周晓吉. 基于多尺度特征融合的单目深度估计算法[J]. 智能计算机与应用,2024,14(9):34-40. DOI:10.20169/j.issn.2095-2163.240905

基于多尺度特征融合的单目深度估计算法

周晓吉
(浙江理工大学 信息科学与工程学院, 杭州 310018)

摘 要: 在当前的单目深度算法中,堆叠的卷积层和过度的下采样操作会造成特征图分辨率和高层信息的损失,影响了深度图整体的精度。针对这一问题,本文提出了一个基于多尺度特征融合的单目深度估计算法。采用了递进式的编-解码结构,由浅到深逐级提取不同尺度的信息,不同层级不同分辨率的特征连接在一起,形成了多尺度特征融合结构;编码器采用 U^2-Net 的设计架构,内部通过 Vision Transformer 模块,使得模型能够在编码过程中拥有全局的感受野,并且避免了下采样操作,从而减少了特征图分辨率和高层信息的损失;解码器中设计了 U 型残差块,能更好地融合不同阶段内的多尺度特征。在 KITTI 和 NYU-Depth V2 数据集上进行了实验,实验结果表明本文所提算法在各项指标上优于大部分同类型算法。

关键词: 单目深度估计; 编-解码器结构; Vision Transformer; U^2-Net
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)09-0034-07

Monocular depth estimation algorithm based on multi-scale feature fusion

ZHOU Xiaoji
(School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In the current Monocular depth estimation algorithms, stacked convolutional layers and excessive downsampling operations lead to the loss of feature map resolution and high-level information, affecting the overall accuracy of the depth map. To address this issue, this paper proposes a monocular depth estimation model based on multi-scale feature fusion. The model adopts a progressive encoder-decoder structure to extract information of different scales from shallow to deep levels. Moreover, the features of different resolutions at different levels are connected to form a multi-scale feature fusion structure. The encoder is inspired by the design of Transformer, which has a global receptive field during encoding, while avoiding downsampling operations to reduce the loss of feature map resolution and high-level information. The decoder incorporates U-shaped residual blocks to better fuse multi-scale features within different stages. Our method was tested on the KITTI and NYU Depth V2 datasets, and the experimental results showed that it exhibited competitive performance on both datasets.

Key words: Monocular depth prediction; encoder-decoder structure; Vision Transformer; U^2-Net

0 引 言

单目深度估计通过拍摄单一视角的单张彩色图像,估计出图像中每个像素到拍摄源的距离,是计算机视觉中的一个重要领域。深度估计应用场景十分广泛,例如增强现实、三维重建、自动驾驶等,能够帮助机器人和自动驾驶车辆感知周边环境,提高增强现实应用的表现效果。

早期处理单目深度估计问题有两种方法:第一种是通过相机的平移和旋转等运动,对同一场景拍摄多帧图片,然后通过计算图像的特征点在不同图像间的位置差异,从而推断相机的运动信息,最后利用三角测量原理计算物体的深度信息。该方法虽然不需要复杂的成像设备,但是对拍摄图像的数量、相机运动等都有一定的限制。第二种是通过分析单目图像中的视觉深度线索,来推断物体的距离和位置。在单目深度估计中,常用的深度线索包括线性透视、聚焦、散焦、大气散射、阴影、纹理、遮挡和相对高度等。然而,单独利用某一种深度线索进行深度估计的精度有限,故通常需要结合多种深度线索以提高估计的准确性和鲁棒性。此外,深度估计的精度还会受到图像质量、场景复杂度、相机参数等因素的影响,因此需要综合考虑多种因素进行考量。

近年来,基于深度学习的方法因其强大的表征

作者简介: 周晓吉(1997-),女,硕士研究生,主要研究方向:计算机视觉。Email:1260752208@qq.com

收稿日期: 2023-05-16

能力在该领域受到广泛关注。Eigen 等[1]提出了第一个单目深度估计网络，该网络分为两个阶段，第一阶段通过输入图像对场景进行全局预测，然后将预测结果与原图像叠加输入第二阶段网络，从而优化深度图的局部细节，但由于网络层数少、感受野小，特征提取能力弱，得到的深度图较为模糊；Chen 等[2]提出了基于注意力模型的聚合网络（ACAN），该网络能够自适应地学习像素间的相似性，以对上下文信息进行建模，但编码器的下采样操作仍损失了特征分辨率和空间信息，导致估计结果不够理想；Ranftl 等[3]引入了密集视觉自注意力模型（Dense Vision Transformer）代替卷积网络作为预测阶段的主干网络，但由于没有对提取到的特征进行足够的处理，导致估计结果边界有些模糊。

本文针对上述方法的不足，提出了一种基于多尺度特征融合的单目深度估计算法。该算法通过多个 Transformer 和小型 U-Net 的堆叠，形成了整体的 U 型结构，能够在不显著增加内存和计算成本的情况下，构造出更深层的网络并获得更高分辨率的输出。编码器采用 Vision Transformer（ViT）的结构，避免在网络中使用下采样操作，从而使得网络编码器能够在每个阶段维持全局的感受野，减少特征图分辨率和高层信息的损失。解码器采用 U 型残差块，能够混合不同大小的感受野，从多尺度特征中恢复出深度信息。

1　基于多尺度特征提取的单目深度估计算法

1.1　基础网络

网络的整体结构借鉴了 U^2-Net 模型，形成了一个两级嵌套的 U 形结构，如图 1 所示。U 型的主体结构能够构造出更加深层的网络，由浅到深逐级提取不同尺度的信息。同时，不同层级不同分辨率的特征连接到一起，形成了多尺度的特征融合结构，实现了全局信息与局部信息的融合、大尺度特征和小尺度特征的融合。

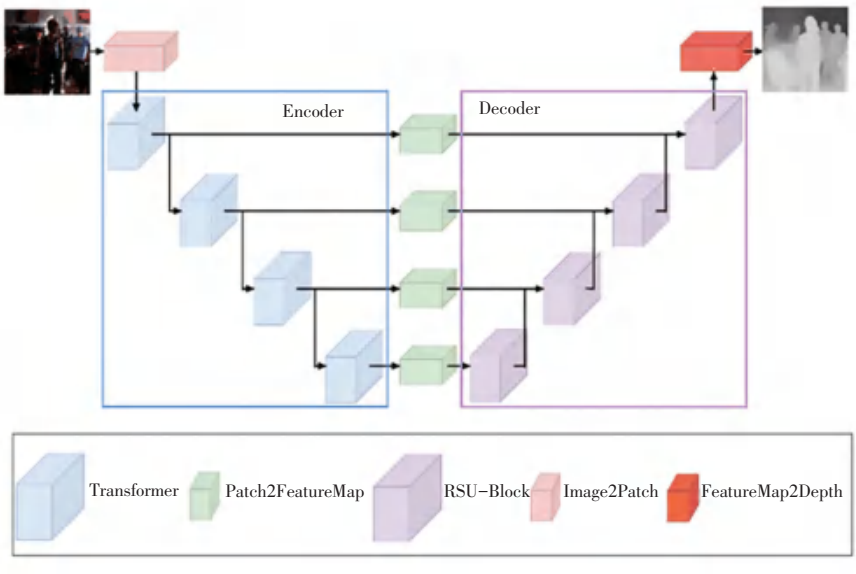

图 1　深度估计网络结构图

Fig. 1　Depth estimation network structure

整个网络可分为 5 个模块，粉色部分是 Image2Patch 模块，用于将图片分割成多个小图块，作为编码器的输入；蓝色的 Transformer 模块基于 ViT 编码器结构，对输入的图块或特征图进行编码；绿色部分为 Patch2FeatureMap 模块，用于将不同分辨率编码阶段的输出，转化为对应解码阶段的特征形式，并与深层解码阶段的信息一起送入同一阶段解码器之中；紫色部分为 RSU-Block 模块，由借鉴 ReSidual U-blocks（RSU）模块设计的 U 型残差块组成，是一个简化版的小型 U-Net 解码器，作用是对图像特征进行融合；红色部分为 FeatureMap2Depth 模块，用于将融合后的特征图转化为深度图。

这样的网络组织方式在底层，ViT 能够在不降低特征图分辨率的情况下更好地提取多尺度特征；在顶层，嵌套的 U 型结构能够更好地聚合阶段间的多尺度特征。

1）编码器结构

本文选取 ViT 网络作为编码器。ViT 使得特征

图在编码处理过程中能够维持较高分辨率,减少特征分辨率和粒度的损失。输入图像首先通过 Image2Patch 模块转化为编码器的输入,该模块将图像切分成 N 个 $p \times p$ 大小的图块(patch),之后 patch 展平为向量并附加上位置编码(token);另外,还在 token 序列前添加了一个可学习的图像类编码(class token)。ViT 的具体结构如图 2 所示,由多头注意力机制(Multi-Head Attention,MHA)和多层感知机(Multi-Layer Perceptron,MLP)组成。

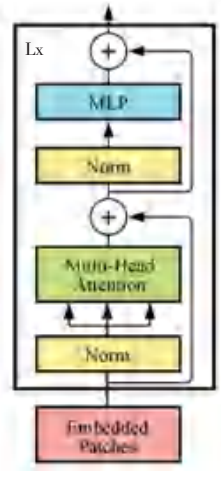

图 2 Vision Transformer 结构图

Fig. 2 Vision Transformer structure

ViT 在编码过程中 token 的数量和大小维持不变,而 token 和图像块是一一对应的关系,所以输入 Transformer 中的图像块的数量也不会发生改变,即特征图的分辨率不会变化。故以 ViT 为主干的编码器在阶段内特征提取过程中维持着输入编码向量的分辨率,避免了以往采用编码—解码结构的单目深度估计任务中容易出现的丢失特征分辨率和粒度的问题,并且 Transformer 的多头注意力机制是一种全局机制,编码器在整个编码过程中都拥有全局感受野,不需要像卷积一样通过堆叠层数来增大感受野。

2)Patch2FeatureMap 模块

算法在残差连接的部分嵌入了 Patch2FeatureMap 模块,用于连接同一阶段的编码器和解码器,保存浅层次编码器的空间信息。Patch2FeatureMap 模块将 ViT 模块编码后输出的特征序列恢复成特征图的表示形式,统一同阶段的编码器和解码器的特征形式,方便实现相同阶段间的信息流动。Patch2FeatureMap 模块包含一个简单的四阶段重组操作,具体结构如图 3 所示。

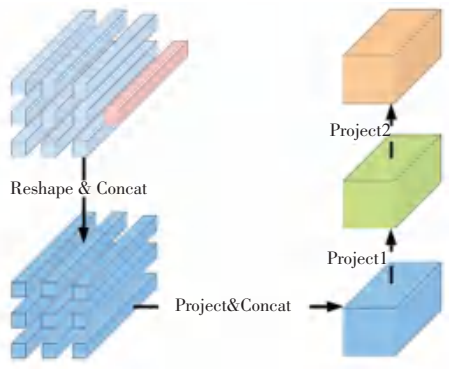

图 3 Patch2FeatureMap 模块结构

Fig. 3 Patch2FeatureMap module structure

Patch2FeatureMap 模块的表达式如下:

$$\text{Patch2FeatureMap}(t) = (\text{Reshape\&concat} \circ \text{Project\&concat} \circ \text{Project1} \circ \text{Project2})(t) \tag{1}$$

第一阶段的 Reshape&concat 操作用于处理 class token 经过编码器处理后形成的特征序列。class token 可以在编码器和解码器之间传递全局信息,解决不同层次的特征间的信息流通问题,提高模型的性能。Reshape&concat 操作将 $N_p + 1$ 个特征序列投影到 N_p 个特征序列上,可以表示如下:

$$\text{ReshapeConcat}(R^{N_p+1\times D}) = R^{N_p\times D} \tag{2}$$

具体做法是把 class token 对应的特征序列连接到其他 N_p 个特征序列上,将信息传递给其他特征序列:

$$\text{ReshapeConcat}(t) = \{mlp(cat(t_0,t_1)),\cdots, mlp(cat(t_0,t_{N_p}))\} \tag{3}$$

第二阶段的 Project&Concat 操作是通过位置编码将特征序列按照图像中初始 patch 的位置摆放,应用空间连接操作,将 N_p 个特征序列转化为特征图,再通过线性层和 GELU(Gaussian Error Linear Unit Layer)层将特征图投影到原始特征维度 D 上,生成 $\frac{H}{p} \times \frac{W}{p} \times D$ 大小的特征图。Project&Concat 操作的表达式如下:

$$\text{ProjectConcat}: R^{N_p\times D} \rightarrow R^{\frac{H}{p}\times\frac{W}{p}\times D} \tag{4}$$

第三阶段的 Project1 操作通过 1 × 1 卷积将特征图缩放到 $\frac{H}{s} \times \frac{W}{s} \times \hat{D}$ 大小:

$$\text{Project1}: R^{\frac{H}{p}\times\frac{W}{p}\times D} \rightarrow R^{\frac{H}{s}\times\frac{W}{s}\times\hat{D}} \tag{5}$$

第四阶段的 Project2 操作分两种情况:当 $s \geqslant p$ 时进行 3 × 3 卷积;当 $s < p$ 时进行 3×3 的转置卷积,

分别实现下采样和上采样操作。

3)解码器结构

解码器用于从 Patch2FeatureMap 模块输出的特征图和上层解码器的输出信息中进一步恢复深度信息。每一个浅层的特征融合层都从之前所有更深的层级获得额外的输入,并将自己的特征图传递给后续更浅的网络层级,最大限度地保证网络中各层级的信息流动,这样的结构使得在重建过程中,编码器深层的语义信息和浅层的空间信息都可以用于特征图的重建,从而恢复出细节更优的深度图。每个解码阶段逐步对特征图进行两倍的上采样,将特征维度减少到原始维度的一半,最终将特征图送入 FeatureMap2Depth 模块,将特征图转化为预测的深度。

解码器中的每个 U 型残差块具体结构如图 4 所示。首先,通过渐进式下采样从任意分辨率的输入特征图中提取特征;其次,通过渐进式上采样、连接操作和卷积重新编码成高分辨率特征图,避免了使用大规模直接上采样可能引起的细节丢失问题。渐进式采样使得 U 型残差块混合了不同大小的感受野,能够从不同尺度捕获更多的上下文信息,并且在增加模型深度的同时,不会显著增加计算成本。

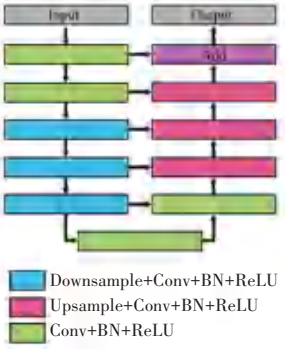

图 4　U 型残差块结构图

Fig. 4　U-shaped residual blocks structure

1.2　损失函数

网络训练时使用了尺度不变损失函数(Scale-invariant loss, L_{si})来计算预测的深度图和真值深度图之间的差距。尺度不变损失函数包含尺度不变损失(Scale-invariant data, L_{data}),多尺度匹配梯度损失(Scale-invariant loss, L_{grad})和鲁棒有序深度损失(Robust ordinal depth loss, L_{ord}),表达式如下:

$$L_{si} = L_{\text{data}} + \lambda L_{\text{grad}} + \gamma L_{\text{ord}} \tag{6}$$

本文中假设 $\lambda = 0.2, \gamma = 0.1$。

1)尺度不变损失

假设 L_i 和 L_i^* 分别为真值深度图和预测深度图中像素 i 的值,n 为具有有效真值的像素个数,并设 $R_i = \log L_i - \log L_i^*$,则:

$$L_{\text{data}} = \frac{1}{n}\sum_{i=1}^{n}(R_i)^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}R_i\right)^2 \tag{7}$$

尺度不变损失通过在对数尺度上计算差异并减去均值平方项,实现了对全局尺度变化的鲁棒性,从而提高了深度估计的准确性和泛化能力。

2)多尺度匹配梯度损失

多尺度匹配梯度损失通过考虑不同尺度的图像信息来提升模型的预测性能,公式如下:

$$L_{\text{grad}} = \frac{1}{n}\sum_{k}\sum_{i}(|\nabla_x R_i^k| + |\nabla_y R_i^k|) \tag{8}$$

其中,R_i^k 为真值深度图和预测深度图中像素 i 和尺度 k 对应的对数深度值的差值。

3)鲁棒有序不变深度损失

选取图像中的一对像素 (i, j),像素(i,j) 分别属于前景区域(Ford)和背景区域(Bord),$P_{ij} = -r^*(L_i - L_j)$,且 r^* 是 i 和 j 之间自动标记的序数深度关系。如果像素 i 比像素 j 远,则 $r_{ij}^* = 1$;反之,$r_{ij}^* = -1$。c 为一个常数集,以保持 L_{ord} 函数的连续性。L_{ord} 函数鼓励方法有序深度点对间的深度差异,表达式如下:

$$L_{\text{ord}} = \begin{cases} \log(1+\exp(P_{ij})), & \text{if} P_{ij} \leqslant \tau \\ \log(1+\exp(\sqrt{P_{ij}})) + c, & \text{if} P_{ij} > \tau \end{cases} \tag{9}$$

其中,c 为保持 L_{ord} 函数连续性的一个固定常数,本文假设 $\tau = 0.25$。

鲁棒有序不变深度损失旨在处理深度预测中的有序性问题,并同时保持对噪声和异常值的鲁棒性。

2　实验结果与分析

2.1　数据集和实验参数设置

实验训练所使用的数据集为 Inria Pose 3d 数据集、Pose Track 数据集和 NYU-Depth V2 数据集。本文将数据集中 50%的数据用于训练集,20%用于评估,30%用于测试,并对图像以 50%的几率进行随机水平翻转、20%的几率进行随机旋转(最大角度值为 10°)以及 30%的几率随机裁剪为 384×384 大小,提高模型对随机扰动的鲁棒性。

使用 Pytorch 深度学习框架来实现模型。硬件环境为 Intel Core i5-6600 处理器,RTX3060 显卡,显存为 12 G。迭代器使用 AdamW 优化器,主干学习率设置为 0.000 05,解码器学习率设置为 0.004。编码器采用在 ImageNet 上预训练的模型参数进行初始化,解码器参数进行随机初始化。模型的批次

大小设置为 8,迭代次数设置为 60。批归一化的衰减率设置为 0.9,Leaky ReLU 中 α 设置为 0.01。

2.2 评价指标

本文采用相对误差、均方根误差、平均 log10 误差、对数均方根误差、精确度等 5 种主流的单目估计算法评价指标来评估模型性能。假设 N 为深度图中像素点个数, T 为具有有效真值的像素个数, d_i 和 f_i 分别表示真实深度图和预测深度图中第 i 个像素点的深度, ↑ 表示指标值越大模型性能越好, ↓ 表示指标值越小模型性能越好。这些指标用公式可以具体表示如下:

1)相对误差(Absolute Relative Error, *AbsRel*) ↓

相对误差通常表示为预测深度与真实深度之间差值的绝对值占真实深度值的百分比,用于量化预测深度值与真实深度值之间的差异:

$$AbsRel = \frac{1}{|T|}\sum \frac{|f_i - d_i|}{di} \tag{10}$$

2)均方根误差(Root Mean Square Error, *RMSE*) ↓

均方根误差是预测值与真实值之间差异的平方和的平均值的平方根:

$$RMSE = \sqrt{\frac{1}{N}\sum (f_i - d_i)^2} \tag{11}$$

3)平均 log10 误差(Mean log10 Error) ↓

平均 log10 误差表示预测深度和真实深度之间的对数差值的绝对值的平均值:

$$\log 10 = \frac{1}{N}\sum_{i=1}^{N} \| \lg f_i - \lg d_i \| \tag{12}$$

4)对数均方根误差(logarithmic Root Mean Square Error, *RMSE*log) ↓

对数均方根误差通过对预测深度和真实深度的对数误差进行平方、平均和开方操作,来衡量两者之间的差异:

$$RMSE\log = \sqrt{\frac{1}{|T|}\sum_{f_i \in T} \| \log f_i - \log d_i \|^2} \tag{13}$$

5)精确度(δ) ↑

精确度用于量化预测深度值与真实深度值之间的接近程度,统计对应位置的深度值比值在 1.25^k ($k = 1,2,3$) 中像素点的个数,得到的像素点个数 M 与 N 的比值即为精确度:

$$\delta = \max\left(\frac{f_i}{d_i}, \frac{d_i}{f_i}\right) \tag{14}$$

2.3 定性和定量分析比较

1)定量分析

本文的算法与近年来发表的最具有代表性的算法在 KITTI 数据集和 NYU-Depth V2 数据集上进行定量对比分析。在 KITTI 数据集上,将预测的深度范围设置为 0~80 m,对于真值深度图和预测深度图中大于 80 m 的像素点,均设置为 80 m;在 NYUv2 depth 数据集上,将预测的深度范围设置为 0~10 m。$\delta < 1.25$、$\delta < 1.25^2$、$\delta < 1.25^3$ 等指标数值越大预示着模型性能越好, *AbsRel*、*RMSE*、*RMSE*log, log10 等指标数值越小预示着模型性能越好。

本文算法与 AdBins、ACAN、MAPUnet、Liu 等算法在 KITTI 数据集和 NYU-Depth V2 数据集上进行对比实验,实验结果见表 1 和表 2。在表 1 所列举的算法中,本文算法在 $\delta > 1.25^2$、$\delta > 1.25^3$、*AbsRel* 和 *RMSE*log 等指标上均取得了最好的效果,其他指标也超过了图中的 ACAN、DAV、MAPUnet 等大部分方法,与最优方法 AdaBins 基本持平。在表 2 所列举的算法中,本文算法在 $\delta > 1.25^2$、$\delta > 1.25^3$ 和 *RMSE* 等指标上均取得了最好的效果。LPF、AdBins 等原文中未披露、无法获取到的数值,在表中以'-'来替代。各项指标上的最好结果以粗体显示。由表 1 和表 2 可知,本文算法在像素级深度估计上取得了较好的性能。

表 1 KITTI 数据集上的评价结果

Table 1 Evaluation results on the KITTI dataset

算法	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	*AbsRel*	*RMSE*	*RMSE* log
VNL[4]	0.938	0.990	0.998	0.072	3.258	0.117
LPF[5]	0.715	0.900	-	0.203	6.561	-
DenseDepth[6]	0.886	0.965	0.986	0.093	2.727	0.120
BTS[7]	0.956	0.993	0.998	**0.060**	2.798	0.096
Wang et al.[8]	0.893	0.963	0.983	0.996	4.327	0.171
Liu et al.[9]	0.942	0.986	0.992	0.070	2.912	0.121
MAPUnet[10]	0.955	0.992	**0.999**	0.061	2.741	0.096
ACAN[11]	0.919	0.982	0.995	0.083	3.599	0.127
AdBins[12]	**0.964**	0.995	**0.999**	-	-	0.088
本文	0.958	**0.997**	**0.999**	**0.060**	2.735	**0.086**

表 2　NYU-Depth V2 数据集上的评价结果

Table 2　Evaluation results on the NYU-Depth V2 dataset

算法	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	*AbsRel*	*RMSE*	log10
VNL	0.875	0.976	0.994	0.111	0.416	0.048
DenseDepth	0.895	0.980	0.996	**0.103**	0.390	0.043
BTS	0.885	0.978	0.994	0.110	0.392	0.047
Godard et al.	0.871	0.975	0.993	0.115	0.519	0.049
Lin et al. [13]	0.866	0.975	0.993	0.115	0.523	0.050
Liu et al.	0.872	0.975	0.993	0.115	0.523	0.049
ACED[14]	0.870	0.974	0.993	0.115	0.528	0.049
ACAN	0.826	0.964	0.990	0.138	0.496	–
MAPUnet	0.888	0.979	0.997	0.109	0.393	**0.040**
DAV[15]	0.882	0.980	0.996	0.108	0.412	–
AdaBins	**0.903**	0.984	0.997	–	–	0.044
本文	0.892	**0.987**	**0.998**	0.107	**0.389**	0.042

2)定性分析

选取了 4 张纹理重复、光线昏暗、环境杂乱的图片,将本文算法与其他算法进行定性分析如图 5 所示,可以看出本文算法生成的深度图物体边缘相对更加清晰、物体与背景区分度更高,即便是较远区域或体积较小的物体,算法也能预测出大致的轮廓。

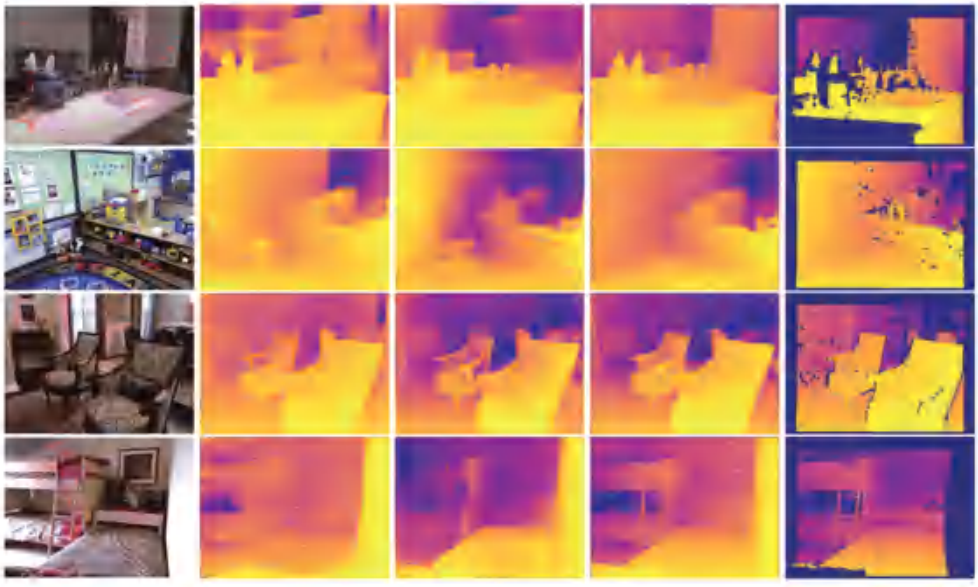

图 5　本文算法与各类基线算法的可视化比较

Fig. 5　Visual comparison of this paper with various baseline algorithms

2.4　消融实验

本文设计了一组消融实验来验证算法中各个组件的有效性,并将本文的最终算法与各个组件的消融实验进行了比较。

w/oTran:代表移除编码器的 ViT 模块,采用 U-Net 网络原本的编码器结构来替代。

w/oRSU:代表移除解码器的 RSU 模块,采用 U-Net 网络原本的解码器结构来替代。

本文在 NYUv2 depth 数据集上测试了 *w/oTran* 和 *w/oRSU* 模型,消融实验的定量结果见表 3。由表 3 可知,本文算法在各个指标上都优于其他用于对照的基线算法。ViT 模块可以带来的强大的特征提取能力和庞大的全局感受野,*w/oTran* 替换 ViT 模块的消融实验印证了这一点。*w/oRSU* 表明当替换解码器的 RSU 模块后,模型性能有所下降,可见 RSU 模块能更好地聚合 Patch2FeatureMap 模块和深层解码器传递的特征图,并进一步提取阶段内的多尺度特征,较大程度地提升了准确率。

表 3 消融实验的定量结果

Table 3 Quantitative results of the ablation experiments

算法	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$	*AbsRel*	*RMSE*	*RMSE* log
w/oTran	0.853	0.954	0.982	0.115	0.407	0.053
w/oRSU	0.889	0.980	0.994	0.112	0.394	0.053
本文	0.892	0.980	0.998	0.107	0.389	0.042

3 结束语

深度估计作为视觉领域的重要任务,为自动驾驶、虚拟现实、增强现实、三维重建等应用场景提供深度信息。以往的单目深度估计算法常采用编码器-解码器结构来解决该问题,然而编码器中频繁的下采样操作会造成特征图分辨率和高层信息的损失,影响深度预测的准确率。基于此,本文借鉴了 U^2-Net 模型的思想,提出了一种基于多尺度特征融合的单目深度估计算法,能够大幅度提高单目深度估计模型的性能。由浅到深的递进式编-解码结构能够提取多尺度特征,实现全局信息与局部信息的融合、大尺度和小尺度信息的融合。以 ViT 为主体模块的编码器避免了下采样操作,使特征图在编码处理过程中能够维持较高分辨率,减少特征图分辨率和高层信息的损失。解码器的设计借鉴了 RSU 模块的思想,通过渐进式采样融合多尺度特征,提高了深度图的清晰度。与最近的一些先进算法相比,本文的算法在定量和定性估计中都展现了良好的性能。

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] CHEN Y, ZHAO H, HU Z, et al. Attention-based context aggregation network for monocular depth estimation [J]. International Journal of Machine Learning and Cybernetics, 2021, 12: 1583-1596.

[3] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12179-12188.

[4] YIN W, LIU Y, SHEN C, et al. Enforcing geometric constraints of virtual normal for depth prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5684-5693.

[5] ZHANG Z, LATHUILIERE S, RICCI E, et al. Online depth learning against forgetting in monocular videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4494-4503.

[6] ALHASHIM I, WONKA P. High quality monocular depth estimation via transfer learning[J]. arXiv preprint arXiv: 1812. 11941, 2018.

[7] LEE J H, HAN M K, KO D W, et al. From big to small: Multi-scale local planar guidance for monocular depth estimation[J]. arXiv preprint arXiv:1907. 10326, 2019.

[8] WANG J, ZHANG G, YU M, et al. Attention-based dense decoding network for monocular depth estimation [J]. IEEE Access, 2020, 8: 85802-85812.

[9] LIU P, ZHANG Z, MENG Z, et al. Joint attention mechanisms for monocular depth estimation with multi-scale convolutions and adaptive weight adjustment[J]. IEEE Access, 2020, 8: 184437-184450.

[10] YANG Y, WANG Y, ZHU C, et al. Mixed-scale UNet based on dense atrous pyramid for monocular depth estimation[J]. IEEE Access, 2021, 9: 114070-114084.

[11] CHEN Y, ZHAO H, HU Z, et al. Attention-based context aggregation network for monocular depth estimation [J]. International Journal of Machine Learning and Cybernetics, 2021, 12: 1583-1596.

[12] BHAT S F, ALHASHIM I, WONKA P. Adabins: Depth estimation using adaptive bins[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4009-4018.

[13] LIN L, HUANG G, CHEN Y, et al. Efficient and high-quality monocular depth estimation via gated multi-scale network[J]. IEEE Access, 2020, 8: 7709-7718.

[14] SWAMI K, BONDADA P V, BAJPAI P K. Aced: Accurate and edge-consistent monocular depth estimation[C]//Proceedings of 2020 International Conference on Image Processing (ICIP). IEEE, 2020: 1376-1380.

[15] HUYNH L, NGUYEN-HA P, MATAS J, et al. Guiding monocular depth estimation using depth-attention volume[J]. Lecture Notes in Computer Science, 2020, 12371: 581-597.