文章编号:1671-4598(2024)04-0174-06 DOI:10.16526/j.cnki.11-4762/tp.2024.04.025 中图分类号:TM93 文献标识码:A

一种弱纹理目标立体匹配网络

刘 泽,姜永利,丁志伟,刘永强
(国能宝日希勒能源有限公司,内蒙古 呼伦贝尔 021500)

摘要:现有深度估计方法在高分辨率图像下存在特征提取不够充分、局部信息特征提取差的问题,为此提出一种面向全局特征的Transformer立体匹配网络;该网络采用编码器-解码器的端到端架构,采用多头注意力机制,允许模型在不同子空间中关注不同特征,从而提高特征提取能力;模型将自注意力机制和特征重构窗口结合,能够提高特征的表征能力,弥补局部特征不足的问题,在减少计算负担的同时有效解决Transformer架构计算复杂度高的问题,将模型的计算复杂度保持在线性范围内;在Scene Flow、KITTI-2015数据集上分别进行实验,与现有方法相比,相关指标得到显著提升,验证了模型的有效性和实用性。

关键词:深度估计;编码器-解码器;自注意力机制;特征重构窗口;全局上下文信息

A Stereo Matching Network for Weak Texture Objects

LIU Ze, JIANG Yongli, DING Zhiwei, LIU Yongqiang
(Guoneng Baorixile Energy Co. Ltd, Hulunbuir 021500, China)

Abstract: Existing depth estimations have the problems of insufficient feature extraction and poor local feature extraction in high-resolution images. Therefore, a Transformer stereo matching network oriented to global features is proposed. The network adopts an encoder-decoder with end-to-end architecture and multi-head attention mechanism, which allows the model to pay attention to different features in different subspaces, thus improving the feature extraction ability. By combining the self-attention mechanism with the feature reconstruction window, the model can improve the representation ability of features to compensate for the shortage of local features, and effectively solve the high computational complexity of Transformer architecture, so that the computational complexity of the model is maintained within a linear range. Experiments on the Scene Flow and KITTI-2015 data sets show that compared with the existing methods, the relevant indicators are significantly improved, which verifies the effectiveness and practicability of the model.

Keywords: depth estimation; encoder-decoder; self attention mechanism; feature reconstruction window; global context information

0 引言

深度估计[1]是计算机视觉领域的一项重要任务,常应用于自动驾驶、增强现实、机器人导航、虚拟现实、三维重建等高级应用中。其本质在于提取图像特征,并通过这些特征来识别目标、匹配对应点、进行像素级的深度估计,通过分析图像或场景中各像素点间的距离或深度信息,实现三维场景的理解和建模。特征提取[2]在深度估计任务中至关重要,传统的计算机视觉方法常使用手工设计特征,如方向梯度直方图(HOG)、尺度不变特征变换(SIFT)和速度不变特征变换(SURF)。这些特征常基于局部纹理和边缘信息,适用于一些传统的深度估计算法。随着深度学习的发展,卷积神经网络在深度估计中得到广泛应用,这类网络能够学习到目标更多特征表示和匹配规则,提高立体匹配的性能。

戴仁月[3]提出一种融合卷积神经网络(CNN, convolutional neural network)与传统即时定位与地图构建算法的深度估计方法,从非结构化视频序列中估计深度,使用当前帧或相邻帧来估计深度,但并未利用全局和几何信息来优化深度图。温静[4]提出一种基于CNN特征提取和加权深度迁移的单目图像深度估计方法,先提取CNN特征,并计算输入图像在数据集中的近邻图像,再获得各候选近邻图像和输入图像间的像素级稠密空间形变函数,将形变函数迁移至候选深度图像集,通过引入基于SIFT的迁移权重SSW,对加权迁移后的候选深度图进行优化,以此获得最终的深度信息。李格[5]提出CNN模型与CRFasRNN相结合的网络结构,该过程先以场景RGB图的超像素块为单元提取局部二进制LBP特征、颜色差异特征、颜色直方图分布差异特征,再归一化3种特征下的特征图,并以此对输出的深度图进行线性滤波,随后将此滤波结果作为联合滤波器的CNN网络输入,进一步提高深度估计精度。

卷积神经网络通过一系列卷积层和池化层学习图像特征,其中卷积层使用不同尺寸卷积核捕获不同感受野信息,

收稿日期:2023-10-20; **修回日期**:2023-11-11。
基金项目:国家自然科学基金(61601213)。
作者简介:刘 泽(1986-),男,大学本科。
引用格式:刘 泽,姜永利,丁志伟,等.一种弱纹理目标立体匹配网络[J].计算机测量与控制,2024,32(4):174-179,187.

有效提取图像中的局部特征，以此获得目标表面结构的重要信息。尽管 CNN 的局部特征提取能力强，但存在以下问题：

1）传统 CNN 模型缺乏长距离依赖建模能力，难以捕获图像中物体间的全局关系和上下文信息。在深度估计任务中，特别在处理高分辨率图像时，全局信息对准确的深度估计至关重要；

2）由于 CNN 关注的是局部区域特征，常无法更好地捕获全局信息，可能导致信息的捕获能力下降。

近年来，基于注意力机制的模型在自然语言处理领域获得广泛应用。Transformer[6-7]的自注意力机制可以突破感受野的限制，使其能够在整个图像上建立关联，实现全局信息捕获，具备较高的泛化性 ，对图像分类、目标检测和分割等任务至关重要。ViT（Vision Transformer）将图像数据转换为序列数据，使用 Transformer 架构来处理序列数据，包括图像块的向量化表示、位置编码、Transformer 编码器结构，以及用于图像分类的分类结构。该架构使得 ViT 能够有效利用多头自注意力机制建模像素间的关联，处理不同尺寸的图像，更好地理解图像中的全局关系，在大规模图像分类任务中表现出色。由于采用了相对模块化结构，使其容易扩展和修改，适用于不同的任务和应用。Swin Transformer[8]通过自注意力机制捕获输入序列中不同元素间的依赖，采用深度分层结构，将输入图像分为不同分辨率图像块，在每个分辨率上应用 Transformer 编码器，有助于模型同时处理全局和局部信息，提高对不同尺度下的特征建模能力。Swin Transformer 引入“Shifted Window”机制，使滑动窗口的方式，允许模型在不同空间尺度下进行全局信息交互 ，同时关注全局和局部信息，有效地捕获了不同位置间的关系，从而进一步增强了模型的特征提取能力。Swin Transformer 可用于各种计算机视觉任务，包括图像分类、对象检测、语义分割和实例分割，多尺度特性使其适用于不同场景和任务。由此可见，视觉 Transformer 的多头注意力机制具有长距离依赖和自适应空间聚合能力，可以从海量数据中学到比 CNN 网络更加强大和鲁棒的表征。

在自动驾驶、增强现实、机器人导航等领域，经常需要在复杂和多样化场景中使用立体匹配技术。这类场景中，常存在弱纹理区域，因此解决弱纹理区域的立体匹配问题对于实际应用至关重要。在传统的立体匹配方法中，弱纹理区域无法获得真实有效的视差，非重叠块的嵌入表达可能导致弱纹理区域匹配歧义。为此，本文提出一种纯粹基于 Transformer 架构的弱纹理目标立体匹配网络。通过引入重叠式块嵌入策略，提升弱纹理区域的匹配性能，使相邻块间的信息有所重叠，从而增加在弱纹理区域表达的一致性，减少歧义并提高深度估计性能，通过借助特征重构窗口策略[9]增强特征的表达能力，以此提高模型在弱纹理区域的立体匹配性能。

1　立体匹配

如图 1 所示，双目相机是由左右两个针孔相机水平拼接而成，当两个针孔相机的光圈中心都在一条线且法向量平行时，光圈中心间的距离为双目相机的基线。

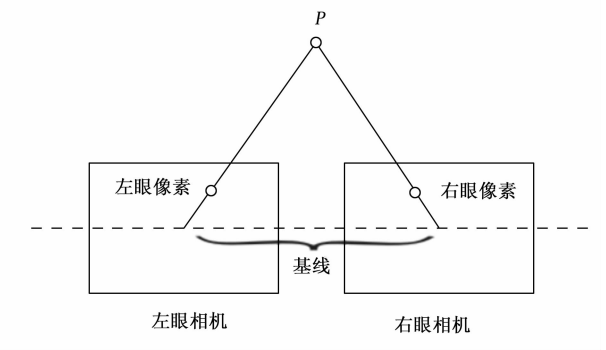

图 1　双目相机模型

利用基线和相机的焦距，存在以下关系：

$$\frac{Z-f}{z}=\frac{b-(U_L-U_R)}{b} \tag{1}$$

其中：z 为 p 点在 Z 轴的投影长度，f 为焦距，b 为基线距离，U_R-U_L 称为视差。

双目测距的匹配问题常称之为立体匹配，其主要目标是找到图像中每个像素间的对应视差，即两个视角下的像素间的距离。视差值可用来估计目标深度，从而还原出三维场景。从采用不同最优化理论方法的角度出发，立体匹配的非学习方法可分为全局立体匹配与局部立体匹配两类方法。从采用不同图像表示基元的角度出发进行分类，可分为区域立体匹配算法、基于特征的立体匹配算法和基于相位立体匹配算法，常见的立体匹配方法[10-13]包括匹配代价计算、代价聚合、视差计算、视差优化 4 个步骤。

匹配代价是指图像中的每个像素与其在另一图像中匹配点间的相似度，可以通过各种方法计算，如灰度值、特征向量的相似性等。匹配代价图通常具有噪声和不确定性，因此需要进行代价聚合，改善深度估计的质量。代价聚合有助于整合匹配代价图的局部信息，以获得更平滑和准确的视差图。视差计算阶段的任务是确定每个像素的最佳匹配点，即匹配代价最小的像素位置，对应于左图像中的像素在右图像中的匹配点。视差值表示两个像素间的距离，可用于估计目标的深度，视差优化阶段旨在进一步改善视差图的质量，常包括使用优化算法，如动态规划、全局优化或半全局匹配等，平滑和修复视差图中的不一致性和噪声。常见的代价计算方法有 SAD（sum of absolute differences）、SSD（sum of squared differences）、AD 算法等，其中 AD 算法是匹配代价计算中最简单的算法之一，其主要思想是不断比较左右相机中两点的灰度值。通过固定左相机中的一个像素点，遍历右相机中的所有像素点，不断比较它们之前的灰度之差，灰度差即为匹配代价，其数学公式为：

$$C_{AD}(p,q) = | I_L(p) - I_R(q) | \tag{2}$$

其中：p 和 q 分别为左右图像中的两点，I_L（ ）为左图像中的灰度值，I_R（ ）为右图像中的灰度值。上式为灰度图像间的匹配代价，彩色图像 AD 算法的计算代价为：

$$C_{AD}(p,q) = \frac{1}{3}\sum_{i=R,G,B} | I_i^L(p) - I_i^R(q) | \tag{3}$$

代价聚合用于处理视差图中的不确定性和噪声，从而改善深度估计的质量。代价聚合的目标是将匹配代价图（Cost Volume）的局部信息进行整合，以获得更平滑和准确的视差图。近年来诸多学者开展了基于深度学习的立体匹配方法研究工作，常采用卷积神经网络构造立体匹配的特征提取器，将特征提取器分解为卷积编码器与卷积解码器。尽管基于卷积的特征提取器获得较好效果，但卷积层的感受野通常是局部的，使得卷积层在处理全局信息或长距离依赖关系时面临挑战。基于 Transformer 的模型可以较好地解决该问题，在解码器模块中，所有注意力计算均采用点积形式，其中输入特征可分为查询（Q）、键（K）和值（V）。查询 Q 借助点积运算得到的注意力权重，可从值 V 中检索相关信息，计算公式如下：

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

2 总体架构

总体架构如图 2 所示，通过编码器和解码器模块协同工作实现图像处理和表示学习。编码器模块对输入特征进行初步处理，增强特征的细粒度，将其传递给 Transformer 块，以进行全局地表征学习。在解码器模块中，输入特征经转置卷积层处理获高分辨率的特征表示，并与编码器中的同级特征进行融合，详细过程如下：

编码器模块通过多层卷积操作增加输入特征的细粒度，卷积层通过一系列卷积核滑动捕捉输入图像的局部特征，逐渐将图像中的细节信息传递到更高级别的表征，这种方式有助于模型更好地理解图像的局部结构。将编码器将处理后的特征输入到 Transformer 结构中进行全局表征学习。Transformer 借助于模型捕获不同位置之间的依赖关系，从而可以更好地理解图像中的全局结构信息。两个组件间的结合使得编码器模块能够在保留细粒度特征的同时，提高对整体图像的特征提取能力。

解码器模块用于恢复高分辨率特征表示，并将其与编码器中的同级特征进行融合，解码器通过转置卷积操作来逆转这种过程。在解码器中，转置卷积层有助于对低分辨率特征进行上采样，从而获得更高分辨率的特征图。这些高分辨率特征图可以帮助模型更好地理解图像细节，如纹理和边缘等，将这些高分辨率特征与编码器中的同级特征进行融合，获得更为全面的特征表示。

这种编码器－解码器架构的优点在于能够从多层次上捕获图像特征，提高模型的特征提取能力，其核心是 Block 模块结构，主要由以下几个部分构成：

1）局部感知单元（Local Perception Unit），将输入图

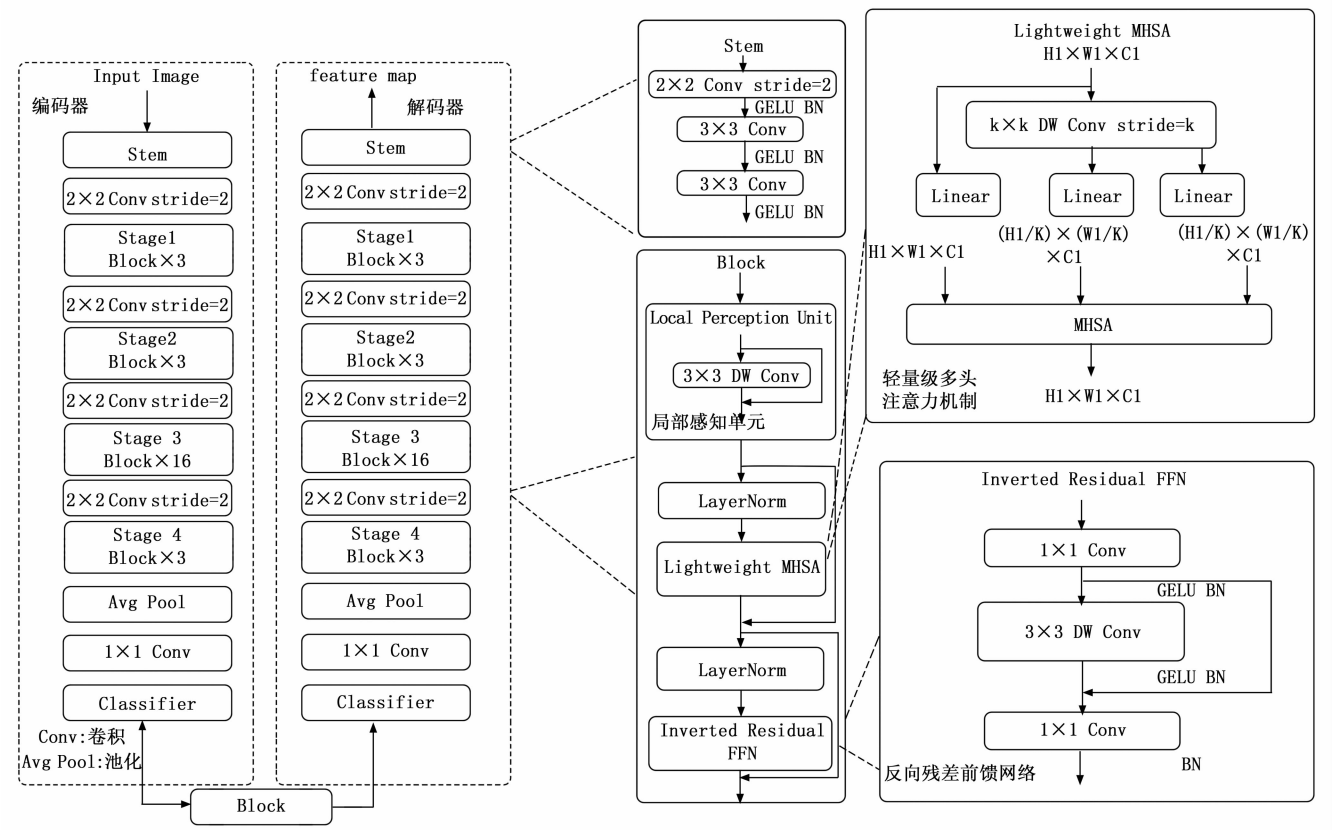

图 2 总体网络架构

片信息，与 3 * 3 的卷积操作相加，旨在增加空间信息提取能力。

2）轻量级多头注意力机制（Lightweight Multi-head Self-attention），使用深度卷积计算代替 key 和 value 的计算，从而减轻计算开销。

3）反向残差前馈网络（Inverted Residual Feed-forward Network），类似于反向残差块，由扩展层、深度卷积和投影层组成。通过改变跳跃连接的位置，提高网络性能。深度卷积用于提取局部信息，其计算成本可以忽略不计，跳跃连接与经典的残差网络相似，可以提高梯度跨层的传播能力。

3　方法

在 Vision Transformer 和 Pyramid Vision Transformer (PVT)[14-16]中，首先将输入图像划分为不相交的图像块，这些图像块被视为模型的“词”或“记号”，类似于自然语言处理中的标记化。每个图像块及其位置编码通过一个线性映射被嵌入到固定维度的向量空间中，以此构成一个序列。将该序列作为输入送入 Transformer 编码器[17]中，用于提取图像中的特征，建模像素之间的关系。与传统的 VIT 和 PVT 方法不同，本文提出一种重叠式词嵌入方法，以更好地处理弱纹理区域和捕获相邻区域的特征信息。采用重叠式词嵌入方法，图像块之间存在重叠，有助于在加强相似像素差异的同时，捕获更为全面的特征信息。以编码器的第一阶段为例，通过卷积操作将输入特征图缩减到较小的尺寸，更好地捕获局部特征。将这些小块特征图转化为词嵌入，加入位置编码输入到 Transformer 中，以便在全局范围内提取图像特征。

Transformer 能够处理序列数据中不同位置的依赖关系，从而有效减轻弱纹理区域缺少特征的问题。全局特征被重新调整为原始大小的特征图，可以获得具有更好表示能力的特征图。如需多尺度的特征图，可将第一阶段的输出再次输入到第二阶段，重复该过程。这种重叠式词嵌入方法可有效捕获每个块区域以及周边邻域的特征信息，从而更好地突出相似像素间的差异。

在处理高分辨率的立体图像对时，使用重叠式词嵌入方法处理整个特征图时，其注意力计算开销较大，庞大的词嵌入数量可能导致计算资源超出范围。为了在处理高分辨率图像时仍能保持计算效率，引入一项特征重构窗口策略，如图 3 所示。该策略允许在提取多尺度特征的同时，使注意力计算具有在线性时间复杂度。具体来说，在不考虑整个特征图的情况下，仅选择一部分窗口进行注意力计算，从而降低计算复杂度。这种方式在处理高分辨率图像时能够节省大量计算资源，通过选择适当的窗口大小和位置，可以在不降低建模质量的情况下，提高模型的计算效率。总的来说，给出的基于重叠式词嵌入和特征重构窗口策略能够更好地处理弱纹理区域和高分辨率图像，同时保持了计算效率，计算公式如下：

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{Q\times FRW(K)^{T}}{\sqrt{d}}\right)FRW(V) \tag{5}$$

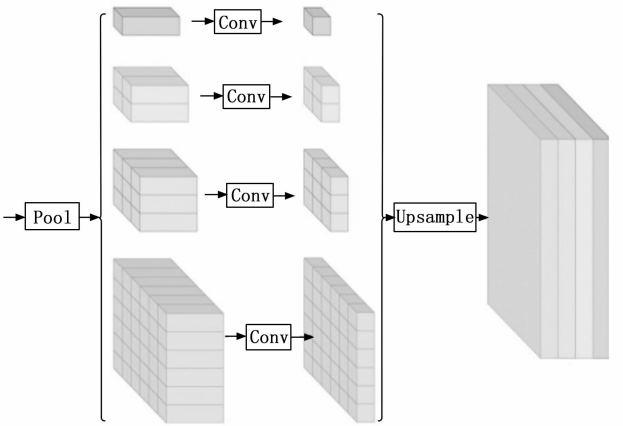

图 3　特征重构窗口

特征重构窗口旨在聚合不同区域的上下文信息，在减少计算复杂度的同时提高网络获取全局信息的能力。这一策略融合了空间信息的提取、Transformer 模型的全局特征学习以及注意力计算的高效性。为确保邻域信息的连续性，将原始高分辨率图像划分为多个重叠块，每个图像块映射为相应数量的词嵌入，引入位置编码来保留图像块的空间位置信息。由于图像块被映射为词嵌入，丢失了原始图像块的空间位置信息。为解决该问题，特征重构窗口引入位置编码，将坐标信息嵌入到词嵌入中，以此保留图像块的空间位置。位置编码使模型能够理解每个块的相对位置，以便更好地捕获全局特征信息。在获得词嵌入后，特征重构窗口进行多尺度的空间聚合，这有助于减小键和值矩阵的尺寸，从而降低注意力计算的复杂度。多尺度的聚合使模型能够在不同尺度上捕获特征，从细节到全局信息都能得到充分考虑，有助于提高网络性能。

经过空间聚合后，通过池化操作将特征合并，并采用卷积操作进行特征重组，该过程有助于将特征信息更好地组织，用于后续的注意力计算。合并池化后的特征，通过卷积操作进行重组，最终转化为可用于注意力计算的词嵌入，将这些词嵌入送入 Transformer 模型，通过多头自注意力[18-19]获取全局信息和长距离依赖，与卷积随着网络加深扩大感受野，在每个阶段均可以提取到全局特征，能尽可能地减少丢失语义信息的情况，为解码器提供更为丰富的全局信息，以此生成高准确度深度图。

以上所述的特征重构窗口[20-21]策略关键点在于充分利用图像的空间信息，将图像分块处理，并引入 Transformer 进行全局特征学习。该策略的优势在于处理高分辨率图像时，仍能保持注意力计算的线性时间复杂度，与传统方法和一些基于词嵌入的方法相比，特征重构窗口为高效深度估计提供了一种新途径，不仅如此，该方法还可在多尺度

条件下提供更为全面的特征信息，从而改善深度估计的性能。

4 实验与分析

4.1 实验设置

实验采用 Pytorch 机器学习库，显卡选用 NVIDIA GTX 3090。为验证所提方法的有效性和不同场景的泛化性，在两个常用的公开数据集 Scene Flow 和 KITTI 上展开实验。KTTTI 数据集是一个广泛用于计算机视觉和自动驾驶研究的数据集，提供了多种类型的传感器数据，包括图像、激光雷达、GPS 和 IMU，以及丰富的标注信息。该数据集用于目标检测、立体视觉、SLAM 和自动驾驶等领域的研究和开发。主要场景有公路，乡村和市区等，为保证实验结果的可比较性，按 Eigen 等人的方法划分数据集，来自 32 个场景的 23 158 张图像作为训练集，652 张来自 29 个不同场景的图像作为测试集，训练时随机裁剪输入图像为 352 像素×704 像素，测试时按 Garg 等人提出的方法做中心裁剪。Scene Flow 为合成的数据集，包含了丰富的图像数据，每个场景都包括 3 个连续帧的图像序列。这些图像序列提供了不同视角下的真实世界场景，用于深度学习模型的训练和评估。此外，数据集还包括了与图像对应的视差地图、光流场、相机参数等附加信息。

采用 3 种标准评估深度估计的 Transformer 架构的性能：

EPE (End-Point-Error)，表示预测值和真实值在视差空间的绝对距离，其中为 *pred* 预测值，*true* 为真实值，计算公式如下：

$$EPE = |\ pred - true\ | \tag{6}$$

3 像素错误 (3*PE*)，表示视差错误大于 3 像素的百分比，其中 *Tr* 表示视差错误大于 3 像素的数量，*L* 表示视差错误像素的数量，计算公式如下：

$$3PE = |\ \frac{Tr}{L}\ | * 100\% \tag{7}$$

遮挡交并比 (OIOU)，用于评估遮挡区域的预测结果，计算公式如下：

$$\mathrm{OIOU} = \frac{A \cap B}{A \cup B} \tag{8}$$

4.2 实验过程及方法

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，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 (Ilya 等，2017)，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 和 λ_2 被设置为 1/3，遮挡概率阈值 θ 设置为 0.05。然而，在混合精度下，训练纯粹的 Transformer 架构并不稳定，模型在少量迭代时会出现损失为空的问题，当注意力分数 $Q^T(\frac{FRW(K)}{\sqrt{d}}$ 超过 16 位浮点数的表示范围上限时，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 $Q^T(\frac{FRW(K)}{\sqrt{d}}$，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。

4.3 Scene Flow 数据集实验结果

合成数据集 Scene Flow 的实验结果如表 1 所示，给出了在 Scene Flow 数据集的实验对比结果。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，且优于其它方法。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。

4.4 KITTI 数据集实验结果

室外数据集 KITTI 的实验结果如表 2 所示，对 KITTI 2015 数据集中的 200 组立体像对进行微调训练，与传统方法相比，本文方法在 KITTI2015 上各指标均得到提升。表 2 中，在前景区域上的平均异常值百分比 D1-fg 指标提升 4%，在背景区域上的平均异常值百分比 D1-bg 指标和整体图像的平均异常值百分比 D1-all 指标也都有显著的提升。

表 1 在 Scene Flow 数据集的对比实验

方法	$D=192$						$D=480$					
	disparity<192			All pixels			disparity<192			All pixels		
	3*PE*	*EPE*	*OIOU*	3*PE*	*EPE*	*OIOU*	3*PE*	*EPE*	*OIOU*	3*PE*	*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
Ours	**0.86**	**0.30**	**0.988**	**0.91**	**0.31**	**0.982**	**0.86**	**0.30**	**0.989**	**0.90**	**0.32**	**0.983**

表 2　KITTI 数据集的对比实验

方法	D1-bg	D1-fg	D1-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
Ours	1.76	3.67 ↑	**2.05**

4.5　弱纹理区域结果分析

目前，大多数公开数据集很少提供对图像中弱纹理程度的定义，本文采用一种基于图像像素聚类的方法来衡量图像的纹理强弱。将每个像素视为一个样本，使用其 RGB 值作为特征维度，利用 K 邻近聚类算法对图像进行聚类，以确定不同像素类别的数量，这个数量可以用来量化图像的纹理强弱程度。在 Scene Flow 测试数据集中，得到了不同类别数目，分别在区间［839，1 500］，［1 500，10 000］和［10 000，15 127］内。这 3 个区间分别代表了“困难”“中等”和“简单”样本。表 3 中，进一步通过使用 EPE 指标进行实验对比，所有方法在“困难”样本上的准确率明显低于“中等”和“简单”样本，这说明弱纹理区域对立体匹配的准确性产生了明显影响。此外，本文方法在 3 种不同样本区间上都获得了较好的结果，而在“困难”样本上的提升尤为显著，表明本文方法对于处理弱纹理区域具有出色的性能，这一优势得益于 Transformer 架构的全局表征学习能力。

表 3　弱纹理区域对比实验

方法	D1-bg	D1-fg	D1-all
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
Ours	**2.66**	**0.40**	**0.23**

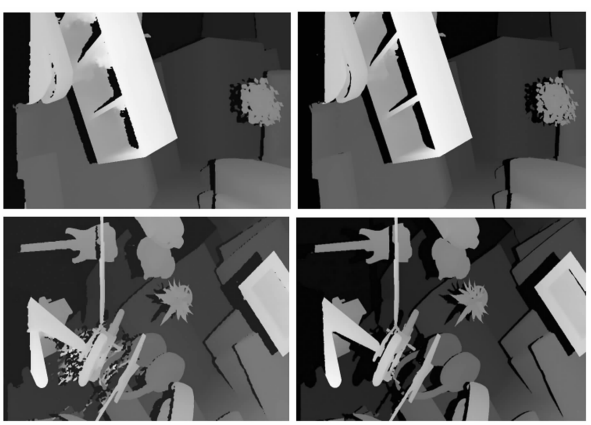

图 4　实验结果对比

由图 4 可见，随着解码器的逐步深入，弱纹理区域和细粒度区域的特征能力得到了明显提升，有助于更好地处理这些具有挑战性的弱纹理区域。

5　结束语

本文提出一种弱纹理目标立体匹配网络，给出一种新的特征提取方法，利用 Transformer 架构，将编码器和解码器结构应用于特征提取器，结合卷积和 Transformer 的优势，利用重叠式词嵌入策略更好地捕获图像中的局部纹理和上下文信息，特别对弱纹理和遮挡区域的深度估计提供显著改进。通过引入特征重构窗口来有效传递信息，减少计算复杂度，并在多个数据集上得到了更准确的深度估计结果。

参考文献：

[1] BADKI A，TROCCOLI A，KIM K，et al. Bi3D：Stereo depth estimation via Binary classifications［C］//Conference on Computer Vision and Pattern Recognition，2020：1597－1605.

[2] 毛静怡，宋余庆，刘　哲. 多尺度深度特征提取的肝脏肿瘤 CT 图像分类［J］. 中国图象图形学报，2021，26（7）：1704－1715.

[3] 戴仁月. 融合 CNN 与视觉 SLAM 的深度图估计技术研究［D］. 上海：上海工程技术大学，2021.

[4] 温　静，安国艳，梁宇栋. 基于 CNN 特征提取和加权深度迁移的单目图像深度估计［J］. 图学学报，2019，40（2）：248－255.

[5] 李　格. 基于深度学习模型的单目图像深度估计［D］. 广州：华南理工大学，2018.

[6] AJINA R，TANKOVI N，IPI I. Multi-task peer-to-peer learning using an encoder-only transformer model［J］. Future Generation Computer Systems，2024，152：170－178.

[7] WANG WE 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.

[8] ZHENG S X，LU J C，ZHAO H S，et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers［C］//Computer Vision and Pattern Recognition，2021，6877－6886.

[9] ZHAO H，SHI J，QI X，et al. Pyramid scene parsing network［C］//2017 IEEE Conference on Computer Vision and Pattern Recognition：CVPR 2017，Honolulu，Hawaii，Institute of Electrical and Electronics Engineers，2017：6230－6239.

[10] TANKOVICH V，HANE C，ZHANG Y，et al. HITNet：hierarchical iterative tile refinement network for real-time stereo matching［C］//Computer Vision and Pattern Recognition，2021：14357－14367.

[11] 管　阳. 室内三维环境重建技术中的立体匹配算法研究与仿真［J］. 电子设计工程，2023，31（19）：186－190.

[12] HUANG Z Y，NORRIS T B，WANG P Q. Es-net：An efficient stereo matching network［EB/OL］.　（2021-01-01）［2023-08-20］. https：//arxiv.org/pdf/2103.03922.pdf.

（下转第 187 页）