计算机视觉中Transformer的应用: 光流估计

张鹏杰

(北京理工大学，计算机学院，3220231268，3220231268@bit.edu.cn）

**摘要**：本文研究了注意力机制在计算机视觉中的应用，特别是光流估计方面。首先，本文简单回顾了Transformer的原理，以及其在计算机视觉领域的应用；随后，本文研究了Transformer与光流估计的结合，探索了已有光流模型FlowFormer的基本框架；最后，本文讨论了注意力机制在光流估计中的独特优势，并展望了其未来的发展方向。

**关键词**：Transformer，计算机视觉，光流估计，FlowFormer

1 引言

注意力机制是一种模拟人类视觉系统的机制，它能够使模型自动学习对输入数据的关注程度，将模型的注意力（计算力）集中在与任务相关的特征上，从而提高模型的效率和准确率。在计算机视觉领域中，注意力机制已经被广泛应用于图像分类[1]、目标检测[2]、人脸识别[3]等任务中。

在计算机视觉领域，光流估计是一项重要的技术，它被广泛应用于运动目标检测、跟踪、三维重建等方面。然而，光流估计面临着诸多挑战，如光照变化、遮挡、复杂背景等。为了解决这些问题，研究者们提出了许多光流估计方法，其中一些方法与注意力机制相结合，取得了较好的结果[4, 5]。

本文研究了注意力机制中Transformer[6]架构在计算机视觉领域，尤其是光流估计领域中的应用。首先，本文简单回顾了Transformer[6]的原理以及其在计算机视觉领域的应用；随后，本文以基于Transformer[6]的光流估计模型FlowFormer[5]为对象，研究了注意力机制与光流估计的结合；最后，本文讨论了注意力机制在光流估计中的独特优势，并展望了其未来的发展方向。

2 计算机视觉中的注意力

2.1 注意力模型的基础框架：Transformer

Transformer[6]是一个完全抛弃RNN、CNN等传统结构，仅用Attention机制实现的模型架构，其在2017年发布并被用于机器翻译任务，取得了很好的性能，这使得注意力机制开始成为研究的热点。

Transformer[6]是一个Seq2Seq的模型，包括编码、解码两部分，编码器将输入序列映射到特征序列，解码器将特征序列解码为输出结果，编码器和解码器都是由若干个注意力层堆叠形成的，如图1所示。

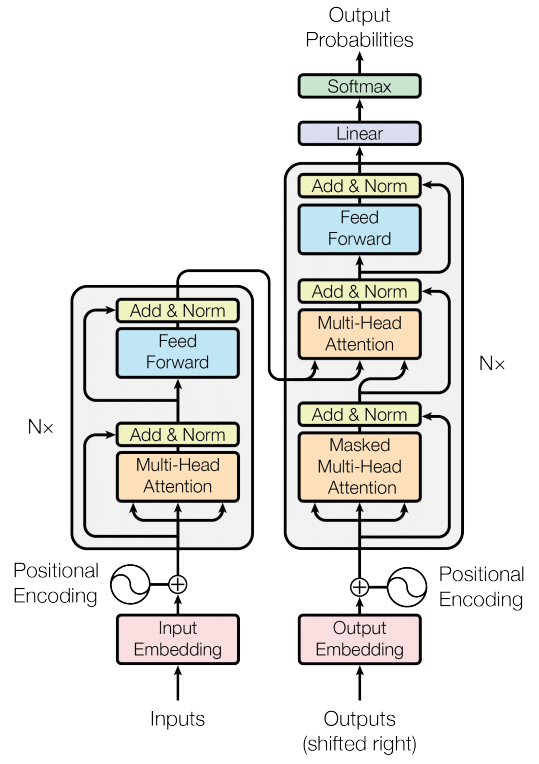


图1 Transformer模型架构[6]

首先模型需要对输入进行词嵌入，以获得个维的Token特征，记为，此外为了编码位置信息，需要额外对Token特征添加位置编码，在原文中作者使用了正余弦函数进行编码。

随后利用编码器对输入进行特征编码，每个编码块都包括多头注意力机制、前馈神经网络两个部分。首先对进行多头自注意力计算，将其进行维度切分后输入到多个注意力头进行计算，对于每个头，利用线性映射获取查询、键和值，并通过注意力机制计算得出特征值：

其中用于将注意力矩阵变为标准正态分布。

随后将特征序列输入前馈神经网络，在原文中前馈网络的结构为两层线性映射加激活函数：

此外，多头自注意力和前馈网络的输出结果需要进行残差、标准化的处理，如图1所示。

对于解码器，其解码块采取与编码块相同的两层结构，所不同的是额外添加了一个注意力块用于从输出序列中获取查询，而键和值通过编码器的中间特征获取。此外，由于编码器每个时刻只能获取当前时刻以前的输出结果，因此需要对输出序列进行Mask，以遮挡尚未获取的结果。

2.2 基于注意力机制的计算机视觉模型：ViT

受到Transformer在自然语言处理领域成功的启发，研究者希望将注意力机制引入计算机视觉领域，提升计算机视觉模型的性能表现，其中首个较为成功的基于注意力的视觉模型是ViT（Vision Transformer）。

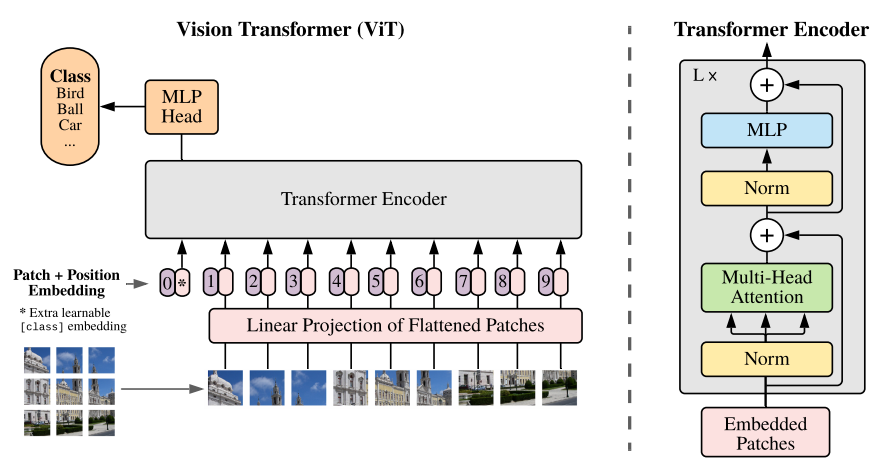


图2 ViT模型架构[1]

ViT的模型架构类似于Bert[7]，即仅利用Transformer的编码块进行特征编码，如图2所示。Transformer的输入是Token序列，因此ViT首先需要将二维图像转化为一维特征序列，具体而言，ViT将图像分为若干Patch块，以的图像为例，假设Patch块尺度设置为，则图像最终生成个Patch块，每个Patch的维度为，将其进行线性映射并拉伸后即可得到的Token序列，由此我们将视觉问题转化为了Seq2Seq的问题。

类似的，ViT同样需要对Token序列进行位置编码，区别于Transformer的是，ViT将位置编码设置为可学习的变量，其编码通过学习获得，由此我们可以获得输入序列：

其中是人为插入的特征，用于输出最终的分类结果。

在Transformer编码块内部，计算单元主要包括多头注意力和MLP两部分，并且结合残差连接、标准化的运算：

编码最终输出特征y用于最终的图像分类，该特征被输入到一个MLP，以计算得到最终的分类结果。

ViT是一个基于Transformer的图像分类模型，是Transformer在CV领域的里程碑成果。实践证明当训练数据足够多时，ViT表现出强于CNN的性能，当然由于Transformer缺乏归纳偏置（inductive bias），当训练数据不够多时，ViT的性能通常比同等规模的ResNets模型要差。

2.3 CNN与Transformer的结合：PVT、Swin

虽然ViT成功将注意力机制引入计算机视觉领域，但是它仅能解决一些输出简单的视觉任务，如图像分类，在目标检测、分割等领域的注意力模型仍有待探索，并且对于多尺度的图像ViT不能很好地进行处理，对高分辨率图像的计算复杂度问题也有待解决。由此，人们提出将传统的CNN架构的思想与Transformer相结合，提出了新的视觉注意力模型，其中最具代表性的两个模型是PVT[8]和Swin[9]。

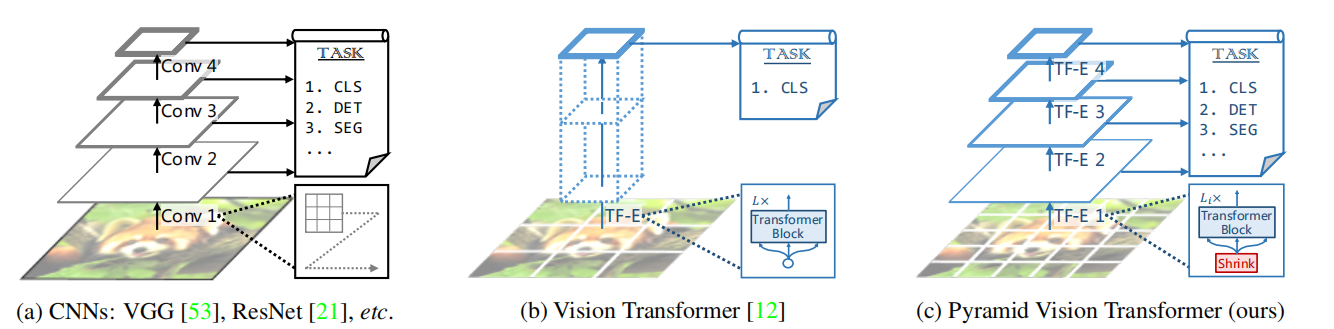


图3 不同模型架构的对比[8]

PVT[8]提出将CNN的金字塔架构与Transformer结合，使模型可以进行各种密集视觉预测任务，如检测、分割等。如图3所示，相比于ViT[1]的低分辨率输出、高计算复杂度、高内存占用，PVT[8]不仅可以对图像进行多粒度划分以输出高分辨率结果，还可以使用金字塔架构降低特征计算的计算量，提高了多种下游视觉任务的性能。

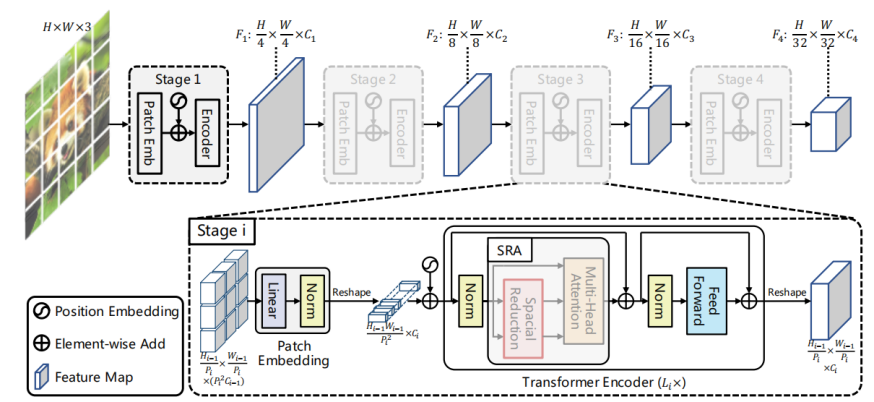


图4 PVT模型架构[8]

PVT[8]的模型架构如图4所示，其基本框架与ViT[1]类似，所不同的是在每层编码块前增加一个对Token的下采样操作，使其形成一个金字塔结构，并且对多头注意力机制进行尺度的调整，以降低计算量。

具体而言，PVT[8]在每个编码块前都执行一次Patch Embedding操作，将特征图的Patch尺度增大、通道数增加，然后重新组织Token块并执行注意力计算，以实现金字塔形式的多尺度运算。并且PVT[8]设计SRA取代普通的多头注意力，将键K和值V的数量减少为原来的，减少了模型的运算量。

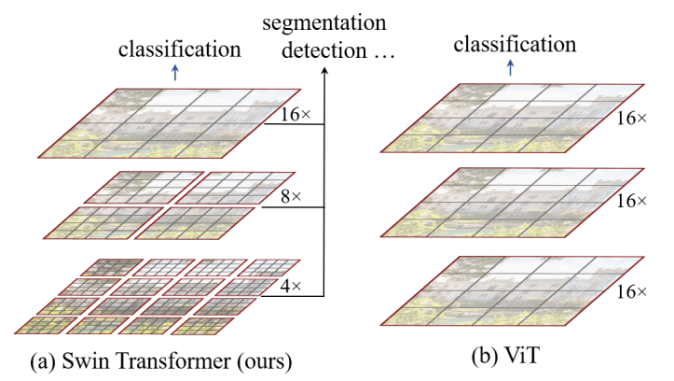


图5 Swin与ViT的模型架构对比[9]

与此同时，Swin[9]也探索了图像多尺度在Transformer中的应用，如图5所示，它将图像划分为不同的窗口，注意力机制只对每个窗口内的Token进行注意力计算，由此降低模型计算量，但这也带来了新的问题，即窗口的视野域较小，且窗口间的Token信息相对独立，这限制了Transformer原有的长程建模能力。

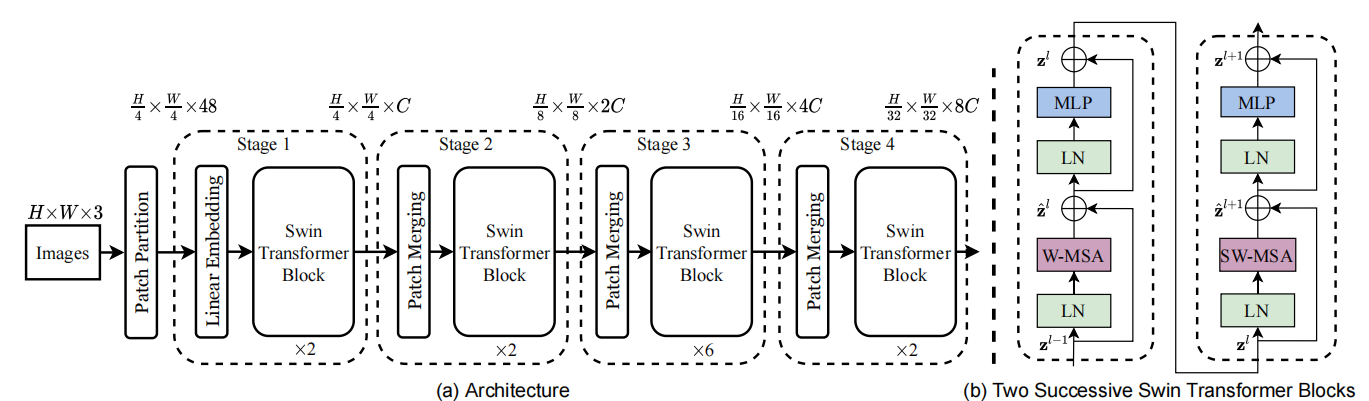


图6 Swin的模型架构[9]

为解决上述问题，Swin[9]设计了两个机制：首先是类似PVT[8]的多尺度下采样，每个Stage设置一个Patch Merging操作，增大Patch的尺度从而增加其感受野；其次是设计SW-MSA配合W-MSA进行窗口间的信息交流，这也是Swin的核心设计。最终的模型架构如图6所示。

2.4 Perceiver与Perceiver IO

Perceiver是基于Transformer的简单改进，其利用Cross Attention，设置一个大小为的Latent array，将Transformer的复杂度从O(MM)降低到O(MN)，从而使得Perceiver的计算量大减，使其可以在多模态融合中以较小的代价完成任务。

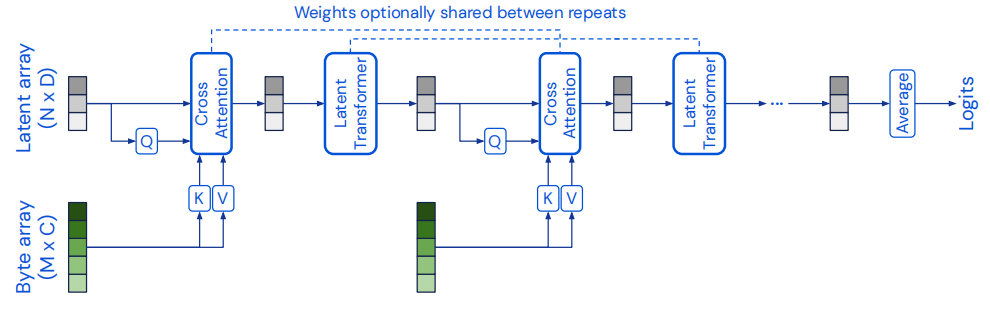


图6 Perceiver的模型架构[10]

Perceiver的模型架构如图6所示，对于的查询Q、的键K和的值V，最终可以计算得到的特征矩阵，即最终输出矩阵的大小由N和C决定，如果减少Q的数目N，就可以降低后续模型计算的复杂度。

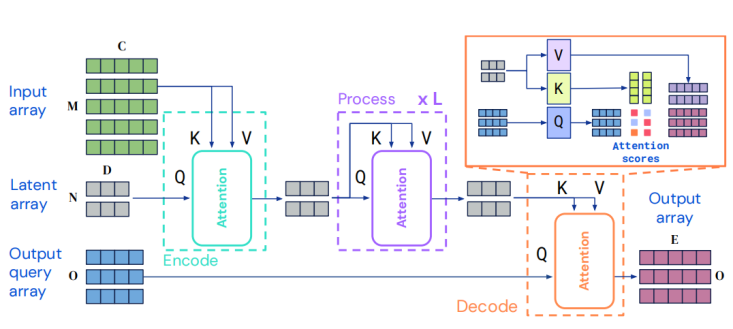


图7 Perceiver IO的模型架构[11]

Perceiver的缺点是其输出较小，只适用于处理简单任务， 因此Perceiver IO在其基础上添加一个Decoder模块，用于输出更复杂的运算结果，模型架构如图7所示。利用Perceiver IO可以处理各种形式的数据，包括文本、图像和音频等，并以此解决不同的多模态任务，例如情感识别、视频分析等。

2.5 应用实例：目标检测模型DERT

面向具体的视觉应用，基于注意力的模型由相应的调整，一种常见的做法是，利用CNN架构作为Backbone提取图像特征，然后在特征域上进行注意力的计算，通常是直接接入Transformer的编码、解码层，最后利用一个前馈神经网络层输出最终的预测结果，一个典型的案例就是目标检测的模型DERT[2]。

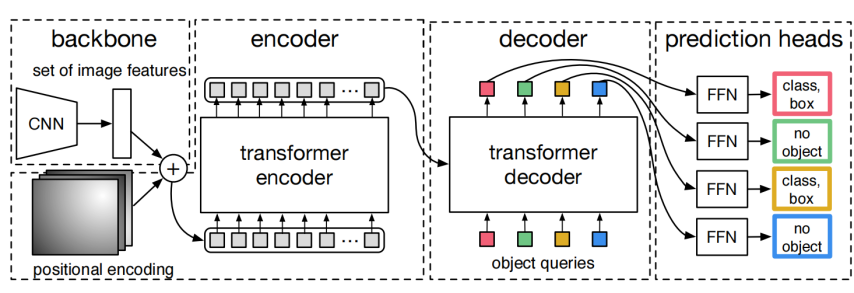


图8 DERT的模型架构[2]

如图8所示，DERT将Transformer架构与CNN架构相结合，利用注意力机制提升了模型在目标检测任务上的最终性能。

3 基于注意力机制的光流估计模型：FlowFormer[5]

光流估计是计算机视觉的重要研究领域，在许多下游视频任务中发挥重要作用。光流是空间运动物体在成像平面上像素运动的瞬时速度，其产生源于场景中前景目标与相机的相对运动。光流能提供帧间像素的空间对应关系，为计算机理解像素级的图像相似性提供依据，其在运动检测、视频压缩、目标分割、帧间插值等任务中发挥着重要作用。因此，研究更加高效准确的光流估计算法是计算机视觉领域的重要课题。

光流估计算法迭代发展至今，从最早的基于梯度的Lucas-Kanade[12]算法、Horn-Schunck[13]算法，到结合深度神经网络的FlowNet[14]、PWCNet[15]、RAFT[16]，光流估计算法的性能得到极大提升。然而，光流估计仍然面临着许多挑战，例如对噪声和遮挡的鲁棒性问题、计算复杂度高等。

Transformer在计算机视觉领域大获成功，因此将Transformer架构的注意力机制引入光流估计模型成为一个十分自然的想法。Perceiver IO[11]中将通用的注意力架构应用于光流估计任务，GMA[4]利用Transformer的长程建模能力重点解决光流估计的遮挡问题，它们都是注意力机制与光流估计结合的探索。

FlowFormer[5]是当前一种最先进的光流估计模型，其将注意力机制适应性地融入光流估计的基本框架RAFT[16]，实现了模型性能的全面提升，是Transformer与光流估计结合的典范，下面本文将具体探索其算法的基本流程。

3.1 基本定义与模型架构

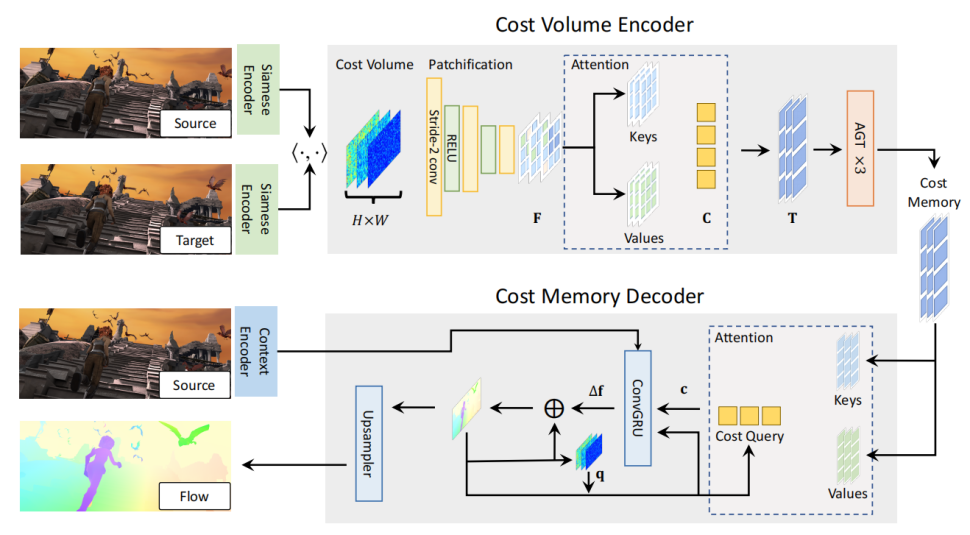


图9 FlowFormer的模型架构[5]

给定相邻帧输入，光流估计的目标是输出一个像素级的位移矩阵，将源图像上的像素坐标映射到目标图像上的像素坐标，通常的深度学习的光流估计方法基于CNN架构实现，而FlowFormer将注意力机制引入光流模型，实现了模型结构的创新。

FlowFormer的基本架构如图9所示，主要包括编码、解码两部分，编码部分利用CNN特征和注意力机制实现对图像的相似度编码，解码部分利用注意力查询机制和GRU实现对光流的迭代解码估计。

3.2 基于Backbone的图像特征提取

与DERT[2]的结构类似，FlowFormer首先利用一个基础骨干网络提取图像的特征，从大小为的RGB图像中提取的特征图。

提取特征图后，参考RAFT[16]的操作，构建四维的代价矩阵（Cost Volume），即利用点积相似度构造出的特征矩阵。

在后续计算中，可以该特征矩阵将其视作大小为的维的特征图。

3.3 特征编码

在利用注意力机制前，我们需要首先将特征转化为Token序列，因此我们需要执行以下两个步骤：分块化（Patchification）、词嵌入（Embedding）。

分块化。参考已有的视觉Transformer方法如ViT[1]，利用卷积操作对输入的特征图进行分块，分块对象为HW个矩阵，设置每个Patch块大小为，最终得到分块后HW个特征图。

词嵌入。为生成Token序列，FlowFormer设置一个隐藏码字作为注意力层的查询，利用线性映射从中提取键和值，利用以下公式计算得到Token值，式中PE为位置编码：

因此我们可以得到最终的Token序列。

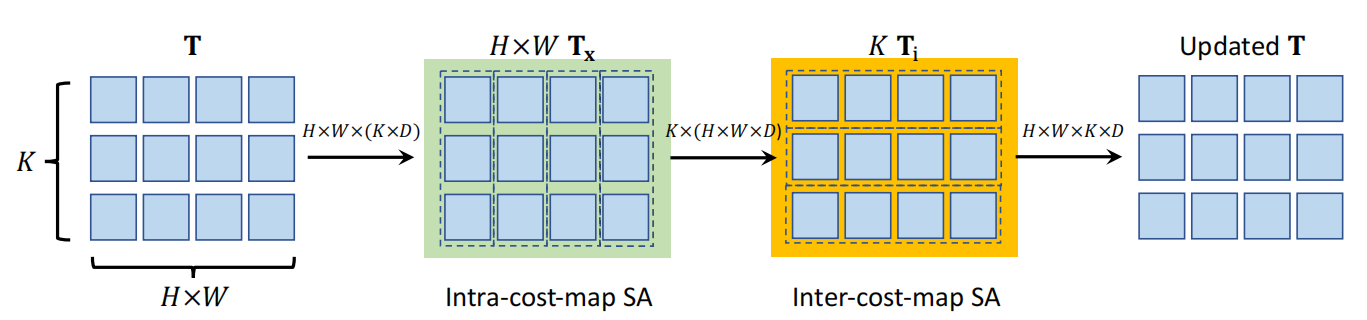


图10 组交替的Transformer层（alternate-group transformer layer, AGT）[5]

利用该Token序列直接进行自注意力编码是一种可行的方案，但是个码字意味着十分庞大的计算量，因此FlowFormer提出使用组交替的Transformer层（alternate-group transformer layer, AGT）进行Token编码，如图10所示。

首先在每个的像素组内执行自注意力计算：

然后将整个分为K组，每个组包括个D维Token，该运算利用Twins[17]中提出的空间可分离的自注意力机制（SS-SelfAttention）进行计算：

经过N次AGT编码，最终可以得到编码后的Token序列。

3.4 特征解码

基于RAFT[16]的基本框架，FlowFormer[5]的特征解码以ConvGRU单元为中心，利用时序单元的迭代优化光流的预测结果，迭代过程中的输入特征通过特征查询获得，与RAFT[16]不同的是，FlowFormer利用交叉注意力机制实现特征查询操作。

给定当前光流，可以得到像素偏移后的坐标，根据相邻帧的相似矩阵，可以获得的局部窗口块，根据可以计算得到查询；同时从中计算得到键和值，利用注意力机制计算得到特征：

获取查询特征后，利用ConvGRU单元计算得到光流更新，其中为上下文特征：

更新光流后完成一次特征解码的迭代，经过若干次迭代后，获得最终的解码光流结果。

4 总结

注意力机制尤其是基于Transformer的注意力模型架构在当前人工智能的发展中占据了一席之地，自Transformer首次被用于自然语言处理，至Transformer在计算机视觉中大放异彩，注意力机制越来越受到研究者的广泛认可，而基于Transformer的各种模型凭借其优秀的性能也越来越被广泛应用。

光流估计作为计算机视觉领域的重要研究课题，近年来有许多研究者探索了其与注意力机制的结合，并取得了一定的成就。注意力机制在光流估计中的优势主要体现在其长程建模能力和选择性的编码能力，这使得光流估计的感受野更广泛、准确，FlowFormer合理地将注意力机制引入光流估计的基本框架，通过发挥注意力机制的优势将光流估计算法性能推上了新的高度，是值得肯定的成就。

当然，可以看到注意力机制在光流估计乃至计算机视觉领域的应用仍有待继续探索，以FlowFormer为例，网络复杂的结构是否可以简化、能否依托注意力机制提出新的基础架构、能否提升模型的可解释性等都是可以研究探索的问题。

因此可以预见，基于注意力机制的光流估计模型将会受到更多的关注和研究，未来也一定会有新的、性能更强的、基于注意力机制的光流估计模型架构出现。

参考文献

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:201011929, 2020.

[2] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers; proceedings of the European conference on computer vision, F, 2020 [C]. Springer.

[3] ZHONG Y, DENG W. Face transformer for recognition [J]. arXiv preprint arXiv:210314803, 2021.

[4] JIANG S, CAMPBELL D, LU Y, et al. Learning to estimate hidden motions with global motion aggregation; proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2021 [C].

[5] HUANG Z, SHI X, ZHANG C, et al. Flowformer: A transformer architecture for optical flow; proceedings of the European Conference on Computer Vision, F, 2022 [C]. Springer.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[7] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:181004805, 2018.

[8] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, 2021 [C].

[9] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, 2021 [C].

[10] JAEGLE A, GIMENO F, BROCK A, et al. Perceiver: General perception with iterative attention; proceedings of the International conference on machine learning, F, 2021 [C]. PMLR.

[11] JAEGLE A, BORGEAUD S, ALAYRAC J-B, et al. Perceiver io: A general architecture for structured inputs & outputs [J]. arXiv preprint arXiv:210714795, 2021.

[12] HORN B K, SCHUNCK B G. Determining optical flow [J]. Artificial intelligence, 1981, 17(1-3): 185-203.

[13] LUCAS B D, KANADE T. An iterative image registration technique with an application to stereo vision; proceedings of the IJCAI'81: 7th international joint conference on Artificial intelligence, F, 1981 [C].

[14] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutional networks; proceedings of the Proceedings of the IEEE international conference on computer vision, F, 2015 [C].

[15] SUN D, YANG X, LIU M-Y, et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].

[16] TEED Z, DENG J. Raft: Recurrent all-pairs field transforms for optical flow; proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, F, 2020 [C]. Springer.

[17] CHU X, TIAN Z, WANG Y, et al. Twins: Revisiting spatial attention design in vision transformers [J]. arXiv preprint arXiv:210413840, 2021, 2(3).