doi:10.11835/j.issn.1000-582X.2024.008
引用格式:邓龙,冯波,葛永新.基于双流互信息联合匹配的小样本行为识别[J].重庆大学学报,2025,48(6):63-73.

基于双流互信息联合匹配的小样本行为识别

邓　龙[1],冯　波[2],葛永新[1]
(1.重庆大学 大数据与软件学院,重庆 400044;2.西南计算机有限责任公司,重庆 400060)

摘要:基于度量学习范式的小样本行为识别方法已经取得了巨大成功,仍有以下问题无法同时解决:1)无法很好地进行动作关系建模,没能充分利用模态信息;2)无法处理好不同长度不同速度视频匹配问题,无法处理好视频子动作错位的视频匹配问题。为了解决以上问题,提出了一种基于双流互信息联合匹配的小样本行为识别方法,该方法分为2个模块:多模态对比学习模块和联合匹配模块。多模态对比学习模块主要用以探索模态间的互信息,利用对比学习将同一个视频的不同模态特征视为正样本对进行拉近。联合匹配模块主要解决上述提到的视频匹配问题,通过动态时间规整算法和二部图匹配算法联合匹配得到最终的匹配结果,以获得良好的小样本行为识别分类准确度。笔者在2个广泛使用的小样本行为识别数据集SSV2和Kinetics上对提出的方法进行评估,同时进行了大量的消融实验验证了所提方法的有效性。

关键词:深度学习;行为识别;多模态;小样本学习
中图分类号:TP181　　**文献标志码:**A　　**文章编号:**1000-582X(2025)06-063-11

Two-stream joint matching based on mutual information for few-shot action recognition

DENG Long[1], *FENG Bo*[2], *GE Yongxin*[1]
(1. School of Big Data and Software Engineering, Chongqing University, Chongqing 400044, P. R. China;
2. Southwest Computer Co., Ltd., Chongqing 400060, P. R. China)

Abstract: Although few-shot action recognition based on the metric learning paradigm has achieved significant success, it fails to address the following issues: 1) inadequate action relation modeling and underutilization of multi-modal information; 2) challenges in handling video matching problems with different lengths and speeds, and misaligned video sub-actions. To address these limitations, we propose a two-stream joint matching (TSJM) method based on mutual information, which consists of two modules: multi-modal contrastive learning module (MCL) and joint matching module (JMM). The MCL extensively explores inter-modal mutual information

收稿日期:2024-04-20　　**网络出版日期:**2024-05-29
基金项目:重庆市技术创新与应用发展专项(CSTB2022TIAD-KPX0100);国家自然科学基金(62176031);中央高校基本科研业务费专项(2023CDJYGRHZD05)。
Supported by the Specialized Project for Technology Innovation and Application Development of Chongqing (CSTB2022TIAD-KPX0100), National Natural Science Foundation of China (62176031), and the Fundamental Research Funds for the Central Universities (2023CDJYGRHZD05).
作者简介:邓龙(1997—),男,硕士研究生,主要从事小样本行为识别方向研究,(E-mail) longdeng@cqu.edu.cn。
通信作者:葛永新,男,教授,博士生导师,(E-mail) yongxinge@cqu.edu.cn。

relationships, and thoroughly extracts modal information to enhance the modeling of action relationships. The JMM is primarily designed to simultaneously solve the aforementioned video matching problems. By integrating dynamic time warping (DTW) and bipartite graph matching, it optimizes the matching process to generate the final alignment results, thereby achieving high few-shot action recognition accuracy. We evaluate the proposed method on two widely used few-shot action recognition datasets (SSV2 and Kinetics), and conduct comprehensive ablation experiments to substantiate the efficacy of our approach.

Keywords: deep learning; action recognition; multimodal; few-shot learning

随着深度学习的迅猛发展,行为识别在各个领域已经取得了巨大的成功[1-8]。然而,这些卓越的成果往往依赖于庞大的训练数据和详细标注数据的支持,需要耗费大量的人力、物力和时间来获取。这一问题在智能交通、视频监控等领域尤为显著,因为获取足够的标注数据十分困难。为解决这一挑战,小样本行为识别应运而生。小样本行为识别的核心目标在于利用有限的标注数据来训练模型,使得模型具备在新的类别上快速应用的能力。相较于传统的行为识别方法,小样本行为识别突破了对海量数据的过度依赖,强调在数据稀缺情况下的模型鲁棒性。这种方法不仅节省了在数据收集和标注上的巨大成本,同时也为在特定领域和任务中快速部署高效的行为识别系统提供了新的可能性。在小样本行为识别的研究中,探索如何在有限数据条件下实现模型的高效学习成为关键挑战。通过采用迁移学习、元学习等技术,小样本行为识别旨在提高模型对新类别的泛化能力,从而使其在实际应用中更具实用性和适应性。

国内对小样本行为识别进行了大量的研究。姚天等[9]提出了一个基于Conv-Involution的红外视频小样本人体行为识别方法,用于解决红外视频的小样本行为识别中颜色信息缺失、背景与主体混淆的问题。宗鹏程等[10]提出了一个元学习网络提取特征,然后基于度量学习,使用正则化及二阶池化等操作对视频特征进行增强,最后用于分类。另外还将目标检测领域常用的通道注意力模块和坐标注意力模块引入到小样本行为识别领域,并根据小样本行为识别领域的特性进行了改进,使其成为一个即插即用的提取特征模块。尹恒等[11]提出利用特征编码网络获取特征,然后基于度量学习计算支持集和查询集之间的距离并进行分类,此外还提出了一个时间切片和多空间中心采样的数据增强策略获得视频样本的完整特征表示,并利用通道注意力机制来优化相似度学习。

上述工作都是单模态的小样本行为识别方法,并不足以构建鲁棒的特征。如图1所示,something-something v2(SSV2)数据集[12]上的大多数视频与行为主体无关,但与动作强相关,因此,对动作关系进行建模尤为重要。在之前的研究方法中,许多学者利用了额外的模态信息进行运动建模,如深度信息[13]、运动向量[14]、时间梯度[15]、帧间差[16]、流形特征[5]和三轴加速度[6]等。然而,在小样本动作识别领域,却很少有人利用光流信息进行运动关系建模。光流信息通过构建运动场来表征物体运动的方向和速度,在运动关系建模方面具有明显的优势,在之前的研究中,只有AMFAR[17]首次通过主动学习方法创新性地将光流信息融入动作关系建模中。但是AMFAR方法需要额外的监督标注,并且使用了计算昂贵的I3D预训练模型进行光流学

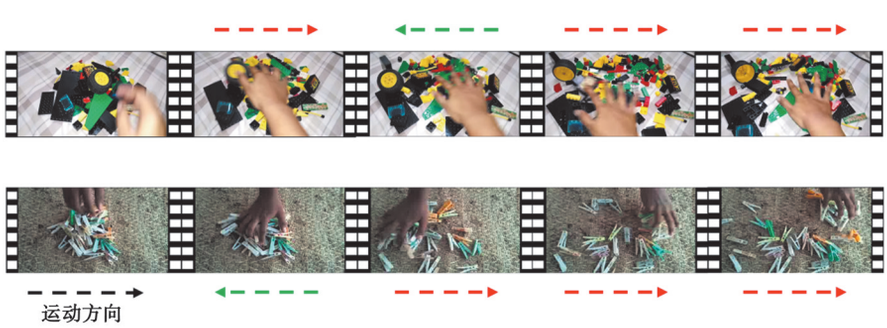

图1 同一类别的2个不同视频

Fig. 1 Two different videos of the same category

习。与AMFAR方法不同，笔者拟利用对比学习探索跨模态的互信息，促进光流信息整合到运动建模中，无需昂贵的标注信息即可提升模型的泛化能力。

在获得有效的特征表示之后，大多数现有的少样本动作识别方法采用度量学习范式评估2个视频之间的相似性。这些方法首先将视频数据映射到新的特征空间，然后利用各种距离度量来量化支持集和查询集之间的距离，如动态时间规整，Sinkhorn距离[18]和双向平均豪斯多夫度量[19]等。尽管这些方法在小样本行为识别领域取得了显著的成功，但仍未解决2个关键挑战：1）具有不同速度和长度的视频匹配问题；2）视频子动作不对齐的问题。为了解决第一个问题，OTAM[20]利用动态时间规整（dynamic time warping，DTW）算法动态对齐2个视频，有效缓解了由于速度和长度变化引起的匹配精度下降问题。然而，随着视频中子动作增多，该方法在解决第二个问题时存在不足。因此，对DTW算法不进行改进而直接应用到视频匹配时，同样会降低视频匹配准确率。

为了克服上述提到的问题并提高小样本行为识别的准确性，提出了一种新的基于双流互信息联合匹配的小样本行为（two-stream joint matching method based on contrastive learning for few-shot action recognition，TSJM）识别方法，包含2个模块：多模态对比学习模块和联合匹配模块。首先使用在ImageNet上预训练好的深度卷积网络ResNet-50提取RGB视频帧和光流图像的特征，利用RGB视频帧捕获物体的外观特征，利用光流信息捕获物体的运动特征。然后将特征输入到多模态对比学习模块中，将同一视频的RGB特征和光流特征视为正样本，使它们更加接近，同时将其他情况视为负样本，使它们相互推远，以此构建模态间互信息。此外，为了解决视频匹配问题，在联合匹配模块中采用DTW算法对查询集和支持集进行距离度量，以实现2个视频序列的动态对齐。为了解决视频子动作错位的问题，本文将带权二部图的完美匹配方法引入小样本行为识别中，并采用Kuhn-Munkres（KM）算法寻找最优匹配。最后，将2个度量方法的得分进行融合，得到最终的视频相似度，并将其用于分类。实验证明，TSJM在2个广泛使用的标志性数据集上都取得了有竞争力的结果。

1　基于双流互信息联合匹配的小样本行为识别方法设计

1.1　问题定义

在小样本行为识别领域主要遵循的是元学习的范式，通过多个元任务的方式进行训练得到具有泛化能力的模型，使其在不可见的类别上有良好的识别能力。元学习主要分为2个阶段：元训练和元测试。在元训练阶段，模型在一系列的训练任务上学习快速适应和泛化到新任务。每个训练任务 $\mathcal{T}$ 通常从训练集 $\boldsymbol{D}_{\text{train}}$ 中采样得到训练样本 $\mathcal{T}_{\text{train}}$，$\mathcal{T}_{\text{train}}$ 中包含支持集 S 和查询集 Q，同时遵循 N-way K-shot 的范式，其中 N-way 表示每个任务中包含 N 个类别数，K-shot 表示每个类别中的支持集样本数为 K。在元测试阶段，将经过元训练的模型用于新的测试任务进行推断和泛化，这些测试任务与元训练任务不同，模型必须根据先前的元训练经验，迅速适应新任务并生成准确的预测。在新的测试数据集 $\boldsymbol{D}_{\text{test}}$ 上进行采样得到新的测试类别 $\mathcal{T}_{\text{test}}$，其中 $\boldsymbol{D}_{\text{train}} \cap \boldsymbol{D}_{\text{test}} = \varnothing$，采用与元训练阶段一样的方法构建支持集和查询集，用于测试模型的泛化能力。

1.2　模型框架

本文的总体框架如图2所示，在训练样本 $\mathcal{T}_{\text{train}}$ 中，分为了支持集和查询集，为方便起见，N-way K-shot 以3-way 1-shot为例，首先将支持集和查询集的视频帧和光流帧进行采帧，采帧的方法与temporal segment networks（TSN）[21]中类似的稀疏采样，得到 T 帧视频图像，然后将图像输入到特征提取器 φ 中，分别得到支持集RGB特征 $\boldsymbol{S}_i^{\text{r}}$，支持集光流特征 $\boldsymbol{S}_i^{\text{f}}$，查询集RGB特征 $\boldsymbol{Q}_i^{\text{r}}$，查询集光流特征 $\boldsymbol{Q}_i^{\text{f}}$，然后将特征输入到一个Adapter模块中对2个模态的特征进行拉近，以便更好的进行对比学习，在多模态对比学习模块（multimodal contrastive learning, MCL）中，同一视频的RGB和光流特征作为正样本，其余作为负样本，采用InfoNCE的损失函数进行学习。另外，为了解决不同速度不同长度的视频匹配问题，以及视频子动作错位的匹配问题，本文设计了一个联合匹配模块（joint matching module, JMM），对于2个模态的特征，输入到匹配模块中进行匹配，分别进行顺序时间匹配和KM算法的二部图匹配，最终得到4个得分 $\text{score}_{\text{ota}}^{\text{r}}$，$\text{score}_{\text{ota}}^{\text{f}}$，$\text{score}_{\text{km}}^{\text{r}}$，$\text{score}_{\text{km}}^{\text{f}}$，经过加权平均得到最终的预测得分。

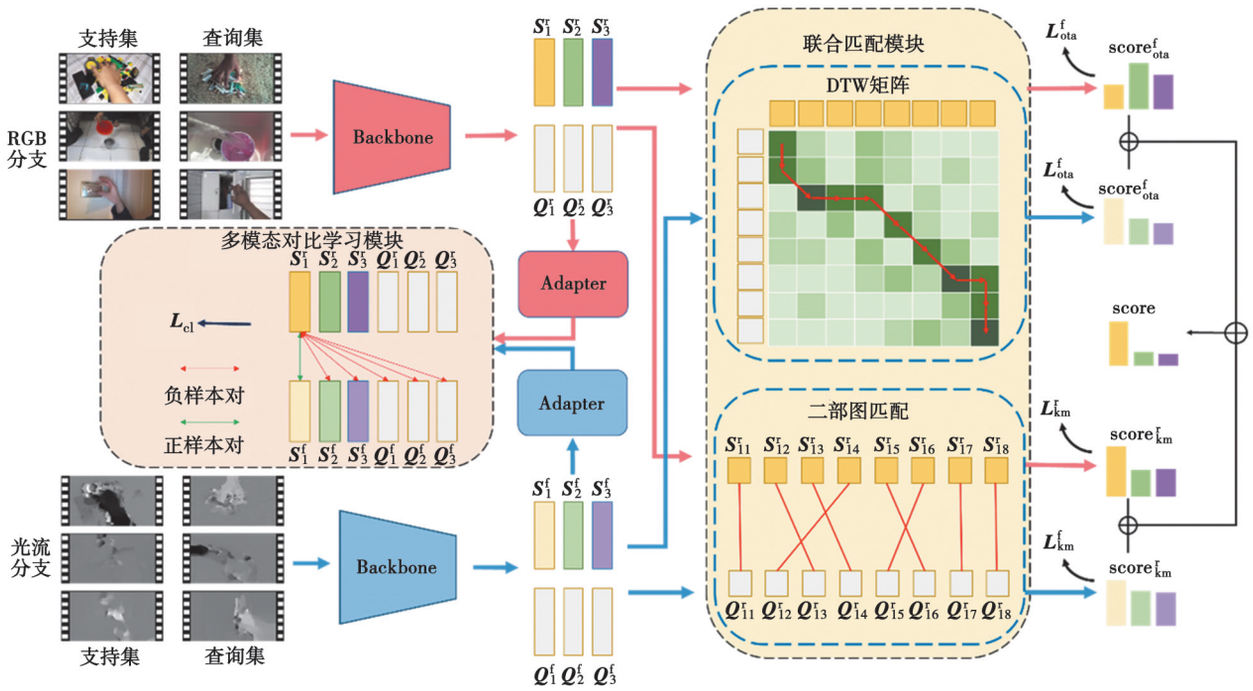

图 2 基于双流互信息联合匹配的小样本行为识别方法流程图

Fig. 2 Framework of two-stream joint matching based on contrastive learning for few-shot action

1.3 多模态对比学习

在同一个视频的2个不同模态中，视频的全局高层次特征表示有一定联系，因此，探索2个模态之间的互信息至关重要，本文中设计了一个多模态对比学习模块来探索RGB模态和光流模态之间的互信息，以达到模态互补的作用。不同模态之间的特征属于不同的特征空间，2个模态的特征差别较大，直接对比学习难以训练，因此，设计了一个Adapter模块（见图3），图中，FC 1（full connection）指的是全连接层1，ReLU（rectified linear unit）线性整流函数是一种人工神经网络中常用的非线性激活函数，FC 2指的是全连接层2。Adapter模块是一个瓶颈体系的结构，由2个全连接层和1个激活层组成，第一个全连接层将输入的特征投影到一个较低的维度，第二个全连接层将其投影回原始维度，这样可以将2个不同域的模态特征投影到一个新的特征空间以便进行对比学习。

通过Backbone得到2个模态的特征后，将其输入到Adapter模块。

$$\boldsymbol{x}_i^{\mathrm{r}}=\mathrm{Adapter}\left(\varphi\left(\boldsymbol{X}_i^{\mathrm{r}}\right)\right),\boldsymbol{X}_i^{\mathrm{r}}=\left\{\boldsymbol{S}_i^{\mathrm{r}},\boldsymbol{Q}_i^{\mathrm{r}}:(1\leqslant i\leqslant N)\right\},\tag{1}$$

$$\boldsymbol{x}_i^{\mathrm{f}}=\mathrm{Adapter}\left(\varphi\left(\boldsymbol{X}_i^{\mathrm{f}}\right)\right),\boldsymbol{X}_i^{\mathrm{f}}=\left\{\boldsymbol{S}_i^{\mathrm{f}},\boldsymbol{Q}_i^{\mathrm{f}}:(1\leqslant i\leqslant N)\right\},\tag{2}$$

式中：φ表示特征提取器；N表示类别数。

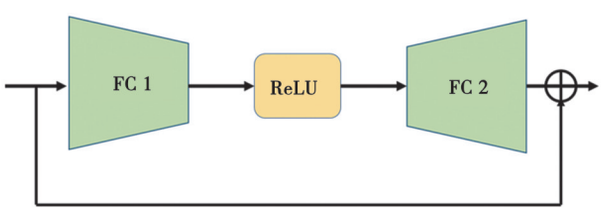

图 3 Adapter示意图

Fig. 3 Schematic diagram of the Adapter

为了衡量全局高层次特征之间的相似度，与TRX[22]方法中类似，研究采用交叉注意力机制来计算相似度，与SimCLR[23]和CMC[24]中类似，采用正样本对和负样本对的对比学习方式，将同一个视频的不同模态的全局高层次特征作为正样本对，其余情况均为负样本对，对比学习的目标是拉近正样本对之间的距离，即同

一个视频2个不同模态之间的特征距离，使其相似度最大化，并同时推远负样本之间的距离，使其相似度最小化。最后，采用InfoNCE的损失函数来优化模型。

$$\boldsymbol{L}_{\mathrm{cl}}=-\log\frac{\sum_{i=1}^{k}\exp\left(\mathrm{sim}\left(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{x}_i^{\mathrm{f}}\right)/\tau\right)}{\sum_{i\neq j}^{k}\exp\left(\mathrm{sim}\left(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{x}_j^{\mathrm{f}}\right)/\tau\right)},\tag{3}$$

式中：sim()表示TRX中计算得到的相似度；τ表示温度系数。得到的相似度进行归一化，再进行正负样本对的构造。

1.4　联合匹配模块

为了同时解决不同长度、不同速度视频匹配问题和视频子动作错位的匹配问题，笔者提出了一个联合匹配模块，包含2个子模块，分别是顺序时间对齐模块和二部图匹配模块。顺序时间对齐模块与OTAM[20]中类似，采用DTW算法对不同长度、不同速度的视频进行匹配，在保持时间顺序不变的同时，动态的对齐2个视频序列。然而它不能解决子动作错位的匹配问题，因此，另外提出了一个二部图匹配的子模块，将支持集和查询集构造为二部图，然后采用二部图的最优匹配算法KM算法来计算视频间的最优匹配。

1.4.1　顺序时间匹配模块

在这个子模块中，主要解决不同长度、不同速度的视频匹配问题，使模型能够在保持视频时间顺序不变的同时，动态的对齐2个视频序列。对于RGB视频帧的支持集特征$\boldsymbol{S}_i^r$和查询集特征$\boldsymbol{Q}_j^r$，与OTAM[20]中类似，计算两者的帧级距离

$$\boldsymbol{D}_{\mathrm{f}}(l,m)=1-\frac{\boldsymbol{S}_{il}^{\mathrm{r}}\cdot\boldsymbol{Q}_{jm}^{\mathrm{r}}}{\left\|\boldsymbol{S}_{il}^{\mathrm{r}}\right\|\cdot\left\|\boldsymbol{Q}_{jm}^{\mathrm{r}}\right\|},\tag{4}$$

式中，$\boldsymbol{D}_{\mathrm{f}}(l,m)$表示$\boldsymbol{S}_i^{\mathrm{r}}$的第$l$帧和$\boldsymbol{Q}_j^{\mathrm{r}}$的第$m$帧的帧级距离。

通过DTW算法计算支持集和查询集帧之间的最短距离，将每一帧的最优匹配的相似度求和，得到2个视频的距离

$$\boldsymbol{D}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)=\sum_{j=1}^{T}f\left(\boldsymbol{D}_{\mathrm{f}}\right),\tag{5}$$

式中：$f()$表示DTW算法；$\boldsymbol{D}_{\mathrm{f}}$表示2个视频的帧级距离矩阵。最后，通过交叉熵损失来优化模型

$$\boldsymbol{L}_{\mathrm{ota}}^{\mathrm{r}}=-\log\frac{\exp\left(-\boldsymbol{D}_{\mathrm{v}}\left(\boldsymbol{S}_j^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)\right)}{\sum_{i=1}^{N}\exp\left(-\boldsymbol{D}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)\right)}。\tag{6}$$

式中，$\boldsymbol{S}_j^{\mathrm{r}}$和$\boldsymbol{Q}_j^{\mathrm{r}}$分别表示支持集和查询集中同一个类别的视频。

对于光流的特征，采用与RGB特征类似的操作，光流特征通过OTA进行优化得到损失函数

$$\boldsymbol{L}_{\mathrm{ota}}^{\mathrm{f}}=-\log\frac{\exp\left(-\boldsymbol{D}_{\mathrm{v}}\left(\boldsymbol{S}_j^{\mathrm{f}},\boldsymbol{Q}_j^{\mathrm{f}}\right)\right)}{\sum_{i=1}^{N}\exp\left(-\boldsymbol{D}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{f}},\boldsymbol{Q}_j^{\mathrm{f}}\right)\right)}。\tag{7}$$

1.4.2　二部图匹配模块

在某些类别上，存在视频子动作错位的问题，DTW算法匹配结果不理想。因此，另外提出一个二部图匹配模块，将二部图的概念映射到小样本行为识别中，并通过二部图领域内的KM算法来获得具有最大权值的完美匹配以得到视频之间的最优匹配。

对于RGB视频帧的支持集特征$\boldsymbol{S}_i^{\mathrm{r}}$的第$l$帧和查询集特征$\boldsymbol{Q}_j^{\mathrm{r}}$的第$m$帧，采用余弦相似度计算两者帧级相似度矩阵

$$\mathrm{sim}_{\mathrm{f}}\left(\boldsymbol{S}_{il}^{\mathrm{r}},\boldsymbol{Q}_{jm}^{\mathrm{r}}\right)=\frac{\boldsymbol{S}_{il}^{\mathrm{r}}\cdot\left(\boldsymbol{Q}_{jm}^{\mathrm{r}}\right)^{\mathrm{T}}}{\left\|\boldsymbol{S}_{il}^{\mathrm{r}}\right\|\cdot\left\|\boldsymbol{Q}_{jm}^{\mathrm{r}}\right\|}。\tag{8}$$

得到相似度后，将支持集特征和查询集特征分别作为2个图的节点，帧与帧之间的相似度作为节点之间边的权值。于是$\boldsymbol{S}_{il}^{\mathrm{r}}$和$\boldsymbol{Q}_{jm}^{\mathrm{r}}$之间构成一个带权的完全二部图，目标是求出具有最大权值的完美匹配。

在图论中，如果找到一种恰当的可行顶点标号，使得对应的相等子图有完美匹配，即可求出带权二部图

的最优匹配。此时设查询集视频和支持集视频的完美匹配结果为$\boldsymbol{M}^*$,最终的匹配得分由匹配结果的相似度加和,然后通过交叉熵损失优化模型。

$$\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)=\sum_{j=1}^{T}\mathrm{sim}_{\mathrm{f}}\left(\boldsymbol{M}^*\right), \tag{9}$$

$$\boldsymbol{L}_{\mathrm{km}}^{\mathrm{r}}=-\log\frac{\exp\left(\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_j^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)\right)}{\sum_{i=1}^{N}\exp\left(\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)\right)}, \tag{10}$$

式中:$\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)$表示支持集视频$\boldsymbol{S}_i^{\mathrm{r}}$和查询集视频$\boldsymbol{Q}_j^{\mathrm{r}}$之间的相似度;$\boldsymbol{S}_j^{\mathrm{r}}$和$\boldsymbol{Q}_j^{\mathrm{r}}$表示支持集和查询集中同一个类别的视频。

对于光流特征,与RGB特征类似,进行相似的操作,得到损失函数

$$\boldsymbol{L}_{\mathrm{km}}^{\mathrm{f}}=-\log\frac{\exp\left(\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_j^{\mathrm{f}},\boldsymbol{Q}_j^{\mathrm{f}}\right)\right)}{\sum_{i=1}^{N}\exp\left(\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_i^{\mathrm{f}},\boldsymbol{Q}_j^{\mathrm{f}}\right)\right)}。 \tag{11}$$

1.5 多模态融合及网络参数优化过程

为了整合2种不同模态和2种匹配方法的分类结果,yanjiu提出了一个多模态融合策略。由于这2种不同的匹配方法和模态在性能上相互补充,因此,需要采用多模态融合方法进行综合处理。同时,鉴于不同模态在各自任务上表现出不同的性能,在进行多模态融合时引入了超参数优化策略,以达到最佳性能配置。

$$\mathrm{score}=\lambda_1\mathrm{score}_{\mathrm{ota}}^{\mathrm{r}}+\lambda_2\mathrm{score}_{\mathrm{ota}}^{\mathrm{f}}+\lambda_3\mathrm{score}_{\mathrm{km}}^{\mathrm{r}}+\lambda_4\mathrm{score}_{\mathrm{km}}^{\mathrm{f}}, \tag{12}$$

式中:$\mathrm{score}_{\mathrm{ota}}^{\mathrm{r}}=-\boldsymbol{D}_{\mathrm{v}}\left(\boldsymbol{S}_j^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)$表示RGB视频特征通过顺序时间对齐模块得到的相似度得分;$\mathrm{score}_{\mathrm{km}}^{\mathrm{r}}=\mathrm{sim}_{\mathrm{v}}\left(\boldsymbol{S}_j^{\mathrm{r}},\boldsymbol{Q}_j^{\mathrm{r}}\right)$表示RGB视频特征通过二部图匹配模块得到的相似度得分;$\lambda_1,\lambda_2,\lambda_3,\lambda_4$是用于控制不同任务得分权重的超参数。

研究提出的损失函数为

$$\mathrm{Loss}=\alpha_1\boldsymbol{L}_{\mathrm{cl}}+\alpha_2\boldsymbol{L}_{\mathrm{ota}}^{\mathrm{r}}+\alpha_3\boldsymbol{L}_{\mathrm{ota}}^{\mathrm{f}}+\alpha_4\boldsymbol{L}_{\mathrm{km}}^{\mathrm{r}}+\alpha_5\boldsymbol{L}_{\mathrm{km}}^{\mathrm{f}}, \tag{13}$$

式中,$\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5$分别表示每个损失函数权重的超参数。

1.6 算法流程总结

算法:基于双流互信息联合匹配的小样本行为识别方法(TSJM)

输入:

等间隔采样的T帧视频片段,模型训练迭代次数nums。

输出:

当前视频的行为类别。

1 for i in nums,do:
2 利用特征提取器φ提取特征$\boldsymbol{S}_i^{\mathrm{r}}$、$\boldsymbol{S}_i^{\mathrm{f}}$、$\boldsymbol{Q}_i^{\mathrm{r}}$、$\boldsymbol{Q}_i^{\mathrm{f}}$;
3 利用公式(1)和(2)将特征映射到新的空间得到$\boldsymbol{x}_i^{\mathrm{r}}$、$\boldsymbol{x}_i^{\mathrm{f}}$;
4 利用公式(3)对特征进行对比学习;
5 利用公式(4)计算顺序时间匹配中支持集和查询集特征距离;
6 利用公式(5)计算2个视频之间的距离;
7 利用公式(6)优化模型;
8 利用公式(8)计算二部图匹配中的特征距离;
9 利用公式(10)优化二部图匹配的模型;
10 利用公式(12)计算查询集视频的置信度分数score并分类;
11 利用公式(13)计算整个模型的损失函数Loss;
12 根据Loss进行反向传播优化参数;
13 end。

2　实验设计与结果分析

2.1　数据集介绍

本文的方法在2个常用的小样本行为识别数据集上评估了方法的有效性：Kinetics[25]和SSV2[12]。对于这2个数据集，首先采用了密集光流算法对视频帧提取光流，类似于Lucas等[30]的方法，从原始视频中生成光流帧序列。在数据集选择方面，从整个数据集中挑选了100个类别，每个类别包含100个视频。与OTAM[20]和CMN[26]中采用的分割方法相一致，将64个类别划分为训练集，12个类别用作验证集，另外24个类别用作测试集。

2.2　实验设置

在小样本行为识别中，通常采用N-way K-shot的实验设置，以随机抽样方式挑选N个类别，每个类别中含有K个实例作为支持集，查询集包含N个实例，每个实例均属于支持集中的N个类别之一。类似于CMN[26]和OTAM[20]，在预处理过程中将视频段均匀地分割成8个片段，然后从每个片段中随机选择1个RGB帧作为视觉数据，并在该帧的位置选择连续的两帧光流数据作为运动数据。同时对所选的视频帧和光流帧进行标准的数据增强，包括随机裁剪、翻转等操作，将视频帧和光流帧都裁剪为224×224。对于RGB视频帧分支和光流分支，使用在ImageNet[31]上预训练的ResNet-50[32]网络作为骨干模型。在训练过程中，采用了Adam[33]优化器，设定学习率为10^{-5}，在测试阶段时，与之前的工作一样，测试轮次是10 000次。

2.3　结果比较

对比实验在小样本行为识别领域中广泛使用的数据集上进行，为了与其他工作保持一致性，采用5-way 1-shot和5-way 5-shot的设置进行实验。为了确保公平性，选择的对比方法均使用在ImageNet上预训练的ResNet-50作为骨干网络，在测试10 000轮次后计算分类准确率。

2.3.1　数据集Kinetics上的性能比较

表1展现了小样本行为识别的经典方法在数据集Kinetics上的结果。从表1可知，TSJM在Kinetics上表现出色，在1-shot实验设置上达到了最佳性能，对比多个单模态小样本行为识别方法均有明显涨幅，而对于多模态小样本行为识别方法，TSJM比之前的最佳方法MTFAN涨0.4%，比AMeFu-Net提高了0.9%，比LSTC提高了1.6%，比MoLo提高了1.0%。在5-shot的实验设置下，TSJM的性能比多个单模态小样本行为识别方法明显更优；与大多数多模态小样本行为识别方法比较，TSJM的性能也优于其他模型，比AMeFu-Net提高了1.2%，比LSTC提高了0.5%，比MoLo提高了1.4%。尽管在5-shot的实验设置中准确率比MTFAN稍低，但MTFAN方法采用了Transformer结构，计算量较大。综上所述，TSJM在Kinetics数据集上性能优异，验证了TSJM的有效性。

表 1　在Kinetics数据集上与现有方法的分类准确率比较

Table 1　Comparison of accuracy with state-of-the-art methods on the Kinetics

模态	方法	来源	1-shot Acc	5-shot Acc
单模态	CMN[26]	ECCV2018	60.5	78.9
	OTAM[20]	CVPR2020	73.0	85.8
	TRX[22]	CVPR2021	63.6	85.9
	STRM[27]	CVPR2022	—	86.7
	HyRSM[19]	CVPR2022	73.7	86.1
	SloshNet[28]	AAAI2023	—	87.0
多模态	AMeFu-Net[13]	ACMMM2020	74.1	85.8
	LSTC[14]	IJCAI2022	73.4	86.5
	MTFAN[16]	CVPR2022	74.6	87.4
	MoLo[29]	CVPR2023	74.0	85.6
	TSJM	本文方法	75.0	87.0

2.3.2 数据集SSV2上的性能比较

在另一个常用的小样本行为视频数据集SSV2上的实验结果如表2所示。由表2可知,TSJM在SSV2数据集上性能优异,在1-shot的实验设置下,对于单模态的小样本行为识别方法,性能明显优于所有单模态方法,与多模态的小样本行为识别方法相比,TSJM也提升明显,比之前的最佳方法MoLo提升1.9%,比LSTC提高了11.8%,比MTFAN提高了12.8%。在5-shot的实验设置下,TSJM优于大部分单模态小样本行为识别方法,在多模态的小样本行为识别方法中对比,TSJM比LSTC提高了1.8%,比MTFAN提高了8.1%。尽管比MoLo的分类准确率略低,但是MoLo在网络框架中加入了Transfomer结构,使得网络复杂度增加,计算量较大。

表2 在SSV2数据集上与现有方法的准确率比较

Table 2 Comparison of accuracy with state-of-the-art methods on the SSV2

模态	方法	来源	1-shot Acc	5-shot Acc
单模态	CMN[26]	ECCV2018	—	—
	OTAM[20]	CVPR2020	42.8	52.3
	TRX[22]	CVPR2021	42.0	64.6
	STRM[27]	CVPR2022	—	68.1
	HyRSM[19]	CVPR2022	54.3	69.0
	SloshNet[28]	AAAI2023	46.5	68.3
多模态	AMeFu-Net[13]	ACMMM2020	—	—
	LSTC[14]	IJCAI2022	46.7	66.7
	MTFAN[16]	CVPR2022	45.7	60.4
	MoLo[29]	CVPR2023	56.6	70.6
	TSJM	本文方法	58.5	68.5

2.4 消融实验

为验证本文中提出的每个模块的有效性,本小节进行了消融实验。

2.4.1 基线方法

本文使用的基线方法从经典的基于度量学习的小样本行为识别方法OTAM而来,使用了ImageNet预训练的Resnet-50模型作为Backbone,采用DTW算法对视频片段进行动态对齐,以得到查询集和原型之间的最佳对齐路径,并计算两者之间的距离,以得到准确的分类结果。

2.4.2 分析各个模块的有效性

分别在SSV2和Kinetics数据集上,采用5-way 1-shot的实验设置进行实验,实验结果如表3所示,基线方法是OTAM方法。从实验数据看出,多模态对比学习模块将光流信息与RGB信息进行互补,可以极大提升准确率,在SSV2数据集上,对比基线方法性能提升了10.1%,在Kinetics数据集上性能提升了0.7%。Adapter模块可以更好的将2个模态的特征投影到新的特征空间,使其更好的进行对比学习。与没有Adapter模块相

表3 在SSV2和Kinetics数据集上的消融实验

Table 3 Ablation study on SSV2 and Kinetics

方法	Kinetics Acc	SSV2 Acc
基线方法	73.0	42.8
基线方法+多模态对比学习	73.7	52.9
基线方法+多模态对比学习+Adapter	74.1	55.3
基线方法+联合匹配模块	73.8	50.9
基线方法+多模态对比学习+Adapter+联合匹配模块	**75.0**	**58.5**

比,在SSV2数据集和Kinetics数据集上分别可以提升2.4%和0.4%。联合匹配模块比基线方法新引入了完美二部图匹配的KM算法,可以更好的解决子动作错位的匹配问题,在SSV2数据集上提升了8.1%,在Kinetics数据集上提升了0.8%。最后,将所有模块组合,达到了最佳性能,证明了本文中提出的多个模块的有效性。

2.4.3　*N*-way小样本行为识别

为了充分研究TSJM在更具挑战性条件下的性能,在SSV2和Kinetics数据集上的*N*-way 1-shot实验设置下,测试了*N*从5到10的性能表现,*N*越大,分类难度越大,准确率越低。如图4所示,在Kinetics数据集上,与OTAM和TRX相比,每一种实验设置均高于之前的方法。如图5所示,在SSV2数据集上,TSJM在5-way到10-way的实验设置下,性能均高于之前的方法。

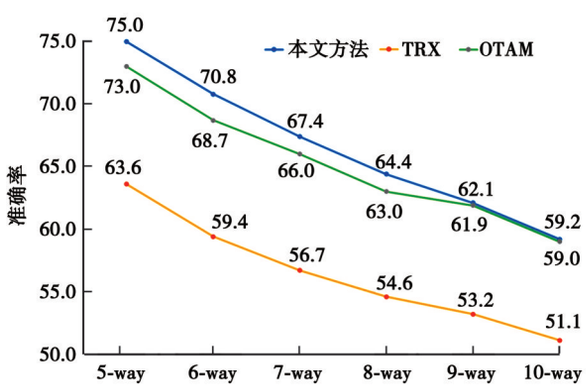

图 4　在Kinetics数据集上的*N*-way 1-shot消融实验

Fig. 4　*N*-way 1-shot on the Kinetics

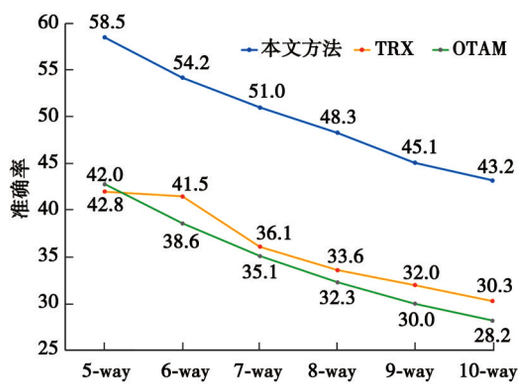

图 5　在SSV2数据集上的*N*-way 1-shot消融实验

Fig. 5　*N*-way 1-shot on the SSV2

3　结束语

提出了一个新的基于双流互信息联合匹配的小样本行为识别方法,在小样本行为识别领域首次将对比学习应用到多模态方法中,以解决小样本行为识别中运动关系建模不足的问题。

1)引入了一个多模态对比学习模块,将光流模态引入小样本行为识别领域,以构建时序动作关系。同时,通过对比学习的方式建立了RGB视频帧特征和光流特征之间的互信息,从而获得更为优秀的视频表示。

2)提出了一个联合匹配模块,将带权二部图的完美匹配问题引入小样本行为识别领域。为了解决视频子动作错位导致的匹配错误问题,采用KM算法来计算视频之间的最优匹配以提高模型鲁棒性。

3)通过大量的实验证明了方法的有效性,在2个广泛使用的标志性数据集SSV2和Kinetics上都取得了具有竞争力的结果。此外,通过大量的消融实验和可视化实验也证实了提出的多个模块的有效性。

参考文献

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.

[2] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV). December 7-13, 2015, Santiago, Chile: IEEE, 2015: 4489-4497.

[3] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA: IEEE, 2018: 6450-6459.

[4] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]// Computer Vision-ECCV 2016: 14th European Conference, October 8-16, 2016, Amsterdam, The Netherlands. Cham, Switzerland: Springer International Publishing, 2016: 20-36.

[5] 蒲瞻星, 葛永新. 基于多特征融合的小样本视频行为识别算法[J]. 计算机学报, 2023, 46(3): 594-608.
Pu Z X, Ge Y X. Few-shot action recognition in video based on multi-feature fusion[J]. Chinese Journal of Computers, 2023, 46(3): 594-608. (in Chinese)

［6］潘金星 . 小样本多模态个体行为识别[D]. 合肥: 合肥工业大学, 2022.
Pan J X. Few-shot multimodal individual activity recognition[D]. Hefei: Hefei University of Technology, 2022. (in Chinese)

［7］李晓旭, 刘忠源, 武继杰, 等 . 小样本图像分类的注意力全关系网络[J]. 计算机学报, 2023, 46(2): 371-384.
Li X X, Liu Z Y, Wu J J, et al. Total relation network with attention for few-shot image classification[J]. Chinese Journal of Computers, 2023, 46(2): 371-384. (in Chinese)

［8］李维刚, 甘平, 谢璐, 等 . 基于样本对元学习的小样本图像分类方法[J]. 电子学报, 2022, 50(2): 295-304.
Li W G, Gan P, Xie L, et al. A few-shot image classification method by pairwise-based meta learning[J]. Acta Electronica Sinica, 2022, 50(2): 295-304. (in Chinese)

［9］姚天, 余磊, 崔帅华, 等 . 基于 Conv-Involution 的红外视频小样本人体行为识别方法[J]. 激光与红外, 2023, 53(2): 246-252.
Yao T, Yu L, Cui S H, et al. Few-shot human behavior recognition method of infrared video based on Conv-Involution[J]. Laser & Infrared, 2023, 53(2): 246-252. (in Chinese)

［10］宗鹏程 . 基于元学习的少样本行为识别方法研究与实现[D]. 杭州: 浙江工业大学, 2022.
Zong P C. Research and implementation of few-shot action recognition method based on meta-learning[D]. Hangzhou: Zhejiang University of Technology, 2022. (in Chinese)

［11］尹恒 . 基于度量学习的少样本行为识别方法研究[D]. 厦门: 厦门大学, 2021.
Yin H. Study on few-shot action recognition method based on metric learning[D]. Xiamen: Xiamen University, 2021. (in Chinese)

［12］Goyal R, Kahou S E, Michalski V, et al. The “something something” video database for learning and evaluating visual common sense[C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017. Venice: IEEE, 2017: 5843-5851.

［13］Fu Y Q, Zhang L, Wang J K, et al. Depth guided adaptive meta-fusion network for few-shot video recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia (MM '20). October 12-16, 2020, Seattle, WA, USA. New York: ACM, 2020: 1142-1151.

［14］Luo W Y, Liu Y F, Li B, et al. Long-short term cross-transformer in compressed domain for few-shot video classification[C]//Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI 2022). July 23-29, 2022, Vienna, Austria. San Francisco: International Joint Conferences on Artificial Intelligence Organization, 2022: 1247-1253.

［15］Xiao J F, Jing L L, Zhang L, et al. Learning from temporal gradient for semi-supervised action recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA: IEEE, 2022: 3242-3252.

［16］Wu J M, Zhang T Z, Zhang Z, et al. Motion-modulated temporal fragment alignment network for few-shot action recognition [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA: IEEE, 2022: 9141-9150.

［17］Wanyan Y Y, Yang X S, Chen C F, et al. Active exploration of multimodal complementarity for few-shot action recognition[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 17-24, 2023, Vancouver, BC, Canada: IEEE, 2023: 6492-6502.

［18］Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport[J]. Advances in Neural Information Processing Systems, 2013, 26: 2292-2300.

［19］Wang X, Zhang S W, Qing Z W, et al. Hybrid relation guided set matching for few-shot action recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA: IEEE, 2022: 19916-19925.

［20］Cao K D, Ji J W, Cao Z J, et al. Few-shot video classification *via* temporal alignment[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA: IEEE, 2020: 10615-10624.

［21］Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//Computer Vision-ECCV 2016: 14th European Conference. Cham, Switzerland: Springer International Publishing, 2016: 20-36.

［22］Perrett T, Masullo A, Burghardt T, et al. Temporal-relational crosstransformers for few-shot action recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021. Nashville, TN, USA: IEEE, 2021: 475-484.

［23］Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings

of the 37th International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2020: 1597-1607.

[24] Tian Y L, Krishnan D, Isola P. Contrastive multiview coding[C]//Computer Vision-ECCV 2020: 16th European Conference, August 23-28, 2020, Glasgow, UK. Cham, Switzerland: Springer International Publishing, 2020: 776-794.

[25] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA: IEEE, 2017: 4724-4733.

[26] Zhu L C, Yang Y. Compound memory networks for few-shot video classification[C]//Computer Vision-ECCV 2018: 15th European Conference, September 8-14, 2018, Munich, Germany. Cham, Switzerland: Springer, 2018: 782-797.

[27] Thatipelli A, Narayan S, Khan S, et al. Spatio-temporal relation modeling for few-shot action recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA：IEEE, 2022: 19926-19935.

[28] Xing J Z, Wang M M, Liu Y, et al. Revisiting the spatial and temporal modeling for few-shot action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 3001-3009.

[29] Wang X, Zhang S W, Qing Z W, et al. MoLo: motion-augmented long-short contrastive learning for few-shot action recognition [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 17-24, 2023, Vancouver, BC, Canada: IEEE, 2023: 18011-18021.

[30] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]//IJCAI'81: Proceedings of the 7th International Joint Conference on Artificial Intelligence, August 24-28, 1981, Vancouver, BC, Canada. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1981, 2: 674-679.

[31] 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: IEEE, 2009: 248-255.

[32] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA: IEEE, 2016: 770-778.

[33] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. 2014: 1412.6980 [2024-01-28]. https://arxiv.org/abs/1412.6980.

（编辑　吕建斌）