Computer Science and Application 计算机科学与应用, 2021, 11(2), 423-434
Published Online February 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.112042

Hans汉斯

视频社会关系识别的多尺度图推理模型

许　飞，张天雨，史俊彪

合肥工业大学计算机科学与信息工程学院，安徽　合肥
Email: 1445258343@qq.com, 528802333@qq.com, 571755138@qq.com

收稿日期：2021年1月22日；录用日期：2021年2月17日；发布日期：2021年2月24日

摘　要

人类社会关系识别作为视频分类中的一个重要问题，逐渐成为计算机视觉领域的一个研究热点。由于视频信息较多，冗余信息过量，关键帧较少，因此如何准确的识别视频中的关键信息进行社会关系推理至关重要。为此，本文提出一种多尺度图推理模型来进行视频社会关系识别。首先我们提取视频中的时空特征和语义对象信息，获得丰富、鲁棒的社会关系表示。接着通过多尺度图卷积利用不同的感受野来进行时间推理，捕捉人物和语义对象间的交互。特别地，我们利用注意力机制来评估每个语义对象在不同场景的效果。在SRIV数据集上的实验结果表明，本文提出的方法优于大多数先进的方法。

关键词

社会关系识别，多尺度图卷积，注意力机制

Multi-Scale Graph Reasoning Model for Video Social Relation Recognition

Fei Xu, Tianyu Zhang, Junbiao Shi

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui
Email: 1445258343@qq.com, 528802333@qq.com, 571755138@qq.com

Received: Jan. 22nd, 2021; accepted: Feb. 17th, 2021; published: Feb. 24th, 2021

Abstract

As an important issue in video classification, human social relationship recognition has gradually become a research hotspot in the field of computer vision. Due to the large amount of video information, excessive redundant information and less key frames, how to accurately identify the key information in the video and carry out social relation reasoning is of great importance. To this end, this paper proposes a multi-scale graph reasoning model to identify video social relationships. First,

文章引用: 许飞, 张天雨, 史俊彪. 视频社会关系识别的多尺度图推理模型[J]. 计算机科学与应用, 2021, 11(2): 423-434. DOI: 10.12677/csa.2021.112042

we extract the temporal and spatial features and semantic object information in the video to obtain a rich and Lupin representation of social relations. Then use different receptive fields to perform temporal reasoning through multi-scale graph convolution, and capture the interaction between characters and semantic objects. In particular, we use the attention mechanism to evaluate the effect of each semantic object in different scenarios. The experimental results on SRIV dataset show that the method proposed in this paper is superior to most advanced methods.

Keywords

Social Relation Recognition, Multi-Scale Graph Convolution, Attention Mechanism

1. 引言

社会关系是多个个体之间的紧密联系，并构成我们社会的基本结构。从图像或视频中识别社交关系可以使机器更好地理解人类的行为或情感。然而，与基于图像的社会关系识别相比，基于视频的场景是一个重要但前沿的话题，常常被社会团体所忽视。它具有许多潜在的应用，例如帮助人们在手机中查找家庭视频[1]，或者向商店中的顾客群推荐合适的产品[2]。

现有的社会关系识别研究主要集中在基于图像的条件下，算法主要识别单个图像中人与人之间的社会关系。为了区分不同的社会关系，研究了人和语境对象的外观和面部属性。尽管在视频或电影中发现了社交网络[3] [4]，社区，角色[5] [6]和群组行为[7]，但从视频片段中明确认识到社会关系的吸引力却远远不足注意。最近的方法仅将基于视频的社交关系识别视为一般的视频分类任务，该任务将 RGB 帧，光流或视频音频作为输入，并将视频片段分类为预定类型[8]。但是，这种模型显然过于简化，从而忽略了人的外观，人与语义对象之间的交互以及带有上下文对象的场景。如何解读场景中的众多特征，视频中的社会关系识别面临着独特的挑战。首先，与社会社区发现相比，在不同的场景中，社会关系更加细化和模糊。模型必须通过视觉内容来区分非常相似的社会关系，例如朋友和同事，即使对于人类来说，这也可能非常困难。此外，与基于图像的社会关系识别相反，人和语境对象可能出现在任意视频帧中，甚至出现在单独的视频帧中。这使得人和语境对象在连续帧中变化极大。因此，基于图像的方法不能直接用于基于视频的场景。此外，视频还提供了人和语义对象的时域特征。对人的动态变化与社会关系之间的潜在关联进行建模仍然具有很大的挑战性。

为此，我们提出了一个多尺度图推理模型(MSGRM)来解决视频中的社会关系理解问题。在特征提取阶段，利用特征提取网络提取场景的时空特征和语义对象特征。然后在多尺度图推理阶段，利用不同的感受野来学习长期和短期信息，以探索场景中人和语义对象之间的交互。此外，利用注意机制，通过测量每个节点的重要性，自适应地选择某一视频场景中最重要的节点进行识别。这样，MSGRM 极大地提高了从视频中获取社会关系的能力。本文主要贡献如下：

1) 本文提出了一种多尺度图推理模型(MSGRM)来识别视频中的社会关系，在端到端的处理过程中，该方法可以准确地捕捉场景中角色的时空信息和交互信息。

2) 为了捕捉视频中的长期和短期时间线索，本文提出了一种基于多尺度时间感受野的 MSGCN 进行

社会关系推理，以捕捉视频中的长期和短期线索。

3) 我们将该方法应用于 SRIV 数据集，并与一些优秀的研究工作进行了比较，取得了较好的识别效果。

2. 相关工作

视频中的社会关系识别。在过去的十年中，社会学和计算机视觉的跨学科研究一直是热门领域。主要的研究主题包括社交网络发现[3] [4]、关键角色检测[5] [6]、多人跟踪[9] [10]和群体行为识别[7]。近年来，基于视觉内容的社会关系识别引起了研究者的关注[11] [12] [13] [14]，现有的方法主要集中在静态图像上。例如，Zhang 等提出通过卷积神经网络(CNN)从人脸图像中学习社会关系特征[12]。Sun 等提出了一种基于社会领域理论的社会关系数据集[15]，并采用 CNN 从一组语义属性中识别社会关系[13]。Li 等提出了一种用于社交关系识别的双视模型，其中第一眼聚焦感兴趣的人，第二眼应用注意力机制发现上下文线索[11]。Wang 等人提出用图来表示图像中的人和物体，并用门控图神经网络[14]进行社会关系推理。而对于基于视频的数据，社交关系识别仅被视为视频分类任务。例如，Lv 等利用时间分段网络[16]，利用视频的 RGB 帧、光流和音频对视频进行分类[8]。他们还建立了一个视频社会关系(SRIV)数据集，其中包含约 3000 个带有多标签注释的视频片段。但是，该方法只考虑全局和粗糙特征，而忽略了视频中的人、对象和场景。因此，我们将视频中的人与对象的时空特征特征嵌入图模型，并在此基础上进行社会关系推理。

计算机视觉中的图模型。在计算机视觉领域，像素，区域，概念和先验知识可以表示为图形，以针对不同任务(例如目标检测[17]，图像分割[18]，图像搜索[19]等)和对它们的关系进行建模。近年来，机器学习的研究人员研究了通过端到端可训练网络在图中进行消息传播，如图卷积网络(GCN) [20] [21]和门控图神经网络(GGNN) [22]。最近，这些模型已被用于计算机视觉任务[14] [23] [24] [25]。例如，Liang 等提出了一个“图形长短期记忆网络”在基于超像素的图形中传播信息，并用于语义对象解析[24]。Qi 等提出了一种 3D 图神经网络在 3D 点云上建立一个 k 近邻图，并预测 RGBD 数据每个像素的语义类别[25]。Wang 等提出用视频中的人物和对象将视频表示为时空图，并采用 GCN 来学习视频级特征以进行动作识别[26]。受以上研究的启发，我们建议将视频中人和物体之间的相互作用用图形来表示，并通过我们提出的多尺度图推理模型网络来进行社会关系识别。

注意力模型。人在观看某物时，总是关注感兴趣的视觉信息。一些研究发现，视觉注意力被信息含量最高的区域[27] [28]所吸引。在深度学习领域，注意力机制已应用于视频描述[29]，图像和动作分类[30] [31]以及文本中的实体歧义消除[26]，以学习数据的更多关键部分。一方面，基于 CNN 的注意力模型已经被提出并应用于不同的领域，这些方法比没有注意力模型取得了更优异的成绩。例如，Yu 等人[32]引入了注视编码注意网络(GEAN)，该网络可以利用注视跟踪信息为视频字幕提供时空关注。Zhu 等[30]提出了一种空间正则化网络，利用注意力机制学习不同标签的更多相关区域。另一方面，注意机制也被应用于序列学习模型中。如 Pei 等人[33]提出了不同的注意力 GRU 模型，可以学习顺序数据的注意得分。然而，这些注意模型忽略了语义对象与特定视频时空特征之间的相关性。因此，在我们的社会关系识别模型中，提出了一种时空注意力机制，从视频中自适应地选择最有区别性的对象来理解社会关系。

3. 多尺度图推理模型

我们的多尺度图推理模型的总体架构主要包含两部分，第一部分是从原始数据提取语义对象以构建图。该框架将一个视频帧作为输入，为了建模人和物体的时空特征和探究人和物体间的交互信息，我们构建了一个人—物图和上下文对象的共存，并用 LSTM 和 ResNet [34]来提取人和物体的时空特征。第二

个部分采用 MSGCN 来进行关系推理，在每个图中进行消息传播。在 MSGCN 中，我们探索多尺度的时间感受野来学习不同时间范围的相互作用。并利用注意力机制来探究场景中的语义对象对社会关系识别的重要性。图 1 给出了所提出模型的总体图解。

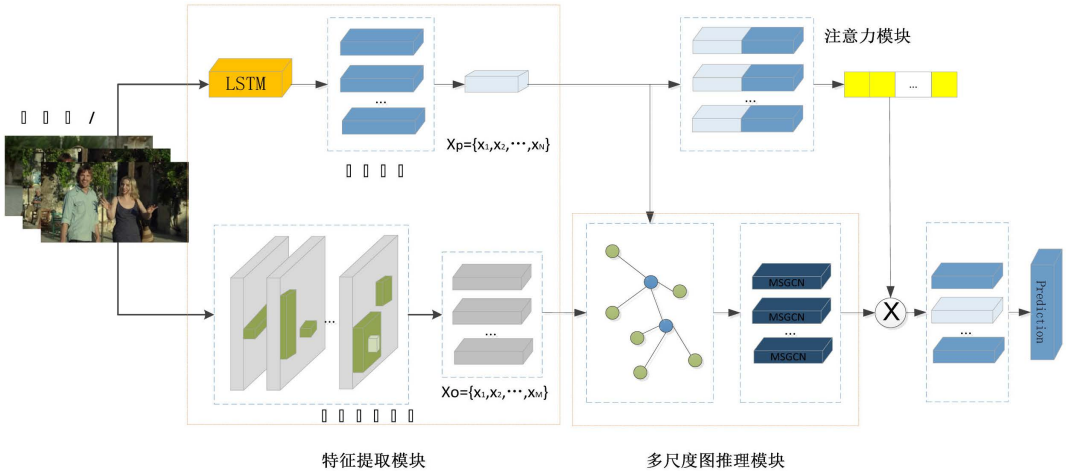

Figure 1. Multi-scale graph reasoning model structure
图 1. 多尺度图推理模型框架

3.1. 特征提取模块

时空特征提取。为了从视频中学习时间特征，我们遵循[35]采用 LSTM 单元对输入视频中采样的 L 个关键帧 $I=\{I_i \mid i=1,2,\cdots,L\}$ 进行处理，来生成具有时间社会关系的特征序列，表示为 $X_t=\{x_t^i \mid 1,2,\cdots,L\}$。然后，将这些时空特征展平并连接起来以形成单个特征向量。

语义对象特征提取。使用预先训练的检测器捕获整个视频中的语义对象区域，并从相应的语义对象中提取特征，我们使用 Faster R-CNN [36]检测器从采样的视频帧中检测视频中的人和物体对象 $P=\{p_i\}, i=1,2,\cdots,N$ 和 $O=\{o_j\}, j=1,2,\cdots,M$，该检测器是在 COCO [37]数据集上训练的。COCO 数据集是一种用于目标检测的大型数据集，涵盖了我们日常生活中经常出现的 80 个目标类别，用于从视频中收集语义对象。Faster R-CNN 使用区域建议网络(RPN)处理输入关键帧 I，生成一组具有高评分语义对象的区域建议。将检测到的置信度高于阈值 ε 的上下文区域 $O_t=\{O_t^j \mid j=1,2,\cdots,C\}$ 作为语义对象，其中 C 表示检测到的类别。为了平衡准确性和效率，我们通过置信度得分将每个视频帧固定为 N 个人和 M 个目标对象。每个边界框的外观特征由 VGG [38]网络来提取的。这些边界框被用作构建人-物图模型的节点，而每个结点的特征将在图卷积中用于社会关系推理。

3.2. 人－物图模型

图形模型可以有效地表示空间视觉内容中对象的时间、空间、概念或者相似性关系[19] [23]，为了捕获视频中不同人之间的交互和探究人物和上下文对象之间的互动，我们构建一个人－物图模型 $G=(V,E)$ 来表示人际之间的交互和人与上下文对象之间的共存,其中 $V=(P,O)$ 是我们场景中的人和目标对象节点，用不同的颜色表示，E 表示节点之间的关系边。

对于建模人与人之间的交互，我们通过估计视频帧及其相邻帧中人的距离来构建图模型。对于人际之间的邻接矩阵 $A_{p-p}\in R^{N*N}$，如果人节点 P_i 和 P_j 是属于同一帧的，我们直接设置 $A_{p-p}(p_i,p_j)=1$。如果人节点 P_i 和 P_j 属于相邻帧，我们设置

$$A_{p-p}\left(p_i,p_j\right)=\begin{cases}1 & dist\left(p_i,p_j\right)\geq\tau\\ 0 & \text{othersize}\end{cases} \tag{1}$$

其中 $dist\left(p_i,p_j\right)=1-\dfrac{f\left(p_i\right)^{\tau}f\left(p_j\right)}{\left\|f\left(p_i\right)\right\|\cdot\left\|f\left(p_j\right)\right\|}$ 是人节点 P_i 和 P_j 之间的余弦距离，τ 是我们设置的超参数。

同样，场景中的上下文对象是社交关系识别的重要信息，为了捕获视频中人物和上下文对象之间的互动。我们通过估计人物和上下文对象在视频帧中的共存来构建图模型。对于人和物之间的邻接矩阵 $A_{p-o}\in R^{(N+M)*(N+M)}$，如果 P_i 和 O_j 来自同一帧，则设置 $A_{p-o}\left(p_i,o_j\right)=1$，否则设置 $A_{p-o}\left(p_i,o_j\right)=0$，公式如下：

$$A_{p-o}\left(p_i,o_j\right)=\begin{cases}1 & \cap\left(p_i,o_j\right)\\ 0 & \text{othersize}\end{cases} \tag{2}$$

其中 $\cap\left(p_i,o_j\right)$ 表示 P_i 和 O_j 来自同一个视频帧。为了方便我们更加直观的进行图推理，我们把人际交互图和人物共存图整合在一个图上，如图 2 所示。

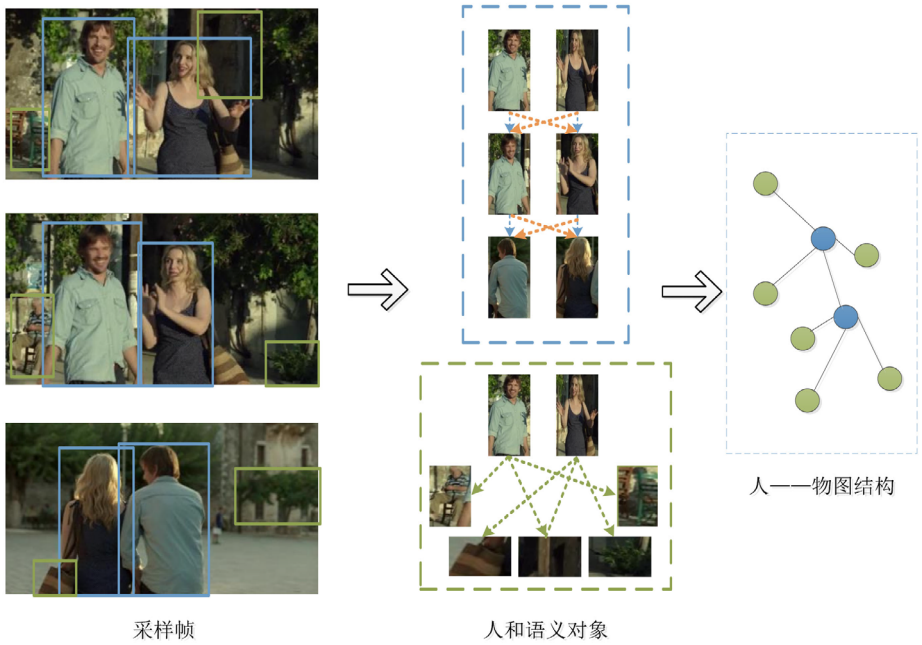

Figure 2. People-objects graph model structure
图 2. 人－物图模型结构

3.3. 多尺度卷积网络

图卷积网络(GCN)通过在图中从节点到其邻居进行消息传播来进行关系推理[20]。因此，我们可以在人－物图模型中应用 GCN 来实现视频帧中的社会关系推理。给定一个有 N 人个节点的图，其中每个节点都有一个 d 长度的特征向量，一个图卷积层的运算可以表示为

$$X^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}W^{(l)}\right) \tag{3}$$

其中 $\tilde{A}\in R^{N*N}$ 是人－物关系图的邻接矩阵，$\tilde{D}\in R^{N*N}$ 是 $\tilde{A}\in R^{N*N}$ 的度矩阵，$X^{(l)}\in R^{N*d}$ 是第 $(l-1)$ 的输

出结果，$W^{(L)} \in R^{d*d}$ 为可学习参数矩阵，$\sigma(\bullet)$ 是一个类似 ReLU 的非线性激活函数。特别说明，在我们的社会关系推理模型中，上式中的邻接矩阵为我们在 3.2 节中定义的 A_{p-p} 和 A_{p-o}。邻接矩阵的索引是按照视频中节点的时间顺序排列的，通过这个顺序，时间信息被隐式地嵌入到构建的图中。初始特征矩阵可表示为 $X^{(0)}=\left[x_p(i)\,|\,i=1,2,\cdots,N;x_o(j)\,|\,j=1,2,\cdots,M\right]^{\mathrm{T}}$，其中 $x_p(i)$ 和 $x_o(j)$ 是从视频中人和物体对象节点中提取的特征向量。GCNs 的最终输出是图中节点的更新特征，这些特征可以聚合成视频级的特征向量用于社会关系预测。

GCN 在一幅图中的所有节点上以及视频的整个时间范围上执行操作，这意味着 GCN 可以在时间域捕获全局视图。然而，社会关系识别的关键因素(如一个人的特定行为)可能出现在被不重要信息淹没的局部时间位置。因此，我们设计了一个多尺度图卷积网络(MSGCN)，通过不同的时间感受野来学习长期和短期信息。如图 3 所示为我们的多尺度卷积网络的一个块结构，每个块包含具有不同感受野的多个平行分支。Scale 1 是标准 GCN，它在整个相邻矩阵上执行图卷积并覆盖图中的所有节点。Scale 2 给出了具有较小时间感受野的图卷积的示例，而 Scale k 是更一般的说明。对于每个 Scale，所有滑动窗口的激活都汇总到一个特征矩阵中，该特征矩阵的形状与标准 GCN 的输出相同。通过沿着相邻矩阵的对角线滑动感受野，模型可以学习从视频的开始到结束的短期特征。最后，对多个尺度的输出进行平均池化合并，以生成下一个 MSGCN 层的特征矩阵 $X^{(l+1)}$。经过多次交互后，节点消息已经通过图进行传播，我们可以得到每个节点最终的状态为

$$Y=\left\{y_1,y_2,\cdots,y_{N+M}\right\} \tag{4}$$

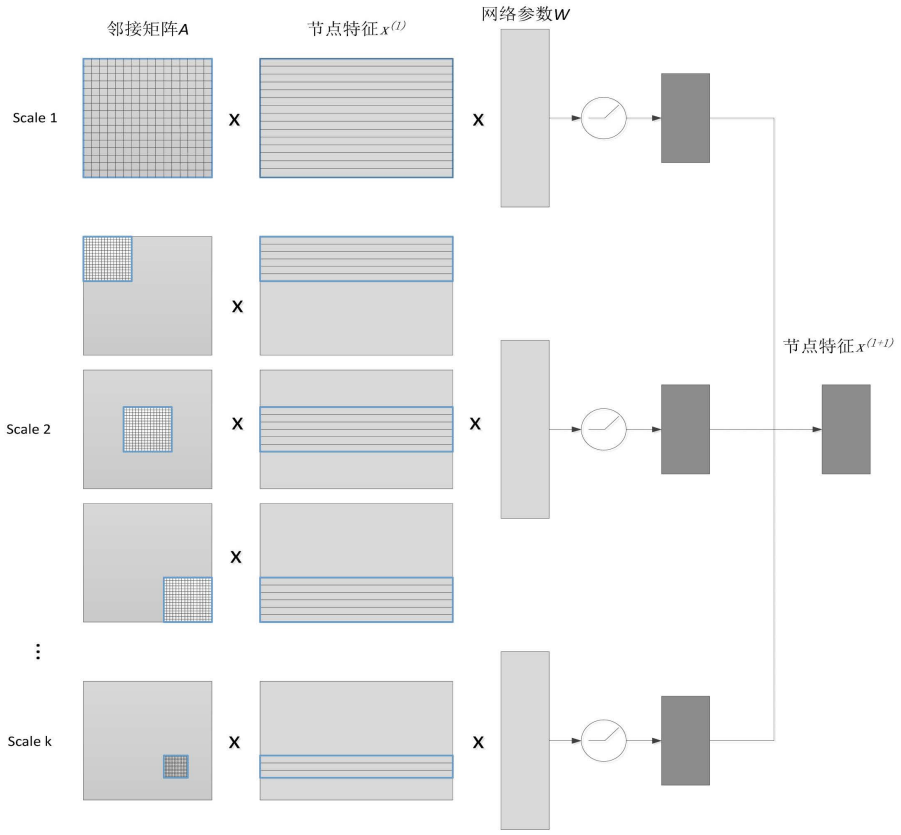

Figure 3. Graph convolution block with multi-scale receptive field
图 3. 具有多尺度感受野的图卷积块

3.4. 注意力机制

计算每个节点的特征后，我们可以直接将它们通过 MSGCN 聚合起来进行关系识别。然而，在不同的视频场景中，语义对象对区分关系的贡献并不相同。为了解决这一问题，我们引入了一种新的注意力机制，根据图形结构和视频特征自适应地推理出最相关的上下文对象。对于每一个社会关系和邻居对象对，该机制将它们的场景外观特征作为输入，并计算出这个对象对关系的重要性。我们首先将每个图中对象节点的外观特征和视频时空特征结合成一个向量 $h_{i,j} \in R^{d*d}$

$$h_{i,j} = RELU\left(x_o + w_t \otimes x_t\right) \tag{5}$$

其中 $w_t \in R^{d*d}$ 是一个权重矩阵，$\otimes$ 表示按矩阵元素相乘。

然后，我们通过 sigmoid 函数来计算每一个对象节点的注意力系数 $a_{i,j} \in [0,1]$，

$$a_{i,j} = \frac{1}{1+\exp\left(-\left(W_{h,a}h_{i,j} + b_a\right)\right)} \tag{6}$$

其中 $W_{h,a} \in R^{1*k}$ 是一个权重矩阵，根据节点 j 对节点 i 的重要性的不同，可将每个特征转换为可用的表达性更强的特征，b_a 是一个偏置项。

对于关系 r_i，我们将其自人物节点的特征与上下文对象节点的加权特征向量连接起来作为其最终特征，

$$f_i = \left[y_{ri}, a_{i1}y_{o1}, a_{i2}y_{o2}, \cdots, a_{iM}y_{iM}\right] \tag{7}$$

然后由最后一层 fc 层对特征向量进行处理，生成关系得分：

$$s_i = Wf_i + b \tag{8}$$

表示视频场景具有社会关系 r_i 的可能性。对所有关系节点重复此过程，计算得分向量 $s = \{s_1, s_2, \cdots, s_N\}$。

整个网络通过交叉熵损失与地面真实标签 $\hat{s}$ 一起训练，

$$F_{loss}(\hat{s}, s) = \sum_{i=1}^{N} \hat{s}_i * \log(s_i) + (1-\hat{s}_i) * \log(1-s_i) \tag{9}$$

其中 s 是预测的类别概率。

4. 实验

4.1. 数据集

SRIV 数据集：本文使用的数据集来自于电影和电视剧，名为 SRIV [8]。SRIV 是第一个从视频中识别社会关系的视频数据集。它包含 3124 个带有多标签的视频，大约 25 个小时，这些视频来自 69 部电视剧和电影。数据集包含 Sub-Relation 和 Obj-Relation 类，其中包括 16 个子类，如表 1 所示。

4.2. 实验细节和评价标准

在特征提取模块，从视频中随机采样的关键帧 L 的数量设置为 128。类似于[11]，我们利用广泛使用的 ResNet-101 [34]提取关键帧的特征，得出的特征向量为 2048 维。对于语义对象区域，我们使用 VGG-16 [38]提取特征，从而得到 4096 维的特征向量。通常用于指导目标检测的阈值为 0.5，而此处语义对象检测的结果将很大程度影响人－物关系图中的特征交互，所以我们的阈值 ε 提高为 0.7，以获取更加准确的检测对象。在整个训练期间，除了 MSGCN 外，我们模型的所有组件使用 SGD 优化，MSGCN 使用 ADAM 优化。对于 SRIV 数据集，学习率 lr 从 0.01 开始，每 20 个 epochs 乘以 0.1，直到训练完 80 个 epochs。

本文采用四个评价标准来评价我们所提出的方法的性能。

Table 1. The statistics of the number for each class on SRIV
表 1. SRIV 上每种类别的统计数量

Sub-Relation			
Dominant	Competitive	Trusting	Warm
770	840	1614	1482
Friendly	Attached	Inhibited	Assured
2221	600	594	810
Obj-Relation			
Supervisor	Peer	Service	Parent
627	469	238	321
Mating	Sibling	Friendly	Hostile
600	141	1073	434

$\boldsymbol{F_1}$**_micro** 和 $\boldsymbol{F_1}$**_macro** 这两个评估基于是 F_1 分数的标签评估，第 i 类的 F_1 表示为

$$F_i(i) = 2*TP(i)/\left(2*TP(i)+FP(i)+FN(i)\right) \tag{10}$$

其中 $TP(i)$、$FP(i)$、$FP(i)$、$FN(i)$分别为第 i 类的正阳性、假阳性、真阴性、假阳性的个数，因此，F_1 _micro 和 F_1 _macro 的计算公式如下

$$F_1\_\text{macro} = \frac{1}{C}\sum_{i=1}^{C} F_1(i) \tag{11}$$

$$F_1\_\text{micro} = 2*\sum_{i=1}^{C} TP(i) \Big/ \left(2*\sum_{i=1}^{C} TP(i)+\sum_{i=1}^{C} FP(i)+\sum_{i=1}^{C} FN(i)\right) \tag{12}$$

其中 C 为类别数。

Accuracy 我们采用了 Zhang 等[12]提出的平衡精度，与以往的 accuracy 计算有所区别，我们充分考虑了样本数据中的不平衡性，使得最终的预测更符合实际结果，具体计算公式如下：

$$\text{Accuracy} = \frac{1}{2}\left(\frac{TP}{N_p}+\frac{TN}{N_n}\right) \tag{13}$$

其中 N_p 和 N_n 为阳性阴性样本数。

Subset Accuracy 由于我们的 sub-relation 类为主观感知的，分类标准更加严格细致，要求预测的标签集与样本真实标签集完全匹配，避免标签集中相似的标签干扰最终的预测，其具体公式如下：

$$\text{Subaccuracy}(s_i) = \frac{1}{n}\sum_{i=1}^{n} I\left(s_i = \hat{s}_i\right) \tag{14}$$

4.3. 消融实验

这里我们探究了我们对尺度图模型中不同模块的效果，实验结果如表 2 所示。从结果中我们发现，MSGRM 的整体准确率要高于 GCN，这表明多尺度感受野能够从长期和短期范围捕捉到有用的特征。此外，在有 Attention 模块辅助下的实验结果要高于没有 Attention 模块的结果，这说明注意力模块可以关注与社会关系识别相关的关键帧。

Table 2. The effect of different feature module
表 2. 不同功能模块的效果

Method	Accuracy	
	Sub-Relation	Obj-Relation
GCN	0.6725	0.6968
MSGRM	07154	0.7326
GCN + Attention	0.7369	0.7531
MSGRM + Attention	0.7756	0.7924

4.4. 与当前主流方法对比

为了验证所提出的多尺度图推理模型框架的有效性，我们在 SRIV 数据集上与几种最先进的方法进行了比较，实验结果如表 3、表 4 所示。具体方法如下：

Table 3. Performance of different methods on sub-relation class
表 3. Sub-relation 类上不同方法的性能

Method	F_{1_micro}	F_{1_macro}	Accuracy	Subaccuracy
C3D [39]	0.3958	0.3018	0.5568	0.1451
LSTM [39]	0.4714	0.4193	0.6547	0.3792
TSN [16]	0.6034	0.4894	0.5412	0.3045
Multi-stream [8]	0.7019	0.6383	0.6136	0.5291
STMV [35]	-	-	0.7535	0.5249
TSM [41]	-	-	0.8274	0.5936
ASRN [42]	**0.7353**	**0.6812**	0.6722	0.5392
MSGRM (Ours)	0.7124	0.6725	**0.7756**	**0.5824**

Table 4. Performance of different methods on Obj-relation class
表 4. Obj-relation 类上不同方法的性能

Method	F_{1_micro}	F_{1_macro}	Accuracy	Subaccuracy
C3D [39]	0.4383	0.3886	0.0557	0.0347
LSTM [40]	0.6780	0.5776	0.6667	0.2797
TSN [16]	0.7142	0.6142	0.7089	0.3482
Multi-stream [8]	0.8119	0.6683	0.7436	0.5213
STMV [35]	-	-	0.6322	0.5311
TSM [41]	-	-	0.7125	**0.6032**
ASRN [42]	**0.8141**	0.6766	0.7692	0.5259
MSGRM(Ours)	0.7945	**0.6941**	**0.7924**	0.5762

C3D [39]：提出了一种基于 3D 卷积的网络结构，该网络结构在视频特征提取中具有良好的性能。
LSTM [39]：基本的 LSTM 模型是一种流行的序列建模技术，具有各种改进。
TSN [16]：TSN 是一种典型的双流 CNN 网络，在许多视频分类数据集上都取得了最先进的性能。

Multi-stream [8]：使用代表社会关系的多个特征来提高识别性能。

STMV [35]：基于多视角(即 RGB，光流和面部)的融合模型，使用多个注意力单元来学习时空信息以进行社会关系理解。

TSM [41]：将语义对象提取、上下文交互和注意机制相结合的模型。

ASRN [42]：一种端到端的可训练模型，融合了多角度特征，如图像、运动、身体、人脸。

MSGRM (Ours)：这是我们所提出的多尺度图推理模型，它采用 MSGCN 学习场景中人物的多尺度动态，并融合了场景的时空注意力，实现社会关系推理。

表 3 和 4 显示了我们的模型与最先进的方法比较的结果。我们的 MSGRM 达到了比较领先的性能。这是因为通过不同尺度的图模型，学习了场景中不同感受野的信息，提取了视频中的关键序列特征，最后融合我们的时空注意力，促进了我们的社会关系识别。C3D、LSTM 和 TSN 的性能很差，这表明这些方法虽然可以更好地描述视频的其他一些特征，但却无法提取社会关系的正确表示。Multi-stream 和 STMV 都只关注视频的时空特征，因此很难获得更好的性能。TSM 和 ASRN 因为融合了场景中的各个角度的特征，这些特征很大程度上能表示场景的社会关系，所以性能有很大提升。

4.5. 实例可视化

注意力机制能为我们的多尺度图推理模型推理出最相关的上下文语义对象，如图 4 给出一些实例。图中左边为我们的原始采样帧，中间为我们的注意力机制生成的一系列热图，右边为我们的热图所对象的语义对象边界框。特征图显示了我们的注意力机制能过准确的捕获场景中重要的语义对象，因此能够进行有效的人－物特征交互，以提升社会关系识别的准确性。

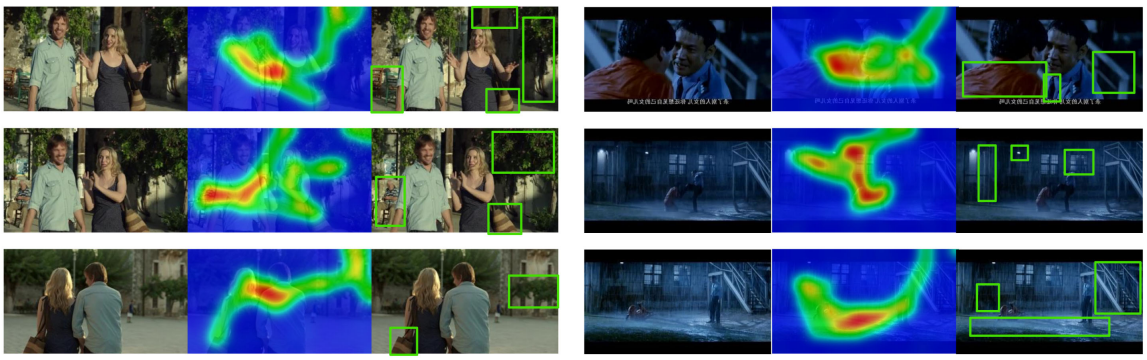

Figure 4. Scene attention visualization example
图 4. 场景注意力可视化实例

5. 结束语

在这项工作中，我们提出了一个多尺度图推理模型来解决视频中的社会关系识别问题，并引入特征提取模块以丰富视频中的时空特征表示。具体来说，我们的模型利用 MSGCN 来探索视频中人物与场景语义之间的交互，并通过不同的时间感受野来学习长期和短期信息。最后融合一种注意力机制，该机制测量场景中每个节点的重要性，以自适应地选择最重要的对象以提高社会关系的性能。在数据集 SRIV 上进行的大量实验证明，我们提出的多尺度图推理模型取得了优秀的表现。

参考文献

[1] Wang, G., Gallagher, A.C., Luo, J.B. and Forsyth, D.A. (2010) Seeing People in Social Context: Recognizing People

and Social Relationships. *European Conference on Computer Vision*, Glasgow, 23-28 August 2010, 169-182. https://doi.org/10.1007/978-3-642-15555-0_13

[2] Park, Y.-J. and Chang, K.-N. (2009) Individual and Group Behavior-Based Customer Profile Model for Personalized Product Recommendation. *Expert Systems with Applications*, **36**, 1932-1939. https://doi.org/10.1016/j.eswa.2007.12.034

[3] Ding, L. and Yilmaz, A. (2011) Inferring Social Relations from Visual Concepts. *IEEE International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 699-706. https://doi.org/10.1109/ICCV.2011.6126306

[4] Yu, T., Lim, S.-N., Patwardhan, K.A. and Krahnstoever, N. (2009) Monitoring, Recognizing and Discovering Social Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 1462-1469. https://doi.org/10.1109/CVPRW.2009.5206526

[5] Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A.N., Murphy, K. and Li, F.-F. (2016) Detecting Events and Key Actors in Multi-Person Videos. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3043-3053. https://doi.org/10.1109/CVPR.2016.332

[6] Ramanathan, V., Yao, B.P. and Li, F.-F. (2013) Social Role Discovery in Human Events. *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 2475-2482. https://doi.org/10.1109/CVPR.2013.320

[7] Bagautdinov, T.M., Alahi, A., Fleuret, F., Fua, P. and Savarese, S. (2017) Social Scene Understanding: End-to-End Multi-Person Action Localization and Collective Activity Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 3425-3434. https://doi.org/10.1109/CVPR.2017.365

[8] Lv, J.N., Liu, W., Zhou, L.L., Wu, B. and Ma, H.D. (2018) Multi-Stream Fusion Model for Social Relation Recognition from Videos. *International Conference on Multimedia Modeling*, Bangkok, 5-7 February 2018, 355-368. https://doi.org/10.1007/978-3-319-73603-7_29

[9] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Li, F.-F. and Savarese, S. (2016) Social LSTM: Human Trajectory Prediction in Crowded Spaces. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 961-971. https://doi.org/10.1109/CVPR.2016.110

[10] Choi, W. and Savarese, S. (2012) A Unified Framework for Multi-Target Tracking and Collective Activity Recognition. *European Conference on Computer Vision*, Florence, 7-13 October 2012, 215-230. https://doi.org/10.1007/978-3-642-33765-9_16

[11] Li, J.N., Wong, Y.K., Zhao, Q. and Kankanhalli, M.S. (2017) Dual-Glance Model for Deciphering Social Relationships. *ICCV* 2017, Palazzo del Cinema, 28 October 2017, 2669-2678.

[12] Zhang, Z.P., Luo, P., Loy, C.C. and Tang, X.O. (2015) Learning Social Relation Traits from Face Images. *ICCV*, Santiago, 7-13 December 2015, 3631-3639. https://doi.org/10.1109/ICCV.2015.414

[13] Sun, Q.R., Schiele, B. and Fritz, M. (2017) A Domain Based Approach to Social Relation Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 435-444.

[14] Wang, Z.X., Chen, T.S., Ren, J.S.J., Yu, W.H., Cheng, H. and Lin, L. (2018) Deep Reasoning with Knowledge Graph for Social Relationship Understanding. *International Joint Conference on Artificial Intelligence*, 1021-1028. https://doi.org/10.24963/ijcai.2018/142

[15] Bugental, D.B. (2000) Acquisition of the Algorithms of Social Life: A Domain-Based Approach. *Psychological Bulletin*, **126**, 187. https://doi.org/10.1037/0033-2909.126.2.187

[16] Wang, L.M., Xiong, Y.J., Wang, Z., Qiao, Y., Lin, D.H., Tang, X.O. and Van Gool, L. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. 14*th European Conference*, Amsterdam, 11-14 October 2016, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[17] Lin, L., Wang, X.L., Yang, W. and Lai, J.-H. (2015) Discriminatively Trained And-Or Graph Models for Object Shape Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 959-972. https://doi.org/10.1109/TPAMI.2014.2359888

[18] Felzenszwalb, P.F. and Huttenlocher, D.P. (2004) Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, **59**, 167-181. https://doi.org/10.1023/B:VISI.0000022288.19776.77

[19] Liu, W., Jiang, Y.-G., Luo, J.B. and Chang, S.-F. (2011) Noise Resistant Graph Ranking for Improved Web Image Search. *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, 20-25 June 2011, 849-856.

[20] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks.

[21] Defferrard, M., Bresson, X. and Vandergheynst, P. (2016) Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Advances in Neural Information Processing Systems*, Barcelona, 5-10 December 2016, 3837-3845.

[22] Li, Y.J., Tarlow, D., Brockschmidt, M. and Zemel, R.S. (2015) Gated Graph Sequence Neural Networks.

[23] Wang, X.L. and Gupta, A. (2018) Videos as Space-Time Region Graphs. *European Conference on Computer Vision*,

Munich, 8-14 September 2018, 413-431. https://doi.org/10.1007/978-3-030-01228-1_25

[24] Liang, X.D., Shen, X.H., Feng, J.S., Lin, L. and Yan, S.C. (2016) Semantic Object Parsing with Graph LSTM. *European Conference on Computer Vision*, Amsterdam, 8-16 October 2016, 125-143. https://doi.org/10.1007/978-3-319-46448-0_8

[25] Qi, X.J., Liao, R.J., Jia, J.Y., Fidler, S. and Urtasun, R. (2017) 3D Graph Neural Networks for RGBD Semantic Segmentation. *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 5209-5218.

[26] Phan, M.C., Sun, A.X., Tay, Y., Han, J.L. and Li, C.L. (2017) NeuPL: Attention-Based Semantic Matching and Pair-Linking for Entity Disambiguation. *Proceedings of the* 2017 *ACM on Conference on Information and Knowledge Management*, Singapore, 6-10 November 2017, 1667-1676. https://doi.org/10.1145/3132847.3132963

[27] Xu, J., Yao, T., Zhang, Y.D. and Mei, T. (2017) Learning Multimodal Attention LSTM Networks for Video Captioning. *Proceedings of the* 25*th ACM International Conference on Multimedia*, Mountain View, 23-27 October 2017, 537-545. https://doi.org/10.1145/3123266.3123448

[28] Li, Y., Miao, Z., He, M., Zhang, Y.F. and Li, H. (2018) Deep Attention Residual Hashing. *IEICE Transactions on Fundamentals of Electronics*, *Communications and Computer Sciences*, **101-A**, 654-657. https://doi.org/10.1587/transfun.E101.A.654

[29] Bin, Y., Yang, Y., Shen, F.M., Xie, N., Shen, H.T. and Li, X.L. (2019) Describing Video with Attention-Based Bidirectional LSTM. *IEEE Transactions on Cybernetics*, **49**, 2631-2641. https://doi.org/10.1109/TCYB.2018.2831447

[30] Zhu, F., Li, H.S., Ouyang, W.L., Yu, N.H. and Wang, X.G. (2017) Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2027-2036. https://doi.org/10.1109/CVPR.2017.219

[31] Girdhar, R. and Ramanan, D. (2017) Attentional Pooling for Action Recognition. *NIPS* 2017, Long Beach, 4-9 December 2017, 34-45.

[32] Rao, T.R., Li, X.X., Zhang, H.M. and Xu, M. (2019) Multi-Level Region-Based Convolutional Neural Network for Image Emotion Classification. *Neurocomputing*, **333**, 429-439. https://doi.org/10.1016/j.neucom.2018.12.053

[33] Pei, W.J., Baltrusaitis, T., Tax, D.M.J. and Morency, L.-P. (2016) Temporal Attention-Gated Model for Robust Sequence Classification.

[34] He, K.M., Zhang, X.Y., Ren, S.Q. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778.

[35] Lv, J.N. and Wu, B. (2019) Spatio-Temporal Attention Model Based on Multi-View for Social Relation Understanding. 25*th International Conference on Multi-Media Modeling*, Thessaloniki, 8-11 January 2019, 1-12.

[36] Ren, S.Q., He, K.M., Girshick, R.B. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems* (*NIPS* 2015), Vol. 28, 91-99.

[37] Lin, T.-Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. (2014) Microsoft COCO: Common Objects in Context. 13*th European Conference*, Zurich, 6-12 September 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[38] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[39] Tran, D., Bourdev, L.D., Fergus, R., Torresani, L. and Paluri, M. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. *ICCV*, Santiago, 13-16 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[40] Findler, N.V. (1972) Short Note on a Heuristic Search Strategy in Long-Term Memory Networks. *Information Processing Letters*, **1**, 191-196. https://doi.org/10.1016/0020-0190(72)90037-3

[41] Dai, P.L., Lv, J.N. and Wu, B. (2019) Two-Stage Model for Social Relationship Understanding from Videos. *ICME* 2019, Shanghai, 8-12 July 2019, 1132-1137. https://doi.org/10.1109/ICME.2019.00198

[42] Lv, J.N., Wu, B., Zhang, Y.L. and Xiao, Y.P. (2019) Attentive Sequences Recurrent Network for Social Relation Recognition from Video. *IEICE Transactions on Information and Systems*, **102-D**, 2568-2576. https://doi.org/10.1587/transinf.2019EDP7104