第 59 卷 第 16 期/2022 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于迁移学习的高效单目图像深度估计

刘佳涛[1]，张亚萍[1*]，杨雨薇[2]

[1]云南师范大学信息学院，云南 昆明 650500；

[2]南通理工学院，江苏 南通 226000

摘要 在进行三维重建、场景理解等计算机视觉任务时，从二维图像中恢复三维空间中的深度信息是一项基本的任务。当前使用深度学习完成该任务时，精确度较高的方法往往需要巨大的数据量，而这些数据的获取通常复杂且开销大。针对这个问题，提出了一种基于迁移学习的全局自注意力编解码网络。所提网络以单张图像作为输入，在编码时的每一个阶段都具有全局性的感受域，解码后把深度回归任务转化为一种分类任务，在保证模型精确度的前提下大大降低所需的训练数据量。实验结果表明，与当前先进的深度估计网络 AdaBins 和 DPT-Hybrid 相比，所提网络在均方根误差上降低了约 2.2% 和 0.3%，在训练数据量上降低了约 80% 和 99.6%。

关键词 成像系统；迁移学习；单目视觉；深度估计；自注意力机制

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP202259.1611002

Efficient Monocular Image Depth Estimation Based on Transfer Learning

Liu Jiatao[1], Zhang Yaping[1*], Yang Yuwei[2]

[1]*School of Information Science and Technology, Yunnan Normal University, Kunming* 650500, *Yunnan, China*;

[2]*Nantong Institute of Technology, Nantong* 226000, *Jiangsu, China*

Abstract When performing computer vision tasks such as three-dimensional reconstruction and scene understanding, it is a basic task to recover depth information in three-dimensional space from two-dimensional images. When deep learning is currently used to complete this task, methods with higher accuracy often require a huge amount of data, and the acquisition of these data is usually complicated and expensive. In response to this problem, this paper based on transfer learning, and proposes a encoder-decoder network using global self-attention. It takes a single image as input and has a global receptive field at each stage of encoding. After decoding, the depth regression task is transformed into a classification task, greatly reducing the amount of training data required while ensuring the accuracy of the model. The experimental results show that compared with the current state-of-the-art depth estimation networks AdaBins and DPT-Hybrid, the designed model reduces the root mean square error by about 2.2% and 0.3%, and reduces the amount of training data by about 80% and 99.6%.

Key words imaging systems; transfer learning; monocular vision; depth estimation; self-attention mechanism

1 引 言

从二维 RGB 图像中进行深度估计具有广泛的应用，例如三维重建、场景理解、自动驾驶、机器人技术等。随着大规模数据集的出现和硬件运算能力的提高，最近关于图像深度估计的研究主要集中在使用深度学习和卷积神经网络进行二维到三维的重建[1-10]。

在图像单目深度估计的神经网络训练上，表现较好的监督学习方法通常需要大量的标注数据，而标注数据是一项枯燥无味且花费巨大的任务。丁萌等[2]利用双目视觉图像设计一个新的损失函数代替真实深度标签，以此实现无监督学习从而解决了场景真实深度标注数据难以获取的问题，但该方法在训练时需要获取大量双目视觉图像。亢超等[3]依据自编码神经网络进行图像重构的思想，设计了一种面向无人机自主飞行的无监督单目深度估计模型。这些无监督学习方法进行深度估计时能达到的准确度始终有限，对于需要高精确度深度信息的应用场景，使用较多的是监督学习方法。

收稿日期：2021-07-30；**修回日期**：2021-09-04；**录用日期**：2021-09-24

基金项目：国家自然科学基金(61863037)、云南省“万人计划”青年拔尖人才专项、南通市科技局项目(JC2019108)

通信作者：*zhangyp@ynnu.edu.cn

近期，迁移学习在许多任务中表现出了有效性。对于单目深度估计任务，迁移学习能够在已有的有限标注数据的条件下，大幅提高监督学习的单目深度估计网络训练效率和预测准确度。Alhashim 等[4]设计了一个简单的单编解码器网络，该网络使用在图像分类任务下预训练的模型参数作为编码器的初始参数，相较于更复杂的多网络结构[5,9]有更好的精确度，且模型易修改和扩展。这种单编解码器结构和迁移学习的思想可以应用到更多的图像深度估计任务当中，但其使用的卷积神经网络通过逐步下采样来扩大感受域，这种操作具有一定的缺陷，即随着网络深度不断加深，特征分辨率和粒度信息可能会有部分丢失。这些信息一旦在编码器中丢失了，那么将很难在解码器中恢复。

在图像深度估计任务使用的编解码器网络中，编码器对于图像的特征提取能力影响模型最终深度估计的精确度。Ranftl 等[11]将在图像分类任务中具有很好精确度的 Vision Transformer（ViT）[12]作为编码器迁移到图像深度估计任务当中，这种编码器基于自注意力机制，不使用下采样操作，因此图像特征分辨率在编码的每一个阶段都不会降低。此外，这种网络结构在每一个阶段都具有全局性的感受域，这种全局性的感受域对于图像深度估计也有积极的作用。但由于 ViT 包含的参数量较大，网络训练难度也随之增大，训练需要的数据量很大。

Fu 等[8]在研究中指出：如果将深度回归任务转化为分类任务，其性能可以得到提高，但其预测的深度信息较为离散。Bhat 等[7]设计了 AdaBins 模块，该模块将深度值范围划分为 256 个区间，每一个区间的中心值为落在该区间的像素的深度值，这样就将深度回归任务转化为了一种分类任务，同时为了解决分类任务导致深度值的离散问题，最终的深度是区间中心深度值的线性组合。该模块移植方便，包含的参数量较少，且能产生较好的效果。

为了解决当前深度估计模型在训练时需要的数据量较大、估计时图像细粒度信息丢失等普遍存在的问题，本文提出了一种基于迁移学习的单目深度估计网络，该网络迁移图像分类任务中在大型数据集下预训练的参数，在编码时不使用下采样操作，且编码的每一个阶段都具有全局性的感受域。为了进一步提高深度估计的精确度并降低模型训练的难度，该网络将深度回归任务转化为分类任务。

2　网络结构

本实验组设计了一种端到端的模型，该模型可根据单张图像较为精确地输出该图像的深度图，模型的整体结构如图 1 所示。使用 ViT-Hybrid[12]作为网络编码器，ViT-Hybrid 以浅层卷积神经网络 ResNet-50 作为特征提取模块，将输入的 RGB 图像编码成 16 个特征向量，此外添加一个独立于图像的标记向量，最后将提取到的特征向量和标记向量输入 12 个 transformer 层中。解码器主要需要实现两个功能：一个是特征重组；另一个是特征融合。经过解码器后，将得到的初步预测的特征图输入 AdaBins 模块，经过 AdaBins 模块后输出最终的深度预测图。

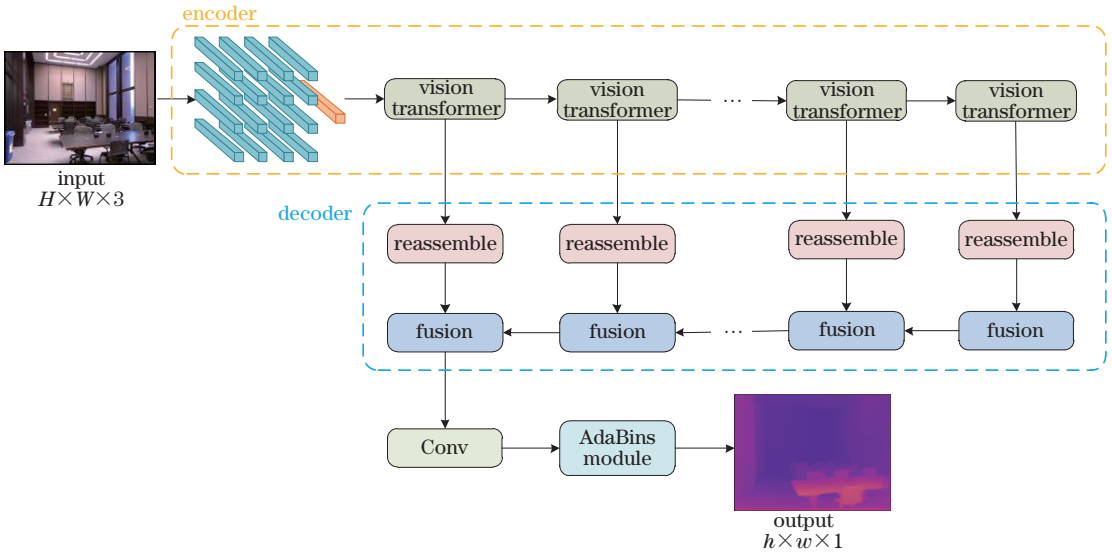

图 1　模型结构

Fig. 1　Model structure

2.1　ViT-Hybrid 编码器

ViT-Hybrid 是 Dosovitskiy 等[12]在图像分类任务上提出的模型，与在自然语言处理中已得到广泛运用的 Transformer[13]相同，ViT-Hybrid 也是一种基于自注意力机制的模型，因此该模型对于图像的处理在每一个阶段都有全局性的感受域，这也是 ViT-Hybrid 与卷积神经网络的最大不同之处。

图 2 展现了编码过程和编码器的结构。在将图片

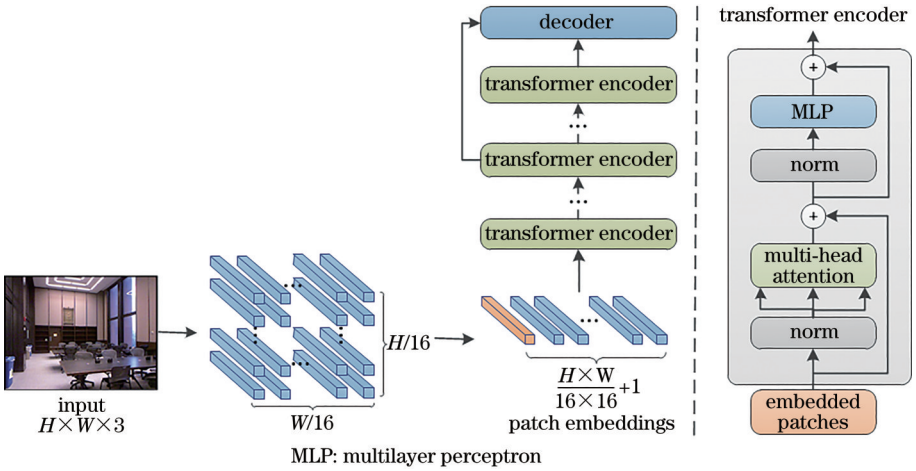

图 2　编码器结构和编码过程

Fig. 2　Encoder structure and encoding process

输入 ViT-Hybrid 前，首先要把一张二维的图片转化成为一维的序列，即

$$\boldsymbol{x}\in\boldsymbol{R}_{H\times W\times C}\rightarrow\boldsymbol{x}\in\boldsymbol{R}_{N\times(p^2\times C)},\tag{1}$$

式中：H、W、C 分别表示输入图像的高、宽和通道数；$N=\dfrac{H\times W}{p^2}$，这个过程通过浅层卷积神经网络提取，且 $p=16$。此外，ViT-Hybrid 也添加了一个独立于图像的标记向量。将提取到的特征向量和标记向量经过扁平化处理后嵌入位置信息，最后输入 Transformer 层中，即最终输入 ViT-Hybrid 中的向量集为

$$\boldsymbol{t}=\left\{\boldsymbol{t}_0,\cdots,\boldsymbol{t}_N\right\},\tag{2}$$

式中：$\boldsymbol{t}_0$ 是添加的标记向量。经过 12 个 Transformer 层编码后，最终将特征序列输出到解码器中。

2.2　解码器

解码器的主要工作是将编码器输入过来的一维特征向量重新组合成二维的特征图，包含重组块和融合块两个部分，其结构和解码过程如图 3 所示。

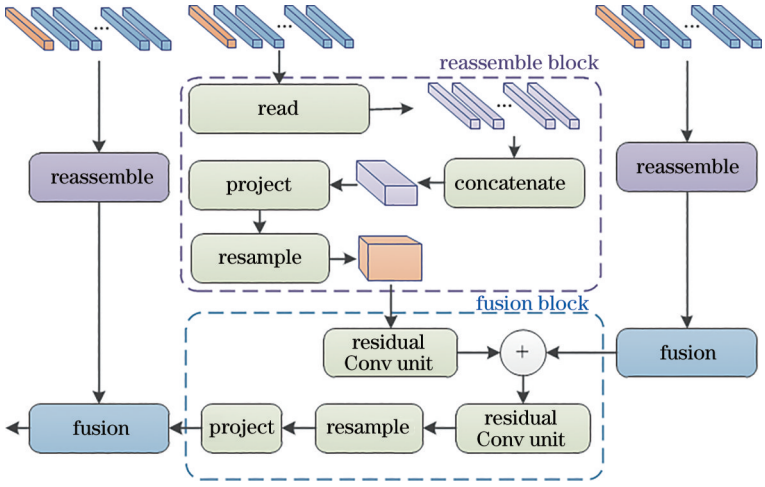

图 3　解码器结构和编码过程

Fig. 3　Decoder structure and decoding process

重组块的第 1 部分读取从编码器输入过来的特征向量，在这一部分中，首先把附加的特征向量 $\boldsymbol{t}_0'$ 与其他特征向量一一拼接（cat），然后经过 MLP 生成最终的特征向量，即

$$\mathrm{Read}\left(\boldsymbol{t}'\right)=\left\{\mathrm{mlp}\left[\mathrm{cat}\left(\boldsymbol{t}_1',\boldsymbol{t}_0'\right)\right],\cdots,\mathrm{mlp}\left[\mathrm{cat}\left(\boldsymbol{t}_N',\boldsymbol{t}_0'\right)\right]\right\}。\tag{3}$$

重组块的第 2 部分需要将读取的特征向量还原成为特征图，本实验组使用一个空间上的连接操作，生成一个分辨率为 $\dfrac{H}{p}\times\dfrac{W}{p}$、通道数为 D 的特征图：

$$\boldsymbol{R}_{N\times D}\rightarrow\boldsymbol{R}_{\frac{H}{p}\times\frac{W}{p}\times D}。\tag{4}$$

重组块的第 3 部分对其重新采样，生成分辨率为 $\dfrac{H}{s}\times\dfrac{W}{s}$、通道数为 $\hat{D}$ 的特征图：

$$\boldsymbol{R}_{\frac{H}{p}\times\frac{W}{p}\times D}\rightarrow\boldsymbol{R}_{\frac{H}{s}\times\frac{W}{s}\times\hat{D}}。\tag{5}$$

在得到每个阶段的特征图后，使用基于 RefineNet[14]

的特征融合块来对其进行融合，再进行2倍上采样即可得到多通道的初步深度预测图。

2.3 AdaBins模块

AdaBins模块有两个作用：1）通过自注意力机制来抑制初步预测中的无用信息；2）将深度范围划分为256个区间来将深度回归任务转化为分类任务，以此提高预测的准确度。该模块的结构和处理的过程如图4所示。

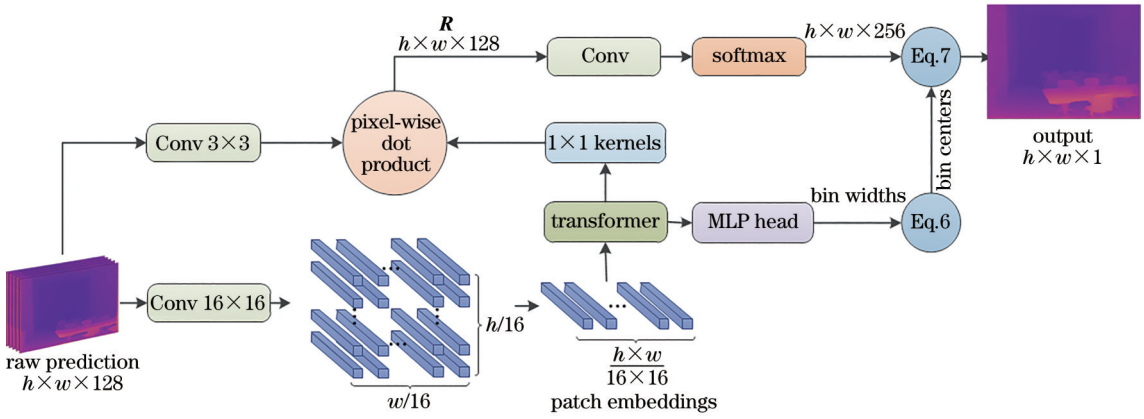

图4　AdaBins模块结构和处理过程

Fig. 4　AdaBins module structure and processing process

AdaBins模块的主体结构仍为Transformer[13]，Transformer的输出包括两部分。在第1部分输出的向量后面添加一个MLP层，可以得到深度范围归一化后划分的区间大小b_i，则区间中心的深度值$c(b_i)$的表达式为

$$c(b_i)=d_{\min}+(d_{\max}-d_{\min})\left(\frac{b_i}{2}+\sum_{j=1}^{i-1}b_j\right),\tag{6}$$

式中：$d_{\max}$和$d_{\min}$分别指场景数据集中真实深度的最大值和最小值，如NYU Depth v2数据集中$d_{\max}=10$，$d_{\min}=0.001$。在第2部分输出的一组向量后连接一组1×1的卷积核，然后与经过了3×3卷积后的解码特征图进行点乘操作，得到深度范围注意力图$\boldsymbol{R}$（range attention maps）。$\boldsymbol{R}$经过卷积和Softmax层，得到具有M个（256）通道的深度概率图，最后深度概率图和区间中心深度线性组合得到最终输出的深度估计图：

$$\hat{\boldsymbol{y}}=\sum_{k=1}^{M}c(b_k)\boldsymbol{P}_k,\tag{7}$$

式中：$\boldsymbol{P}_k$指深度概率图中的第k通道，其中每个像素的数值为该像素的深度值落在第k个区间的概率。

3 实　　验

3.1 损失函数

像素级的深度值损失L_{depth}：首先引入Eigen等[5]提出的尺度不变（SI）损失，并在数值上对其进行了缩放：

$$L_{\text{depth}}(\boldsymbol{y},\hat{\boldsymbol{y}})=\alpha\sqrt{\frac{1}{n}\sum_{p}d_p^2-\frac{\lambda}{n^2}\left(\sum_{p}d_p\right)^2},\tag{8}$$

式中：$d_p=\ln y_p-\ln\hat{y}_p$，y_p和$\hat{y}_p$分别是真实深度图$\boldsymbol{y}$和模型预测的深度图$\hat{\boldsymbol{y}}$中像素点p的深度值；n是深度图的像素总数；同时为了以此损失项的初始值为基准控制其他损失项所占的权重，实验中将α设置为10；λ设置为0.5。

多尺度结构相似度量$L_{\text{ms-ssim}}$：MS-SSIM损失函数是基于多层，即图片按照一定规则由大到小缩放的结构相似性度量，相当于考虑了图像的分辨率且保留了图像中的高频信息[15]：

$$L_{\text{ms-ssim}}(\boldsymbol{y},\hat{\boldsymbol{y}})=\beta\sqrt{1-\text{MS-SSIM}(\boldsymbol{y},\hat{\boldsymbol{y}})}。\tag{9}$$

式中：β设置为10。

区间中心分布损失L_{bins}：为了提高AdaBins模块中深度区间划分的准确度，引入Bhat等[7]提出的区间中心分布损失：

$$L_{\text{bins}}(\boldsymbol{y},\hat{\boldsymbol{y}})=\text{chamfer}[\boldsymbol{X},\boldsymbol{c}(b)]+\text{chamfer}[\boldsymbol{c}(b),\boldsymbol{X}],\tag{10}$$

$$\text{chamfer}(\boldsymbol{S}_1,\boldsymbol{S}_2)=\sum_{\boldsymbol{x}\in\boldsymbol{S}_1}\min_{\boldsymbol{y}\in\boldsymbol{S}_2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2+\sum_{\boldsymbol{y}\in\boldsymbol{S}_2}\min_{\boldsymbol{x}\in\boldsymbol{S}_1}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,\tag{11}$$

式中：$\boldsymbol{X}$表示真实深度的区间中心深度值集合；$\boldsymbol{c}(b)$表示模型预测的区间中心深度值集合。chamfer是Fan等[16]提出来用于描述点集对之间差异的损失，其中$\boldsymbol{x}$和$\boldsymbol{y}$分别表示两个点集$\boldsymbol{S}_1$和$\boldsymbol{S}_2$中的点。

最终损失函数的表达式为

$$L(\boldsymbol{y},\hat{\boldsymbol{y}})=L_{\text{depth}}(\boldsymbol{y},\hat{\boldsymbol{y}})+L_{\text{ms-ssim}}(\boldsymbol{y},\hat{\boldsymbol{y}})+\eta L_{\text{bins}}(\boldsymbol{y},\hat{\boldsymbol{y}}),\tag{12}$$

式中：η设置为0.1。

3.2 数据集

NYU Depth v2是在图像深度估计任务中一个较常用的数据集，它提供了不同室内场景的RGB图像和其对应的深度图，图像深度范围为0～10 m，分辨率为480×640。该数据集包含1.2×10^5个训练样本，为了便于与当前的先进方法进行横向对比，本实验组以与文献[4]和文献[7]相同的5×10^4个样本作为训练集。

由于采集设备的硬件原因，数据集中原始深度图里部分数据点的深度值是缺失的，本实验组对于缺失的部分使用 Levin 等[17]提出的方法进行填充，此外也同样使用了图像随机水平翻转和随机通道交换的方法来扩展训练数据，以减少过拟合并提高模型的泛化性能。最终输出深度图的分辨率是输入图像的一半。评估模型时，以官方划分的 654 张图片作为模型的输入，并将输出上采样到原来的 2 倍以匹配真实深度的分辨率，此外采用了 Eigen 等[5]预定义的裁切边界方法将原始图像和水平翻转后的图像作为模型输入，取二图预测深度的平均值作为模型最终输出并进行评估。

3.3　实验设置

使用 PyTorch[18]框架实现了所提网络结构，并在 NVIDIA GeForce RTX 3090 上训练了该模型。使用 Adam[19]作为优化器，编码器部分参数的初始学习速率是 1×10^{-5}，AdaBins 模块参数的初始学习速率是 2×10^{-4}，其余部分参数的初始学习速率是 1×10^{-4}。将 epoch 总数设置为 20，随着训练进行，学习速率将会在第 4、10、16、19 个 epoch 降低为原来的 0.8，每一个 epoch 训练时长大约为 140 min，每一次迭代的批量大小设置为 2。由于 ViT 具有包含参数量较大的特点，所提模型整体参数也相对较大，其一共包含了 1.2903×10^8 个参数，其中 ViT-Hybrid 编码器包含 1.078×10^8 个参数，解码器包含 1.531×10^7 个参数，AdaBins 模块包含 5.92×10^6 个参数。受 Dosovitskiy 等[12]工作的启发，ViT-Hybrid 在更大型的数据集上训练后具有更好的图像分类准确度，因此本实验组将在 ImageNet[20]和 JFT-300M[21]上完成图像分类训练的 ViT-Hybrid 参数进行迁移，加载到编码器部分，其余部分采用了 Glorot 等[22]提出的随机初始化方法进行初始化。

3.4　评价指标

对于图像深度估计，有 6 个标准的精确度评价指标[5]，这些指标包括：阈值准确度（$\delta_i, i=1,2,3$）、平均相对误差（Rel）、对数平均误差（log10）、均方根误差（RMS），各指标的计算表达式分别为

$$\delta_i=\max\left(\frac{y_p}{\hat{y}_p},\frac{\hat{y}_p}{y_p}\right)<t_{\text{threshold}},\tag{13}$$

$$E_{\text{Rel}}=\frac{1}{n}\sum_p^n\frac{\left|y_p-\hat{y}_p\right|}{y_p},\tag{14}$$

$$E_{\log10}=\frac{1}{n}\sum_p^n\left|\log_{10}\left(y_p\right)-\log_{10}\left(\hat{y}_p\right)\right|,\tag{15}$$

$$E_{\text{RMS}}=\sqrt{\frac{1}{n}\sum_p^n\left(y_p-\hat{y}_p\right)^2},\tag{16}$$

式中：当 $i=1,2,3$ 时，$t_{\text{threshold}}$ 分别为 $1.25,1.25^2,1.25^3$。

除精确度比较外，本实验组还引入了关于模型训练代价的比较，训练代价以进行深度估计训练总使用的数据量来衡量。

4　实验结果与分析

4.1　定性结果

4.1.1　深度图

图 5 展现了所提模型对某些场景下的单张 RGB 图像的深度预测结果，并且与文献[7]和文献[11]中提供的当前较先进的两个模型预测结果进行了对比。从

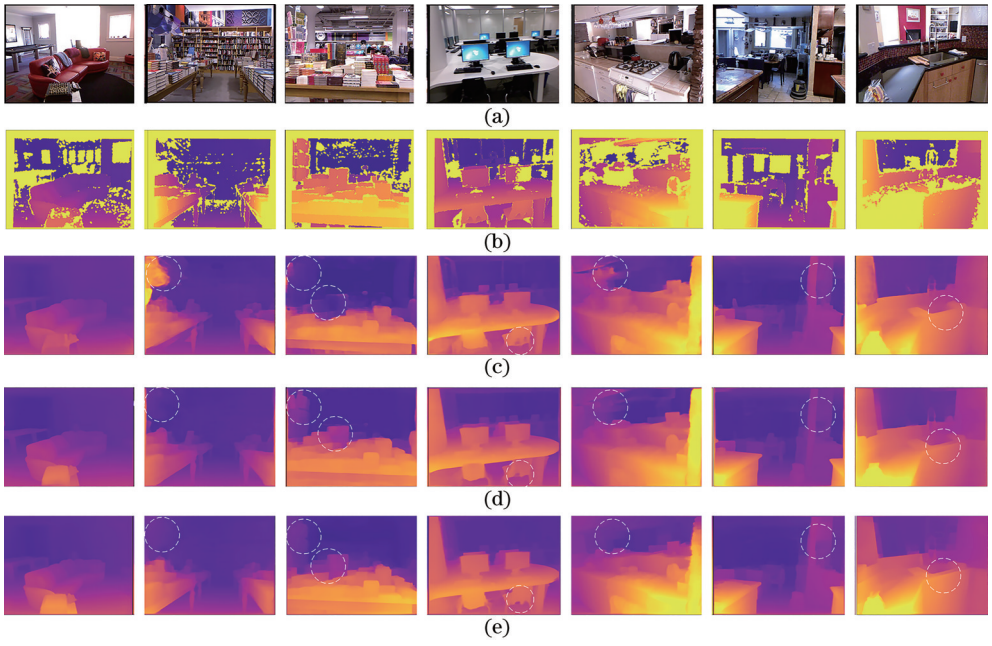

图 5　不同模型预测结果。(a)输入的 RGB 图像；(b)真实深度图；(c)文献[7]模型预测的深度图；(d)文献[11]模型预测的深度图；(e)所提模型预测的深度图

Fig. 5　Prediction results of different models. (a) Input RGB images; (b) ground truth depth maps; (c) depth map predicted by model in literature [7]; (d) depth map predicted by model in literature [11]; (e) depth map predicted by proposed model

图 5 不难发现，所提模型能够较为准确地预测图像的深度信息，且在某些复杂场景下有更好的预测结果，主要表现在对于图像中不重要信息的有效抑制和重要细节信息的有效保留上。

4.1.2　三维点云

将二维图像转化成为三维点云是深度估计的一个重要应用，图 6 展现了所提模型通过预测深度从单张 RGB 图像生成三维点云的示例。

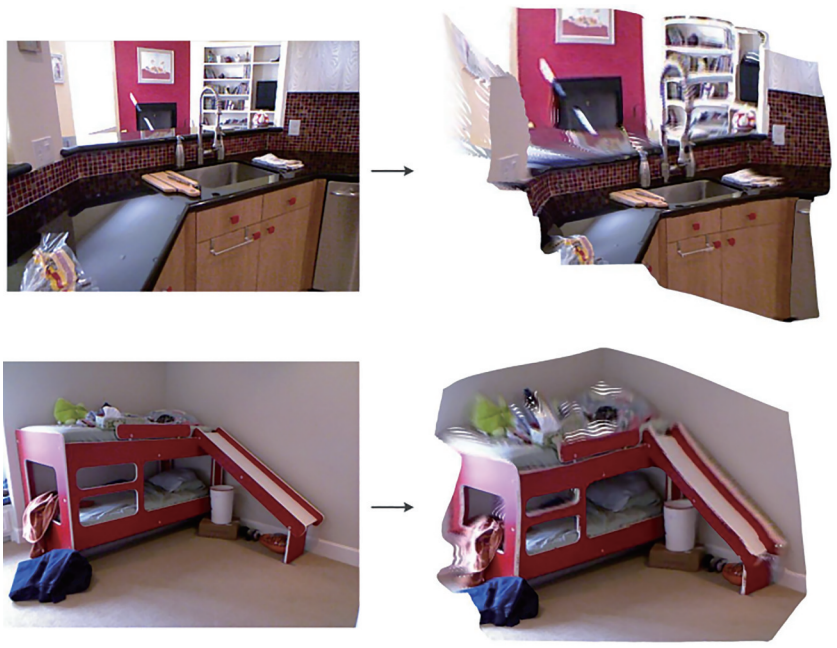

图 6　所提模型从单张 RGB 图像生成三维点云的示例

Fig. 6　Examples of proposed model generating a three-dimensional point cloud from a single RGB image

4.2　定量结果

将 RMS 测试值最小的模型参数作为训练的最优结果，在训练了 5 个 epoch 后，即迭代约 1.26×10^5 次达到了最优。表 1 列出了各种模型在 NYU Depth v2 数据集上的表现，其中其他模型的结果均来自于相应的原论文，最佳结果以粗体显示，表中训练代价以进行深度估计训练使用的总数据量来衡量。从表 1 可以看出，所提模型在 RMS 上比文献[7]和文献[11]提供的当前较先进模型分别减少了约 2.2%、0.3%，且训练代价分别减少了约 80%、99.6%。

表 1　与其他模型预测结果的定量比较

Table 1　Quantitative comparison with prediction results of other models

Model	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	REL ↓	RMS ↓	$\log_{10}\downarrow$	Train cost /10^5
Model in literature [5]	0.769	0.950	0.988	0.158	0.641		
Model in literature[8]	0.828	0.965	0.992	0.115	0.509	0.051	
Model in literature[4]	0.895	0.980	0.996	**0.103**	0.390	0.043	80
Model in literature[7]	0.903	0.984	0.997	**0.103**	0.364	**0.044**	12.67
Model in literature[11]	**0.904**	**0.988**	**0.998**	0.110	0.357	0.045	691.2
Proposed model	0.902	0.987	**0.998**	**0.103**	**0.356**	**0.044**	**2.53**

4.3　消融研究

4.3.1　损失函数

在研究中发现，损失函数的选取和比例控制对于所提模型最终预测的精确度和训练的代价有一定的影响，图 7 展现了使用不同损失项和损失项参数分别训练 20 个 epoch 时的 RMS 变化情况折线图，[图 7(c)、(d)]中 gradient 损失项的表达式为

$$L_{\text{gradient}}(\boldsymbol{y},\hat{\boldsymbol{y}})=\omega\sqrt{\frac{1}{n}\sum_{p}^{n}\left|g_x(y_p,\hat{y}_p)\right|+\left|g_y(y_p,\hat{y}_p)\right|},\tag{17}$$

式中：g_x、g_y 分别计算 y_p 和 $\hat{y}_p$ 在 x、y 分量上的差值。[图 7(b)、(c)、(d)]中 SI 项的数值为式(8)中 λ 的取值，MS_SSIM 项的数值为式(9)β 的取值，gradient 项的数值为式(17)中 ω 的取值。

从[图 7(a)]可以看出，在前几个 epoch 中使用 L_{bins} 对于模型的精确度有积极作用。从[图 7(b)、(c)、(d)]可以看出，将损失函数中 SI 损失项的 λ 设置为 0.5，MS_SSIM 损失项中 β 的设置为 10.0，且不使用梯度损失，模型能有较低的误差和较小的训练数据量。

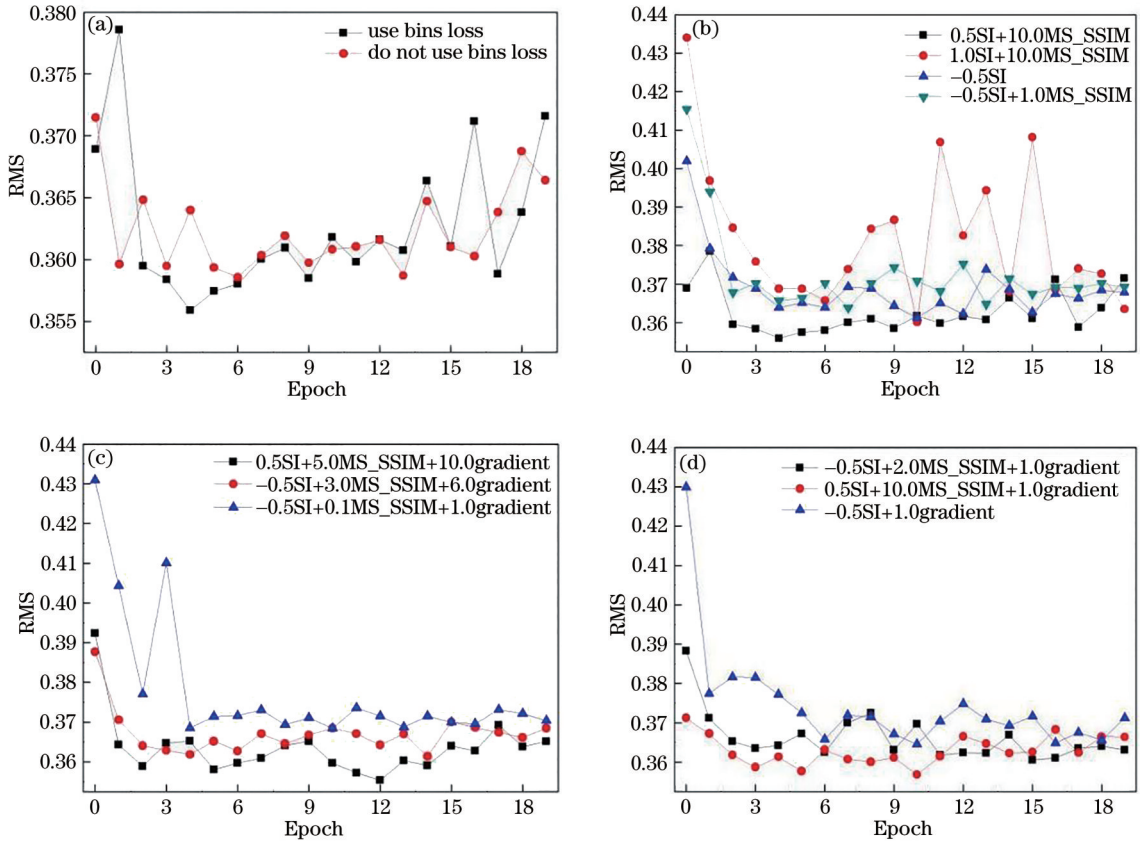

图 7　不同损失函数的消融研究。(a)使用 L_{bins} 与否的比较；(b)(c)(d)使用 L_{bins} 与不同损失项不同权重组合的比较

Fig. 7　Ablation study of different loss functions. (a) Comparison of using L_{bins} or not; (b)(c)(d) comparison of using L_{bins} combined with different loss items and weight values

4.3.2　编码器

依据 ViT 获取输入向量的方式和 Transformer 的层数不同，ViT 具有多种不同的变体[12]。考虑到计算机硬件与模型参数量限制，本实验组选择 ViT-Hybrid 与 ViT-Base 来进行编码器的消融研究。这两种编码器在提取特征向量上有差别，与 ViT-Hybrid 的浅层卷积神经网络提取不同，ViT-Base 将图像直接进行空间上的分割来提取。图 8 为使用这两种编码器的网络结构分别训练 20 个 epoch 时的 RMS 变化情况折线图。从图中可以看出，使用 Vit-Hybrid 作为网络中的编码器时，模型预测深度信息的误差更小。

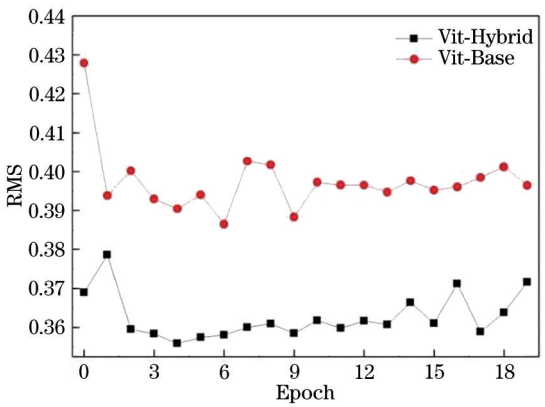

图 8　不同编码器的消融研究

Fig. 8　Ablation study of different encoders

4.3.3　预训练模型

所提基于迁移学习的模型将在图像分类任务下有较高准确度模型的参数在训练前加载到编码器当中，为了研究所提模型使用这种迁移学习相较于随机初始化的优势，将这两种初始化方法进行一组消融实验，实验结果如图 9 所示。显然，所提模型使用的这种基于迁移学习的方法能够大幅提高模型精确度，同时进行训练所使用的数据量也能够大幅降低。

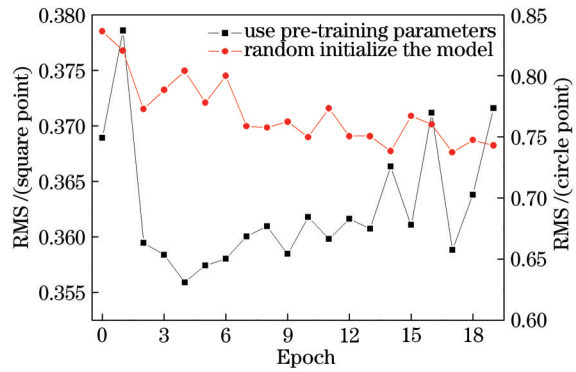

图 9　不同初始化方法的消融研究

Fig. 9　Ablation study of different initialization methods

4.3.4　AdaBins模块

基于编码-解码器结构，在解码器后添加AdaBins模块作为后处理模块来设计模型网络，为了研究这种后处理模块对于模型的训练和预测精确度是否有积极作用，对AdaBins模块进行一组消融实验，实验结果如图10所示，其中在不使用AdaBins的实验中也同时不使用L_{bins}作为损失项。从图中可以看出，在出现过拟合之前，AdaBins模块在所提模型当中有降低训练所需数据量和预测误差的作用。

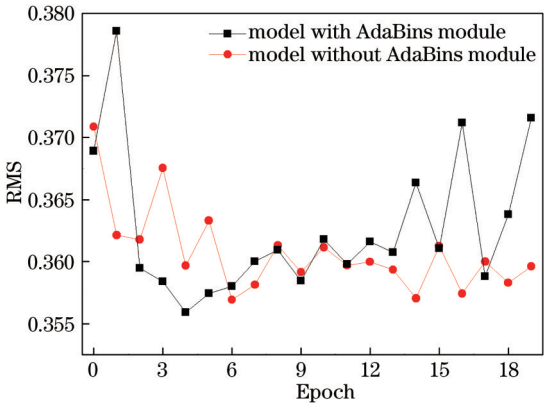

图10　AdaBins模块的消融研究

Fig. 10　Ablation study of AdaBins module

5　结　　论

针对图像深度恢复任务，提出了一个基于迁移学习的单目深度估计网络。该网络使用ViT-Hybrid编码器，可以同时兼顾图像全局感受域与图像细粒度特征信息，同时基于自注意力机制的Transformer结构对深度估计也有积极作用；针对当前先进方法的模型训练需要的数据量较大的问题，使用迁移学习和任务转化的方法以降低训练数据量。实验结果表明，所提模型能够较为准确恢复单张RGB图像的深度信息，相较于当前先进方法在重要细节信息保留和不重要信息抑制上有更好表现，且训练模型时所需数据量大大降低。但ViT-Hybrid结构复杂且包含的参数量较大，模型容易过拟合；对图像不使用下采样操作导致训练时占用硬件资源较大。所提模型在更大的数据集上或许能够有更好的精确度，在模型精确度表现和训练数据量大小的平衡或许是下一步研究的问题。

参　考　文　献

[1] 王一同，周宏强，闫景逍，等. 基于深度学习算法的计算光学研究进展[J]. 中国激光，2021，48(19)：1918004.
Wang Y T，Zhou H Q，Yan J X，et al. Advances in computational optics based on deep learning[J]. Chinese Journal of Lasers，2021，48(19)：1918004.

[2] 丁萌，姜欣言. 先进驾驶辅助系统中基于单目视觉的场景深度估计方法[J]. 光学学报，2020，40(17)：1715001.
Ding M，Jiang X Y. Scene depth estimation based on monocular vision in advanced driving assistance system [J]. Acta Optica Sinica，2020，40(17)：1715001.

[3] 亢超，李文祥，黄岫，等. 基于深度学习的主动光学校正算法研究[J]. 光学学报，2021，41(6)：0611004.
Kang C，Li W X，Huang S，et al. Research on active optical correction algorithm based on deep learning[J]. Acta Optica Sinica，2021，41(6)：0611004.

[4] Alhashim I，Wonka P. High quality monocular depth estimation via transfer learning[EB/OL]. (2018-12-31) [2021-07-21]. https://arxiv.org/abs/1812.11941.

[5] Eigen D，Puhrsch C，Fergus R. Depth map prediction from a single image using a multi-scale deep network[EB/OL]. (2014-06-09) [2021-07-21]. https://arxiv. org/abs/1406.2283.

[6] Laina I，Rupprecht C，Belagiannis V，et al. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth International Conference on 3D Vision (3DV)，October 25-28，2016，Stanford，CA，USA. New York：IEEE Press，2016：239-248.

[7] Bhat S F，Alhashim I，Wonka P. AdaBins：depth estimation using adaptive bins[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 20-25，2021，Nashville，TN，USA. New York：IEEE Press，2021：4008-4017.

[8] Fu H，Gong M M，Wang C H，et al. Deep ordinal regression network for monocular depth estimation[C]// Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition，June 18-23，2018，Salt Lake City，UT，USA. New York：IEEE Press，2018：2002-2011.

[9] Li J，Klein R，Yao A. A two-streamed network for estimating fine-scaled depth maps from single RGB images[C]//Proceedings of the IEEE International Conference on Computer Vision，October 22-29，2017，Venice，Italy. New York：IEEE Press，2017：3372-3380.

[10] 伍俊龙，郭正华，陈先锋，等. 基于深度学习的光场成像三维测量方法研究[J]. 中国激光，2020，47(12)：1204005.
Wu J L，Guo Z H，Chen X F，et al. Three-dimensional measurement method of light field imaging based on deep learning[J]. Chinese Journal of Lasers，2020，47(12)：1204005.

[11] Ranftl R，Bochkovskiy A，Koltun V. Vision transformers for dense prediction[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV)，October 10-17，2021，Montreal，QC，Canada. New York：IEEE Press，2021：12159-12168.

[12] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16x16 words：transformers for image recognition at scale[EB/OL]. (2020-10-22) [2021-07-21]. https://arxiv.org/abs/2010.11929.

[13] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30：Annual Conference on Neural Information Processing Systems 2017，December 4-9，2017，Long Beach，CA，USA. [S.l.：s.n.]，2017：5998-6008.

[14] Lin G S，Milan A，Shen C H，et al. RefineNet：multi-

path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5168-5177.

[15] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]// The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, November 9-12, 2003, Pacific Grove, CA, USA. New York: IEEE Press, 2003: 1398-1402.

[16] Fan H Q, Su H, Guibas L. A point set generation network for 3D object reconstruction from a single image [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2463-2471.

[17] Levin A, Lischinski D, Weiss Y. Colorization using optimization[C]//ACM SIGGRAPH 2004 Papers on-SIGGRAPH'04, August 8-12, 2004. AngelesLos, California. New York: ACM Press, 2004: 689-694.

[18] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library [C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, December 8-14, 2019, Vancouver, BC, Canada. [S.l.: s.n.], 2019: 8024-8035.

[19] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22) [2021-07-21]. https://arxiv.org/abs/1412.6980.

[20] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[21] Sun C, Shrivastava A, Singh S, et al. Revisiting unreasonable effectiveness of data in deep learning era [C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 843-852.

[22] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy. Cambridge: JMLR, 2010: 249-256.