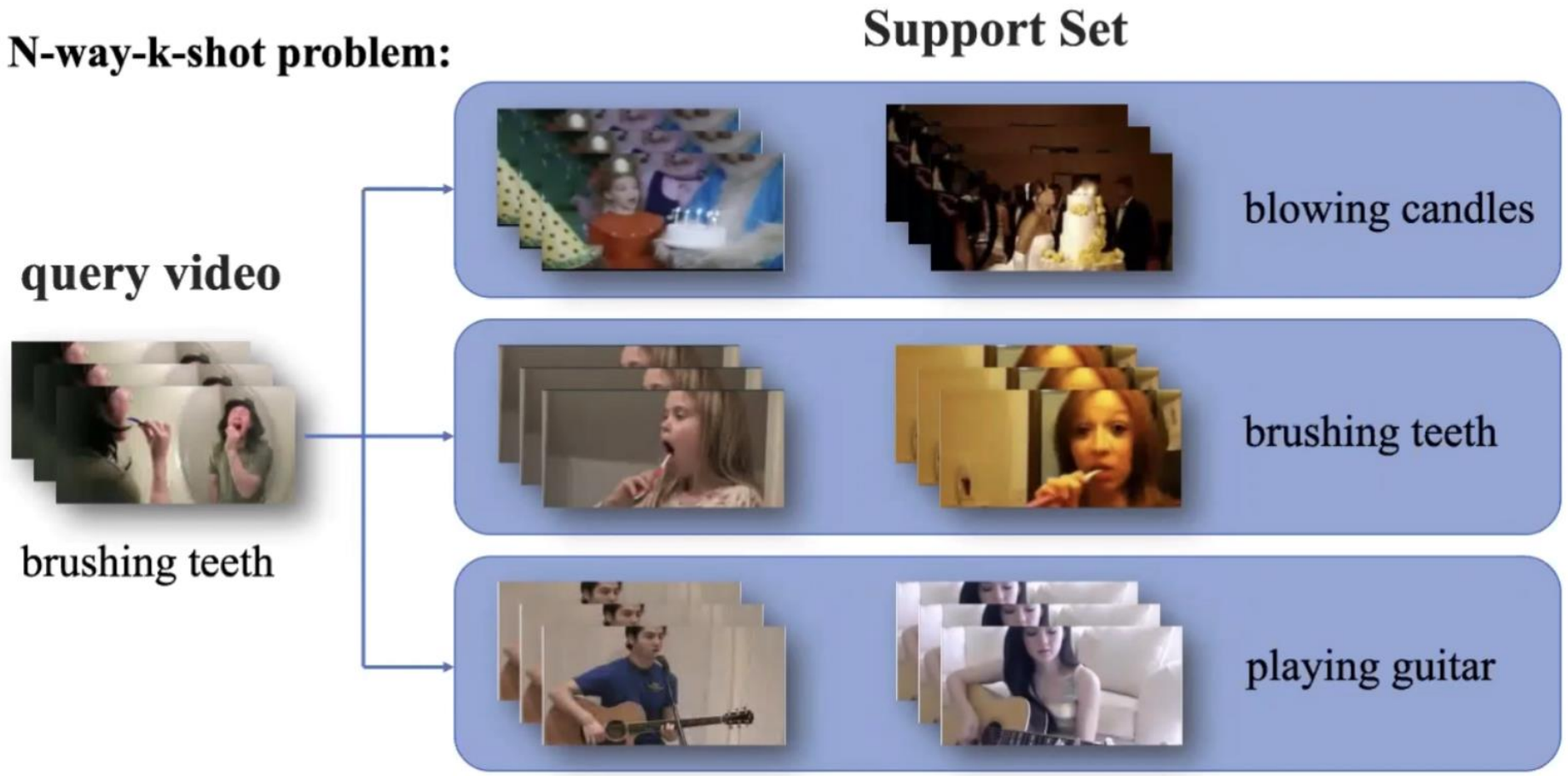
Active Exploration of Multimodal Complementarity for Few-Shot Action Recognition

Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, Changsheng Xu
wanyanyuyang2021@ia.ac.cn, xiaoshan.yang@nlpr.ia.ac.cn, chencfbupt@gmail.com, csxu@nlpr.ia.ac.cn

State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA)

Few-shot action recognition

Recognize the action of the query video with the help of only a few annotated samples.



Background

- **Few-shot action recognition (FSAR)** receives increasing attention and achieves remarkable progress.
- These methods can be roughly divided into two groups: **data augmentation-based** methods and **alignment-based** methods.

Data augmentation:
- ARN (ECCV 2020)
- AMeFu-Net (MM 2020)
- MTFAN (CVPR 2022)
- … …

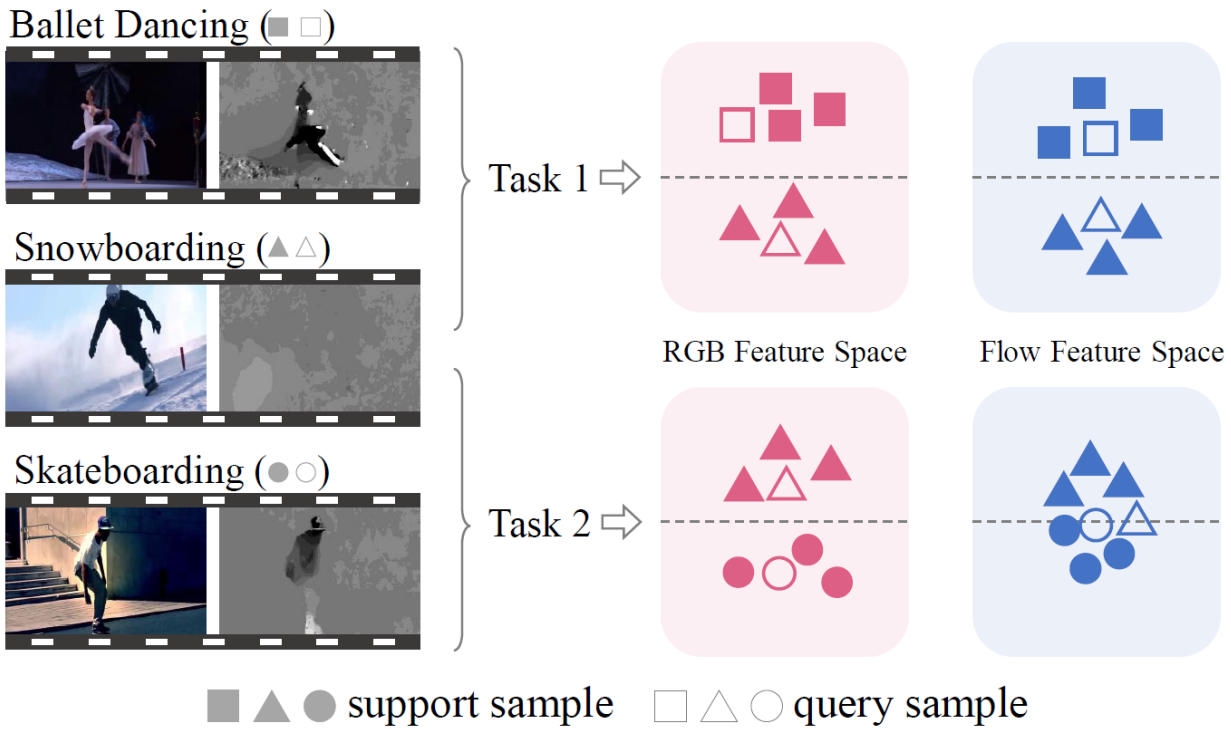
Alignment-based:
- OTAM (CVPR 2020)
- TRX (CVPR 2021)
- STRM (CVPR 2022)
- … …

- Previous methods mainly rely on limited **unimodal data** (e.g., RGB frames) while the multimodal information remains relatively underexplored.

Motivation

- The **contribution of a specific modality** is not consistent for different query samples and it highly depends on the **contextual information** in each few-shot task.

- It requires a **task-dependent strategy** for exploring the complementarity between different modalities in few-shot action recognition.



Active Multimodal Few-shot Action Recognition

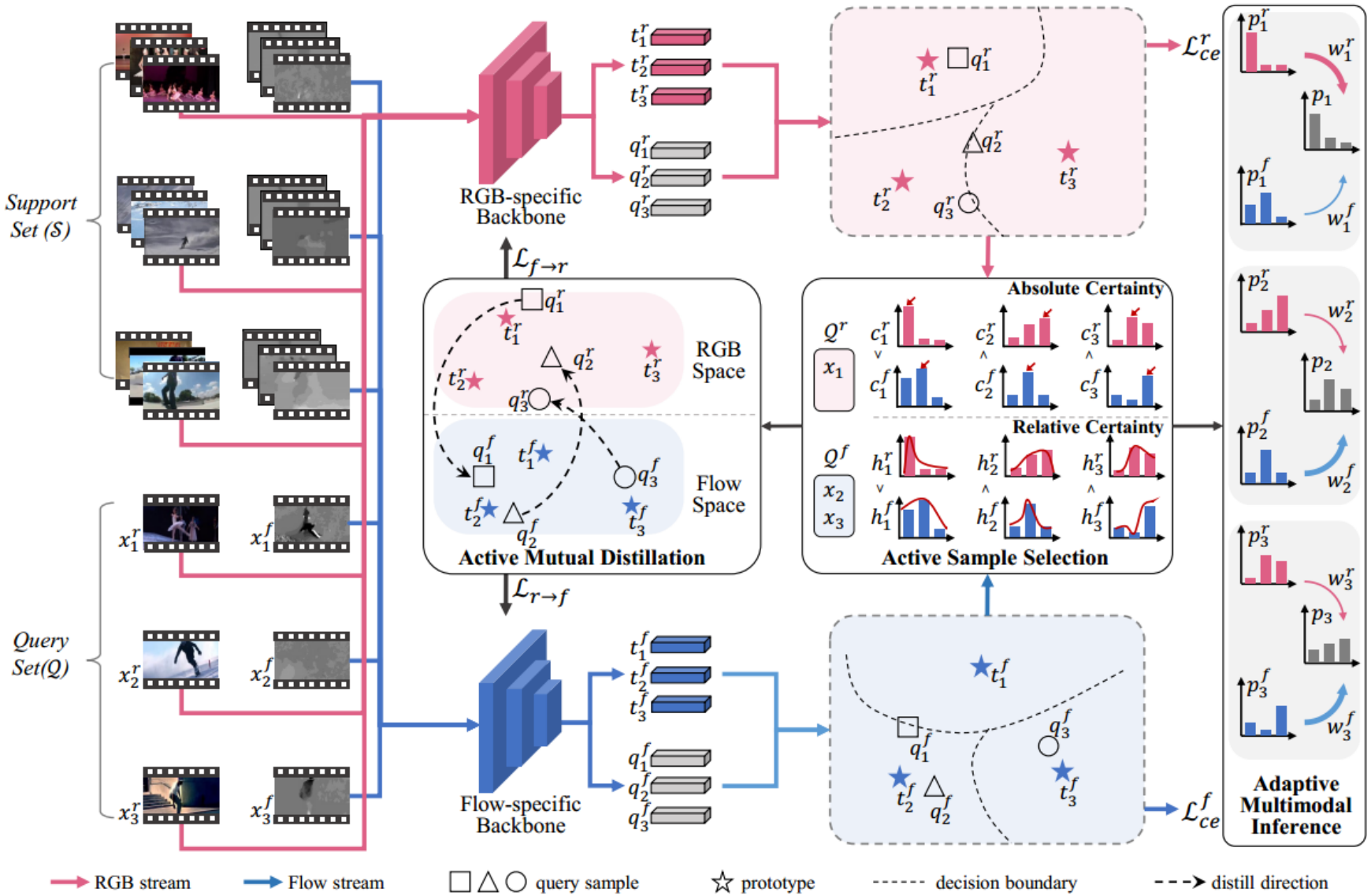


Illustration of the proposed AMFAR framework in the 3-way 3-shot setting.

Active Multimodal Few-shot Action Recognition

Active Sample Selection (ASS)

organizes query samples with large differences in the **reliability** of modalities into different groups.

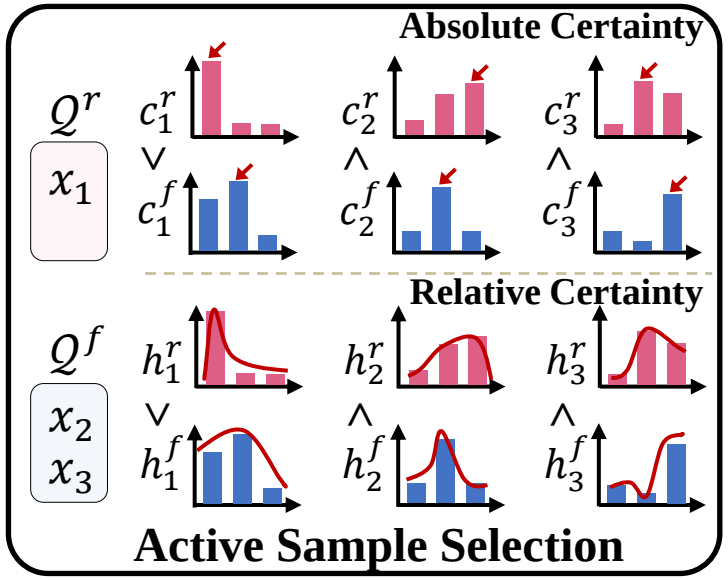
Absolute Certainty

$$c_i^m = \max_k \mathcal{P}(\hat{y}_i = k | x_i^m). \quad (1)$$

Relative Certainty

$$h_i^m = \sum_{k=1}^{N} \mathcal{P}(\hat{y}_i = k | x_i^m) \log \mathcal{P}(\hat{y}_i = k | x_i^m). \quad (2)$$

$$\begin{aligned} \mathcal{Q}^r &= \left\{ (x_i^r, x_i^f) \mid (x_i^r, x_i^f) \in \mathcal{Q}, c_i^r > c_i^f, h_i^r > h_i^f \right\}, \\ \mathcal{Q}^f &= \left\{ (x_i^r, x_i^f) \mid (x_i^r, x_i^f) \in \mathcal{Q}, c_i^f > c_i^r, h_i^f > h_i^r \right\}, \end{aligned} \quad (3)$$



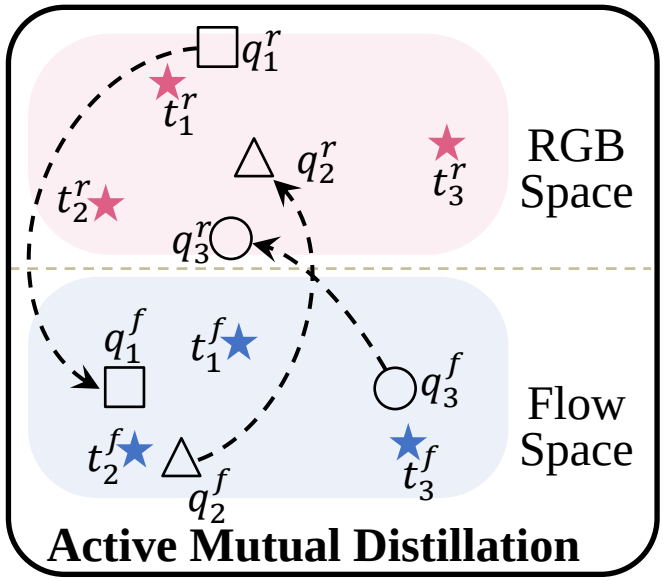
Active Multimodal Few-shot Action Recognition

Active Mutual Distillation (AMD)

captures discriminative **task-specific knowledge** from the reliable modality to improve the representation learning of unreliable modality by **mutual distillation**.

$$\mathcal{L}_{f\to r}(\theta^r) = \frac{1}{\sum_{(x_i^r,x_i^f)\in\mathcal{Q}^f} c_i^f} \sum_{(x_i^r,x_i^f)\in\mathcal{Q}^f} c_i^f \mathcal{D}_{KL}(p_i^f, p_i^r),$$

$$\mathcal{L}_{r\to f}(\theta^f) = \frac{1}{\sum_{(x_i^r,x_i^f)\in\mathcal{Q}^r} c_i^r} \sum_{(x_i^r,x_i^f)\in\mathcal{Q}^r} c_i^r \mathcal{D}_{KL}(p_i^r, p_i^f), \quad (5)$$



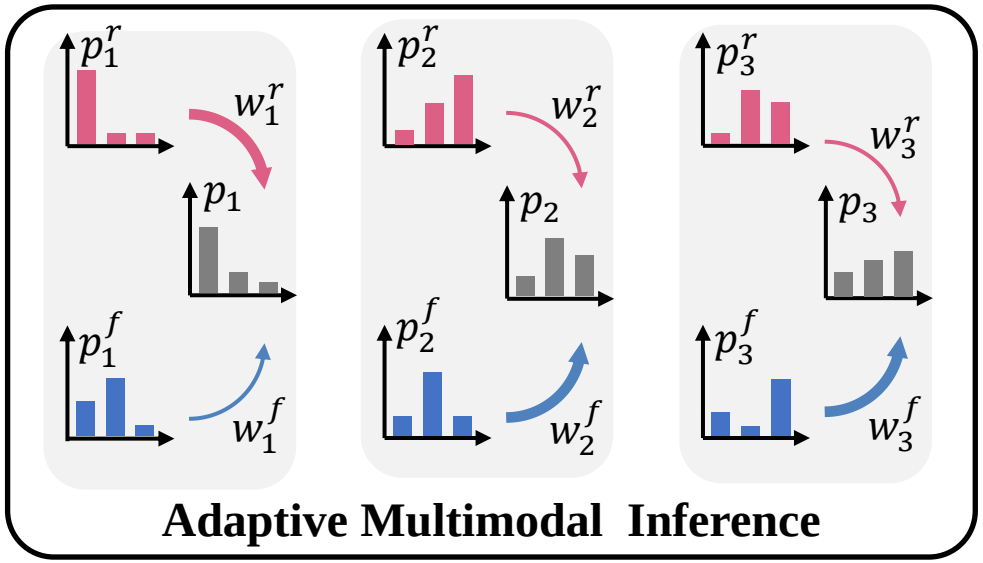
Active Multimodal Few-shot Action Recognition

Adaptive Multimodal Inference (AMI)

adaptively fuses the modality-specific posterior distributions with a larger weight on the **reliable modality**.

$$\mathcal{P}(\hat{y}_i = k | x_i^r, x_i^f) = \frac{\exp\left(-w_i^r \psi(q_i^r, t_k^r) - w_i^f \psi(q_i^f, t_k^f)\right)}{\sum_{k'=1}^{N} \exp\left(-w_i^r \psi(q_i^r, t_{k'}^r) - w_i^f \psi(q_i^f, t_{k'}^f)\right)}, \tag{6}$$

$$w_i^r = \frac{c_i^r}{c_i^r + c_i^f}, \quad w_i^f = \frac{c_i^f}{c_i^f + c_i^r}. \tag{7}$$



Experiments

Table 1. Comparison with state-of-the-art few-shot action recognition methods. We use † to mark methods that are re-implemented by ourselves. For multimodal approaches extended from existing unimodal methods, "EC" denotes the early fusion scheme of concatenation, "EA" denotes the early fusion scheme of Co-Attention, and "LF" denotes late fusion. "-" means the result is not available in published works.

Modality	Method	Kinetics		SSv2		HMDB51		UCF101	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
RGB	Matching Net [74]	53.3	78.9	-	-	-	-	-	-
	ProtoNet † [53]	55.5	84.6	26.7	53.3	45.2	71.9	70.9	94.4
	MAML [74]	54.2	78.9	-	-	-	-	-	-
	CMN [74]	60.5	78.9	-	-	-	-	-	-
	TARN [5]	66.6	78.5	-	-	-	-	-	-
	ARN [72]	63.7	82.4	-	-	44.6	59.1	62.1	84.8
	OTAM [8]	73.0	85.8	42.8	52.3	-	-	-	-
	TRX [44]	63.6	85.9	42.0	64.6	-	75.6	-	96.1
	TA2N [33]	72.8	85.8	47.6	61.0	59.7	73.9	81.9	95.1
	HyRSM [66]	73.7	86.1	54.3	69.0	60.3	76.0	83.9	94.7
	STRM [58]	-	86.7	-	68.1	-	77.3	-	96.9
Flow	ProtoNet-F [53]†	45.2	69.5	32.9	51.1	43.7	65.0	69.7	89.6
	TRX-F [44]†	44.8	69.7	30.7	52.4	43.0	67.6	65.6	90.6
	STRM-F [58]†	47.8	69.7	36.3	55.7	52.2	67.9	79.7	91.6
Multimodal	ProtoNet-EC [53]†	63.8	84.1	33.0	49.5	56.9	73.8	78.3	93.9
	ProtoNet-EA [53]†	61.7	83.9	31.1	50.5	53.2	76.3	76.7	94.3
	ProtoNet-LF [53]†	58.5	86.9	33.3	59.5	52.0	78.0	81.5	97.4
	AmeFu-Net [22]	74.1	85.8	-	-	60.2	75.5	85.1	95.5
	MTFAN [69]	74.6	87.4	45.7	60.4	59.0	74.6	84.8	95.1
	TRX-LF [44]†	65.9	86.8	37.2	61.1	57.4	78.2	81.6	94.1
	STRM-EC [58]†	68.3	87.4	45.5	66.7	59.3	78.3	87.4	96.3
	STRM-EA [58]†	68.4	87.0	44.1	62.4	60.3	76.3	85.4	94.7
	STRM-LF [58]†	66.9	87.7	41.4	70.4	55.0	81.3	83.8	98.4
	AMFAR(ours)	**80.1**	**92.6**	**61.7**	**79.5**	**73.9**	**87.8**	**91.2**	**99.0**

Experiments

Table 2. Ablation results on Kinetics and SSv2.

ASS		AMD	AMI	Kinetics		SSv2	
AC	RC			1-shot	5-shot	1-shot	5-shot
✘	✔	✔	✔	72.9	89.9	57.9	73.9
✔	✘	✔	✔	77.8	89.5	58.8	78.6
✔	✔	✘	✔	77.2	90.4	55.1	78.2
✔	✔	✔	✘	72.9	89.1	50.4	73.8
✔	✔	✔	✔	80.1	92.6	61.7	79.5

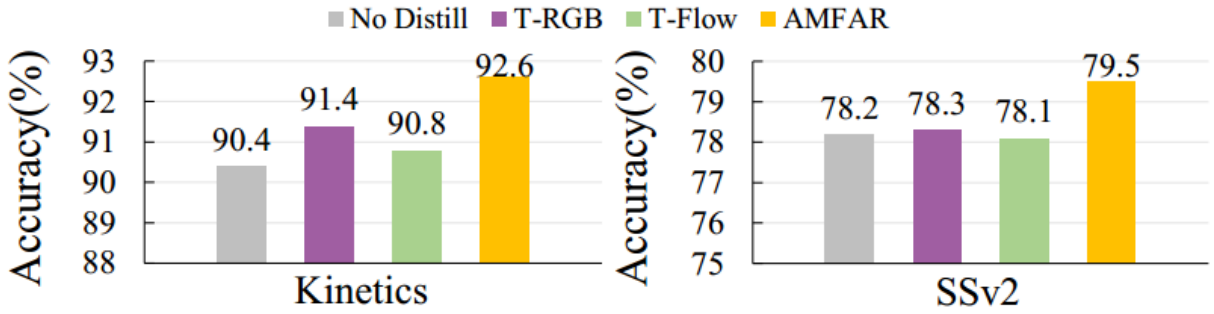


Figure 3. Comparison with conventional distillation strategies in 5-way 5-shot setting. T-RGB (or T-Flow) denotes distillation where RGB (or optical flow) is consistently regarded as teacher.

Conclusion

- We are the first to adopt **active learning** to explore the multimodal complementarity in few-shot learning.
- We propose an **active mutual distillation strategy** to transfer task-dependent knowledge learned from the reliable modality.
- We propose an **adaptive multimodal few-shot inference** approach to fuse modalityspecific results.
- Extensive **experiments** on four challenging datasets.


CASIA


CHINESE ACADEMY OF SCIENCES


JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks!