The 40th Annual AAAI Conference on Artificial Intelligence

Task-Specific Distance Correlation Matching for Few-Shot Action Recognition

Fei Long*[1], YaoZhang*[1], Jiaming Lv[1], Jiangtao Xie[1], Peihua Li†[1]
[1]Dalian University of Technology
*Equal contribution,†Corresponding author.

Paper Code Team

Motivation

Limitations of current matching-based metrics in FSAR

- Rely on cosine similarity to construct a distance matrix that captures inter-frame relationships between query and support.
- Perform matching based on instance-level information without considering task-specific cues.

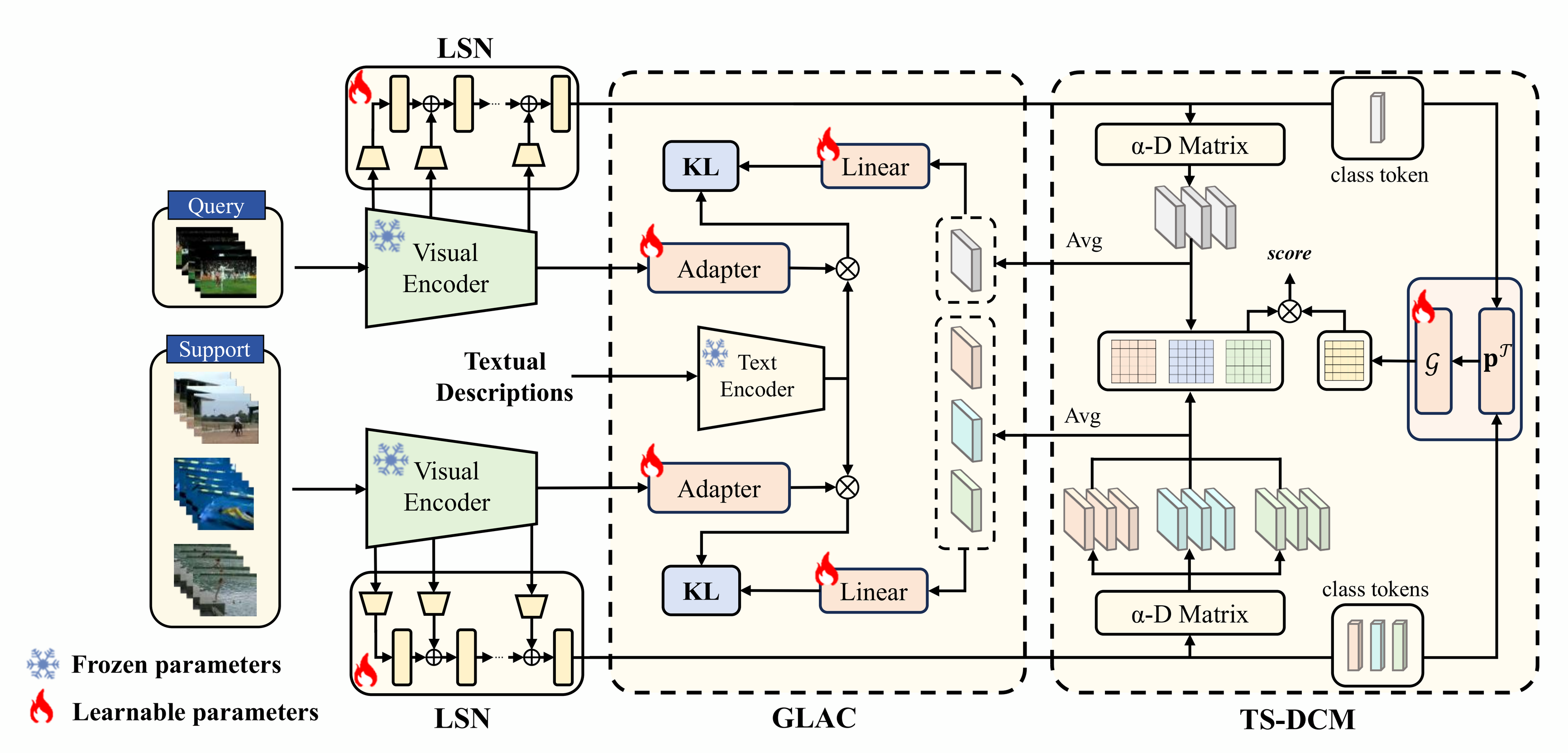
Solution: Perform Task-Specific Distance–Correlation Matching (TS-DCM).

Limitations of existing efficient methods for adapting CLIP to FSAR

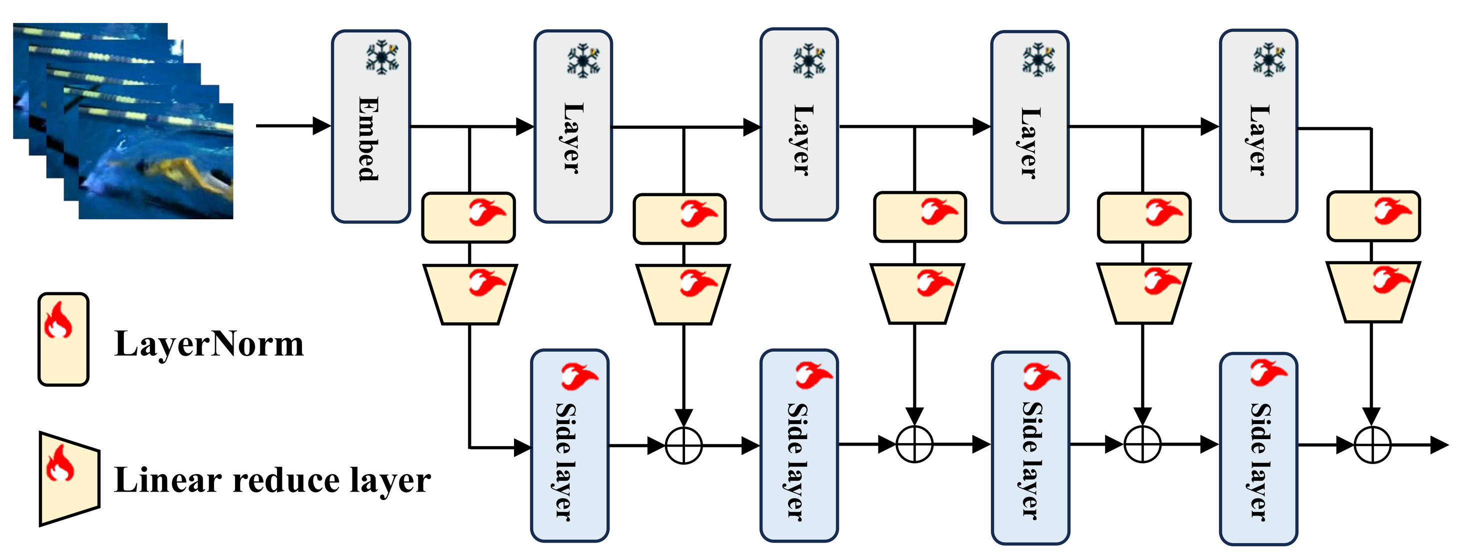
- Optimizing skip-fusion layers (also called LSN) under limited data remains challenging, especially on static datasets that depend heavily on pretrained knowledge

Solution: Guiding the LSN training using Adapted CLIP (GLAC).

Method Overall



Ladder Side Network (LSN)



$$x_l = (1 - a_{l-1}) \cdot r_{l-1}(ln_{l-1}(f_{l-1}(x))) + a_{l-1} \cdot g_{l-1}(x)$$

$ln_l(\cdot)$: ℓ-th layernrom

$r_l(\cdot)$: ℓ-th linear reduce layer

$f_l(x)$: ℓ-th frozen CLIP layer

$a_l \in [0, 1]$: ℓ-th learnable scalar

$g_l(x)$: ℓ-th LSN layer

x_l : the output of ℓ-th layer

Task-Specific Distance Correlation Matching (TSDCM)

α-Distance Correlation (α-DC)

① Definition

$$\mathrm{DCov}^{2(\alpha)}(\mathbf{X}, \mathbf{Y}) = \|\varphi_{\mathbf{X},\mathbf{Y}}(t,s) - \varphi_{\mathbf{X}}(t)\varphi_{\mathbf{Y}}(s)\|^2_\alpha$$

② Discrete form

Give two random variables , each with m i.i.d. samples:

$$\mathbf{X} \quad \mathbf{Y} \quad \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$$

Get pairwise Euclidean distances to obtain matrices:

$$\widehat{\mathbf{A}} = (\widehat{a}_{kl}) \quad \widehat{\mathbf{B}} = (\widehat{b}_{kl})$$

$$\widehat{a}_{kl} = \|\mathbf{x}_k - \mathbf{x}_l\|^\alpha, \widehat{b}_{kl} = \|\mathbf{y}_k - \mathbf{y}_l\|^\alpha$$

The α-distance covariance and correlation are defined as:

$$\mathrm{DCov}^{2(\alpha)}(\mathbf{X}, \mathbf{Y}) = \frac{1}{m^2}\mathrm{tr}(\mathbf{AB}) \quad \mathrm{DCorr}^{2(\alpha)}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{AB})}{\sqrt{\mathrm{tr}(\mathbf{AA})}\sqrt{\mathrm{tr}(\mathbf{BB})}}$$

Inter-Frame α-Distance Correlation (IF-D$^\alpha$C)

Extract the features of the i-th frame from a support video and a query video, respectively:

$$\mathbf{V}_S^i \in \mathbb{R}^{(P+1)\times d} \longrightarrow \mathbf{A}^i, \quad \mathbf{V}_Q^j \in \mathbb{R}^{(P+1)\times d} \longrightarrow \mathbf{B}^j \quad \longrightarrow \mathbf{M}^{\text{IF-D}^\alpha\text{C}} = (m_{ij}) \in \mathbb{R}^{T\times T}$$

Task-Specific Matching (TSM)

A query specific task prototype: $\mathbf{p}^{\mathcal{T}} = \widetilde{\mathbf{v}}^{\mathcal{Q}} + \frac{1}{N_{\mathcal{S}}}\sum_{x_i \in \mathcal{S}} \widetilde{\mathbf{v}}_i^{\mathcal{S}}$

A matching matrix produced by the learnable generator: $\mathbf{M}^{\text{task}} = \mathcal{G}(\mathbf{p}^{\mathcal{T}})$, $\mathbf{M}^{\text{task}} \in \mathbb{R}^{T\times T}$

Similarity score: $score = \langle \mathbf{M}^{\text{task}}, \mathbf{M}^{\text{IF-D}^\alpha\text{C}} \rangle$

Guiding LSN with Adapted CLIP (GLAC)

① Output of LSN

Let $\widetilde{\mathbf{A}}_{\alpha-\mathrm{D}} \in \mathbb{R}^{d\times d}$ denote the video-level α-D representation. For each class, we introduce a learnable weight matrix $\widetilde{\mathbf{W}}_i$. The LSN prediction is computed as the inner products between $\widetilde{\mathbf{A}}_{\alpha-\mathrm{D}}$ and $\widetilde{\mathbf{W}}_i$: $\mathbf{p} = [p_1, p_2, \ldots, p_C]$

② Output of Adapted CLIP

Let $\widetilde{e}$ denote the video-level class token produced by the adapted CLIP. We compute the cosine similarities between $\widetilde{e}$ and the text features to obtain the prediction vector: $\mathbf{q} = [q_1, q_2, \ldots, q_C]$

③ Guiding the LSN by using KL divergence:

$$\mathcal{L}_{\text{GLAC-KL}} = \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = \sum_{i=1}^{C} p_i \log \frac{p_i}{q_i}$$

$$\mathcal{L}_{\text{GLAC-CE}} = -\sum_{i=1}^{C} y_i \log(p_i) + (-\sum_{i=1}^{C} y_i \log(q_i))$$

Training Objective

The training loss of our TS-FSAR is composed of the three components : the vision-language alignment loss for the LSN, the TS-DCM loss, and the GLAC loss. Accordingly, the total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{LSN}} + \lambda_1 \mathcal{L}_{\text{TS-DCM}} + \lambda_2 \mathcal{L}_{\text{GLAC}}$$

Experiments

Method	Backbone	SSv2-Full		SSv2-Small		HMDB51		UCF101		Kinetics	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
OTAM (Cao et al. 2020)	IN-RN50	42.8	52.3	36.4	48.0	54.5	68.0	79.9	88.9	73.0	85.8
TRX (Perrett et al. 2021)	IN-RN50	42.0	64.6	36.0	56.7	54.9	75.6	81.0	96.1	65.1	85.9
STRM (Thatipelli et al. 2022)	IN-RN50	43.1	68.1	37.1	55.3	57.6	77.3	82.7	96.9	65.1	86.7
HyRSM (Wang et al. 2022)	IN-RN50	54.3	69.0	40.6	56.1	60.3	76.0	83.9	94.7	73.7	86.1
HCL (Zheng, Chen, and Jin 2022)	IN-RN50	47.3	64.9	38.7	55.4	59.1	76.3	82.6	94.5	73.7	85.8
Nguyen (Nguyen et al. 2022)	IN-RN50	43.8	61.1	–	–	59.6	76.9	84.9	95.9	74.3	87.4
SloshNet (Xing et al. 2023a)	IN-RN50	46.5	68.3	–	–	59.4	77.5	86.0	97.1	70.4	87.0
GgHM (Xing et al. 2023b)	IN-RN50	54.5	69.2	–	–	61.2	76.9	85.2	96.3	74.9	87.4
TEAM (Lee et al. 2025)	IN-RN50	–	–	–	–	62.8	78.4	87.2	96.2	75.1	88.2
CLIP-FSAR (Wang et al. 2024)	CLIP-ViT-B/16	62.1	72.1	54.6	61.8	77.1	87.7	97.0	99.1	94.8	95.4
EMP-Net (Wu et al. 2024)	CLIP-ViT-B/16	63.1	73.0	57.1	65.7	76.8	85.8	94.3	98.2	89.1	93.5
MVP-shot (Qu et al. 2025)	CLIP-ViT-B/16	-	-	55.4	62.0	77.0	88.1	96.8	99.0	91.0	95.1
MA-FSAR (Xing et al. 2025)	CLIP-ViT-B/16	63.3	72.3	59.1	64.5	83.4	87.9	97.2	99.2	95.7	96.0
D^{2}ST-Adapter (Pei et al. 2025)	CLIP-ViT-B/16	66.7	81.9	55.0	69.3	77.1	88.2	96.4	99.1	89.3	95.5
TSAM (Li et al. 2025)	CLIP-ViT-B/16	65.8	74.6	**60.5**	66.7	84.5	**88.9**	98.3	99.3	96.2	97.1
TS-FSAR (Ours)	CLIP-ViT-B/16	**75.1**	**83.5**	**60.5**	**70.3**	**85.0**	**88.9**	**98.7**	**99.3**	**96.3**	96.6

Ablation on key components

LSN	IF-D$^\alpha$C	TSM	GLAC	SSv2-Full		HMDB51	
				1-shot	5-shot	1-shot	5-shot
				37.0	37.0	75.9	75.9
✓				67.1	77.2	77.7	79.6
✓	✓			71.4	81.7	82.1	84.2
✓	✓	✓		73.8	82.8	83.4	85.6
✓	✓	✓	✓	75.1	83.5	85.0	88.9

Combine IF-D$^\alpha$C with existing metrics

Metric	SSv2-Full		HMDB51	
	w/o	w/	w/o	w/
GAP	68.6	72.0 (+3.4)	80.5	81.8 (+1.3)
OTAM	71.5	72.4 (+0.9)	81.7	83.7 (+2.0)
BiMHM	72.3	73.2 (+0.9)	82.5	84.5 (+2.0)
OT	71.8	73.5 (+1.7)	82.1	82.9 (+0.8)

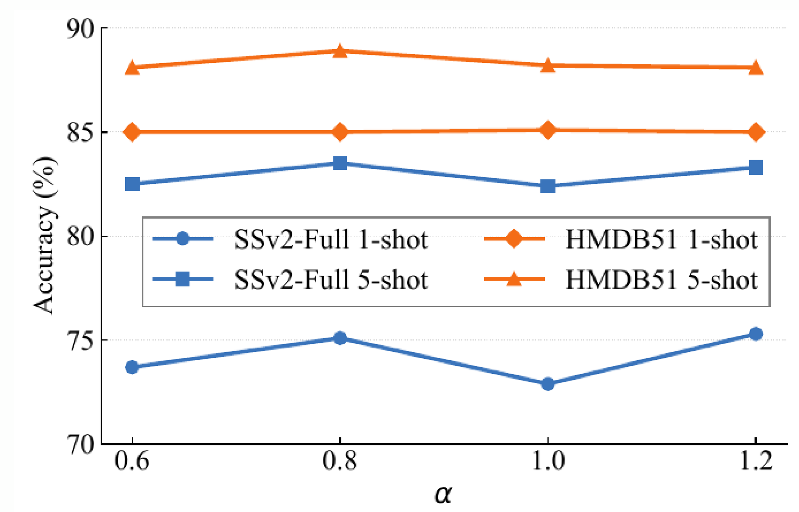
Different task prototype

Query-Specific	Task Prototype	SSv2-Full	HMDB51
w/o	Average	74.1	84.5
w/	Average	75.1	85.0
	Concatenation	73.6	84.6
	Cross-Attention	74.1	84.6

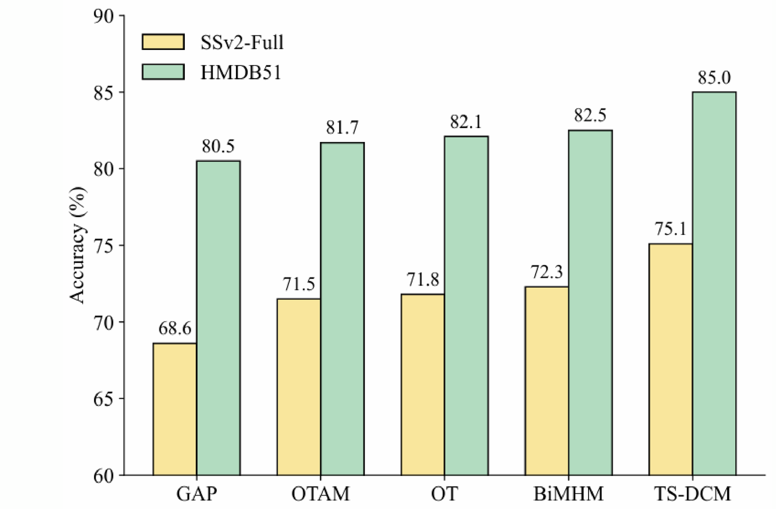
Effect of LSN Depth

Depth	HMDB51	SSv2-Full
3	84.5	67.0
6	84.7	69.4
12	85.0	75.1

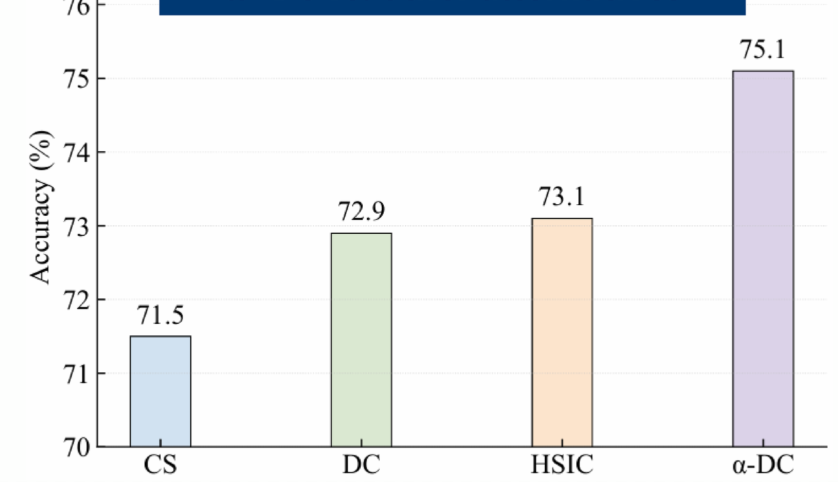
Impact of α



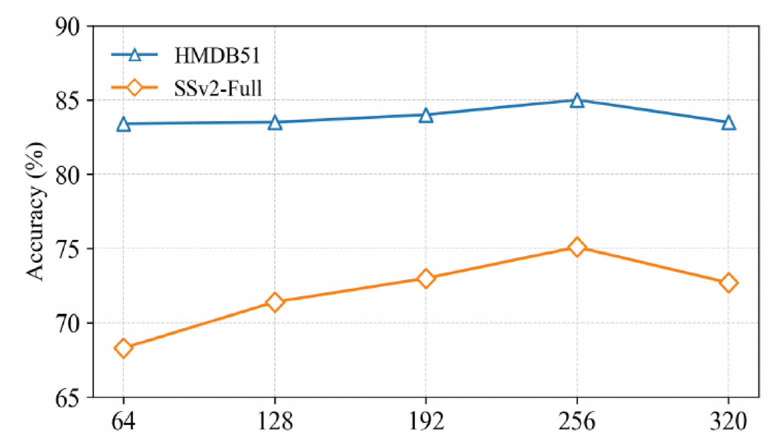
Different metrics



α –DC vs. other alternatives



Impact of LSN dim



Conclusion

In this paper, we:

- We propose a novel metric, termed Task-specific Distance Correlation Matching (TS-DCM), for few-shot action recognition.
- We propose a GLAC module that leverages the adapted frozen CLIP's output distribution to better guide LSN training under limited data.
- Our proposed TS-FSAR achieves leading performance across several few-shot action recognition benchmarks, with particularly significant improvements on temporally challenging datasets like SSv2-Full.