Computer Science and Application 计算机科学与应用, 2024, 14(4), 1-12
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144071

Hans汉斯

# 跨模态时空交叉注意力下机器人抓取滑动检测

**谷　鑫**

广东工业大学自动化学院，广东　广州

收稿日期：2024年3月3日；录用日期：2024年4月2日；发布日期：2024年4月9日

## 摘　要

**在机器人领域，滑动检测是一个关键的任务。机器人需要利用多模态信息进行特征提取、信息融合交互与灵巧操作。为此，提出一个基于跨模态时空交叉注意力机制的多模态融合模型，用于滑动检测。该模型利用时空注意力学习多模态传感器反馈的物理特征，将学习到的视触觉时空特征通过跨模态交叉注意力进行交互融合。最后，通过多层感知机(MLP)预测滑动检测结果。使用7自由度XArm机械臂、D455摄像头和XELA触觉传感器进行数据采集、模型训练和验证。结果表明，该模型的滑动检测准确率高达97.8%，所提出的模型在可靠、顺利执行机器人抓取任务方面具有较高的研究和应用价值。**

## 关键词

**跨模态交叉注意力，时空注意力，多模态融合，滑动检测**

# Robot Grasping Slip Detection Based on Cross-Modal Spatiotemporal Cross-Attention Mechanism

**Xin Gu**

School of Automation, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 3rd, 2024; accepted: Apr. 2nd, 2024; published: Apr. 9th, 2024

## Abstract

**In the field of robotics, slip detection is a crucial task. Robots need to utilize multimodal information for feature extraction, information fusion interaction, and dexterous manipulation. For this, a multimodal fusion model based on cross-modal spatiotemporal attention mechanism is proposed for slip detection. The model uses spatiotemporal attention to learn the physical features reflected by multimodal sensor feedback, and the learned visuotactile spatiotemporal features are interac-**

文章引用: 谷鑫. 跨模态时空交叉注意力下机器人抓取滑动检测[J]. 计算机科学与应用, 2024, 14(4): 1-12.
DOI: 10.12677/csa.2024.144071

**tively fused through cross-modal attention. Finally, slip detection results are predicted using a Multi-layer perceptron (MLP). Data collection, model training, and validation are carried out using a 7-DOF XArm robotic arm, a D455 camera, and XELA tactile sensors. The results indicate that the slip detection accuracy of this model reaches up to 97.8%, demonstrating the high research and practical value of the proposed model in ensuring reliable and smooth execution of robotic grasping tasks.**

## Keywords

**Cross-Modal Cross Attention, Spatiotemporal Attention, Multi-Modal Fusion, Slip Detection**



## 1. 引言

多模态集成感知能力有利于提高感知精度、减少模糊性，是智能体认知世界的必要条件。人们在抓取物体时，通过视触力觉融合，可事先估计并实时调整抓取力，以防止物体滑动或过度变形。当滑动时人们会自觉加大力度并短暂轻微举升以调整姿态，直至抓稳[1] [2]。视觉可提供目标对象的外观、形状和其他可见的、全局的、不太精确的动/静态空间特征，而触力觉则可以提供更加精确的诸如纹理、粗糙度、接触强度等局部细节信息，表征为动态的时间模态[2] [3]，可对视觉起到重要的补充作用。因此，多模态感知对于增强机器人感知能力、提高抓取质量具有重要的研究价值。滑动检测在机器人操作中起着至关重要的作用，滑动检测任务中对于多模态信息的处理也是一个具有挑战性的难题。滑动是一种常见的失接触状态，发生在抓握力不足或抓握策略不当。检测滑动现象和初始滑动状态可以帮助机器人自动调整抓取力并选择合适的动作规划。

近年来，相关学者已设计开发出多种触觉传感器和相应的检测方法，用于滑动检测[4]。在视触觉多模态融合的滑动检测任务中，LI 等[1]使用卷积神经网络(convolutional neural network, CNN)和长短期记忆网络(long short-term memory networks, LSTM)用于机器人视触觉融合下的滑动检测。YAN，黄兆基等[5] [6]将时序卷积网络(temporal convolutional network, TCN)和多尺度时序卷积网络(multi-scale temporal convolution network, MS-TCN)引入滑动检测任务，用来提取视触觉的时序特征。上述工作都取得了不错的效果，但是仅局限于对特定模态特征的提取，没有引入模态之间的信息交互。如何更加有效的利用多模态信息进行特征融合，实现精准而高效的滑动检测，以防止抓取物体时发生滑动或者过度形变，具有重要的研究意义和研究价值。

注意力机制因其更关注输入数据中的关键信息在深度学习方面具有显著的优势。注意力机制通过对输入的不同部分分配不同的权重，可以自动学习到所输入数据中的重要特征，并集中关注对任务具有决定性影响的部分，已在自然语言处理等领域得到广泛应用[7]。注意力机制是一种在序列数据中学习元素之间关系的技术，它通过将每个元素与其他元素进行关联建模输入序列。这种注意力机制已被证明在机器翻译等任务中取得了显著的性能提升[8]。注意力机制能够有效处理输入相互作用并产生准确的注意力输出。通过查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 之间的交互计算，动态地计算每个输入的注意权重，从而使模型能够聚焦于输入数据的关键信息。这种输入相互作用的方式赋予模型更强大的表达能力，并且使模型能够更好地捕捉上下文的关联性。最终，通过注意力机制的输出，模型能够提供丰富、准确的信息表示。CUI 等

[9]通过引入注意力机制，用于学习有效的视触觉融合特征以预测抓取结果。该方法虽然采用了注意力机制，但还是通过预训练 CNN 网络的方式提取特征，并未采用性能更佳的端到端的基于注意力模型的设计。

最近，研究人员通过将图像分块方式设计了视觉 Transformer (ViT) [10]，并在图像分类任务中取得了突破性进展。BERTASIUS [11]，ARNAB [12]等，进一步将 ViT 扩展到视频分析领域，并考虑了时空关系的重要性。CAO [13]和 KIM [14]等，将 Transformer 编码器与 CNN 网络相结合，使用 CNN 网络提取每帧的特征，再通过 Transformer 编码器对这些特征进行处理。Transformer 已在有关图像和文本的多模态学习方面得到广泛应用，并取得非常好的性能[15] [16]。然而，机器人抓取检测涉及更多的动态交互，与常见的图像学习任务有所不同，它更关注捕捉跨模态、跨时空的交互特征和提取/融合两种模态特征的能力。BERTASIUS 等[11]所采用的时空注意力分解策略使得将时空注意力模型引入机器人任务成为可能。因为联合的时空注意力模型的参数量巨大，容易导致 GPU 内存溢出，训练成本太高。相比之下，这种分解策略的时空注意力模型为无法访问数百个 GPU 的实验室提供了更高效的替代方案。基于时空注意力分解策略的 Transformer 模型在几类动作识别基准任务上均有优秀表现[11]，针对 Kinetics-400 数据集所取得的最佳准确率为 78%。CUI 等[17]提出使用特定模态的 Transformer 结构和跨模态的 Transformer 结构分别用于提取模态特定特征和跨模态特征。结果表明，这种跨模态的方法优于传统融合方法。

目前，将触觉和视觉图像引入到机器人抓取任务中，并采用端到端的 Transformer 模型方面的研究还鲜有报道。本文通过对 Transformer 结构的研究，利用不同的注意力模型对机器人抓取领域的视触觉多模态架构进行优化设计。采用时空自注意力分解策略进行视触觉数据的时空特征提取。提出一种跨模态的时空交叉注意力分解策略，用于视触觉多模态特征融合，从而提高机器人抓取物体时的滑动检测性能。

## 2. 多模态感知模型

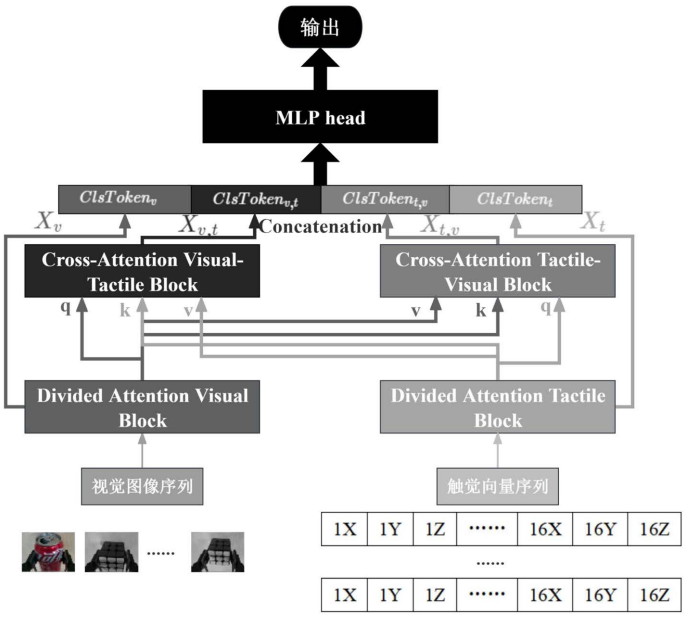


**Figure 1.** The structure of multi-modal perception model based on spatio-temporal attention
**图 1.** 基于时空注意力的多态模感知模型结构

本文所提出视触觉多模态感知模型 CrossTSformer 结构如图 1 所示，视觉和触觉两种模态作为模型的输入。其中，视觉输入序列 $x_v^T$ 和触觉输入序列 $x_t^T$ 分别经过视觉时空自注意力分解模块(Divided Attention

Visual Block, DAVB)和触觉时空自注意力分解模块(Divided Attention Tactile Block, DATB)提取各自的模态时空特征，得到视觉和触觉时空特征 $X_v$ 和 $X_t$。将视觉特征 $X_v$ 作为视觉跨模态交叉注意力模块(Cross-Attention Vistal-Tactile Block, CAVTB)的查询输入，触觉特征 $X_t$ 作为视觉跨模态交叉注意力模块(CAVTB)的键值输入，得到视觉跨模态特征 $X_{v,t}$。对应的将触觉特征 $X_t$ 作为触觉跨模态交叉注意力模块(Cross-Attention Tactile-Vistal Block, CATVB)的查询输入，视觉特征 $X_v$ 作为触觉跨模态交叉注意力模块(CATVB)的键值输入，得到触觉跨模态特征 $X_{t,v}$。最后提取视觉和触觉跨模态特征的分类标签 $X_{v,t}^{ClsToken}$ 和 $X_{t,v}^{ClsToken}$，以及视觉和触觉特征的分类标签 $X_v^{ClsToken}$ 和 $X_t^{ClsToken}$。将这四个分类标签向量在嵌入特征维度上进行拼接，得到最终的特征 $F_{v,t}$，再将其输入至 *MLP* 分类头，得到一个二分类结果 $y$。

## 2.1. 多模态时空特征提取

多模态数据通常包含时序信息和空间信息，为了更好地利用这些数据的特点，需要考虑时序和空间的关系。时序特征和空间特征在其特征表示和特征交互方式上存在差异，为了充分利用不同特征类型的特点，针对不同的特征类型设计不同的注意力模块，以更好地建模它们之间的关系。采用分离的时空注意力机制可以显著降低模型的参数量和计算复杂度。通过分别对时序和空间特征进行建模，避免了同时处理时序和空间信息时的冗余计算。还可以根据具体的任务和多模态数据的特点，对时间和空间注意力的权重分配进行灵活调整。该分离机制使得模型的注意力权重更容易被解释和分析，模型也具有更好的灵活性和可解释性。为此，本文视触觉多模态的时空特征提取采用的是分离的时空自注意力机制。

如图 1 所示，将视觉序列 $X \in \mathbb{R}^{H\times W\times 3\times F}$ 作为输入，每一帧分解为 $N$ 个不重叠图像块，每个图像块大小为 $P\times P$。$N$ 个图像块就可包含整个画面，即 $N = HW/P^2$。图像块 $x_{(p,t)} \in \mathbb{R}^{3P^2}$，$p=1,\cdots,N$ 表示空间位置，$t=1,\cdots,F$ 表示对帧的索引，即时间位置。通过一个可学习矩阵 $\boldsymbol{E} \in R^{D\times 3P^2}$，将每个图像块 $x_{(p,t)} \in \mathbb{R}^{3P^2}$ 线性映射到嵌入向量 $\boldsymbol{z}_{(p,t)}^{(0)} \in \mathbb{R}^D$，作为分离的时空自注意力块输入。

$$\boldsymbol{z}_{(p,t)}^{(0)} = \boldsymbol{E}x_{(p,t)} + \boldsymbol{e}_{(p,t)}^{pos} \tag{1}$$

其中，$\boldsymbol{e}_{(p,t)}^{pos} \in \mathbb{R}^D$ 代表一个可学习的位置嵌入，用于编码每个图像块的时空位置。在序列的第一个位置添加一个特殊的可学习向量 $\boldsymbol{z}_{(0,0)}^{(0)} \in \mathbb{R}^D$，表示分类标签。

对于每个编码块 $\ell$，从前一个编码块的嵌入表示 $\boldsymbol{z}_{(p,t)}^{(\ell-1)}$ 中计算每个图像块的查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 向量：

$$\boldsymbol{q}_{(p,t)}^{(\ell,\alpha)} = \boldsymbol{W}_Q^{(\ell,\alpha)} LN\left(\boldsymbol{z}_{(p,t)}^{(\ell-1)}\right) \in \mathbb{R}^{D_h} \tag{2}$$

$$\boldsymbol{k}_{(p,t)}^{(\ell,\alpha)} = \boldsymbol{W}_K^{(\ell,\alpha)} LN\left(\boldsymbol{z}_{(p,t)}^{(\ell-1)}\right) \in \mathbb{R}^{D_h} \tag{3}$$

$$\boldsymbol{v}_{(p,t)}^{(\ell,\alpha)} = \boldsymbol{W}_V^{(\ell,\alpha)} LN\left(\boldsymbol{z}_{(p,t)}^{(\ell-1)}\right) \in \mathbb{R}^{D_h} \tag{4}$$

其中，$LN(\ )$表示 LayerNorm，$\alpha=1,\cdots,A$ 为多个注意力头的索引，$A$ 表示注意力头的总数。每个注意力头的的维度为 $D_h = D/A$。

自注意力权重通过点乘计算，查询图像块$(p,t)$的自注意力权重 $\alpha_{(p,t)}^{(\ell,a)} \in \mathbb{R}^{NF+1}$ 由式(5)得出：

$$\alpha_{(p,t)}^{(\ell,a)} = SM\left(\frac{\boldsymbol{q}_{(p,t)}^{(\ell,\alpha)T}}{\sqrt{D_h}} \cdot \left[\boldsymbol{k}_{(0,0)}^{(\ell,\alpha)} \left\{\boldsymbol{k}_{(p',t')}^{(\ell,\alpha)}\right\}_{\substack{p'=1,\cdots,N \\ t'=1,\cdots,F}}\right]\right) \tag{5}$$

其中，*SM* 表示 softmax 激活函数。

当注意力只在一个维度上计算时，即只在空间上或时间上，计算量会大大减少。例如，在空间注意力的情况下，只使用 $N+1$ 个查询–键进行比较，并且只使用与查询来自同一帧的键：

$$\alpha_{(p,t)}^{(\ell,a)space}=SM\left(\frac{\boldsymbol{q}_{(p,t)}^{(\ell,\alpha)T}}{\sqrt{D_h}}\cdot\left[\boldsymbol{k}_{(0,0)}^{(\ell,\alpha)}\left\{\boldsymbol{k}_{(p',t')}^{(\ell,\alpha)}\right\}_{p'=1,\cdots,N}\right]\right) \tag{6}$$

编码块 $\ell$ 的编码 $\boldsymbol{z}_{(p,t)}^{(\ell)}$ 通过使用每个注意力头的自注意力系数计算值向量的加权和得到：

$$s_{(p,t)}^{(\ell,a)}=\alpha_{(p,t),(0,0)}^{(\ell,a)}v_{(0,0)}^{(\ell,a)}+\sum_{p'=1}^{N}\sum_{p'=1}^{N}\alpha_{(p,t),(p',t')}^{(\ell,a)}v_{(p',t')}^{(\ell,a)} \tag{7}$$

然后，将这些向量串联，并使用可学习矩阵 $\boldsymbol{W_O}$ 投影到与输入维度相同的矩阵，通过 *MLP*，每次操作后使用残差连接：

$$\boldsymbol{z}'^{(\ell)}_{(p,t)}=\boldsymbol{W}_o\begin{bmatrix}\boldsymbol{s}_{(p,t)}^{(\ell,1)}\\ \vdots\\ \boldsymbol{s}_{(p,t)}^{(\ell,A)}\end{bmatrix}+\boldsymbol{z}_{(p,t)}^{(\ell-1)} \tag{8}$$

$$\boldsymbol{z}_{(p,t)}^{(\ell)}=MLP\left(LN\left(\boldsymbol{z}'^{(\ell)}_{(p,t)}\right)\right)+\boldsymbol{z}'^{(\ell)}_{(p,t)} \tag{9}$$

时空注意力的分解就可以通过将式(5)的时空注意力替换为式(10)中相同位置不同帧的时间注意力和式(6)中每一帧内的空间注意力。其中，时间注意力和空间注意力是单独分别计算。

在每个编码块内，比较每个图像块$(p,t)$与其他帧中相同空间位置的所有图像块来计算时间注意力：

$$\alpha_{(p,t)}^{(\ell,a)time}=SM\left(\frac{q_{(p,t)}^{(\ell,\alpha)T}}{\sqrt{D_h}}\cdot\left[\boldsymbol{k}_{(0,0)}^{(\ell,\alpha)}\left\{\boldsymbol{k}_{(p',t')}^{(\ell,\alpha)}\right\}_{t'=1,\cdots,F}\right]\right) \tag{10}$$

时间注意力通过式(7)和式(8)得到编码 $\boldsymbol{z}'^{(\ell)time}_{(p,t)}$，反馈给空间注意力进行计算，即新的查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 向量由 $\boldsymbol{z}'^{(\ell)time}_{(p,t)}$ 得出。然后，用式(6)~(8)计算空间注意力。最后，得到向量 $\boldsymbol{z}'^{(\ell)space}_{(p,t)}$ 传递给式(9)的 *MLP*，计算出视觉的时空特征 $\boldsymbol{z}_{(p,t)}^{(\ell)visual}$。对于此注意力模型，在时间和空间上分别学习不同的查询、键和值矩阵：

$$\left\{\boldsymbol{W}_{Qtime}^{(\ell,\alpha)},\boldsymbol{W}_{Ktime}^{(\ell,\alpha)},\boldsymbol{W}_{Vtime}^{(\ell,\alpha)}\right\}\left\{\boldsymbol{W}_{Qspace}^{(\ell,\alpha)},\boldsymbol{W}_{Kspace}^{(\ell,\alpha)},\boldsymbol{W}_{Vspace}^{(\ell,\alpha)}\right\}$$

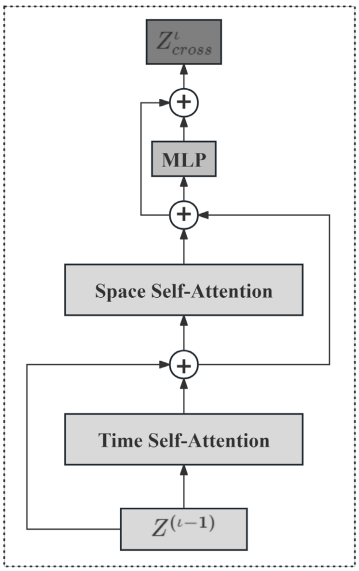


**Figure 2.** The separated spatial-temporal self-attention block
**图 2.** 分离的时空自注意力块

本文所采用的分离式时空注意力整体结构如图 2 所示，在时间和空间维度上按顺序处理数据。对于每一个自注意力块，先计算同一空间位置的时间自注意力，再计算同一帧的空间自注意力。

触觉数据通过 XELA 触觉传感器进行采集。该传感器以 4 × 4 的形式分布着 16 个感应点，每个感应点可以检测 3D 力触觉数据。其中，Z 方向是垂直于传感器表面的法向力，X 和 Y 方向是平行与传感器表面两个方向的切向力。当物体沿某一轴方向移动时，在另一轴方向检测到的力触觉数据变化相对较小。当物体被稳定抓取时，X 和 Y 方向的力触觉数据都保持相对稳定。将每个方向的触觉数据拼接在一起，得到的触觉序列为 $X \in \mathbb{R}^{4\times4\times3\times F}$。类似于视觉数据的处理方式，将触觉序列映射为嵌入向量 $\boldsymbol{z}_{(p,t)}^{(0)} \in \mathbb{R}^{D}$，并通过分离的时空自注意力块得到触觉的时空注意力特征 $\boldsymbol{z}_{(p,t)}^{(\ell)tactile}$。

## 2.2. 多模态特征融合

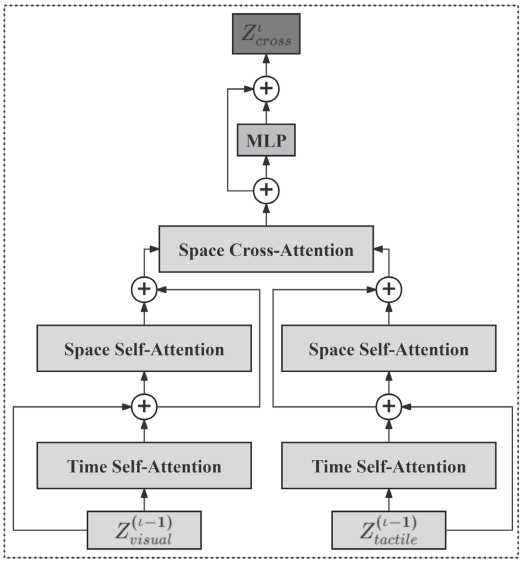


**Figure 3.** The cross-modal Spatial-Temporal Cross-Attention block
**图 3.** 跨模态时空交叉注意力块

机器人在执行抓取任务时，对于不同环境和不同物体所关联的视觉和触觉的重要程度并不相同，如何确定该程度显得至关重要。视觉信息帮助机器人识别物体的外观和位置，触觉信息则提供关于物体形状、材质和力反馈等信息。通过视觉和触觉感知信息的融合，机器人能够更好地理解不同物体的特征。受此启发，本文引入交叉注意力进行跨模态特征交互，使多模态感知模型能够学习到视触觉跨模态的交互信息。所设计的跨模态时空交叉注意力整体结构如图 3 所示。

将网络前一层所提取的视触觉时空特征 $\boldsymbol{z}_{(p,t)}^{(\ell-1)visual}$ 和 $\boldsymbol{z}_{(p,t)}^{(\ell-1)tactile}$ 作为跨模态时空交叉注意力的输入。先分别通过时间自注意力(Time Self-Attention, TSA)和空间自注意力(Space Self-Attention, SSA)计算得到当前层的视触觉时空特征 $\boldsymbol{z}_{(p,t)}^{(\ell)visual}$ 和 $\boldsymbol{z}_{(p,t)}^{(\ell)tactile}$，将其作为空间交叉注意力(Space Cross-Attention, SCA)的输入，最终通过 MLP 层得到当前层的跨模态特征 $\boldsymbol{z}_{(p,t)}^{(\ell)cross}$。视触觉时空特征跨模态交互具体过程如下：

计算查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 向量：

$$\boldsymbol{q}_{(p,t)}^{(\ell,\alpha)vusal} = \boldsymbol{W}_{Q}^{(\ell,\alpha)} LN\left(\boldsymbol{z}_{(p,t)}^{(\ell-1)vusal}\right) \in \mathbb{R}^{D_h} \tag{11}$$

$$\boldsymbol{k}_{(p,t)}^{(\ell,\alpha)tactile} = \boldsymbol{W}_{K}^{(\ell,\alpha)} LN\left(\boldsymbol{z}_{(p,t)}^{(\ell-1)tactile}\right) \in \mathbb{R}^{D_h} \tag{12}$$

$$\boldsymbol{v}_{(p,t)}^{(\ell,\alpha)tactile} = \boldsymbol{W}_{V}^{(\ell,\alpha)} LN\left(\boldsymbol{z}_{(p,t)}^{(\ell-1)tactile}\right) \in \mathbb{R}^{D_h} \tag{13}$$

通过点乘计算交叉注意力权重：

$$\alpha_{(p,t)}^{(\ell,a)cross}=SM\left(\frac{\boldsymbol{q}_{(p,t)}^{(\ell,\alpha)vT}}{\sqrt{D_h}}\cdot\left[\boldsymbol{k}_{(0,0)}^{(\ell,\alpha)}\left\{\boldsymbol{k}_{(p',t)}^{(\ell,\alpha)tactile}\right\}_{p'=1,\cdots,N}\right]\right) \tag{14}$$

由交叉注意力权重计算与值向量的加权和得到：

$$\boldsymbol{s}_{(p,t)}^{(\ell,a)cross}=\sum_{p'=0}^{N}\sum_{t'=0}^{N}\alpha_{(p,t),(p',t')}^{(\ell,a)cross}\boldsymbol{v}_{(p',t')}^{(\ell,a)tactile} \tag{15}$$

使用可学习矩阵 $\boldsymbol{W_O}$ 投影到与输入维度相同的矩阵，然后进入 *MLP*，并且每次计算都使用残差连接：

$$\boldsymbol{z}'^{(\ell)cross}_{(p,t)}=\boldsymbol{W_o}\begin{bmatrix}\boldsymbol{s}_{(p,t)}^{(\ell,1)cross}\\ \vdots\\ \boldsymbol{s}_{(p,t)}^{(\ell,A)cross}\end{bmatrix}+\boldsymbol{z}_{(p,t)}^{(\ell-1)cross} \tag{16}$$

$$\boldsymbol{z}_{(p,t)}^{(\ell)cross}=MLP\left(LN\left(\boldsymbol{z}'^{(\ell)cross}_{(p,t)}\right)\right)+\boldsymbol{z}'^{(\ell)cross}_{(p,t)} \tag{17}$$

最后，从视觉和触觉跨模态特征中分别提取分类标签 $\boldsymbol{z}_{(0,0)}^{(0)cross\_vusal}$ 和 $\boldsymbol{z}_{(0,0)}^{(0)cross\_tactile}$。从视觉和触觉单模态特征分别提取分类标签 $\boldsymbol{z}_{(0,0)}^{(0)self\_vusal}$ 和 $\boldsymbol{z}_{(0,0)}^{(0)self\_tactile}$，在嵌入特征维度上拼接得到最终分类标签特征 $\boldsymbol{X}_{v,t}^{Class\_token}$，并将其输入到 *MLP* 分类头用于预测最终滑动检测类别。

## 3. 数据采集与处理

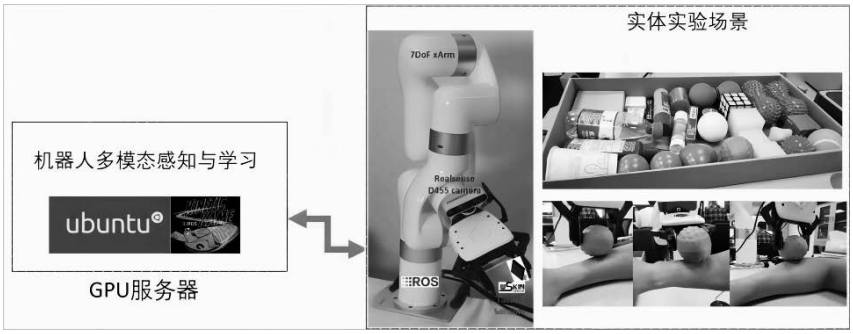


**Figure 4** The research platform and experimental scenarios
**图 4.** 研究平台及实验场景

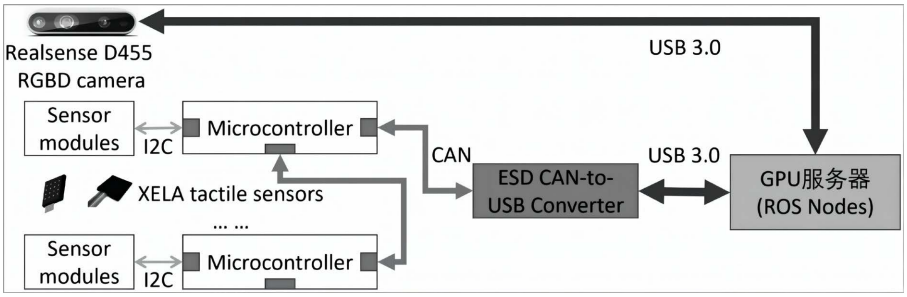


**Figure 5.** The multi-modal sensor network
**图 5.** 多模态传感器网络

所构建的研究平台及实验场景如图 4 所示，主要包括 GPU 服务器、机器人操作系统(Robot Operation System, ROS)、xArm 7DoF 机械臂、二指夹爪、D455 相机和 XELA 触觉传感器。基于该实验场景所构建

的视触力觉多模态传感器网络如图 5 所示，视觉传感器采用 Realsense D455 相机，获取机器人待抓取物体的 RGB 图片，分辨率为 640 × 480 像素。触力觉传感器采用 XELA 3D 4 × 4 阵列触力觉传感器，每一点阵可用于测量接触区域的 3D 形变，XYZ 每轴 16 位测量分辨率，每帧 CAN 消息由接触式测量的三轴分量(8Bytes)所组成。XELA 触觉传感器贴装在二指夹抓上，用于采集待抓取物体的力触觉数据。

设置二指夹爪的最大有效行程为 77 mm。考虑到 XELA 触觉传感器的测量范围，待抓取物体的重量限定在 1 kg 之内。XELA 触觉传感器和 D455 摄像头通过该数据采集平台进行数据的同步采集和记录。预先放置物体在既定位置上，通过安置在机械臂上的二指夹爪进行抓取。在每次抓取过程中，机械臂根据给定的抓取位置和抓取宽度进行抓取动作。开始时，将夹爪置于抓取起始位置，并闭合夹爪，然后启动抬升动作将物体抬起。当物体顶端到达所预设高度时，自动停止数据采集。然后，机械臂返回到初始位置，将物体松开，完成一次抓取动作。在实验中，对 50 个具有不同大小、形状、材料和重量的每一个物体重复抓取 15~20 次，共进行 952 次抓取，其中发生滑动和稳定抓取的比例为 1:1，每次抓取有效数据为 21 帧。若每组数据取 14 帧，则可采用数据增广进行 8 倍增广，最终得到 7616 组有效数据。将我们自主制作的 Gdut_Xela 数据集分为训练集和测试集，其中，训练集包含 40 种物体的 6112 组有效抓取数据，测试集包含另外 10 种物体的 1504 组有效抓取数据。图 6 为 Gdut_Xela 数据集中部分物体，通过该数据集，可以对机械臂抓取不同物体时的性能进行评估和分析，对所提出的多模态感知网络进行训练、测试和优化。

**Figure 6.** Partial objects in the dataset
**图 6.** 数据集中的部分物体

## 4. 实验结果与分析

在 Pytorch 环境中搭建本文所提出的模型，对网络参数进行随机初始化。损失函数和优化器分别使用交叉熵函数和 Adam，batch_size 设置为 4，学习率设置为 $1 \times 10^{-4}$。服务器平台配置主要包括 Intel Xeon E5-2620 CPU，48 GB 内存以及两块 GTX1080Ti GPU。数据集采用公开 Gelsight 数据集[1]以及我们自己的数据集 Gdut-Xela。下面将从模型参数设定、数据输入模态以及消融对比实验来分析模型 CrossTSformer 的性能。

### 4.1. 模型参数设定

首先，针对数据输入序列长度进行测试。虽然较长的输入序列可能包含更多的信息，但并不意味着这些信息对于检测效果一定更好。如果增加的信息是无用的或者噪声，反而可能会对模型的性能产生负面影响，导致模型难以准确地学习有效的特征和模式，从而降低检测效果。对于本文所提出的多模态感知模型，设置输入序列为 8~14，基于视触觉多模态数据进行机器人抓取准确率测试。如图 7 所示，随着输入序列值的增加，准确率呈上升趋势，且在序列值为 14 时接近最优。考虑到该数据集大小容量的限制，在后续实验中使用 14 作为输入序列长度值。

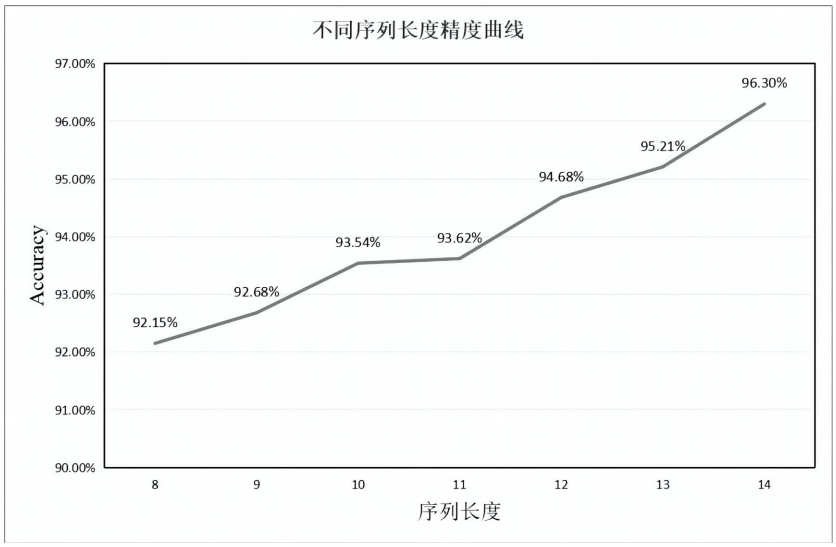


**Figure 7.** Precision curve of input sequences of different lengths
**图 7.** 不同长度输入序列的精度曲线

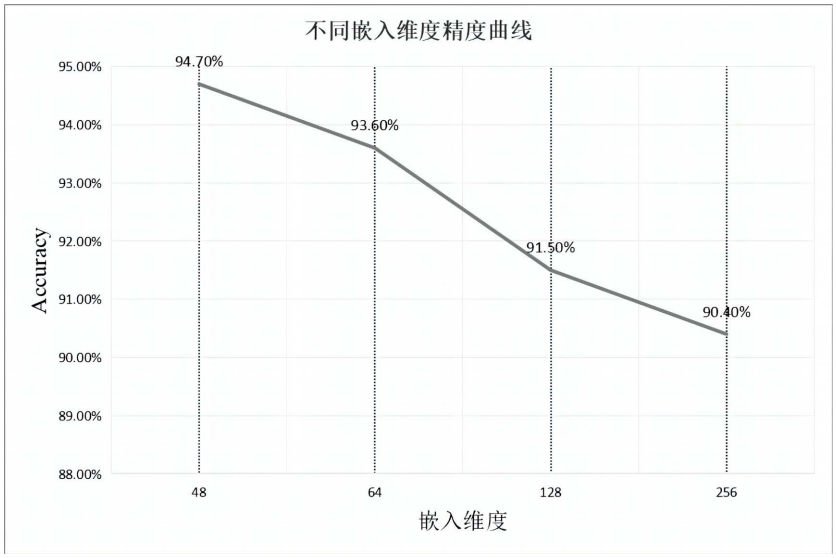


**Figure 8.** The precision curve of different tactile embedding dimensions
**图 8.** 不同触觉嵌入维度大小精度曲线

进而，针对嵌入维度进行测试。由于触觉数据比较稀疏，若使用与视觉同等嵌入维度大小，无法充分表示原始数据的信息，势必影响模型性能。在此，分别设置嵌入维度为 48、64、128 和 256 进行机器人抓取准确率测试。

如图 8 所示，当触觉数据的嵌入维度与数据本身维度(48)相近时，可以更好表示原始数据的信息。因此，后续实验中触觉序列的嵌入维度大小设为 48。

## 4.2. 数据输入模态

在机器人感知领域，视觉和触觉模态是两种重要的感知方式。视觉模态通过摄像头获取物体的外观和位置信息，而触觉模态则通过力触觉传感器获取与物体接触时的力信息。为了充分利用这两种模态信息，将视觉和触觉信息进行融合。通过视触觉融合，可以综合利用视觉和触觉模态的优势，提高抓取任

务的性能和准确度。为了验证本文所设计的跨模态时空交叉注意力分解融合方法的性能，通过与单一模态或其他融合方法进行对比，评估视触觉融合在抓取任务中的优势和效果。测试结果如表 1 所示，视触觉多模态融合比两种单模态输入的测试准确率都要高。

**Table 1.** The test results of different modal inputs
**表 1.** 不同模态输入的测试结果

| 输入模态 | 准确率 |
|---|---|
| 单触觉 | 94.7% |
| 单视觉 | 95.3% |
| 视觉 + 触觉 | **97.8%** |

### 4.3. 消融实验

首先，对三种不同融合模式进行对比测试。第一种分别提取视触觉时空特征的分类标签，将其拼接，送入 *MLP* 分类头进行输出；第二种将视触觉时空特征在嵌入特征维度上拼接，输入到分离的时空自注意力模型进行融合；第三种，将视触觉时空特征进行跨模态的交叉注意力计算，增加跨模态信息交互。实验结果如表 2 所示，引入跨模态的交叉注意力能够帮助模型有效地捕捉和利用不同模态之间的关联性，从而提高任务的性能。通过跨模态交叉注意力，模型可以自动学习和调整不同模态之间的关联权重。这使得模型能够更加注重那些对任务结果影响较大的模态，减少受无关信息的干扰，可以更全面地利用多模态信息。

**Table 2.** The comparison of different fusion modes
**表 2.** 不同融合模式对比结果

| 模型 | 准确率 |
|---|---|
| 直接融合 | 96.3% |
| Concat 融合 | 96.8% |
| Cross 融合 | **97.8%** |

进而，针对不同的模型进行对比测试。包括文献[1]的 CNN + LSTM、文献[5]的 CNN + TCN 和本文所提出的 CrossTSformer 模型。分别针对我们的 Gdut_Xela 数据集和文献[1]所采用的 Gelsight 数据集进行对比实验，以验证模型在多模态感知任务中有关时空特征提取的有效性和泛化性能。其中，Gelsight 数据集使用 Gelsight 触觉传感器，其触觉数据是高分辨率的图片。为此，使用视觉图像时空特征提取网络提取触觉特征。

**Table 3.** The comparative results of different models
**表 3.** 不同模型对比结果

| 方法 | 准确率 | |
|---|---|---|
| | Gdut-Xela 数据集 | Gelsight 数据集[1] |
| CNN + LSTM [1] | 93.1% | 82.3% |
| CNN + TCN [5] | 96.1% | 84.1% |
| CrossTSformer | 97.8% | 88.2% |

测试结果如表3所示，本文所提出的CrossTSformer模型针对两种数据集都获得了最高的测试准确率。CrossTSformer 模型通过注意力机制对不同模态的特征进行建模，通过交叉注意力融合这些特征，对不同模态的特征进行交互。它可以学习到不同模态特征之间的时空依赖关系，并通过注意力机制动态地调整特征的重要性。这种方式能够更全面地捕捉视觉和触觉之间的相关性，提高模型的表达能力，因而具有灵活的特征交互和更佳的多模态建模能力。而 CNN + 时序网络类模型通常在时序网络的末端将所有模态的特征进行拼接，无法对不同模态之间的特征交互进行灵活的建模。再者 CrossTSformer 模型对于使用阵列触觉传感器或光学触觉传感器的数据集，相对于 CNN + 时序网络类模型都有着更高的预测准确率，因而具有更好的泛化能力。

## 5. 结论

针对机器人抓取任务中的滑动检测、多模态感知问题，对多模态感知架构进行了深入研究，提出一种跨模态交叉时空注意力模型 CrossTSformer，用于机器人的视触觉多模态感知、抓取任务中的滑动检测。为了验证该模型的有效性，进行了不同融合模式和不同模型的对比实验，并针对 Gelsight 数据集和 Gdut_Xela 数据集进行了较为全面的测试。结果表明，我们的 CrossTSformer 模型在感知和判断目标物体的滑动状态方面表现出较好的稳定性和可靠性。该模型能够有效地整合视觉和触觉信息，实现多模态信息的高效交互和关联。

本研究成果对于机器人多模态感知、抓取任务的可靠执行具有积极的促进作用。为实现更高检测精度和抓取成功率的机器人抓取任务提供新的可能性和解决方案。诚然，我们也意识到仍然存在一些问题需要进一步深入研究。比如，如何更好地将多模态感知网络部署在机械臂上，实现可靠的实时在线抓取等。这也是我们进一步的主要研究方向之一。

## 参考文献


[1] Li, J., Dong, S. and Adelson, E. (2018) Slip Detection with Combined Tactile and Visual Information. 2018 *IEEE International Conference on Robotics and Automation* (*ICRA*), Brisbane, 21-25 May 2018, 7772-7777. https://doi.org/10.1109/ICRA.2018.8460495

[2] Cui, S., Wang, R., Wei, J., *et al.* (2020) Grasp State Assessment of Deformable Objects Using Visual-Tactile Fusion Perception. 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*), Paris, 31 May-31 August 2020, 538-544. https://doi.org/10.1109/ICRA40945.2020.9196787

[3] Zhang, W., Sun, F., Wu, H., *et al.* (2017) A Framework for the Fusion of Visual and Tactile Modalities for Improving Robot Perception. *Science China Information Sciences*, **60**, Article No. 12201. https://doi.org/10.1007/s11432-016-0158-2

[4] Francomano, M.T., Accoto, D. and Guglielmelli, E. (2013) Artificial Sense of Slip—A Review. *IEEE Sensors Journal*, **13**, 2489-2498. https://doi.org/10.1109/JSEN.2013.2252890

[5] Yan, G., Schmitz, A., Tomo, T.P., *et al.* (2022) Detection of Slip from Vision and Touch. 2022 *International Conference on Robotics and Automation* (*ICRA*), Philadelphia, 23-27 May 2022, 3537-3543. https://doi.org/10.1109/ICRA46639.2022.9811589

[6] 黄兆基, 高军礼, 唐兆年, 等. 基于注意力机制和视触融合的机器人抓取滑动检测[J/OL]. 信息与控制: 1-9. https://doi.org/10.13976/j.cnki.xk.2023.2598, 2024-04-06.

[7] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. arXiv: 1409.0473.

[8] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. In: Guyon, I., Von Luxburg, U., *et al.*, Eds., *Advances in Neural Information Processing Systems* 30, Long Beach, 4-9 December 2017, 1-15.

[9] Cui, S., Wang, R., Wei, J., *et al.* (2020) Self-Attention Based Visual-Tactile Fusion Learning for Predicting Grasp Outcomes. *IEEE Robotics and Automation Letters*, **5**, 5827-5834. https://doi.org/10.1109/LRA.2020.3010720

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[11] Bertasius, G., Wang, H. and Torresani, L. (2021) Is Space-Time Attention All You Need for Video Understanding? *ICML*, **2**, 1-12.

[12] Arnab, A., Dehghani, M., Heigold, G., *et al.* (2021) Vivit: A Video Vision Transformer. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 6836-6846. https://doi.org/10.1109/ICCV48922.2021.00676

[13] Cao, G., Zhou, Y., Bollegala, D., *et al.* (2020) Spatio-Temporal Attention Model for Tactile Texture Recognition. 2020 *IEEE*/*RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Las Vegas, 24 October 2020-24 January 2021, 9896-9902. https://doi.org/10.1109/IROS45743.2020.9341333

[14] Kim, H., Ohmura, Y. and Kuniyoshi, Y. (2021) Transformer-Based Deep Imitation Learning for Dual-Arm Robot Manipulation. 2021 *IEEE*/*RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Prague, 27 September-1 October 2021, 8965-8972. https://doi.org/10.1109/IROS51168.2021.9636301

[15] Li, J., Selvaraju, R., Gotmare, A., *et al.* (2021) Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. *Advances in Neural Information Processing Systems*, **34**, 9694-9705.

[16] Bao, H., Wang, W., Dong, L., *et al.* (2022) Vlmo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *Advances in Neural Information Processing Systems*, **35**, 32897-32912

[17] Cui, S., Wei, J., Li, X., *et al.* (2020) Generalized Visual-Tactile Transformer Network for Slip Detection. *IFAC-PapersOnLine*, **53**, 9529-9534. https://doi.org/10.1016/j.ifacol.2020.12.2430