文章编号：2095-2163(2024)01-0168-08 中图分类号：TP391.41 文献标志码：A

卷积神经网络与 Transformer 混合的自监督单目深度估计算法

方 俊[1]，刘红喜[2]，穆晓飞[1]
（1 吉林化工学院 信息与控制工程学院，吉林 吉林 132022；2 长春工程学院 电气与信息工程学院，长春 130012）

摘 要：自监督深度估计是一种具有广阔前景的研究领域，无需来源困难的真实深度标签即可实现对模型的训练。模型通常采用卷积神经网络和 Transformer 两种方法进行特征提取，卷积神经网络能够有效地提取局部特性，但由于感受野小而缺乏全局特性提取的局限性。而 Transformer 能保证全局特征的提取，但计算量较大。针对于此，提出一种卷积神经网络与 Transformer 相结合的自监督深度估计模型进行深度特征提取，以确保该模型既能加强全局特征提取同时又不会损失局部特征。该模型在 KITTI 数据集上进行验证，从实验结果可以看出所提模型的绝对相对误差和平方相对误差分别下降至 0.100 和 0.698，精确度达到 89.7%，与其他先进的算法对比有着较好的提升，并在 Cityscapes 数据集中有着不错的泛化性。

关键词：深度估计；自监督学习；Transformer；通道注意力

Self-supervised monocular depth estimation based on convolution neural network hybrid transformer

FANG Jun[1], LIU Hongxi[2], MU Xiaofei[1]
(1 School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin Jilin 132022, China; 2 School of Electrical and Information Engineering, Changchun Institute of Engineering, Changchun 130012, China)

Abstract: Self-supervised depth estimation is a promising research area that enables training of models without the need to source difficult real depth labels. The model usually uses two methods for feature extraction, convolutional neural network and Transformer, the convolutional neural network can effectively extract local features, but lacks the limitation of global feature extraction due to the small sensory field. While Transformer can ensure the extraction of global features, but the amount of computation is large. To address this, a self-supervised depth estimation model combining convolutional neural network and Transformer is proposed for deep feature extraction to ensure that the model can enhance global feature extraction without losing local features. The model is validated on the KITTI dataset, and from the experimental results, it can be seen that the absolute relative error and the squared relative error of the proposed model are reduced to 0.100 and 0.698, respectively, with an accuracy of 89.7%, which is a good improvement compared with other state-of-the-art algorithms, and it has a good generalization in the Cityscapes dataset.

Key words: depth estimation; self-supervised learning; Transformer; channel attention

0 引 言

在计算机视觉领域中，从二维图像还原三维场景信息，是重要的研究方向。由二维图像还原三维场景信息是为获取二维图像上每一个像素点与相机之间的距离，即深度信息。如今，场景中深度信息已被运用于 3D 人脸识别，辅助驾驶，机器人定位导航以及其它日常生活。

前期基于深度学习的深度估计算法研究多聚焦于有监督，依据具有真实深度标签实现单幅图像深度回归。Eigen 等[1-2]首先提出利用卷积神经网络和多尺度框架相结合的方法，对单目图像进行深度估计。但具有真实深度的标签较难获得，造成实验成本较高。因此，一些研究人员开始对自监督深度估计算法进行研究。自监督深度估计算法是由深度回归问题转化为图像重建问题，Grad[3]等将双目图像用作训练

基金项目：吉林省教育厅科学技术研究项目(JJKH20210691KJ)。

作者简介：方 俊(1999-)，男，硕士研究生，主要研究方向：深度学习、计算机视觉、图像处理；穆晓飞(1997-)，男，硕士研究生，主要研究方向：计算机视觉、图像处理。

通讯作者：刘红喜(1977-)，男，硕士，副教授，主要研究方向：计算机视觉、机器学习。Email：liuhongxi@ccit.edu.cn

收稿日期：2022-12-25

集，通过对两幅影像间差值与双目相机间位置关系的预测，估算出影像真实深度值，从而实现无监督单目深度的估算。Zhou[4] 等提出利用视频序列对模型进行训练，使得自监督深度估计模型的训练不再局限于双目数据集上，并提出一种位姿网络，该网络能够计算出两幅输入图像之间的相对位置。

近年来，在计算机视觉领域，卷积神经网络通过局部连接、权重共享及汇聚等特性，极大地提升了神经网络的学习能力。在深度估计方面，也有很多研究通过变换的卷积神经网络结构，来进一步提升网络性能。Yin 等人[5] 利用 ResNet 替代 VGG 网络编码器[4]，以增加网络深度增强网络学习能力。PackNet[6] 模型中，通过使用 3D 卷积来增强网络对细节特征的提取能力。然而卷积神经网络受限于模型参数，大部分网络主体使用的是 3×3 卷积，导致网络感受野较小而造成不利于全局特征抽取。而对于单目深度估计任务，有效的全局特征可以保证网络学习到更加精确的深度信息。

Transformer 最早出现于自然语言处理领域，但随着研究的不断深入与完善，近来 Transformer 已在计算机视觉方向上获得不错的结果，Transformer 能够对像素间的长期依赖关系进行建模并生成全局范围内的接受域。此外，在深度估计方面，Transformer 亦有涉及。梁水波等[7] 使用 Transformer 作为编码器，为深度估计网络提供全局特征。然而由 Ttransformer 的复杂性导致训练计算量大，可以通过卷积神经网络与 Transformer 混合模型来减小模型的计算量并且保证网络学习到全局特征和局部特征。刘佳涛等[8] 在有监督的深度估计算法中，使用卷积神经网络与 Transformer 混合模型，提升模型对深度预测的准确性。混合模型主要在有监督单目深度估计中应用。

深度估计模型算法是以 U-Net 为框架[9]，解码器环节直接使用跳过连接与普通卷积，实现浅层与深层特征融合和提取。但该模型无法保留更多有效的浅层信息，将造成不同特征间融合效率较低以及在物体边界上出现模糊伪影等问题。

为解决上述问题，将卷积神经网络和 Transformer 的混合模型运用到自监督深度估计网络中，并提出一种新深度估计网络模型，充分利用 Transformer 和卷积神经网络各自特点，较好地感知全局场景结构以及局部细节，并将通道注意力机制引入到该模型解码器中，给通道方向上的特征赋予权重，从而强化关键的浅层特征信息。

1　单目深度模型

1.1　自监督深度估计算法原理

自监督深度估计算法和有监督深度估计算法不同，是在没有真实深度数据的情况下来训练模型，从单一的彩色图像来预测深度图像。自监督深度估计算法的本质是图像重建，从相邻帧图像来重建当前帧图像，由原当前帧图像作为监督来对重建图像进行优化。将两帧视频图像输入进模型进行训练，当前帧表示为 I_s，相邻帧表示为 I_t，重投影的当前帧表示为 I_r。根据相机坐标转换公式，对于图像像素的重投影可表示为

$$P_r = K\ [R,T]^{-1} K^{-1} D P_t \tag{1}$$

式中：P_r 表示重投影中的二维像素点，P_t 是相邻帧图像中的二维像素点，K 为相机的内参，$[R,T]$ 代表相机的外参，是通过将 I_s 与 I_t 两张图像输入到位姿估计网络进行预测，输出带有 6 个自由度的图像相对位姿信息，D 为深度信息，由深度估计网络输出得到。

由式(1)可知重建图像与相邻帧图像的像素坐标对应关系，即可通过像素坐标对应关系，来对相邻帧图像进行双线性插值采样重建当前帧图像。

1.2　深度估计网络框架

深度估计网络整体结构如图 1 所示，由编码器和解码器组成的网络，能够通过一张彩色图片较为准确地预测出对应的深度图。对于编码器的设计参考 Next -ViT[10]，使用其中新提出的卷积神经网络与 Transformer 混合模块。通过一个卷积模块和 4 个混合模块来对图像进行特征提取和 5 次下采样。在解码器部分通过 SE(Squeeze-and-Excitation)[11] 模块增强浅层特征后，与上采样的深层特征进行融合，最后在预测模块预测 4 个不同尺度的深度图。

1.2.1　基于混合模型的特征提取编码器

传统混合模型，一般在浅层将卷积进行堆叠，而在网络最后两层中堆叠 Transformer。然而传统的模型不能很好适用在深度估计这类需要在解码器部分提取特征信息，并在编码器部分中进行特征融合的实验，这是由于对于浅层网络中仅存在卷积捕获的局部特征信息。因此，该设计的编码器分为 5 个阶段，第一个阶段是由 3 个普通卷积组成，将高宽为 $H \times W$ 的图片输入到第一阶段通过第一个步距为 2、卷积核大小为 3 的卷积对图片进行一个下采样操作，生成高宽分别为 $H/2$ 和 $W/2$ 的特征图。第二阶段到第五阶段都是先对输入特征图进行下采样，然

后使用混合模块,如图 2 所示,保证每个阶段输出的图像大小分别为原图的 1/4、1/8、1/16、1/32。对于每个混合模块先是将 MHCA(多头卷积注意力)和 MLP(多层感知机)模块进行堆叠,提取局部细节特征并避免多次使用 Transformer 导致模型的复杂度与计算量增加,在混合模块的最后一层加入 Transformer,保证每个阶段的输出特征都带有全局和局部信息。

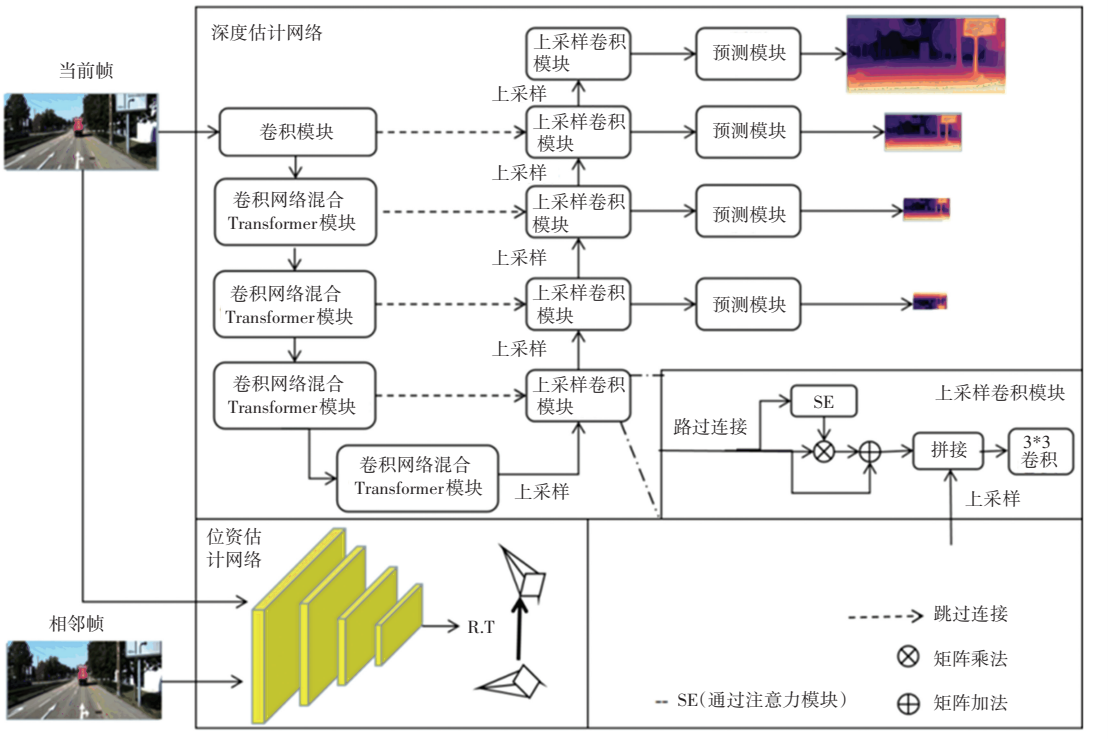


图 1 自监督单目深度估计网络框架

Fig. 1 Network framework of self-supervised monocular depth estimation

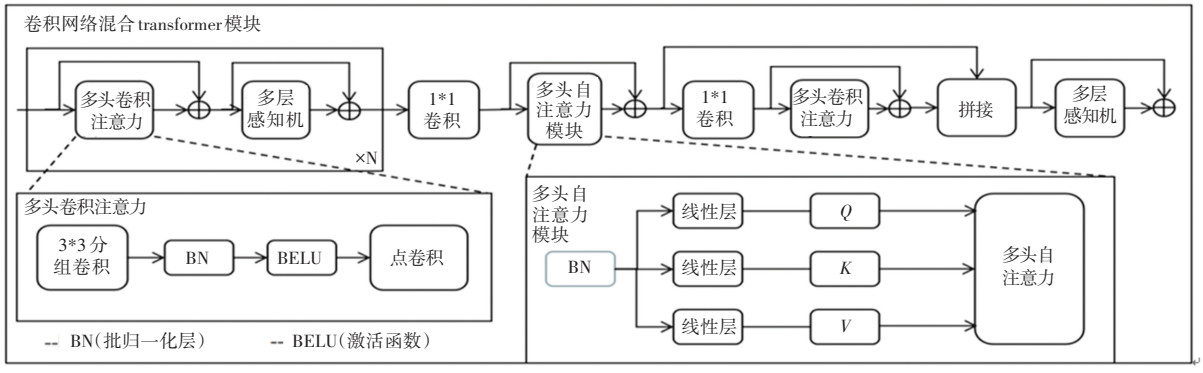


图 2 卷积神经网络与 transformer 混合模块示意图

Fig. 2 Convolution neural network and transformer hybrid module

为减少模型的参数量实现高效的局部特征提取,使用 MHCA 模块替代普通卷积。MHCA 模块由 GConv(分组卷积)、BN(批归一化)、ReLU(激活函数)、PWConv(点卷积)组成。MHCA 模块通过分组卷积将特征图在通道维度上分成多个子空间,对其分别进行卷积大大减小参数量,再通过 PWConv 来增加多个子空间之间的联系,通过该方法来加强对不同子空间的特征关注,实现卷积高效的局部特征信息学习。MHCA 模块的结构可表示为

$$\mathrm{MHCA}(f) = \mathrm{GC}(f) \otimes W_p \tag{2}$$

$$\mathrm{GC}(f) = \mathrm{concat}(f_1 \otimes W_1, f_2 \otimes W_2, \cdots, f_h \otimes W_h) \tag{3}$$

输入模块的特征图为 $f \in R^{H\times W\times C}$,且 $\{f_1, f_2, \cdots, f_h\} \in f$。$h$ 为 GConv 的分组数,该处设置为 32。公式(3)中的 $\{W_1, W_2 \cdots W_h\} \in W_g$,且 W_p、$W_g \in R^{C\times C}$ 分别为 PWConv 的训练权重和 GConv 的训练权重,$\otimes$ 表示卷积操作。使用普通 3×3 卷积的参数量为

$$C_{\text{in}} \times C_{\text{out}} \times K^2 + C_{\text{out}} \tag{4}$$

式中：C_{in} 为输入的通道数量，C_{out} 为输出的通道数量，由于此处不改变通道数即 $C_{\text{in}} = C_{\text{out}}$。$K$ 为卷积核的尺寸，该处 K 设置为 3。而 MHCA 的参数数量为

$$\frac{C_{\text{in}}}{h} \times C_{\text{out}} \times K^2 + C_{\text{in}} \times C_{\text{out}} + 2 \times C_{\text{out}} \tag{5}$$

对比式(4)和式(5)，可以明显比较出使用 MHCA 模块可以大大减小参数量，且再使用 PWConv 亦保证特征提取的效果。

对于通过堆叠的 MHCA 和 MLP 提取局部信息后，再通过 MHSA(多头自注意力模块)实现对全局信息进行捕获。由于 MHSA 模块的输入是一维向量，因此对于输入的二维图片信息需要压缩成一维，再分别通过 3 个不同线性层的训练参数 W_k、W_q、$W_v \in R^{C\times C}$ 来分别映射得到 K、Q、V 3 个同源的数据。K、Q、V 分别代表的意思为 key、query、value，后续操作对 K 与 Q 计算两者的相关性来作为 V 的权重，最后通过一个线性层来对增强后的全局特征进行提取，最终实现自注意力对全局特征的学习。MHSA 模块可表示为：

$$\text{MHSA}(X) = \text{concat}(\text{SA}(X_1),\text{SA}(X_2),\cdots,\text{SA}(X_h)) \tag{6}$$

$$\text{SA}(X_h) = \text{softmax}\left(\frac{X_h W_q \ (X_h W_k)^{\text{T}}}{\sqrt{d}}\right)(X_h W_v) \tag{7}$$

原输入的 $X \in R^{H\times W\times C}$，将输入的特征图压缩成一维数据，即 $X \in R^{H\times W\times C}$。$d$ 是缩放因子，用来防止 Q 与 K 中计算相关度进行点乘后的数值过大。softmax 计算 Q 与 K 中的相关度并将值限定在[0,1]范围作为 V 的权重。

将 MHSA 提取到的特征再通过 MHCA 进一步加强局部信息，并在通道维度将 MHSA 的输出和 MHCA 的输出进行拼接，通过 MLP 层来提取更重要和更明显的特征信息。

1.2.2　基于 SE 模块的解码器

解码器的一个重要作用，是将编码器中浅层特征和解码器中的深层特征进行融合，防止特征图在上采样后丢失浅层特征。本文研究设计的编码器结构中包含上采样卷积模块和预测两个模块，其中上采样卷积模块的主要作用是在编码器中融合浅层与深层特征并提取进一步特征信息。在该特征融合部分加入 SE[11] 通道注意力机制，保证特征融合时能有效利用提取到的特征信息。对输入 SE 模块的特征信息 $x \in R^{H\times W\times C}$，通过 AvgPool(Average Pooling)将 $H \times W$ 大小的特征图压缩到 1×1，再通过 FC(Fully Connected)对通道注意力进行学习，最后通过 sigmoid 将输出值映射到[0,1]，最终输出 $y \in R^{1\times 1\times C}$：

$$\begin{aligned} \hat{x} &= \frac{1}{W \times H}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} x_c(i,j) \\ \hat{x} &= \max(\hat{x}W_1, 0) \\ y &= \text{sigmoid}(\hat{x}W_2) \end{aligned} \tag{8}$$

式中：W_1 与 W_2 是两个全连接的学习权重。SE 模块输出 y 作为原特征信息 x 的权重，有效地增强 x 中的重要特征信息。在上采样卷积模块中，最后通过边缘填充、卷积核为 3 的普通卷积，对增强后的特征进行特征融合、特征提取。在卷积前对特征图进行边缘填充，是为了防止特征图丢失边缘信息。该深度估计模型采用的是多尺度预测结构，即预测的深度图像分辨率分别为原图的 1/8、1/4、1/2、1。因此，在经过上采样卷积模块特征图的分辨率为原图的 1/8、1/4、1/2、1 时，需要利用预测模块将特征信息聚合起来进行深度预测，并最终输出对应大小的相对深度图像。

1.3　损失函数

对于自监督深度估计网络训练的损失函数，采用 Godard 等[12-13] 的相关工作。深度估计模型的损失函数主要由最小光度误差损失 L_p 和平滑误差损失 L_{smooth} 两部分组成。最小光度误差损失 L_{photo}，由原图像和重建图像的 $L1$ 损失函数与两张图像的结构相似性损失函数[14]($SSIM$) 组成，最小光度误差损失函数为

$$\begin{aligned} L_p = \min\Big(\sum_{p\in I_r}\lambda \ \| I_s(p) - I_r(p) \|_1 + \\ (1-\lambda)\frac{1 - SSIM(I_s, I_r)}{2}\Big) \end{aligned} \tag{9}$$

式中：p 为重建图像 I_r 中的像素点，λ 与 $(1-\lambda)$ 是 $L1$ 损失和相似性损失的权重，该处 $\lambda = 0.15$。

由于无纹理或低纹理区域光度重建损失的影响将减弱，需添加平滑误差损失对整个深度估计模型优化方向进行约束，平滑误差损失函数公式为

$$L_{smooth} = e^{-|\partial x I_s|}|\partial x \bar{D}_s| + e^{-|\partial y I_s|}|\partial y \bar{D}_s| \tag{10}$$

式中：∂_y、∂_x 表示对 y 与 x 方向上求偏导数，$\bar{D}_s$ 表示经过归一化的深度图，方便计算梯度。$e^{-|\partial_x I_s|}$ 表示作为 x 方向上深度值梯度的权重，对于偏导数在低纹理区域时，得到的梯度值小，从而深度值的整个权重值增大，以此增大损失函数的值，保证在低纹理区域模型的优化。

模型采用多尺度预测,网络模型输出 4 个不同分辨率下的深度图像,最终自监督深度估计损失函数为

$$L_{total} = \frac{1}{S}\sum_{0}^{S-1}(\alpha L_{photo} + \beta L_{smooth}) \tag{11}$$

式中:S 为深度估计网络预测不同尺度深度图的数量,α、β 是两个损失函数的超参数。

2 实验与分析

2.1 数据集和实验设置

KITTI 是一个用于深度和姿态估计的视觉数据集。该数据集包含 5 天内采集的 200 个 RGB 摄像机拍摄的市区、郊区和高速公路视频,以及激光扫描仪拍摄的深度图像,采集到的图像分辨率大小为 1 242×375。为验证算法的有效性,方便与其他算法进行比较,实验采用 Eigen 对 KITTI 数据集分割的方法,将39 810张图像作为训练集,4 424 张图像作为验证集,进行模型的训练。实验测试中,分别从原始激光雷达采集的 697 幅图像数据和经过改进[15]的带有真实深度信息的 652 张图片等分别测试进行模型性能评估。

实验通过显存为 11 G 的 NVIDIA GeFo rce RTX 2080TI 来对网络模型进行训练,采用 PyTorch 深度学习框架。模型的优化器采用 Adam,初始学习率设置为1×10^{-4},学习率在第 15 个 epoch 时会下降到1×10^{-5},以该学习率完成最后的训练。将数据设置为 4 个一组的批量输入到模型中训练,位姿估计网络采用 resnet50。对于从 KITTI 数据集输入的图片分辨率全部缩放到 640 * 192 大小,且编码器部分均使用在 ImageNet 数据集上预先训练过的权重作为模型的初始化参数。

2.2 评价指标

实验评价指标采用绝对相对误差(Abs-Rel)、平方相对误差(Sq Rel)、均方根误差(RMSE)以及均方根对数误差(RMSE Log)等深度评价的标准指标。此外,使用不同阈值时,对估算深度的准确性进行计算。

2.3 定量分析

对比当前几个较为先进的自监督深度估计模型,见表 1。表 1 内都是通过分辨率为 640 * 192 的单目视频来进行训练的模型,训练方式处 M 代表使用视频序列进行训练。表格的上半部分是在 Eigen 分割的原始数据上进行测试,下半部分是在经过改进的真实深度数据中进行测试。

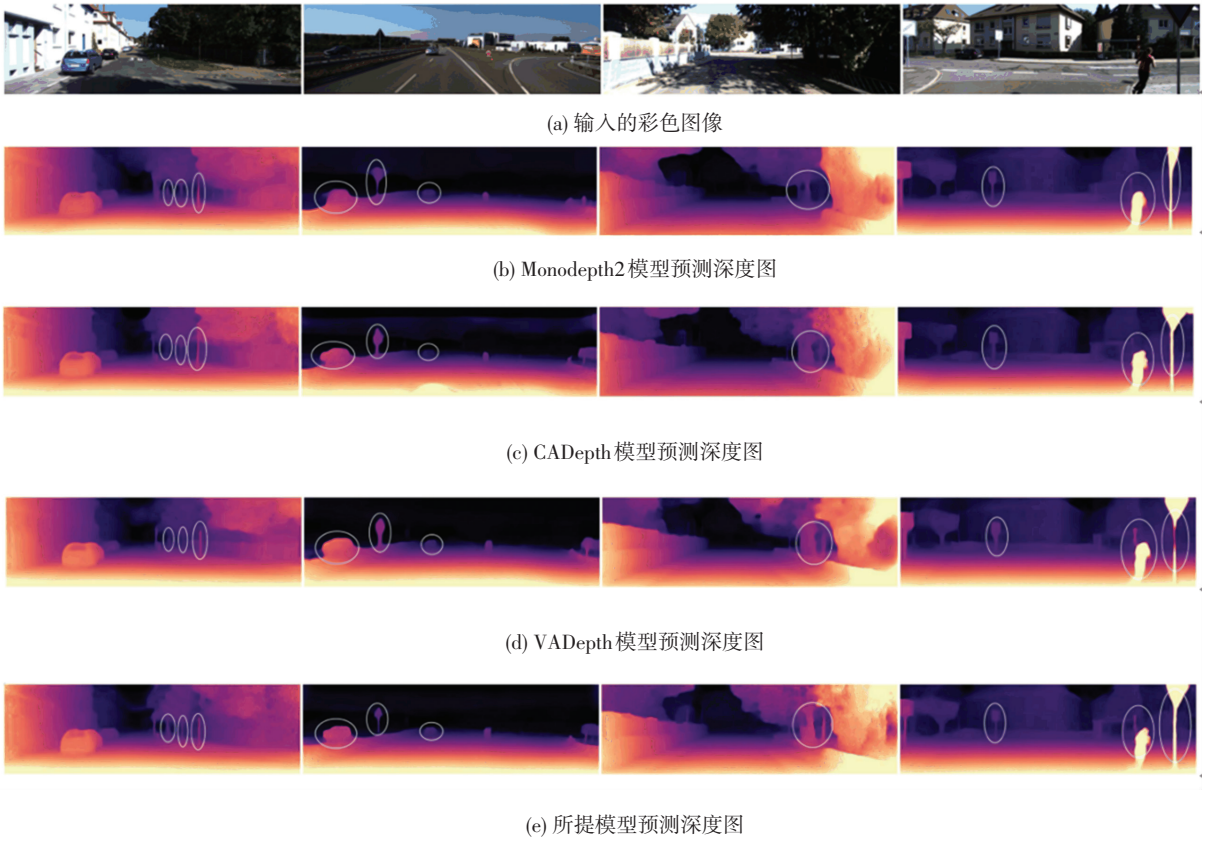
由测试结果中可以看出,所提模型在误差和精度上有一定的提升。在 Eigen 测试集中,所提模型的绝对相对误差相比 VADepth 下降 0.004,由于通过使用 SE 模块加强浅层中重要的特征,物体边界更加清晰,减少了边界模糊现象,导致实验结果的误差值下降。在阈值为 $\delta < 1.25$ 的精确度上,相比 VADepth 上升 0.5%,该模型能有效的提取全局和局部的深度信息,丰富网络可学习的深度特征,来实现更加精确的深度预测。

表 1 与不同深度估计算法预测结果的定量比较

Table 1 Quantitative comparison with prediction results of different depth estimation algorithms

测试集	方法	训练方式	误差(越低越好)				精度(越高越好)		
			Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen split	Monodepth2[13]	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	VC-Depth[16]	M	0.112	0.816	4.715	0.190	0.880	0.960	0.982
	HR-Depth[17]	M	0.109	0.792	4.632	0.185	0.884	0.962	0.983
	GCNdepth[18]	M	0.104	0.720	4.494	0.181	0.888	0.965	0.984
	VADepth[19]	M	0.104	0.774	4.552	0.181	0.892	0.965	0.983
	CADepth[20]	M	0.105	0.769	4.535	0.181	0.982	0.964	0.983
	本文方法	M	0.100	0.698	4.376	0.176	0.897	0.967	0.984
Improved ground truth[15]	Monodepth2	M	0.090	0.546	3.940	0.137	0.914	0.983	0.995
	CADepth	M	0.080	0.452	3.649	0.125	0.927	0.986	0.996
	VADepth	M	0.078	0.430	3.593	0.121	0.931	0.988	0.997
	本文方法	M	0.077	0.395	3.472	0.118	0.933	0.988	0.997

本文所提模型与 Monodepth2、CADepth 及 VADepth 等模型预测深度图的效果比较结果如图 3 所示。由于使用 SE 通道注意力模块，加强特征的融合，增强物体的边界信息。因此与其他几个模型进行比较，可以看到对人、路标、车辆和树干等对象的边界将更加锐利。自监督模型最大测试深度是 80 m，深度图将在 80 m 以上的距离处出现黑色。从图片中可以看到 Monodepth2 模型和 CADepth 模型对天空等 80 m 以上的距离处出现较多的错误深度估计，而本文模型使用卷积神经网络和 Transformer 混合来提取全局和局部深度信息，因此能够获取更多的深度信息，亦能对更远的物体进行更准确的预测，减小模型对远处深度的错误估计。此外，相对于 VADepth模型还降低了部分深度丢失的问题，相比其他几种先进的自监督模型具有更好的视觉效果。



(a) 输入的彩色图像

(b) Monodepth2模型预测深度图

(c) CADepth模型预测深度图

(d) VADepth模型预测深度图

(e) 所提模型预测深度图

图 3　在 KITTI 数据集上不同模型预测的深度图

Fig. 3　Depth maps predicted by different models on KITTI dataset

为测试模型的泛化性，将在 KITTI 数据集中训练的模型，通过 Cityscapes 数据集中的 1 500 张图像来测试模型效果。Cityscapes 数据集中采集了不同季节的 50 个城市内的街景，并且用来测试的图片有部分是在阴天这类光线并不充足的情况下，这表明该数据集更具有多样性和挑战性。表 2 中，所提模型对比 CADepth 模型在阈值为 $\delta < 1.25$、$\delta < 1.25^2$ 的精确度上分别高出 1.8% 和 1.2%，表明混合模型在更具有挑战性的数据集中亦能捕获到更多有效的深度信息。

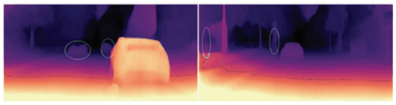
从预测的深度图中可以看出，对比其他两个模型，所提模型的物体边界预测要更锐利，证明网络通过 SE 模块从浅层特征中学习到重要特征，对目标边界预测更加准确。在图 4 中，从不同模型对于第一张图片的预测结果中可以看出，即使在阴天的情况下，亦能较好的捕获图片中的深度信息，进行更精确的深度预测。

表 2　在城市景观的数据集中泛化性比较

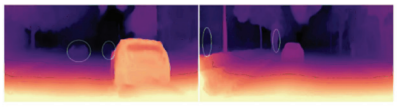
Table 2　Comparison of generalization in Cityscapes dataset

方法	误差(越低越好)		精度(越高越好)	
	Abs rel	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$
Monodept2	0.201	0.318	0.681	0.874
CADepth	0.186	0.307	0.714	0.881
本文方法	0.178	0.289	0.732	0.893

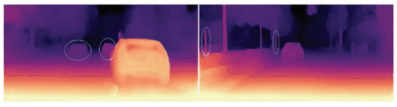
(a) 输入的彩色图像



(b) Monodepth2模型预测深度图



(c) CADepth模型预测深度图



(d) 所提模型预测深度图

图 4 Cityscapes 数据集不同模型预测深度图

Fig. 4 Depth maps predicted by different models on Cityscapes dataset

2.4 消融研究

为进一步证明卷积神经网络和 Transformer 的混合模型与 SE 模块的有效性,进行了相关的消融实验,实验结果见表 3。表 3 中第一行为基础模型,仅选中 CNN 部分表示模型采用 ResNet50 作为编码器进行特征提取,当同时选中 CNN 和 Transformer 时,表示模型的编码器部分使用混合模型来提取特征;当选择 SE 部分,表示解码器部分加入通道注意力机制。当使用混合模型作为编码器可以捕获全局特征和局部特征,网络可以捕获到更多重要的深度特征信息。实验结果表明,对比基础模型使用混合模型的 $\delta < 1.25$ 精度提升 1%,主要由于网络从更丰富的深度信息中进行学习,能更精确进行深度预测。当在解码器部分使用 SE 模块更改通道方向的权重,增强浅层特征中重要的特征,网络能更有效学习到重要特征,减少学习过程偏差的出现。在实验结果中,对比基础模型增加 SE 模块平方相对误差下降 0.035。

表 3 网络消融分析

Table 3 Ablation analysis of networks

方法			误差(越低越好)				精度(越高越好)		
CNN	Transformer	SE	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
√			0.110	0.792	4.629	0.186	0.880	0.962	0.983
√	√		0.103	0.735	4.469	0.179	0.890	0.964	0.984
√		√	0.106	0.757	4.593	0183	0.885	0.963	0.983
√	√	√	0.100	0.698	4.376	0.176	0.897	0.967	0.984

3 结束语

本文提出了一种新的自监督单目深度估计模型。通过在网络编码器内使用卷积神经网络和 Transformer 混合,可以联合建模图像的全局特征和局部特征;同时在解码器部分引入通道注意力机制,进一步有效利用提取到的浅层特征。在 KITTI 数据集上测试证明,所提模型可以获取到更多的深度信息,与当前先进的模型相比能对更远的深度进行更准确的预测。此外,在 Cityscapes 数据集上的实验表明,模型具有不错的泛化性能。

参考文献

[1] 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, 2014: 2366-2374.

[2] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi - scale convolutional architecture [C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 2650-2658.

[3] GARG R, BG V K, CARNEIRO G, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C] // European Conference on Computer Vision. Springer, Cham, 2016: 740-756.

[4] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1851-1858.

[5] YIN Z, SHI J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1983-1992.

[6] GUIZILINI V, AMBRUS R, PILLAI S, et al. 3d packing for self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2485-2494.

[7] 梁水波,刘紫燕,孙昊堃,等. Transformer 与多尺度注意力的自监督单目图像深度估计[J]. 小型微型计算机系统,2023,44(4): 825-831.

(下转第 179 页)