Static and Dynamic Concepts for Self-supervised Video Representation Learning

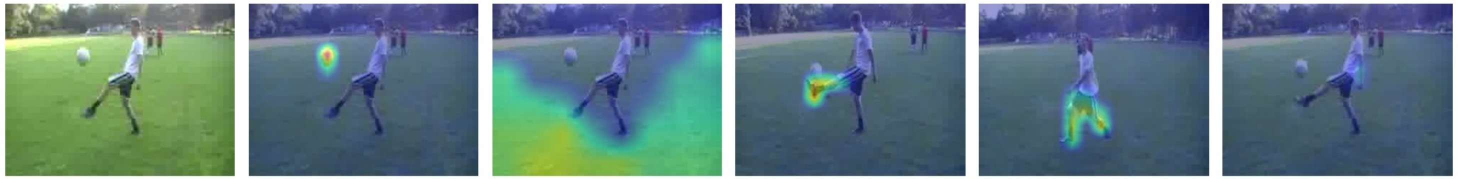
Rui Qian[1], Shuangrui Ding[2], Xian Liu[1], Dahua Lin[1,3]

[1] The Chinese University of Hong Kong
[2] Shanghai Jiao Tong University
[3] Shanghai Artificial Intelligence Laboratory

Motivation

- Humans can conclude general basic concepts from detailed observations for visual perception

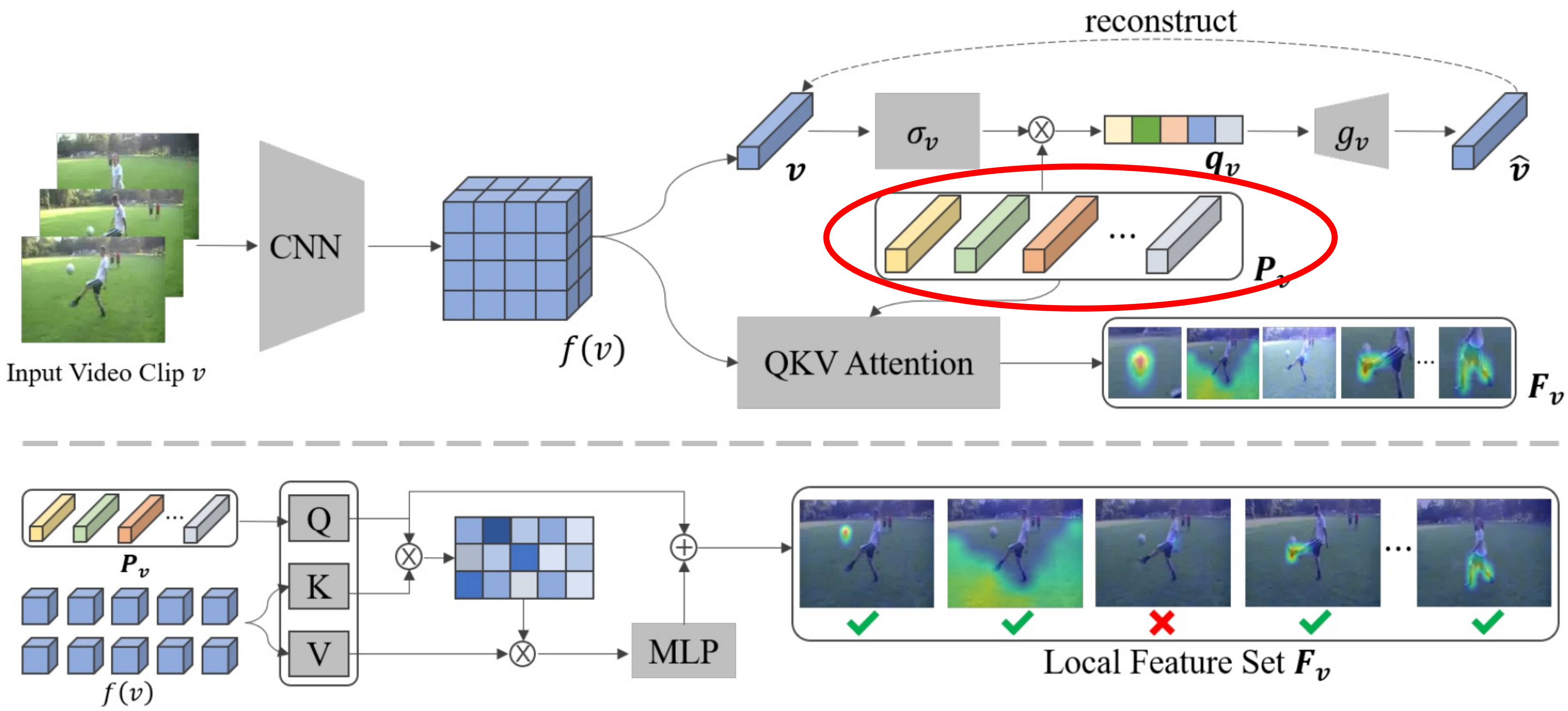
- Videos typically contain static and dynamic concepts that facilitate video understanding



(a) Soccer Juggling

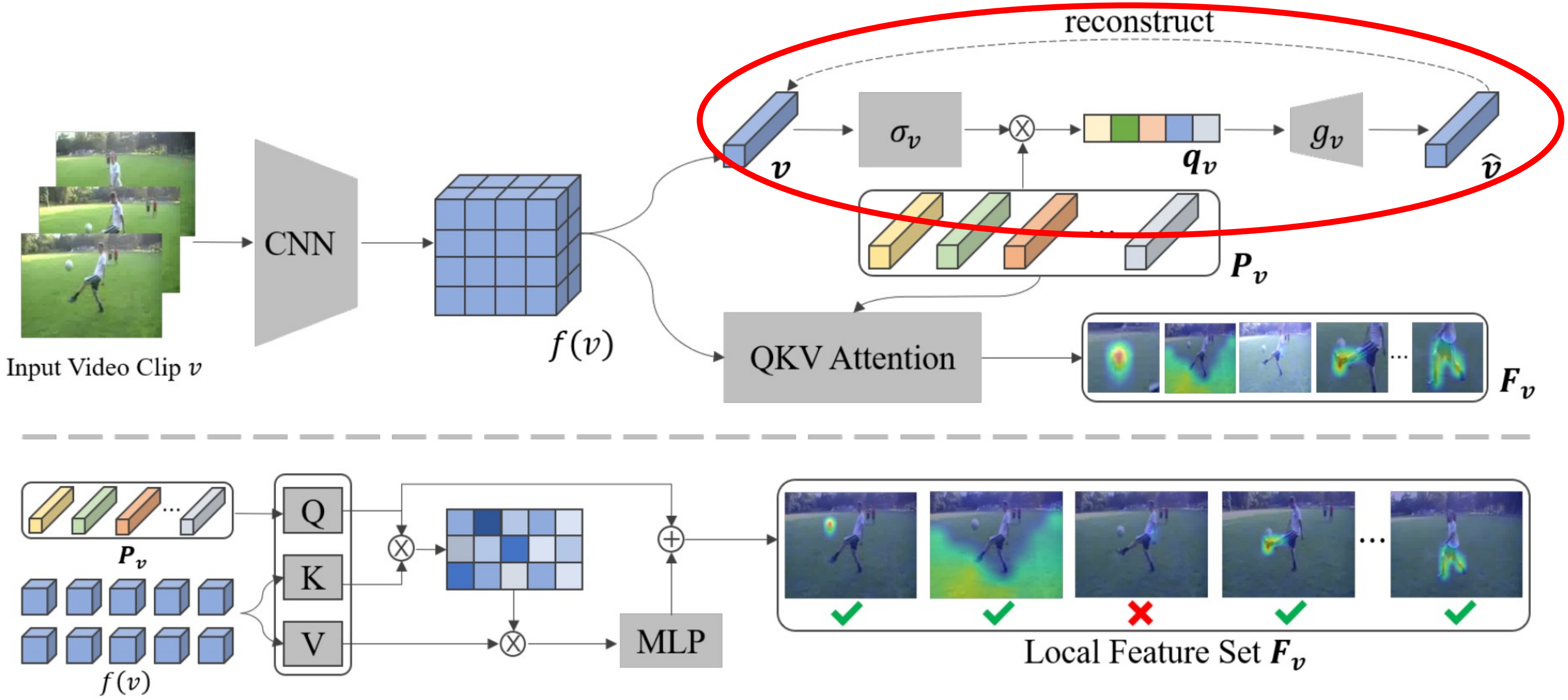
(b) Basketball

Method



reconstruct
σ_v
q_v
g_v
v
$\hat{v}$
CNN
$f(v)$
P_v
QKV Attention
F_v
Input Video Clip v
P_v
Q
K
V
MLP
$f(v)$
Local Feature Set F_v

Method


reconstruct
Input Video Clip v
CNN
$f(v)$
v
σ_v
q_v
g_v
$\hat{v}$
P_v
QKV Attention
F_v
P_v
Q
K
V
$f(v)$
MLP
Local Feature Set F_v

Method

- Concept prototype definition

$$\boldsymbol{P_s} \in \mathbb{R}^{K_s \times C}, \quad \boldsymbol{P_d} \in \mathbb{R}^{K_d \times C}, \quad \boldsymbol{P_v} \in \mathbb{R}^{(K_s+K_d) \times C}$$

- Feature extraction

$$\boldsymbol{s} = GAP(f(s)), \quad \boldsymbol{d} = GAP(f(d)), \quad \boldsymbol{v} = GAP(f(v))$$

- Latent concept code formulation

$$\boldsymbol{q_s}^{(k)} = \frac{\boldsymbol{P_s}^{(k)} \sigma_s(\boldsymbol{s})^T}{||\boldsymbol{P_s}^{(k)}||_2 ||\sigma_s(\boldsymbol{s})||_2}, \quad \boldsymbol{q_s} \in \mathbb{R}^{K_s}$$

Method

- Decoupled concept alignment

$$\mathcal{L}_{aln} = -\sum_{k=1}^{K_s}\left(\overline{\boldsymbol{q_s}}^{(k)}\log\frac{\exp(\boldsymbol{q_v^s}^{(k)}/\tau)}{\sum_{k'}\exp(\boldsymbol{q_v^s}^{(k')}/\tau)} + \overline{\boldsymbol{q_v^s}}^{(k)}\log\frac{\exp(\boldsymbol{q_s}^{(k)}/\tau)}{\sum_{k'}\exp(\boldsymbol{q_s}^{(k')}/\tau)}\right)$$
$$-\sum_{k=1}^{K_d}\left(\overline{\boldsymbol{q_d}}^{(k)}\log\frac{\exp(\boldsymbol{q_v^d}^{(k)}/\tau)}{\sum_{k'}\exp(\boldsymbol{q_v^d}^{(k')}/\tau)} + \overline{\boldsymbol{q_v^d}}^{(k)}\log\frac{\exp(\boldsymbol{q_d}^{(k)}/\tau)}{\sum_{k'}\exp(\boldsymbol{q_d}^{(k')}/\tau)}\right)$$

Method

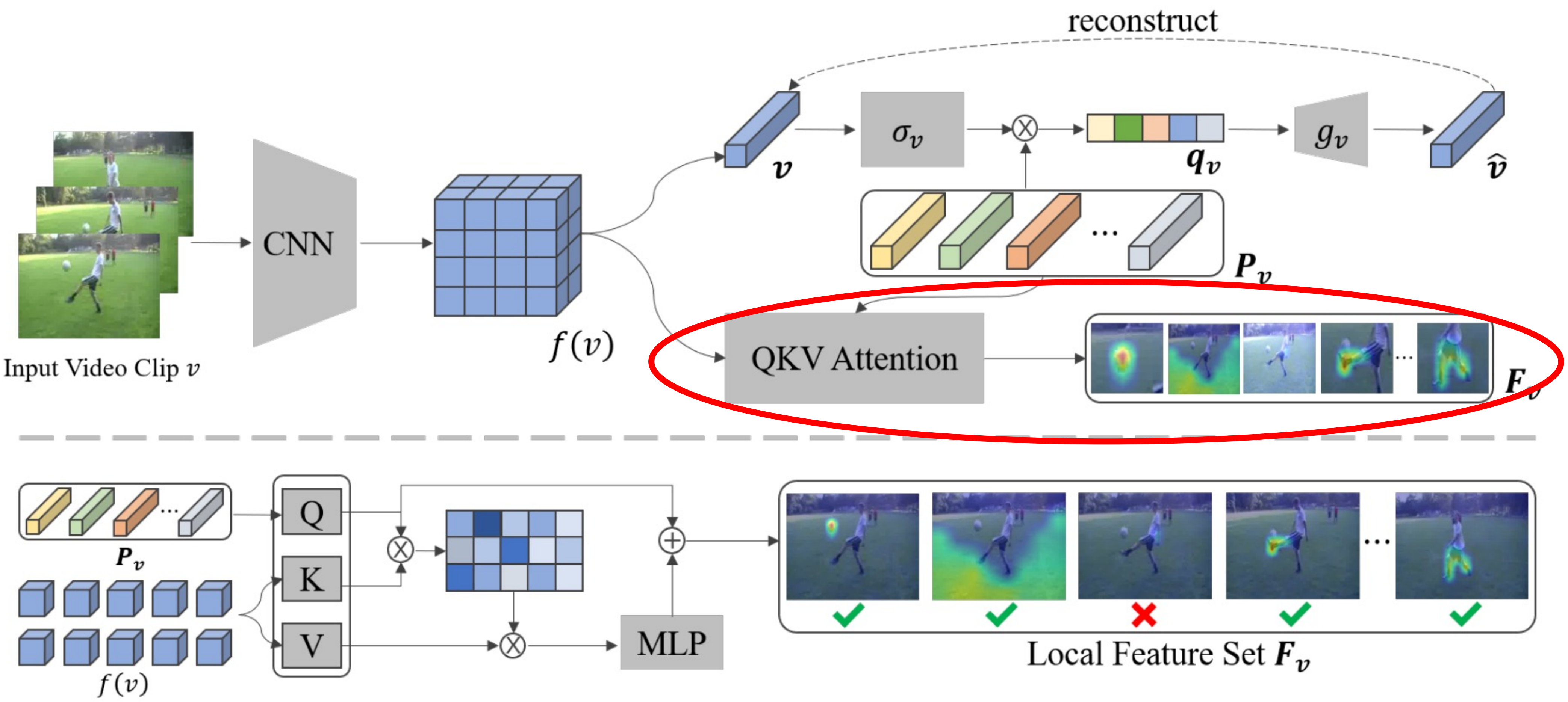
- Diversity regularization

$$\mathcal{L}_{div} = \|\boldsymbol{q_s}\|_1 + \|\boldsymbol{q_d}\|_1 + \|\boldsymbol{q_v}\|_1$$

- Fidelity regularization

$$\mathcal{L}_{fid} = \|g_s(\boldsymbol{q_s}) - \boldsymbol{s}\|_2^2 + \|g_d(\boldsymbol{q_d}) - \boldsymbol{d}\|_2^2 + \|g_v(\boldsymbol{q_v}) - \boldsymbol{v}\|_2^2$$

Method



reconstruct
CNN
Input Video Clip v
$f(v)$
v
σ_v
q_v
g_v
$\hat{v}$
P_v
QKV Attention
F_v
P_v
Q
K
V
$f(v)$
MLP
Local Feature Set F_v

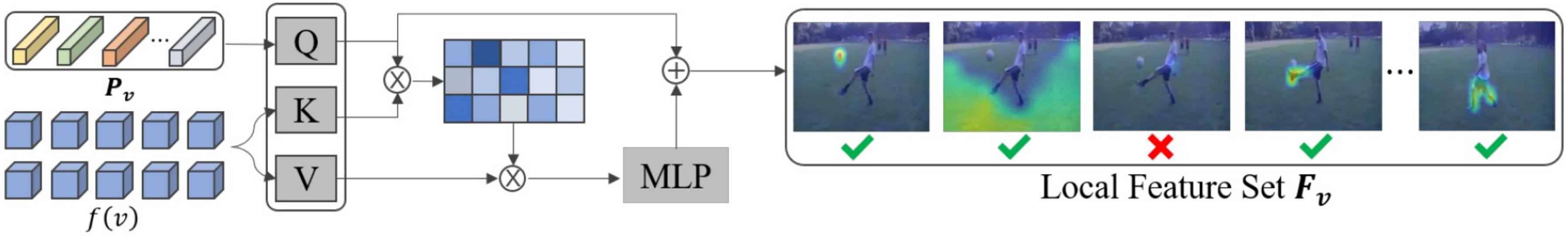
Method

- Local concept attention

$$\boldsymbol{F_s} = QKV(\boldsymbol{P_s}, f(\boldsymbol{s}), f(\boldsymbol{s})), \quad \boldsymbol{F_s} \in \mathbb{R}^{K_s \times C}$$

- Valid concept selection

$$\boldsymbol{idx_s} = \text{top-k}(\boldsymbol{q_s}, K) \cap \text{top-k}(\boldsymbol{q_v^s}, K)$$



Method

- Local concept contrast

$$l(\boldsymbol{F_s}, \boldsymbol{F_v^s}) = \sum_{k \in \boldsymbol{idx_s}} \left[\left\| \boldsymbol{F_s}^{(k)} - \boldsymbol{F_v^s}^{(k)} \right\|_2^2 + \sum_{\widetilde{\boldsymbol{F}} \in \mathcal{N}} \max\left(\lambda - \left\| \boldsymbol{F_s}^{(k)} - \widetilde{\boldsymbol{F}}_{\boldsymbol{v}}^{\boldsymbol{s}(k)} \right\|_2, 0 \right)^2 \right]$$

- Local contrast loss

$$\mathcal{L}_{loc} = l(\boldsymbol{F_s}, \boldsymbol{F_v^s}) + l(\boldsymbol{F_v^s}, \boldsymbol{F_s}) + l(\boldsymbol{F_d}, \boldsymbol{F_v^d}) + l(\boldsymbol{F_v^d}, \boldsymbol{F_d})$$

- Overall training loss

$$\mathcal{L} = \mathcal{L}_{aln} + \alpha\mathcal{L}_{loc} + \beta\mathcal{L}_{fid} + \gamma\mathcal{L}_{div}$$

Experiment

Video action recognition

- Linear probe
- End-to-end finetune

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF-101	HMDB-51
CBT [64]	S3D	Kinetics-600	16	112	✓	54.0	29.5
RSPNet [11]	R3D	Kinetics-400	16	112	✓	61.8	42.8
MLRep [57]	R3D	Kinetics-400	16	112	✓	63.2	33.4
CoCLR† [28]	S3D	Kinetics-400	32	128	✓	74.5	46.1
Ours	R(2+1)D	UCF-101	16	112	✓	67.4	40.7
Ours	R(2+1)D	Kinetics-400	16	112	✓	72.1	45.9
Ours	S3D	Kinetics-400	16	128	✓	75.1	47.4
TempTrans [35]	R(2+1)D	UCF-101	16	112	✗	81.6	46.4
LSFD [3]	R3D	UCF-101	32	112	✗	77.2	53.7
STS† [68]	R(2+1)D	UCF-101	16	112	✗	77.8	40.7
CoCLR† [28]	S3D	UCF-101	32	128	✗	81.4	52.1
Ours	R(2+1)D	UCF-101	16	112	✗	82.1	49.7
Ours	S3D	UCF-101	32	128	✗	83.7	53.8
ASCNet [31]	R3D	Kinetics-400	16	112	✗	80.5	52.3
Pace [70]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
VideoMoCo [53]	R(2+1)D	Kinetics-400	32	112	✗	78.7	49.2
RSPNet [11]	R(2+1)D	Kinetics-400	16	112	✗	81.1	44.6
TCLR [15]	R(2+1)D	Kinetics-400	16	112	✗	84.3	54.2
TimeEq [34]	S3D-G	Kinetics-400	32	128	✗	86.9	63.5
STS† [68]	S3D-G	Kinetics-400	64	224	✗	89.0	62.0
CoCLR† [28]	S3D	Kinetics-400	32	128	✗	87.9	54.6
Ours	R(2+1)D	Kinetics-400	16	112	✗	86.1	54.8
Ours	S3D	Kinetics-400	16	128	✗	88.3	56.4

Experiment

Ablation study

- Training loss
- Number of concepts

$\mathcal{L}_{aln}$	$\mathcal{L}_{fid}$	$\mathcal{L}_{div}$	$\mathcal{L}_{loc}$	UCF-101		HMDB-51	
				Linear	Finetune	Linear	Finetune
✓				61.4	76.3	40.3	44.7
✓	✓	✓		68.1	80.1	43.2	47.9
✓			✓	67.4	78.9	43.3	46.4
✓	✓	✓	✓	72.1	82.1	45.9	49.7

Experiment

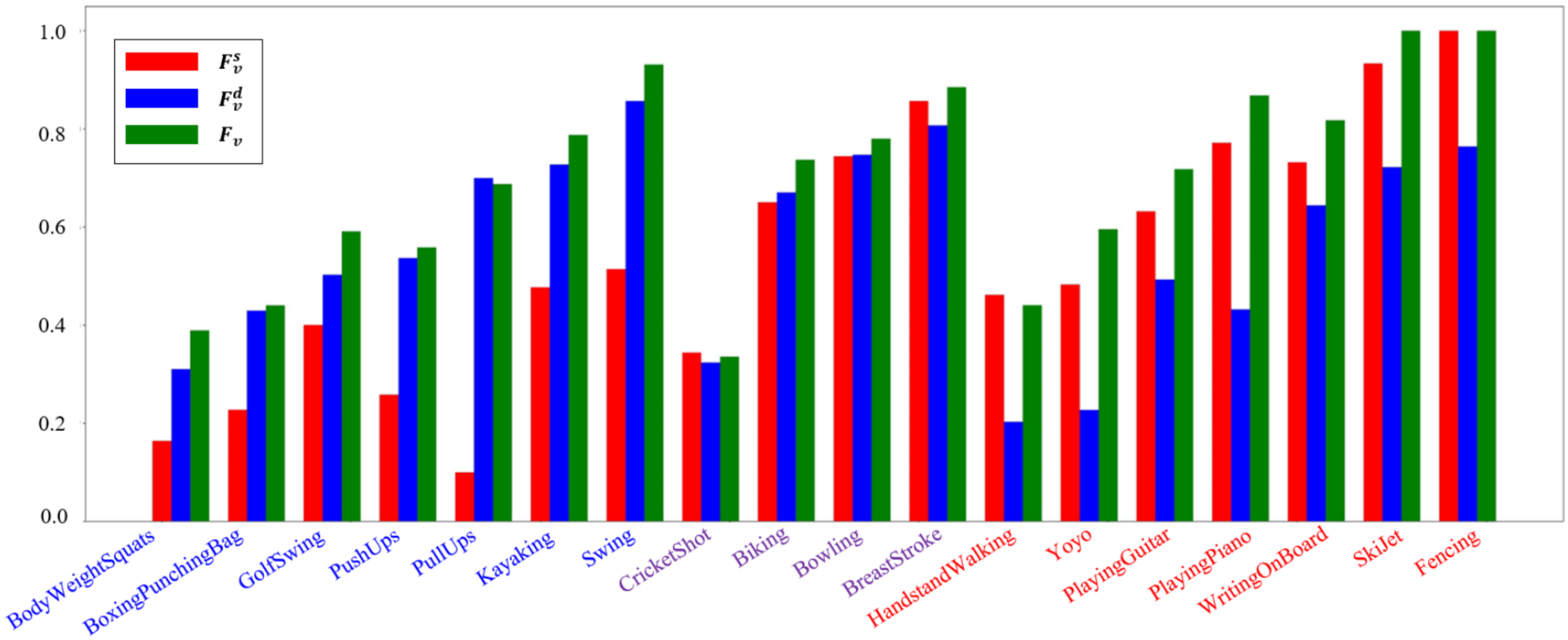
Ablation study

- Training loss
- Number of concepts

K_s	K_d	UCF-101		HMDB-51	
		w/ $\mathcal{L}_{loc}$	w/o $\mathcal{L}_{loc}$	w/ $\mathcal{L}_{loc}$	w/o $\mathcal{L}_{loc}$
25	25	70.3	61.2	43.0	39.4
25	50	71.7	66.3	44.1	40.8
50	25	71.3	65.2	44.8	42.4
50	50	72.1	68.1	45.9	43.2
100	100	72.3	68.8	45.8	44.3
200	200	72.3	69.4	45.6	44.1

Experiment

Per-class Static Dynamic and Joint Feature Analysis

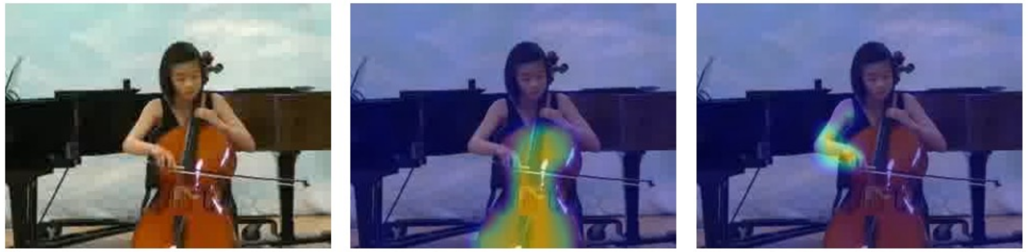


Experiment

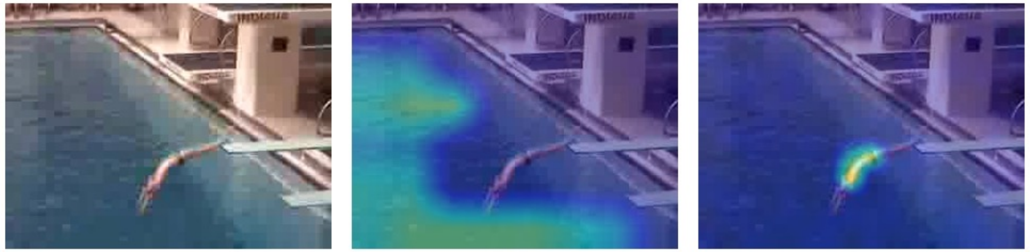
Visualization of static and dynamic concept attention map

(a) Playing Violin

(b) Breast Stroke



(c) Playing Cello



(d) Diving