中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)08-2413-13

论文引用格式:Jia D, Cai P, Wu S, Wang Q and Song H L. 2024. Transformer network for stereo matching of weak texture objects. Journal of Image and Graphics, 29(08):2413-2425(贾迪, 蔡鹏, 吴思, 王骞, 宋慧伦. 2024. 面向弱纹理目标立体匹配的Transformer网络. 中国图象图形学报, 29(08):2413-2425)[DOI:10.11834/jig.230575]

面向弱纹理目标立体匹配的Transformer网络

贾迪[1],蔡鹏[1*],吴思[2],王骞[1],宋慧伦[1]

1. 辽宁工程技术大学电子与信息工程学院,葫芦岛　125105; 2. 国网葫芦岛供电公司,葫芦岛　125000

摘　要:**目的**　近年来,采用神经网络完成立体匹配任务已成为计算机视觉领域的研究热点,目前现有方法存在弱纹理目标缺乏全局表征的问题,为此本文提出一种基于Transformer架构的密集特征提取网络。**方法**　首先,采用空间池化窗口策略使得Transformer层可以在维持线性计算复杂度的同时,捕获广泛的上下文表示,弥补局部弱纹理导致的特征匮乏问题。其次,通过卷积与转置卷积实现重叠式块嵌入,使得所有特征点都尽可能多地捕捉邻近特征,便于细粒度匹配。再者,通过将跳跃查询策略应用于编码器和解码器间的特征融合部分,以此实现高效信息传递。最后,针对立体像对存在的遮挡情况,对固定区域内的匹配概率进行截断求和,输出更为合理的遮挡置信度。**结果**　在Scene Flow数据集上进行了消融实验,实验结果表明,本文网络获得了0.33的绝对像素距离,0.92%的异常像素占比和98%的遮挡预测交并比。为了验证模型在实际路况场景下的有效性,在KITTI-2015数据集上进行了补充对比实验,本文方法获得了1.78%的平均异常值百分比,上述指标均优于STTR(stereo Transformer)等主流方法。此外,在KITTI-2015、MPI-Sintel(max planck institute sintel)和Middlebury-2014数据集的测试中,本文模型具备较强的泛化性。**结论**　本文提出了一个纯粹的基于Transformer架构的密集特征提取器,使用空间池化窗口策略减小注意力计算的空间规模,并利用跳跃查询策略对编码器和解码器的特征进行了有效融合,可以较好地提高Transformer架构下的特征提取性能。

关键词:立体匹配;弱纹理目标;Transformer;空间池化窗口;跳跃查询;截断求和;Scene Flow;KITTI-2015

Transformer network for stereo matching of weak texture objects

Jia Di[1], Cai Peng[1*], Wu Si[2], Wang Qian[1], Song Huilun[1]

1. School of Electronic and Information Engineering, Liaoning Technical University, Huludao　125105, China;

2. State Grid Huludao Electric Power Supply Company, Huludao　125000, China

Abstract: **Objective** In recent years, the use of neural networks for stereo matching tasks has become a major topic in the field of computer vision. Stereo matching is a classic and computationally intensive task in computer vision. It is commonly used in various advanced visual processing applications such as 3D reconstruction, autonomous driving, and augmented reality. Given a pair of distortion-corrected stereo images, the goal of stereo matching is to match corresponding pixels along the epipolar lines and compute the horizontal disparity, also known as disparity. In recent years, many researchers have explored deep learning-based stereo matching methods, which achieving promising results. Convolutional neural networks are often used to construct feature extractors for stereo matching. Although convolution-based feature extractors have

收稿日期:2023-08-18;**修回日期**:2023-11-07;**预印本日期**:2023-11-14

*通信作者:蔡鹏　lntu_caipeng@163.com

基金项目:国家自然科学基金项目(61601213);辽宁省教育厅项目(LJ2020FWL004)

Supported by: National Natural Science Foundation of China (61601213); Research Foundation of Education Bureau of Liaoning Province (LJ2020FWL004)

yielded significant improvements in performance, neural networks are still constrained by the fundamental operation unit of "convolution". By definition, convolution is a linear operator with a limited receptive field. Achieving sufficiently broad contextual representation requires stacking layers of convolutions in deep architectures. This limitation becomes particularly pronounced in stereo matching tasks. In stereo matching tasks, captured stereo image pairs inevitably contain large areas of weak texture. Substantial computational resources are required to obtain comprehensive global feature representations through repeated convolutional layer stacking. We build a dense feature extraction Transformer for the stereo matching tasks, which incorporates Transformer and convolution blocks, to address the abovementioned issue. **Method** In the context of stereo matching tasks, FET exhibits three key advantages. First, by addressing high-resolution stereo image pairs, the inclusion of a pyramid pooling window within the Transformer block allows us to maintain linear computational complexity while obtaining a sufficiently broad context representation. This way addresses the issue of feature scarcity caused by local weak textures. Second, we utilize convolution and transposed convolution blocks for implementing subsampling and upsampling overlapping patch embeddings, which ensures that all points nearby features are captured as comprehensively as possible to facilitate fine-grained matching. Third, we experiment with employing a skip-query strategy for feature fusion between the encoder and decoder to efficiently transmit information. Finally, we incorporate the attention-based pixel matching strategy of stereo Transformer (STTR) to realize a purely Transformer-based architecture. This strategy truncates the summation of matching probabilities within fixed regions to output more reasonable occlusion confidence values. **Result** In the experimental section, we implemented our model using the PyTorch framework and trained it on an NVIDIA GTX 3090. We employed mixed precision during the training process to reduce GPU memory consumption and improve training speed. However, training a pure Transformer architecture in mixed precision proved to be unstable. The model experienced loss divergence errors after only a few iterations. We modified the order of computation for attention scores to suppress related overflows for addressing this issue. We also restructured the attention calculation method based on the additivity invariance of the softmax operation. Ablation experiments were conducted on the Scene Flow dataset. Results show that the proposed network achieves an absolute pixel distance of 0.33, an outlier pixel ratio of 0.92%, and a 98% overlap prediction intersection over union. Additional comparative experiments were conducted on the KITTI-2015 dataset to validate the effectiveness of the model in real-world driving scenarios. In these experiments, the proposed method achieved an average outlier percentage of 1.78, which outperformed mainstream methods such as STTR. Moreover, in tests on the KITTI-2015, MPI-Sintel, and Middlebury-2014 datasets, the proposed model demonstrated strong generalization capabilities. Subsequently, considering the limited definition of weak texture levels in currently available public datasets, we employed a clustering approach to filter images from the Scene Flow test dataset. Each pixel in the images was treated as a sample, with RGB values serving as the feature dimensions. This clustering process resulted in quantifying the number of different pixel categories within each image, which provided a measure of the texture strength or weakness in the images. The images were then categorized into "difficult", "moderate", and "easy" cases based on the number of clusters. Through comparative analysis, our approach consistently outperformed existing methods across the three sample categories, with a particularly notable improvement observed in the "difficult" case category. **Conclusion** For the stereo matching task, we propose a feature extractor based on the Transformer architecture. First, we transplant the architecture of the encoder and decoder of the Transformer into the feature extractor, which effectively combines the inductive bias of convolutions with the global modeling capabilities of the Transformer. In addition, the Transformer-based feature extractor can capture a broader range of contextual representations, which partially alleviates region ambiguity issues caused by local weak textures. Furthermore, we introduce a skip-query strategy between the encoder and decoder to achieve efficient information transfer, which mitigates semantic discrepancies between them. We also design a spatial pooling window strategy to reduce the significant computational burden resulting from overlapping block embeddings, which keeps the attention computation of the model within linear complexity. Experimental results demonstrate a significant improvement in weak texture region prediction, occluded region prediction, and domain generalization when compared with relevant methods.

Key words: stereo matching; low-texture target; Transformer; spatial pooling windows; jump queries; truncated summation; Scene Flow; KITTI-2015

0 引 言

立体匹配是计算机视觉中的经典密集型估计任务，常应用于三维重建、自动驾驶和增强现实等各类高级视觉处理中。给定一对畸变矫正后的立体像对，立体匹配的目的是沿着极线匹配真值像素并计算像素间的水平位移，即视差。近几年诸多学者开展了基于深度学习的立体匹配方法，并获得了较好的结果。一些方法采用卷积神经网络（convolutional neural network，CNN）构造立体匹配的特征提取器。（Chang 和 Chen，2018；Badki 等，2020；Tankovich 等，2021；Huang 等，2021；Hussain 等，2022；Li 等，2022）。将特征提取器分解为卷积编码器与卷积解码器，卷积编码器对输入图像进行下采样处理，以提取多尺度特征；卷积解码器聚合编码器的多尺度特征，并将其转化为最终的密集特征。

尽管基于卷积的特征提取器获得了较好效果，但神经网络依然受到“卷积”这一基本操作单元的限制。根据定义，卷积是一种有限感受野的线性算子，单个卷积的有限感受野和表达能力需要层层堆叠到深层架构中，从而获得足够广泛的上下文表达。上述限制在立体匹配任务中表现得尤为突出，在立体匹配任务中，实际拍摄的立体像对不可避免地存在大面积弱纹理区域，若通过重复堆叠卷积层以得到充分的全局特征表达，需要消耗大量计算资源。卷积的有限感受野导致弱纹理区域在特征空间中难以区分，增加了后续特征匹配的难度。有研究（Emlek 和 Peker，2023；Lipson 等，2021）尝试通过分层视差的方式获取弱纹理区域的全局信息，但从实验结果看，模型配准率提高有限，并未获得足够的竞争力。此外，很多工作（Bendig 等，2022；Chen 等，2018；Chen 等 2021；Peng 等，2017；Zhao 等，2017；Wang 等，2018；毛静怡 等，2021）都在致力于扩大 CNN 的感受野，以提高其上下文建模能力，但仍难以解决卷积操作的有限感受野问题。

基于注意力机制的模型在自然语言处理领域获得了显著成功，推动了其在计算机视觉处理任务中的发展。一些工作表明，Transformer 可以被视为密集特征提取器的一种可选择编码器架构，并在语义分割（Zheng 等，2021）、单目深度估计（Ranftl 等，2021）等密集型估计任务中获得理想的结果。视觉 Transformer（vision Transformer，ViT）（Dosovitskiy 等，2021）采用一个纯粹的 Transformer 替代卷积作为视觉主干网络，固定大小的输入图像被划分为固定个数的非重叠块嵌入表达，最终送入 Transformer 层来捕获全局信息。随后，Pyramid ViT（Wang 等，2021）利用金字塔结构来优化。Swin Transformer（Liu 等，2021）提出分层式 ViT，借助可移动窗口分区实现局部自注意力层间的信息交互。以上这些工作更多地关注特征提取中的编码器设计，并未给出用于密集预测任务的合理设计。DPT（dense prediction Transformer）（Ranftl 等，2021）提出一种基于编码器/解码器架构的密集特征提取器。它将 ViT（Dosovitskiy 等，2021）提供的块嵌入表达重新组合成多尺度特征表示，并采用卷积解码器将多尺度特征逐步融合成最终的密集特征图。这种策略导致立体匹配任务呈现出以下 3 个缺点：1）非重叠的块嵌入表达会增加匹配过程中的区域歧义，如弱纹理区域和重复区域；2）边缘像素无法通过区域以外的邻近像素捕获特征，影响后续细致化的匹配；3）固定大小且数量有限的补丁块在有限数据集下常常难以收敛，对立体像对数据集的规模要求较高。

对于立体深度估计任务，遮挡区域无法获得真实有效的视差。此前的工作往往通过分段平滑假设来推断遮挡区域的视差，这种做法难以保证预测精度。提供遮挡区域置信度估计将有利于配准等下游场景的分析，实现对遮挡和低置信度估计进行加权或拒绝。然而，大多数先前的方法不提供此类信息，STTR（stereo Transformer）（Li 等，2021）是首先提出计算遮挡区域置信度的工作。区别于 STTR 引入数据集标签中遮挡信息的做法，本文调整了计算匹配概率的方式，提供了更为合理的遮挡区域预测。

本文提出了一种纯粹的基于 Transformer 架构的立体匹配网络模型，贡献点如下：1）为克服基于卷积神经网络的立体匹配模型难以应对弱纹理区域特征匮乏的问题，通过引入跳跃查询思想，将 Transformer 的编码器/解码器架构移植到特征提取器中，利用 Transformer 架构的全局建模能力来增强弱纹理区域的特征表示。2）在 Transformer 架构下，非重叠块的嵌入表达会导致匹配过程中弱纹理区域匹配歧义，本文采用重叠式块嵌入策略，在保证多尺度特征的基础上，通过构建空间池化窗口减少重叠式块嵌入

所带来的巨大计算量，并维持线性复杂度的注意力计算。3）在立体匹配任务中，输入的双目像对不可避免地存在遮挡区域。参考STTR（Li等，2021）的工作，提出一种基于Transformer架构的特征匹配器。区别于STTR引入先验遮挡信息，本文通过调整其计算匹配概率的方式，提供了更为合理的遮挡区域预测结果，便于下游扩展任务的分析（例如场景理解和配准工作）。如图1所示，本文方法在Scene Flow数据集上优于STTR，且表现出较强的泛化性。图中纵坐标EPE（end-pint-error）表示绝对像素距离，横坐标3PE（3-px error）表示视差错误大于3像素的百分比。

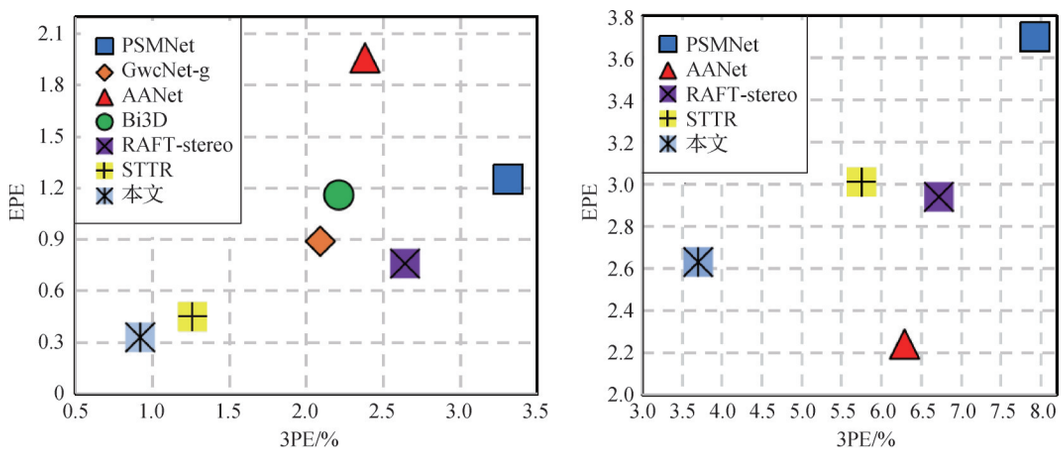


图1　可视化对比实验

Fig. 1　Visual comparison experiment((a)Scene Flow dataset;(b)MIP-Sintel dataset)

1　方　法

图2为本文网络的总体结构。在编码器中，输入的特征通过卷积层增加特征细粒度，随后被送入Transformer块进行全局表征学习。在解码器模块中，输入特征经转置卷积层捕获高分辨率特征表示，并与编码器中的同级特征进行融合。特征融合方式有3种：特征相加、特征合并和跳跃查询。前两种方法为常见的特征融合方式，广泛应用于各类网络架

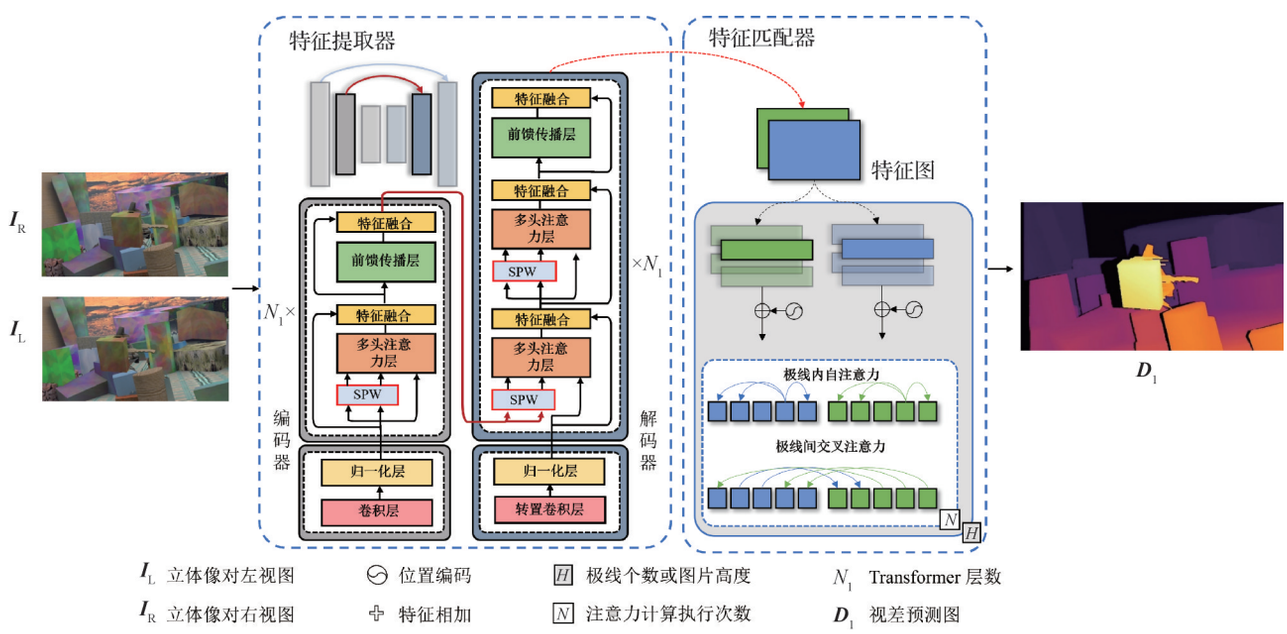


图2　总体网络架构

Fig. 2　Overall network architecture

构中，第3种方式为将编码器的输出特征视为键和值，并与同级的解码器特征执行多次多头注意力计算。此外，利用卷积/转置卷积实现了重叠式的块嵌入，并通过构建空间池化窗口策略来减小注意力计算中的词向量规模。

1.1 特征提取编码器

与 ViT（Dosovitskiy 等，2021）和 Pyramid ViT（Wang 等，2021）的非重叠式的词嵌入方法不同，本文使用单个卷积实现重叠式图像块划分，并将图像块映射成同等数量的词嵌入。重叠式的词嵌入方法可以提取固定块区域外的邻域特征，强化相似像素之间的差异。以编码器的第 i 阶段为例：给定输入特征图的长宽为 H_i 和 W_i，通过步长为 S_i、核大小为 $2S_i-1$、零填充大小为 S_i-1 的卷积处理，将特征图采样至 $\frac{H_i}{S_i}\times\frac{W_i}{S_i}$ 大小，特征维度为 C_i。将下采样后的特征图拉伸为水平排列的图像块，并映射为对应的词嵌入。带有位置编码的词嵌入经归一化操作后送入 Transformer（Vaswani 等，2017）中，以捕获全局特征信息。由于 Transformer 架构可以捕获序列中不同位置间的依赖关系，可有效缓解局部弱纹理导致的特征退化。最终输出的全局特征将重新调整为 $\frac{H_i}{S_i}\times\frac{W_i}{S_i}$ 的特征图。如果将第 $i-1$ 阶段输出的特征图作为第 i 阶段的输入，迭代上述操作可以获得多尺度特征图 $\{\boldsymbol{E}_1,\boldsymbol{E}_2,\boldsymbol{E}_3\}$。在解码器中，多尺度特征图将作为输入，实现特征间的跳跃查询。

在特征提取编码器中，所有的注意力计算均采用点积形式，将输入特征分组或映射为查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$。查询 $\boldsymbol{Q}$ 借助点积运算所得到的注意力权重从值 $\boldsymbol{V}$ 中检索相关信息，计算式为

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

式中，d 为 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的特征维度。softmax 为归一化指数函数。

由于立体匹配过程中的立体像对存在分辨率较大的情况，在整体特征图上进行注意力计算时间消耗较大，如采用重叠式的词嵌入方式，其数量往往是数万级别，远超计算能够接受的范围。而在 ViT 等工作中词嵌入个数往往只能保持固定数量，例如 ViT 保持 256 个词嵌入。为此，如图 3 所示，本文给出一种空间池化窗口策略（spatial pooling window，SPW），它能够在提取多尺度特征的同时，维持线性注意力计算，图中 h 和 w 表示经过 SPW 操作后的缩减尺度，具体为

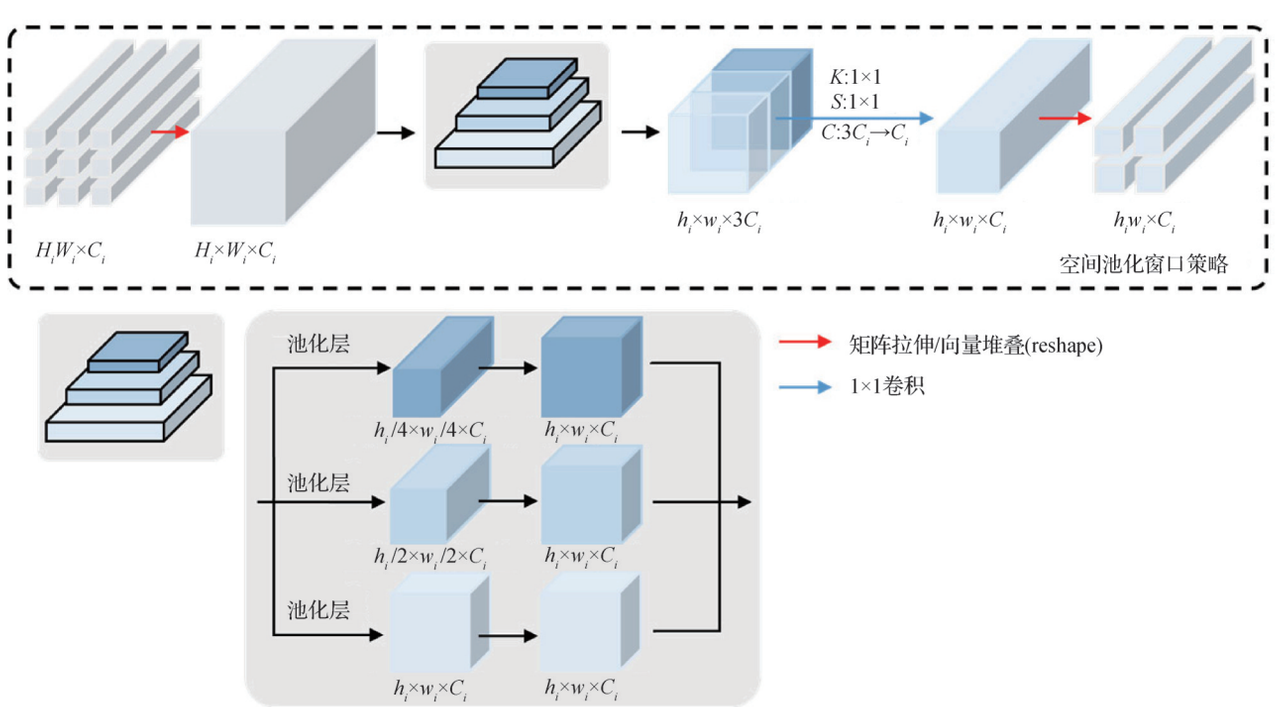


图3 空间池化窗口

Fig. 3 Spatial pooling window

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\cdot SPW(\boldsymbol{K})^{\text{T}}}{\sqrt{d}}\right)SPW(\boldsymbol{V}) \tag{2}$$

式中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{H_iW_i\times C_i}$分别为查询矩阵、键矩阵和值矩阵。

在SPW模块中，输入信息将进行多尺度的空间聚合，以缩小的比例呈现键与值矩阵，以此减少注意力计算的复杂度，SPW计算为

$$SPW(x)=Ft(Ct(SP_1(x),SP_2(x),SP_3(x))) \tag{3}$$

式中，$Ct(\cdot)$为特征合并，$Ft(\cdot)$为特征还原，$SP_i(\cdot)$为池化分支。具体而言，$Ct(\cdot)$用于融合不同尺度下的特征信息，将合并后的特征送入核大小为1的卷积进行重组，最终通过$Ft(\cdot)$将特征还原为可用于注意力计算的词嵌入。若给定最大池化的大小为p_0，则最终输出结果的大小为$p_0^2\times C_i$，其中$SP_i(\cdot)$计算为

$$SP_i(x)=f_{\text{Interpolate}}(AvgPool(Rs(x),P_i),P_0) \tag{4}$$

式中，$Rs(\cdot)$表示改变变量的空间维度，$AvgPool(\cdot)$表示对变量进行平均池化操作，Interpolate操作将对不同尺度的池化特征进行上采样。

1.2 特征提取解码器

特征提取解码器使用转置卷积实现重叠式的词嵌入，并对输入图像进行上采样操作。与编码器相似，划分后的图像块将被映射为多个词嵌入。以解码器的第i个阶段为例：给定输入特征图的长宽为$H_i'\times W_i'$，经过步长为S_i'、核大小为$2S_i'-1$、零填充大小为S_i'、外填充大小为$S_i'+1$的转置卷积，转化成数量为$S_i'^2H_i'W_i'$的图像块。将沿水平方向延伸的图像块映射为同等数量的词嵌入，通过归一化操作处理，与来自同级编码器的跳跃键值共同完成迭代式多头注意力计算，以此获得大小为$S_i'H_i'\times S_i'W_i'$的特征图$\boldsymbol{D}_i$。将上一阶段输出的特征图作为当前阶段的输入，重复上述操作获得多尺度的特征图$\{\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{D}_3\}$。

特征相加与特征合并的方式已广泛用于融合特征，如Unet（Ronneberger等，2015），Swin-Unet（Cao等，2022），DPT（Ranftl等，2021）等通用网络模型。为高效地融合来自编码器和解码器的特征，给出一种特征融合方法：跳跃查询（skip query，SQ），如图4（c）所示，将编码器/解码器的特征图转换为可学习的词嵌入，并将编码器的词嵌入作为多头注意力计算的键和值，将前一阶段解码器的词嵌入作为查询条件，并在跳跃查询中执行空间池化窗口策略。经典Transformer解码器由多头自注意力模块、多头交叉注意力模块和前馈网络构成，本文通过引入跳跃查询机制，将解码器的结构调整为跳跃查询模块、多头自注意力模块和前馈网络。

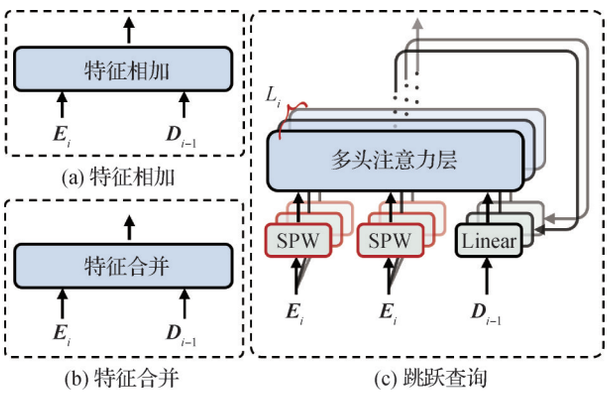


图4 特征融合方法

Fig. 4 Feature fusion method

((a)feature addition; (b)feature concatenation; (c)skip query)

1.3 特征匹配器

特征提取器从左右视图中分别提取密集特征图$\boldsymbol{F}_{\text{L}}$和$\boldsymbol{F}_{\text{R}}$。在特征匹配器中，将沿极线从特征图$\boldsymbol{F}_{\text{L}}$和$\boldsymbol{F}_{\text{R}}$中获得$H$个序列对$\langle\boldsymbol{e}_{\text{IL}},\boldsymbol{e}_{\text{IR}}\rangle$，对极线内注意力计算而言，键和值来自同一个序列$\boldsymbol{e}_{\text{IL}}$或$\boldsymbol{e}_{\text{IR}}$。对极线间注意力计算而言，键和值的组合选择为（$\boldsymbol{e}_{\text{IL}}$和$\boldsymbol{e}_{\text{IR}}$）或（$\boldsymbol{e}_{\text{IR}}$和$\boldsymbol{e}_{\text{IL}}$）。若将注意力分数的负值视为最优传输理论（Sarlin等，2020；Sun等，2021）的成本矩阵，输出最优耦合矩阵便可作为匹配概率P_{c}。由于特征匹配器沿极线进行注意力计算，摆脱了视差上限的约束，匹配概率P_{c}可用于计算全像素的视差估计值。

STTR（Li等，2021）采用调整计算匹配概率的方式，当定位到最大匹配概率P_{c}的位置x时，STTR在x周围建立一个尺寸大小为k的窗口$\boldsymbol{N}_k$，并对其中的匹配概率进行归一化。窗口内的相对偏移的加权和是视差值，具体为

$$D(x)=\sum_{I\in\boldsymbol{N}_k(x)}d_IP_I' \tag{5}$$

式中，d_I为窗口内的相对偏移量，$D(x)$为位置x的视差加权值。

当窗口内的概率总和过低时，当前分配区域的置信度应处于低水平，盲目地对所有窗口归一化将导致低置信度区域获得高匹配率，从而影响遮挡区域匹配。与STTR策略不同，本文通过引入非先验遮挡知识对固定区域内的匹配概率进行截断求和，具

体为

$$P_I' = \begin{cases} \dfrac{P_I}{\sum\limits_{j \in \boldsymbol{N}_k(x)} P_J} & \sum\limits_{j \in \boldsymbol{N}_k(x)} P_J > \theta \\ 0 & \sum\limits_{j \in \boldsymbol{N}_k(x)} P_J \leqslant \theta \end{cases} \tag{6}$$

式中，若P_I'取0时，将对应像素点预测为遮挡。θ为概率阈值，在训练过程中设置为0.05。

1.4 损失函数

采用相对响应损失L_{rr}作为监督项，当给定匹配像素集$\boldsymbol{M}_{\mathrm{c}}$和非匹配像素集$\boldsymbol{M}_{\mathrm{nc}}$，通过最小化$P_{\mathrm{c}}'$和$P_{\mathrm{nc}}'$的最大似然估计的负值回传损失，具体为

$$L_{\mathrm{rr}} = \frac{1}{|\boldsymbol{M}_{\mathrm{c}}|}\sum_{i \in \boldsymbol{M}_{\mathrm{c}}} \log(P_{\mathrm{c}}'(i)) + \frac{1}{|\boldsymbol{M}_{nc}|}\sum_{i \in \boldsymbol{M}_{nc}} \log(P_{\mathrm{nc}}'(i)) \tag{7}$$

采用$L1$距离监督最终的视差值，给定像素点i的视差估计值$\boldsymbol{D}_i$和真值$\boldsymbol{D}_i'$，视差值L_{dp}的监督损失计算为

$$L_{\mathrm{dp}} = \sum_{i=1}^{T} \left\| \boldsymbol{D}_i - \boldsymbol{D}_i^{\mathrm{gt}} \right\|_1 \tag{8}$$

给定像素点i的遮挡估计值O_i和真值O_i^{gt}，遮挡置信度值的监督损失表示为

$$L_{\mathrm{occ}} = -\frac{1}{T}\sum_{i=1}^{T}(O_i^{\mathrm{gt}} \log O_i + (1 - O_i^{\mathrm{gt}})\log(1 - O_i^{\mathrm{gt}})) \tag{9}$$

通过λ平衡各类损失项，最终的整体损失函数为

$$L = \lambda_1 L_{rr} + \lambda_2 L_{\mathrm{dp}} + (1 - \lambda_1 - \lambda_2)L_{\mathrm{occ}} \tag{10}$$

式中，λ_1，λ_2的作用为平衡不同损失项之间的强弱关系。在训练过程中，λ_1，λ_2均设置为1/3。

1.5 算法复杂度分析

在特征提取器中，空间池化窗口策略解决了采用重叠词嵌入的方式导致的计算开销问题。若不使用空间池化窗口，注意力机制的内存消耗常与图像分辨率成二次方关系。在特征提取器中以32位浮点运算精度为例，内存消耗（memory consumption，MC）为

$$MC = 2\sum_{i=1}^{4} 32 \frac{(I_{\mathrm{h}} I_{\mathrm{w}})^2}{s^2(i)} N_{\mathrm{h}}(i) N_{\mathrm{l}}(i) \tag{11}$$

式中，以Scene Flow数据集为例，图像高$I_{\mathrm{h}} = 540$、宽$I_{\mathrm{w}} = 960$。令每层相对步长$s = \{4, 8, 16, 32\}$，注意力头的个数$N_{\mathrm{h}} = \{1, 2, 4, 8\}$，Transformer的层数$N_{\mathrm{l}} = \{3, 4, 6, 3\}$，训练一个4层编码器/解码器结构需要225 GB，内存消耗过大。本文采用两种策略来解决该问题：1）使用混合精度减少实验过程中的内存消耗；2）在特征提取器中采用空间池化窗口策略减小词嵌入的空间规模，最终的内存消耗量为

$$MC = 2\sum_{i=1}^{4} 16 \frac{I_{\mathrm{h}} I_{\mathrm{w}}}{s^2(i)} P^2 N_{\mathrm{h}}(i) N_{\mathrm{l}}(i) \tag{12}$$

式中，若将p_0设置为8，所需的内存将从225 GB降低至0.42 GB，因此本文网络适用于边缘设备。

2 实验与结果分析

2.1 评价指标

本文采用3种标准进行方法评估：1）绝对像素距离（EPE），表示预测值和真实值在视差空间的绝对距离；2）3像素错误（3PE），表示视差错误大于3像素的百分比；3）遮挡交并比（occlusion intersection over union，OIOU），用于评估遮挡区域的预测结果。

2.2 实验环境

在特征提取器中，将Transformer的层数置为{3，4，6，3}，隐藏层的特征通道大小设置为{64，128，320，512}，特征提取器的输出通道数设为128，将空间池化窗口的池化尺度p_0设置为16，非重叠词嵌入的上/下采样倍率设置为2。在特征匹配器中，使用自注意力与交叉注意力交叉计算6次，交叉注意力计算步长为4，优化器采用权重衰减系数为0.001的AdamW方法（Loshchilov和Hutter，2019），批量大小为2。将特征提取器和特征匹配器的初始学习率设置为{0.000 1，0.000 2}，损失项系数λ_1、λ_2设置为1/3，遮挡概率阈值θ设置为0.05。本文网络采用PyTorch框架实现，在NVIDIA GTX 3090上进行训练。训练过程中采用混合精度减少GPU的内存消耗，并提升训练速度。然而，在混合精度下，训练一个纯粹的Transformer架构并不稳定，模型在少量迭代时会出现损失为空的错误。当注意力分数$\dfrac{\boldsymbol{Q}^{\mathrm{T}} SPW(\boldsymbol{K})}{\sqrt{d}}$超过16位浮点数的范围上限时，将导致训练溢出问题。本文采用如下方式抑制相关溢出：1）改变注意力分数的运算顺序，调整为$\boldsymbol{Q}^{\mathrm{T}}\left(\dfrac{SPW(\boldsymbol{K})}{\sqrt{d}}\right)$，以此缓解矩阵乘法的溢出；2）基于softmax操作的加法不变性，通过设置系数c将注意力分数约束在16位精度范围内，可将式（2）改写为

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = f_{\text{softmax}}\left(\left(\boldsymbol{Q}\cdot\frac{SPW(\boldsymbol{K})^{\text{T}}}{c\sqrt{d}} - \max\left(\boldsymbol{Q}\frac{SPW(\boldsymbol{K})^{\text{T}}}{c\sqrt{d}}\right)\right)\times c\right)SPW(\boldsymbol{V}) \tag{13}$$

2.3 实验对比与分析

实验部分采用4个公共数据集评估模型。Scene Flow(Mayer等,2016)为合成的数据集,该数据集的双目视图由沿着随机轨迹飞行的3D日常物品构成;KITTI(Menze和Geiger,2015)是一个具有200个广角立体像对组成的街景数据集;Middlebury(Scharstein等,2014)是一个室内场景数据集,它提供了23对不同光照条件下的高分辨率立体像对;MPI-Sintel(Butler等,2012)是一个基于动画电影的合成数据集,包含了大量逼真的光线特性:镜面反射、运动模糊和散焦模糊等。

2.3.1 基线指标对比

选用两种公开数据集(Scene Flow和KITTI-2015),表1给出了在Scene Flow数据集的实验对比结果。由于大部分模型(Laga等,2022)需要构建"代价体",并预先设置指定视差范围,为客观对比实验,在训练和评估阶段,将最大视差值分别设为192和480。由表1可见,本文方法在Scene Flow数据集上的指标获得了显著提升。由于注意力计算不受像素间的距离约束,本文方法在D = 480时依然能够保持D = 192的性能,且优于其他方法。在表1中,Oom表示在相同的实验条件下,对应模型无法处理Scene Flow数据集中高分辨率和大视差范围的图像。

为验证模型在真实街道的有效性,在Scene Flow数据集上进行20轮预训练,并对KITTI-2015数据集中的200组立体像对进行微调训练,结果如表2所示。与STTR相比,本文方法在KITTI-2015上各指标均得到了提升。表2中,D1-fg指标为在前景区域上的平均异常值百分比,D1-bg指标为在背景区域上的平均异常值百分比,D1-all指标为整体图像的平均异常值百分比。

表1 在Scene Flow数据集上的对比实验

Table 1 Comparative experiments on the Scene Flow dataset

方法	D = 192						D = 480					
	局部评估(D < 192)			全像素评估			局部评估(D < 192)			全像素评估		
	3PE/%	EPE	OIOU	3PE/%	EPE	OIOU	3PE/%	EPE	OIOU	3PE/%	EPE	OIOU
PSMNet	2.87	0.95	N/A	3.31	1.25	N/A	3.09	0.92	N/A	3.60	1.03	N/A
GwcNet-g	1.57	0.48	N/A	2.09	0.89	N/A	1.60	0.50	N/A	1.72	0.53	N/A
AANet	1.86	0.49	N/A	2.38	1.96	N/A	Oom		N/A	Oom		N/A
GANet-11	1.60	0.48	N/A	2.19	0.97	N/A	Oom		N/A	Oom		N/A
Bi3D	1.70	0.54	N/A	2.21	1.16	N/A	Oom		N/A	Oom		N/A
P3SNet	3.24	1.09	N/A	3.65	1.16	N/A	3.24	1.09	N/A	3.65	1.16	N/A
RAFT-stereo	1.43	0.49	0.965	2.64	0.76	0.961	1.60	0.52	0.955	2.21	0.81	0.958
STTR	1.13	0.42	0.961	1.26	0.45	0.954	1.13	0.42	0.961	1.26	0.45	0.954
本文	**0.88**	**0.31**	**0.987**	**0.92**	**0.33**	**0.982**	**0.88**	**0.31**	**0.987**	**0.92**	**0.33**	**0.982**

注:加粗字体表示各列最优结果,Oom表示对应模型无法处理高分辨率图像,N/A表示对应模型不支持输出相关指标。

2.3.2 泛化性实验

为了验证模型的域泛化性,在Scene Flow合成数据集上进行训练,在其余3个非同源数据集上进行评估,对比结果如表3所示。为公平起见,所有模型都将采用相同的数据增强技术,并保持一致的训练超参数。考虑到部分模型的最大视差限制,统一评估最大视差范围内的像素。对不支持预测遮挡区域的模型方法,采用N/A进行标记。由实验结果可见,本文方法在所有数据集的泛化性指标均获得了较高分数,图5给出了本文方法在各个数据集上的稠密深度估计结果。

2.3.3 弱纹理区域预测分析

目前,现有公开数据集鲜有对弱纹理程度进行定义,本文基于聚类思想对Scene Flow测试数据集

表2　在KITTI-2015数据集上的对比实验

Table 2　Comparison of evaluation results on KITTI-2015 dataset

/%

方法	*D*1-bg	*D*1-fg	*D*1-all
PSMNET	1.71	4.31	2.14
GwcNet-g	**1.61**	3.49	1.92
AANet	1.80	4.93	2.32
BI3D	1.79	**3.11**	2.01
RAFT-stereo	1.88	3.03	2.07
STTR	1.80	4.10	3.15
本文	1.78↑	3.68↑	**2.07**↑

注：加粗字体表示各列最优结果，“↑”表示优于STTR模型。

的图像进行筛选，将图像中每个像素视为一个样本，像素RGB值作为样本的特征维度，最终聚类出每个图像中不同像素类别的个数，通过像素类别数目量化图像的纹理强弱程度。在Scene Flow测试数据集中，利用K近邻聚类算法得到839～15 127不等类别数目，并按[839，1 500]，[1 500，10 000]，[10 000，15 127]的区间进行划分，分别对应“困难”样本、“中等”样本和“简单”样本，同时基于EPE指标进行实验对比，结果如表4。所有方法在“困难”样本上的准确率远低于“中等”和“简单”样本，表明弱纹理区域明显影响了立体匹配的准确率。其次，本文方法在3种不同样本区间上均获得较优结果，且在“困难”样本上的提升更为明显。可视化结果如图6所示，其中图6(a)(b)的纯色矩形框为典型弱纹理区域，与STTR相比，本文方法受益于Transformer架构的全局表征学习能力，对弱纹理目标具有较强的响应。

表3　泛化性实验

Table 3　Cross-domain generalization experiment

方法	MPI-Sintel			KITTI-2015			Middlebury-2014		
	3PE/%	EPE	OIOU	3PE/%	EPE	OIOU	3PE/%	EPE	OIOU
PSMNet	7.93	3.70	N/A	7.43	1.39	N/A	10.24	2.02	N/A
GwcNet-g	5.83	**1.32**	N/A	6.75	1.59	N/A	6.60	1.95	N/A
AANet	6.29	2.24	N/A	7.06	**1.31**	N/A	9.57	1.71	N/A
GANet-11	9.82	4.78	N/A	8.41	1.71	N/A	8.12	1.86	N/A
Bi3D	5.61	2.73	N/A	7.81	1.69	N/A	6.56	2.97	N/A
RAFT-stereo	6.72	2.94	0.87	7.59	1.91	0.95	8.42	1.64	0.93
STTR	5.75	3.01	0.86	6.74	1.50	**0.98**	**6.19**	2.33	0.95
本文	**3.70**↑	2.63↑	**0.91**↑	**6.65**↑	1.44↑	**0.98**↑	7.27	**1.58**↑	**0.96**↑

注：加粗字体表示各列最优结果，“↑”表示优于STTR模型，N/A表示对应模型不支持输出相关指标。

为直观展示本文方法对弱纹理区域的特征提取能力，对特征提取解码器的输入输出进行PCA(principal component analysis)降维，将特征图的维度降低至三维。图7(a)中的弱纹理区域面积较大，图7(b)为特征提取编码器的输出结果(即E_3)。图7(c)—(e)分别为特征提取解码器的各阶段输出，即$\{D_1, D_2, D_3\}$。图7(f)为最终视差预测值。由图7可见，随着解码器的逐步深入，弱纹理和细粒区域的特征能力获得了显著提升。

2.3.4　遮挡条件下的实验分析

对于立体深度估计任务而言，立体像对不可避免地存在遮挡现象。根据立体匹配任务的几何约束和匹配唯一性，遮挡区域无法获得真实有效的视差。区别于以往工作通过分段平滑假设推断遮挡区域视差的做法，本文通过对遮挡位置进行预测，输出像素级遮挡置信度，并用遮挡交并比作为评价指标。由表1和表3中的指标可见，本文方法对遮挡区域的预测结果优于STTR和RAFT-stereo方法。其中，N/A表示相关方法不支持进行遮挡区域预测。

为直观展示本文方法对遮挡区域处理的有效性，采用黑色区域表示遮挡区域，图6(c)(d)给出了可视化结果。在图6(c)中，STTR未能准确地预测出刀柄遮挡区域，而本文方法可以清晰地保留更多遮挡区域细节信息。在图6(d)中，本文方法能够关注到更多摩托车细节，且对摩托车的轮毂和车架附件遮挡区域进行了细致区分与预测。

(a) Middlebury-2014　(b) MPI-Sintel　(c) KITTI-2015

图 5　泛化性验证

Fig. 5　Generalization verification ((a) Middlebury-2014; (b) MPI-Sintel; (c) KITTI-2015)

表 4　弱纹理区域的绝对误差对比

Table 4　Comparison of absdute errors in weakly textured regions

方法	困难样本	中等样本	简单样本
Bi3D	4.56	1.09	0.79
GANet-11	3.35	0.93	0.46
RAFT-stereo	4.98	0.78	0.56
STTR	3.53	0.58	0.31
本文	**2.67** ↑	**0.41** ↑	**0.22** ↑

注：加粗字体表示各列最优结果，“↑”表示优于STTR模型。

2.4　性能评估

采用 Vladislav Sovrasov 测量模型的参数量(parameters)与执行时间(runtime)评估网络性能，结果如表 5 所示。相关实验在 NVIDIA GTX 1080Ti GPU 上进行，数据集选择 Scene Flow，对应图像的分辨率为 920 × 560 像素。由于本文方法使用了空间池化窗口策略，极大减少了 Transformer 架构的空间开销，在 920 × 560 像素分辨率的图像上，单组估计时间为 0.29 s，与同类方法相比，具有较高的准确率和较快的处理速度。此外，进一步将空间池化窗口 p_0 调整为 4 或 6，执行速率进一步提高，且参数量更少，EPE 错误率也保持在一个较低的水平上。

2.5　消融实验

消融实验选用 Scene Flow 数据集，实验过程与结果如表 6 所示，主要包括 6 组实验：A、B、C 组实验

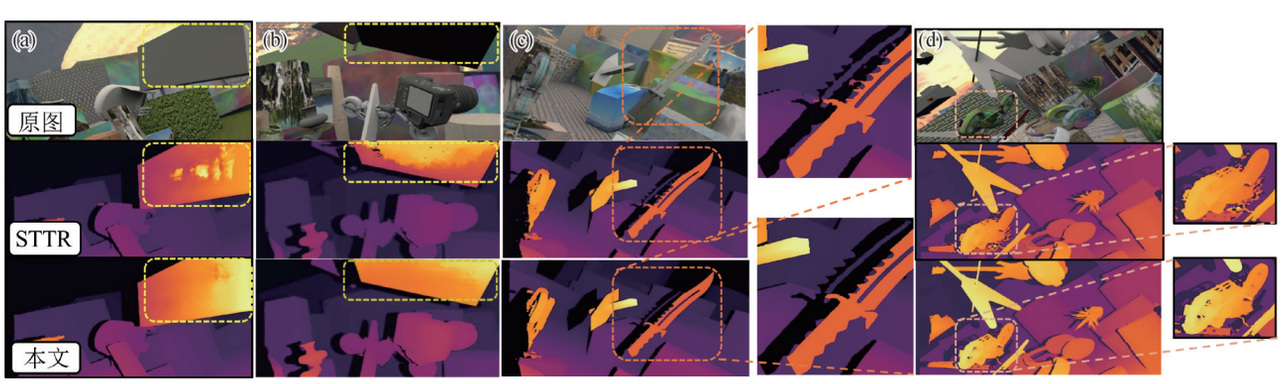
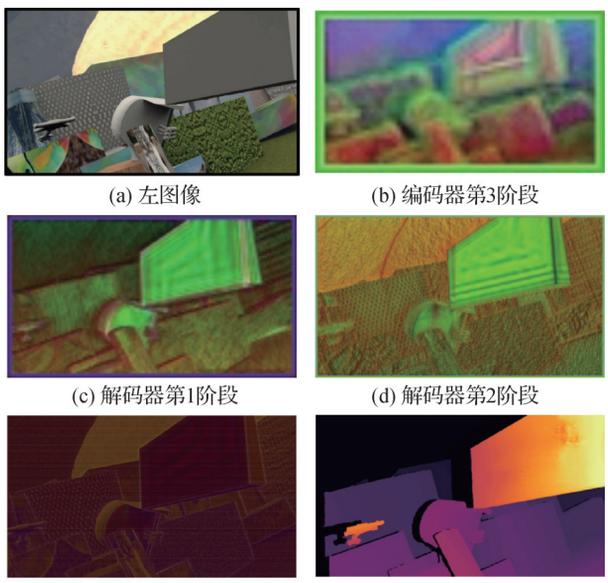


图 6　对 Scene Flow 数据集的深度图可视化对比

Fig. 6　Comparison of depth map visualization for Scene Flow dataset



(a) 左图像　(b) 编码器第3阶段
(c) 解码器第1阶段　(d) 解码器第2阶段
(e) 解码器第3阶段　(f) 预测视差图

图 7　多阶段特征可视化结果

Fig. 7　Results of multi-stage feature visualization
((a)left image;(b)encoder stage 3;(c)decoder stage 1;
(d)decoder stage 2;(e)decoder stage 3;
(f)predicted disparity map visualization)

表 5　性能评估对比

Table 5　Performance evaluation comparison

方法	参数量/M	执行时间/s	EPE
PSMNet	5.22	0.45	0.95
GwcNet-g	6.91	0.38	0.48
AANet	**3.81**	0.22	0.49
GANet-11	8.64	1.80	0.48
Bi3D	10.54	1.56	0.54
RAFT-stereo	13.34	0.23	0.49
STTR	19.26	0.42	0.45
本文_8	13.40	0.29	**0.31**
本文_6	11.49	0.24	0.35
本文_4	9.32	**0.21**	0.40

注:加粗字体表示各列最优结果。

用于对比特征融合方法,SC(skip concatenate)为特征合并,SA(skip add)为特征相加,SQ(skip query)为跳跃查询;D组实验对3种特征融合架构采用更多的训练轮数;E组实验验证了空间池化窗口策略的有效性;F组实验选用ResNet34作为主干网络。由于跳跃查询方法引入了额外的Transformer层,其具有更大的模型规模,为公平开展实验,为A、B两组实验

表 6　不同组件的消融实验

Table 6　Different component ablation experiments

实验组	方法					3PE/%	EPE	OIOU
	SC	SA	SQ	SPW	ME			
A	√	×	×	√	×	1.04	0.40	0.96
B	×	√	×	√	×	1.11	0.43	0.96
C	×	×	√	√	×	0.97	0.38	0.97
D	×	×	√	√	√	**0.92**	**0.33**	**0.98**
E	√	×	×	×	×	Oom	Oom	Oom
F	ResNet					1.26	0.45	0.95

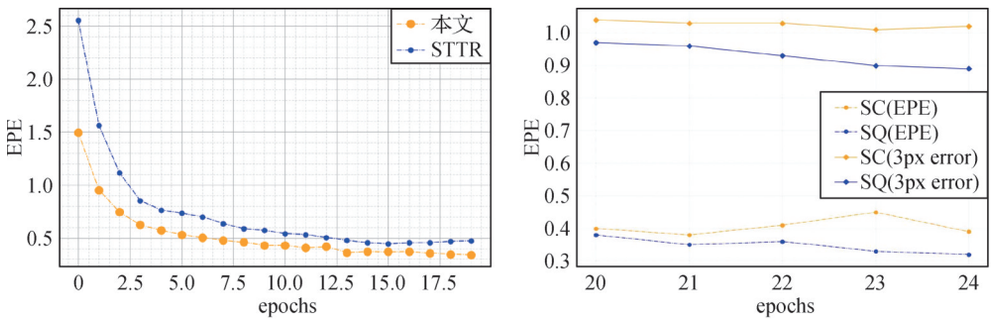
注:加粗字体表示各列最优结果。Oom表示对应模型无法处理高分辨率图像,“√”表示采用,“×”表示未采用。

增加了特征解码器中的自注意力层数,以达到同等量的模型规模。由A、B、C三组实验可见,当训练轮数保持一致时,3种连接方式均表现良好,其中跳跃查询获得了最佳结果。

由于纯粹的Transformer架构不易完全收敛,因此在初始训练轮数为20的基础上,D组实验对跳跃查询和跳跃合并方法增加额外训练轮数(ME表示额外训练轮数)。D组实验结果表明,更多的训练轮数可以显著提高跳跃查询结构的性能,且特征合并结构已经达到饱和状态。图8给出了训练过程中实验指标的变化曲线。随着训练轮数的增加,跳跃查询方法在EPE指标和3像素错误率上有所降低。E组实验中,Oom表示内存超出限制,若不使用空间池化窗口策略,一个纯粹的视觉Transformer架构难以处理Scene Flow数据集中的高分辨率图像。F组实验中,将卷积神经网络ResNet34作为主干网络,主要原因是ResNet34的参数量与本文模型接近,便于进行实验对比。综上所述,在采用空间池化窗口和跳跃查询策略后,本文模型获得了最佳实验结果。

3　结　论

针对立体匹配任务中现存的弱纹理目标缺乏全局表征的问题,本文提出一种面向弱纹理目标立体匹配的Transformer网络。通过空间池化窗口策略,在维持线性计算复杂度的同时,解决局部弱纹理导致的特征匮乏问题;其次,在编码器和解码器间使用跳跃查询策略,以实现高效的信息传递,缓解信息在编码器和解码器间存在的语义差异;最后,在特征匹



(a) 前20轮训练曲线　(b) 后5轮训练曲线

图8　收敛曲线

Fig. 8　Convergence curves ((a) the first 20 rounds of training curves; (b) the last 5 rounds of training curves)

配阶段，提出截断概率求和的方法，输出更为准确的遮挡区域预测。在4个公开数据集上分别进行了对比实验、域泛化性实验和弱纹理区域对比实验。实验结果表明，本文方法在弱纹理区域预测、遮挡区域预测和域泛化方面均获得了显著提升，而且在处理速度上具备较高的实时性。尽管本文模型在立体匹配任务中获得了效果的提升，但未涉及到其他的密集型估计任务，后续工作将在更多的密集型任务中验证本文方法的有效性。

参考文献(References)

Badki A, Troccoli A, Kim K, Kautz J, Sen P and Gallo O. 2020. Bi3D: stereo depth estimation via binary classifications//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 1597-1605 [DOI: 10.1109/CVPR42600.2020.00167]

Bendig K, Schuster R and Stricker D. 2022. Self-superflow: self-supervised scene flow prediction in stereo sequences//Proceedings of 2022 IEEE International Conference on Image Processing (ICIP). Bordeaux, France: IEEE: 481-485 [DOI: 10.1109/ICIP46576.2022.9897832]

Butler D J, Wulff J, Stanley G B and Black M J. 2012. A naturalistic open source movie for optical flow evaluation//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 611-625 [DOI: 10.1007/978-3-642-33783-3_44]

Cao H, Wang Y Y, Chen J, Jiang D S, Zhang X P, Tian Q and Wang M N. 2022. Swin-Unet: Unet-like pure Transformer for medical image segmentation//Proceedings of 2022 European Conference on Computer Vision. Tel Aviv, Israel: Springer: 205-218 [DOI: 10.1007/978-3-031-25066-8_9]

Chang J R and Chen Y S. 2018. Pyramid stereo matching network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5410-5418 [DOI: 10.1109/CVPR.2018.00567]

Chen C F R, Fan Q F and Panda R. 2021. CrossViT: cross-attention multi-scale vision Transformer for image classification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 347-356 [DOI: 10.1109/ICCV48922.2021.00041]

Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 833-851 [DOI: 10.1007/978-3-030-01234-2_49]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2023-08-18]. https://arxiv.org/pdf/2010.11929.pdf

Emlek A and Peker M. 2023. P3SNet: parallel pyramid pooling stereo network. IEEE Transactions on Intelligent Transportation Systems, 24(10): 10433-10444 [DOI: 10.1109/TITS.2023.3276328]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Huang Z Y, Norris T B and Wang P Q. 2021. ES-Net: an efficient stereo matching network [EB/OL]. [2023-08-18]. https://arxiv.org/pdf/2103.03922.pdf

Hussain M I, Rafique M A and Jeon M. 2022. RVMDE: radar validated monocular depth estimation for robotics [EB/OL]. [2023-08-18]. https://arxiv.org/pdf/2109.05265.pdf

Laga H, Jospin L V, Boussaid F and Bennamoun M. 2022. A survey on deep learning techniques for stereo-based depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4): 1738-1764 [DOI: 10.1109/TPAMI.2020.3032602]

Li J K, Wang P S, Xiong P F, Cai T, Yan Z W, Yang L, Liu J Y, Fan H Q and Liu S C. 2022. Practical stereo matching via cascaded recurrent network with adaptive correlation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16242-16251 [DOI: 10.1109/CVPR52688.2022.01578]

Li Z S, Liu X T, Drenkow N, Ding A, Creighton F X, Taylor R H and Unberath M. 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with Transformers//Proceedings

of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 6177-6186 [DOI: 10.1109/ICCV48922.2021.00614]

Lipson L, Teed Z and Deng J. 2021. RAFT-stereo: multilevel recurrent field transforms for stereo matching//Proceedings of 2021 International Conference on 3D Vision. London, UK: IEEE: 218-227 [DOI: 10.1109/3DV53792.2021.00032]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization [EB/OL]. [2023-08-18]. https://arxiv.org/pdf/1711.05101.pdf

Mao J Y, Song Y Q and Liu Z. 2021. CT image classification of liver tumors based on multi-scale and deep feature extraction. Journal of Image and Graphics, 26(7): 1704-1715 (毛静怡, 宋余庆, 刘哲. 2021. 多尺度深度特征提取的肝脏肿瘤CT图像分类. 中国图象图形学报, 26(7): 1704-1715) [DOI: 10.11834/jig.200041]

Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A and Brox T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4040-4048 [DOI: 10.1109/CVPR.2016.438]

Menze M and Geiger A. 2015. Object scene flow for autonomous vehicles//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3061-3070 [DOI: 10.1109/CVPR.2015.7298925]

Peng C, Zhang X Y, Yu G, Luo G M and Sun J. 2017. Large kernel matters——improve semantic segmentation by global convolutional network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1743-1751 [DOI: 10.1109/CVPR.2017.189]

Ranftl R, Bochkovskiy A and Koltun V. 2021. Vision Transformers for dense prediction//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 12159-12168 [DOI: 10.1109/ICCV48922.2021.01196]

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Sarlin P E, DeTone D, Malisiewicz T and Rabinovich A. 2020. SuperGlue: learning feature matching with graph neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 4937-4946 [DOI: 10.1109/CVPR42600.2020.00499]

Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X and Westling P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth//Proceedings of the 36th German Conference on Pattern Recognition. Münster, Germany: Springer: 31-42 [DOI: 10.1007/978-3-319-11752-2_3]

Sun J M, Shen Z H, Wang Y A, Bao H J and Zhou X W. 2021. LoFTR: detector-free local feature matching with Transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 8918-8927 [DOI: 10.1109/CVPR46437.2021.00881]

Tankovich V, Häne C, Zhang Y D, Kowdle A, Fanello S and Bouaziz S. 2021. HITNet: hierarchical iterative tile refinement network for real-time stereo matching//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14357-14367 [DOI: 10.1109/CVPR46437.2021.01413]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, Lu T, Luo P and Shao L. 2021. Pyramid vision Transformer: a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 548-558 [DOI: 10.1109/ICCV48922.2021.00061]

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zheng S X, Lu J C, Zhao H S, Zhu X T, Luo Z K, Wang Y B, Fu Y W, Feng J F, Xiang T, Torr P H S and Zhang L. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 6877-6886 [DOI: 10.1109/CVPR46437.2021.00681]

作者简介

贾迪，男，教授，主要研究方向为立体匹配与三维重建、摄影测量、视觉空间定位和视觉机械臂作业。

E-mail: lntu_jiadi@163.com

蔡鹏，通信作者，男，硕士研究生，主要研究方向为立体匹配与三维重建。E-mail: lntu_caipeng@163.com

吴思，女，硕士研究生，主要研究方向为电力视觉技术。

E-mail: lntu_wusi @163.com

王骞，男，硕士研究生，主要研究方向为视觉机械臂作业。

E-mail: lntu_wangqian@163.com

宋慧伦，女，硕士研究生，主要研究方向为三维重建。

E-mail: lntu_shl@163.com