Domain-Aware Knowledge Debiasing for Generalizable Video Understanding in CLIP

用于CLIP视频理解的去偏领域知识引导方法

朱庆猛，吴琪欢，于志鹏，李懿，顾子茵，何灏

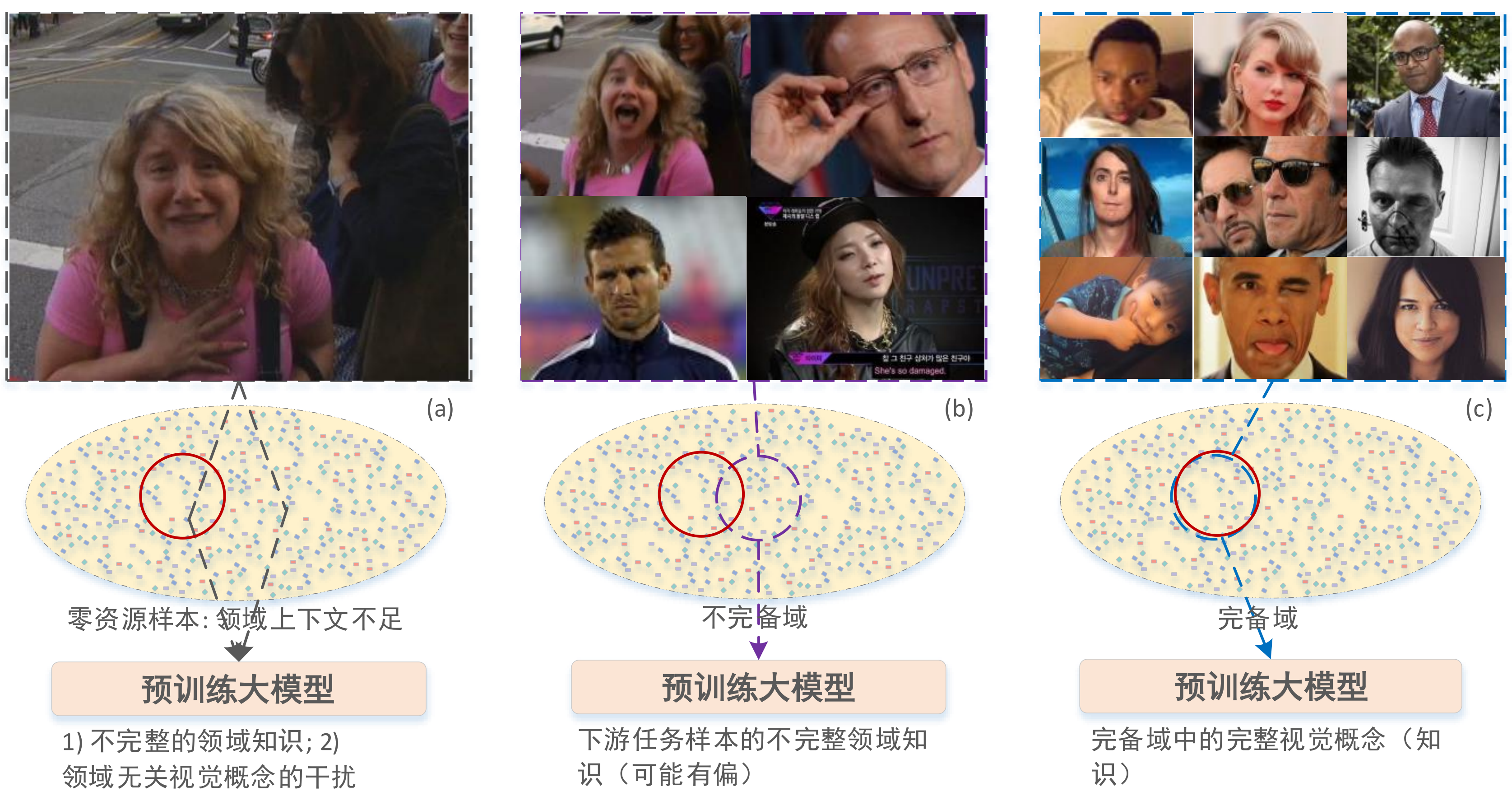
In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025), Hyderabad, India. IEEE, 2025: 1-5. 20250406-0411.

Contact Email: hehao21@iscas.ac.cn

背景介绍

CLIP作为多模态预训练模型的代表，在图像-文本匹配任务上取得了巨大成功。然而在实际应用中，CLIP经常出现混淆错误，例如将“跳伞”视频识别为“游泳”。

由于预训练模型中的海量知识和某特定领域的不完备知识，两者可能共同导致CLIP在通用知识表示空间与特定领域对应的知识分布匹配发生错误。通过领域数据微调CLIP，由于数据有偏或知识不完备，可能导致微调模型中的知识歧义。如果能利用完整的情感领域信息，则能有效地应对该问题。



方法原理

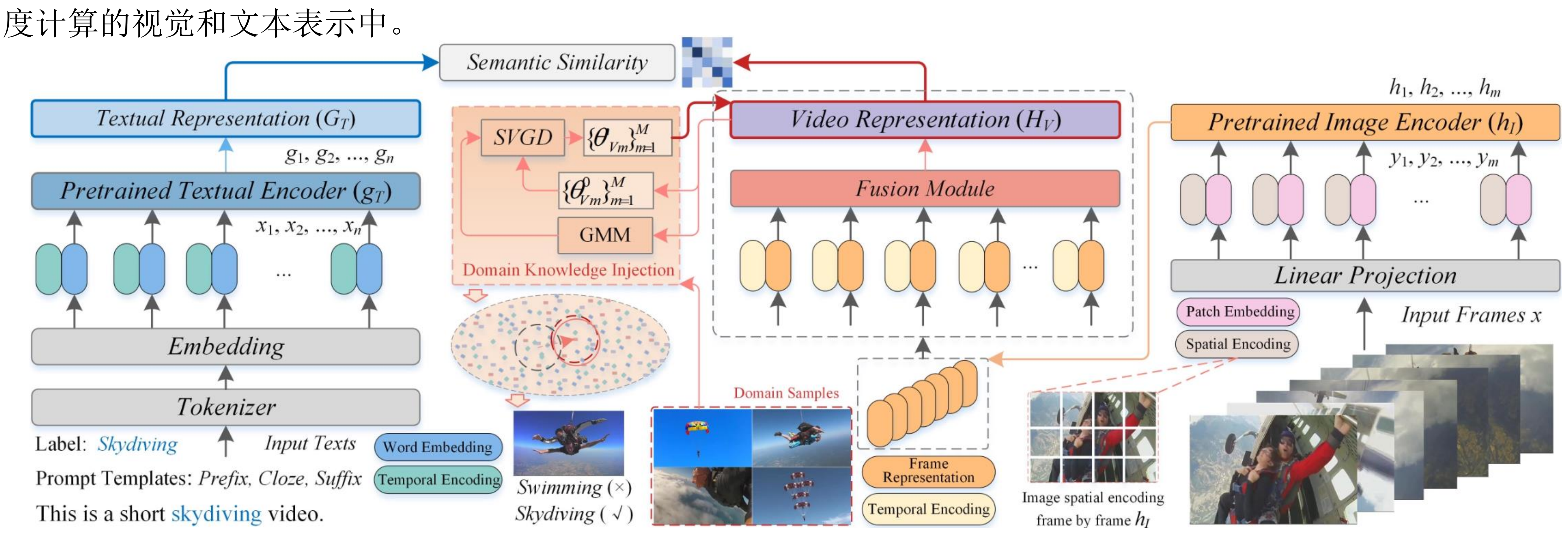
模型主要由预训练视觉编码器、预训练文本编码器和两个模态的领域知识注入模块构成。整体实现流程如下：

（1）通过预训练模型对图像分类领域的未标注数据进行推理，编码器输出表示记为θ。

（2）该输出θ用于获取下游域的去偏真实分布，即某特定领域图像分类的GMM，以得到混合高斯分布p(θ)的参数，包括均值μ、方差σ和混合系数λ。

（3）使用SVGD方法，迭代近似该去偏高斯混合分布，以得到包含图像分类领域知识的表示θ'。

（4）用θ'替换编码器的输出表示（包括视觉编码器和文本编码器），将这些情感相关知识注入到参与语义相似度计算的视觉和文本表示中。



实验结果

TABLE I
State-of-the-Art Zero-Shot Learning: Top-1 Accuracy Comparison on UCF101 and HMDB51 Benchmarks

	Method	Video	Class	UCF101	HMDB51
Intra	IAP [9]	FV	A	16.7	-
	HAA [10]	FV	A	14.9	-
	SJE [11]	FV	W_N	9.9	13.3
	MTE [12]	FV	W_N	15.8	19.7
	GA [13]	C3D*	W_N	22.7	-
	O2A [14]	Obj†	W_N	30.3	15.6
	CEWGAN [15]	I3D	A	38.3	-
	TS-GCN [14]	Obj+	W_N	34.2	23.2
	TARN [16]	C3D*	W_N	19.0	28.9
	PS-GNN [17]	Obj	W_N	36.1	29.5
	DASZL [18]	TSM	A	48.9	-
	ED [19]	(st+Obj)†	ED	51.8	35.3
Cross	URL [20]	R200	W_N	42.5	28.9
	E2E [21]	R(2+1)_18*	W_N	46.1	33.1
	DeCalGAN [22]	(R(2+1)_18+Obj)*	W_N	51.4	36.1
	DKD	**VisualTransformer**	**W_N**	**67.64**	**45.34**

TABLE II
State-of-the-Art Fine-Tune Learning: Top-1 Accuracy Comparison on UCF101 and HMDB51 Benchmarks

	Method	Pretrain	Frames×Crops ×Clips	HMDB 51	UCF 101
Compressed	CoViAR [23]	ImageNet	25×10×1	59.1	90.4
	CoViFocus+Flow [23]	ImageNet	25×10×1	70.2	94.9
	Refined-MV [24]	ImageNet	25×10×1	59.7	89.9
	TTP [25]	ImageNet	25×10×1	58.2	87.2
	IP TSN [26]	ImageNet	16×1×1	69.1	94.3
	DMC-Net [27]	ImageNet	25×1×1	71.8	92.3
	MFCD-Net [28]	Kinetics	12×1×15	66.9	93.2
	SIFP-Net [29]	Kinetics	32×3×10	72.3	94.0
	CoViFocus [30]	Kinetics	8×1×1	74.4	95.5
	CoViFocus [30]	Kinetics	8×10×1	74.8	95.8
Image	TSN [31]	ImageNet	25×10×1	68.5	94.0
	AdaFocus [32]	Kinetics	16×1×1	69.2	94.5
	13D [33]	Kinetics	32×3×10	74.8	95.6
	R(2+1)D [34]	Sports-IM	32×1×10	74.5	96.8
	ARTNet [35]	Kinetics	16×10×25	70.9	94.3
	STM [36]	Kinetics	16×3×10	72.2	96.2
	TSM [37]	Kinetics	16×3×10	73.2	96.0
	TEINET [38]	Kinetics	16×3×10	72.1	96.7
	TEA [39]	Kinetics	16×3×10	73.3	96.9
	DKD	**Kinetics**	**8×1×1**	**76.26**	**96.96**

本文在视频理解数据集UCF101和HMDB51上进行了全面评估。UCF101包含13,320个视频，涵盖101类人类动作；HMDB51包含6,766个视频，涵盖51类日常动作。

零样本学习实验：在不使用任何训练数据的情况下，DKD方法在UCF101上达到67.64%的准确率，比第二名DeCalGAN高出16.24%；在HMDB51上达到45.34%，比第二名高出8.74%。这种显著提升证明了领域知识注入的有效性，特别是在处理“跳伞-游泳”这类易混淆领域场景时表现突出。

微调学习实验：当进行微调训练时，DKD使用更少的计算资源就达到了96.96%和76.26%的准确率，超越了使用更密集采样的复杂模型。这表明DKD不仅性能优异，还具有更高的计算效率。

创新点和意义

即插即用设计：DKD模块可以以即插即用方式插入预训练模型，无需修改原始架构。

领域知识注入：通过高斯混合模型(GMM)建模领域知识分布，使用SVGD算法动态注入。

多模态增强：同时优化文本和视觉两个分支，利用跨模态信息增强理解。

自适应去偏：根据输入样本的特征，动态调整领域知识注入强度。

轻量级实现：推理时几乎不增加额外计算开销，适合实时应用。