176	0.902	0.965	0.983
4	SEM(w/o)(LKA)	0.096	0.761	4.440	0.176	0.901	0.965	0.983
5	SEM	0.096	0.759	4.437	0.176	0.901	0.965	0.983

5　结　论

由于深度估计网络对场景结构信息感知不够充分，导致输出深度图存在物体边缘失真、模糊等缺陷。本文提出了一种多尺度特征增强的多帧自监督深度估计网络。该网络编码器在多尺度残差网络的基础上，加入了大核注意力的激活模块，增强了网络对全局特征和长程依赖关系的感知能力；在编码器最小尺度处加入了结构增强模块，通过注意力机制增强了对关键特征的语义表达；在解码器加入动态上采样方法进行连续上采样，逐步恢复空间分辨率，减少恢复的深度特征图伪影。

实验结果表明，相比于之前的工作，本文方法在室外自动驾驶场景公共数据集上取得了更优秀的评价结果；在室外自动驾驶场景公共数据集KITTI上的正确预测深度像素的比例测试指标($\delta<1.25$)达到了90.3%。同时本文也在KITTI和CityScapes数据集上进行了可视化验证，体现了大核注意力的激活模块对图像全局场景结构的关注以及结构增强模块对关键区域特征的提取的有效性，相比基线方法[13]，本文方法在基于视频帧监督方式训练后测试的绝对相对误差、平方相对误差、均方根误差、均方根对数误差分别减少了3.1%，3.5%，1.9%，0.6%，在阈值为$\delta<1.25$的准确度指标上提升了0.33%。可视化图也表明本文预测的物体边缘更加贴合实际，不同深度平面区域预测伪影更少，不同深度区域的边缘深度信息区分明显。虽然本文单目深度估计工作获得了更高的精度，但同时由于卷积计算的增多导致网络模型复杂臃肿，与Manydepth相比，本文网络模型的参数量多出7.6 M。在以后的工作中，将考虑对单目深度估计网络进行轻量化，使之具有更小的参数量和更快的计算速度。

参考文献：

[1] 伍锡如，薛其威. 基于激光雷达的无人驾驶系统三维车辆检测[J]. 光学 精密工程，2022，30(4)：489-497.
WU X R，XUE Q W. 3D vehicle detection for unmanned driving systerm based on lidar[J]. *Opt. Precision Eng.*，2022，30(4)：489-497.(in Chinese)

[2] 史晓刚，薛正辉，李会会，等. 增强现实显示技术综述[J]. 中国光学，2021，14(5)：1146-1161.
SHI X G，XUE Z H，LI H H，*et al*. Review of augmented reality display technology[J]. *Chinese Optics*，2021，14(5)：1146-1161.(in Chinese)

[3] 鄢化彪，徐方奇，黄绿娥，等. 基于深度学习的多视图立体重建方法综述[J]. 光学 精密工程，2023，31(16)：2444-2464.
YAN H B，XU F Q，HUANG L /LÜ) E，*et al*. Review of multi-view stereo reconstruction methods based on deep learning[J]. *Opt. Precision Eng.*，2023，31(16)：2444-2464.(in Chinese)

[4] GARG R，B G V K，CARNEIRO G，*et al*. Unsupervised CNN for single view depth estimation：geometry to the rescue[C]. *European Conference on Computer Vision. Cham*：Springer，2016：740-756.

[5] ZHOU T H，BROWN M，SNAVELY N，*et al*. Unsupervised learning of depth and ego-motion from video[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI，USA*. IEEE，2017：6612-6619.

[6] GODARD C，AODHA OMAC，BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI，USA*. IEEE，2017：6602-6611.

[7] BIAN J W，LI Z C，WANG N，*et al*. Unsupervised scale-consistent depth and ego-motion learning from monocular video[C]. 33*rd Conference on Neural Information Processing Systems (NeurIPS). La Jolla*：Vancouver，2019：1-12.

[8] GODARD C，AODHA OMAC，FIRMAN M，*et al*. Digging into self-supervised monocular depth estimation[C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV). Seoul，Korea (South)*. IEEE，2019：3827-3837.

[9] YAO Y，LUO Z X，LI S W，*et al*. MVSNet：depth inference for unstructured multi-view stereo[C]. *European Conference on Computer Vision. Cham*：Springer，2018：785-801.

[10] WIMBAUER F，YANG N，VON STUMBERG L，*et al*. MonoRec：semi-supervised dense reconstruction in dynamic environments from a single

moving camera[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA.* IEEE, 2021: 6108-6118.

[11] SCHÖPS T, SCHÖNBERGER J L, GALLIANI S, *et al.* A multi-view stereo benchmark with high-resolution images and multi-camera videos [C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA.* IEEE, 2017: 2538-2547.

[12] KNAPITSCH A, PARK J, ZHOU Q Y, *et al.* Tanks and temples: Benchmarking large-scale scene reconstruction [J]. *ACM Transactions on Graphics*, 2017, 36(4): 1-13.

[13] WATSON J, AODHA OMAC, PRISACARIU V, *et al.* The temporal opportunist: self-supervised multi-frame monocular depth [C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA.* IEEE, 2021: 1164-1174.

[14] FENG Z Y, YANG L, JING L L, *et al.* Disentangling object motion and occlusion for unsupervised multi-frame monocular depth[C]. *AVIDAN S, BROSTOW G, CISSÉ M, et al. European Conference on Computer Vision. Cham*: Springer, 2022: 228-244.

[15] SHAO S W, PEI Z C, CHEN W H, *et al.* SMUDLP: Self-Teaching Multi-Frame Unsupervised Endoscopic Depth Estimation with Learnable Patchmatch [EB/OL]. 2022: *arXiv*: 2205.15034. http://arxiv.org/abs/2205.15034

[16] WANG X F, ZHU Z, HUANG G, *et al.* Crafting monocular cues and velocity guidance for self-supervised multi-frame depth learning[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(3): 2689-2697.

[17] LI R, GONG D, YIN W, *et al.* Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada.* IEEE, 2023: 21539-21548.

[18] GUO M H, LU C Z, LIU Z N, *et al.* Visual attention network [J]. *Computational Visual Media*, 2023, 9(4): 733-752.

[19] LIU W Z, LU H, FU H T, *et al.* Learning to upsample by learning to sample [C]. 2023 *IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France.* IEEE, 2023: 6004-6014.

[20] HE K M, ZHANG X Y, REN S Q, *et al.* Deep residual learning for image recognition [C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA.* IEEE, 2016: 770-778.

[21] DOSOVITSKIY A, FISCHER P, ILG E, *et al.* FlowNet: learning optical flow with convolutional networks [C]. 2015 *IEEE International Conference on Computer Vision (ICCV). Santiago, Chile.* IEEE, 2015: 2758-2766.

[22] WANG W H, XIE E Z, LI X, *et al.* PVT v2: improved baselines with pyramid vision transformer [J]. *Computational Visual Media*, 2022, 8(3): 415-424.

[23] YAN J X, ZHAO H, BU P H, *et al.* Channel-wise attention-based network for self-supervised monocular depth estimation[C]. 2021 *International Conference on 3D Vision (3DV). London, United Kingdom.* IEEE, 2021: 464-473.

[24] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C]. 2015 *IEEE International Conference on Computer Vision (ICCV). Santiago, Chile.* IEEE, 2015: 2650-2658.

[25] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]. *Advances in Neural Information Processing Systems. Montreal*: NIPS, Montreal, Canada, 2014.

[26] JOHNSTON A, CARNEIRO G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA.* IEEE, 2020: 4755-4764.

[27] GUIZILINI V, AMBRUS R, PILLAI S, *et al.* 3D packing for self-supervised monocular depth estimation[C]. 2020 ieee/cvf conference on computer vision and pattern recognition (cvpr). seattle, wa, usa. IEEE, 2020: 2482-2491.

[28] XIANG J, WANG Y, AN L F, *et al.* Visual attention-based self-supervised absolute depth estimation using geometric priors in autonomous driving [J]. *IEEE Robotics and Automation Letters*,

2022, 7(4): 11998-12005.

[29] SURI Z K. Pose Constraints for consistent self-supervised monocular depth and ego-motion [C]. *GADE R, FELSBERG M, KÄMÄRÄINEN JK. Scandinavian Conference on Image Analysis. Cham*: Springer, 2023: 340-353.

[30] BOULAHBAL H, VOICILA A, COMPORT A. STDepthFormer: Predicting Spatio-Temporal Depth from Video with A Self-Supervised Transformer Model [EB/OL]. 2023: *arXiv*: 2303.01196. http://arxiv.org/abs/2303.01196

[31] PATIL V, VAN GANSBEKE W, DAI D X, *et al*. Don't forget the past: recurrent depth estimation from monocular video[J]. *IEEE Robotics and Automation Letters*, 2020, 5(4): 6813-6820.

[32] SAUNDERS K, VOGIATZIS G, MANSO L J. Self-supervised monocular depth estimation: Let'S talk about the weather[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *Paris*, *France*. IEEE, 2023: 8873-8883.

[33] WANG J R, ZHANG G, WU Z Y, *et al*. Self-supervised Joint Learning Framework of Depth Estimation Via Implicit Cues[EB/OL]. 2020: *arXiv*: 2006.09876. http://arxiv.org/abs/2006.09876

[34] SHU C, YU K, DUAN Z X, *et al*. Feature-metric Loss for self-supervised learning of depth and egomotion[C]. *European Conference on Computer Vision. Cham*: Springer, 2020: 572-588.

[35] LI H H, GORDON A, ZHAO H, *et al*. Unsupervised monocular depth learning in dynamic scenes [C]. *Conference on Robot Learning*. PMLR, 2021: 1908-1917.

作者简介:

寇旗旗(1988—),男,河南襄城人,博士,副教授,硕士生导师。主要研究方向为图像处理/AI视频分析、智能检测与模式识别、图像增强与复原。E-mail: kouqiqi@cumt.edu.cn

通讯作者:

姬玉成(1987—),男,河南商丘人,博士,主要研究方向为矿山智能化、机器视觉分析。E-mail: j.yc@outlook.com