第二十届中国图象图形学学会青年科学家会议

Not All Pairs are Equal: Hierarchical Learning for Average-Precision-Oriented Video Retrieval

Yang Liu[1], Qianqian Xu[2,*], Peisong Wen[1,2], Siran Dai[4], Qingming Huang[1,2,3,*]

1 School of Computer Science and Technology, University of Chinese Academy of Sciences 3 BDKM, University of Chinese Academy of Sciences

2 Institute of Computing Technology, Chinese Academy of Sciences 4 Institute of Information Engineering, Chinese Academy of Sciences

This paper was presented with the Honourable Mention Award at the ACM MultiMedia 2024 Conference.

Alignment of the Objective with the Metric

Evaluation Metric: Average Precision (AP)

- Evaluates the overall rankings of relevant videos
- Assigns larger weights on higher-ranked instances

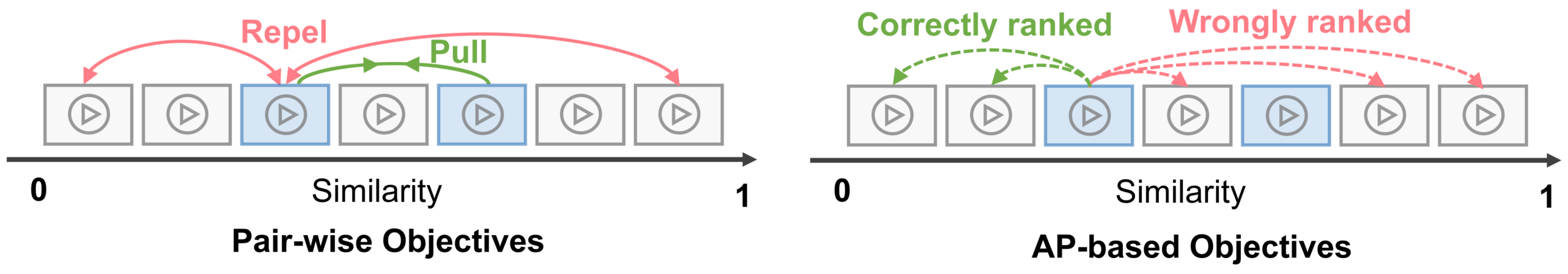
$$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{r_i}$$

Previous Training Objectives: Pair-wise Objectives

- Pull the positive instances closer and repel the negative ones
- ✗ Treat all mis-ranked pairs equally
- ✗ Mismatch with the evaluation metric

New Training Objectives: AP-based Objectives

- Rectify the wrongly ranked positive-negative pairs in the list
- ✓ Fill the gap between training objectives and evaluation metrics



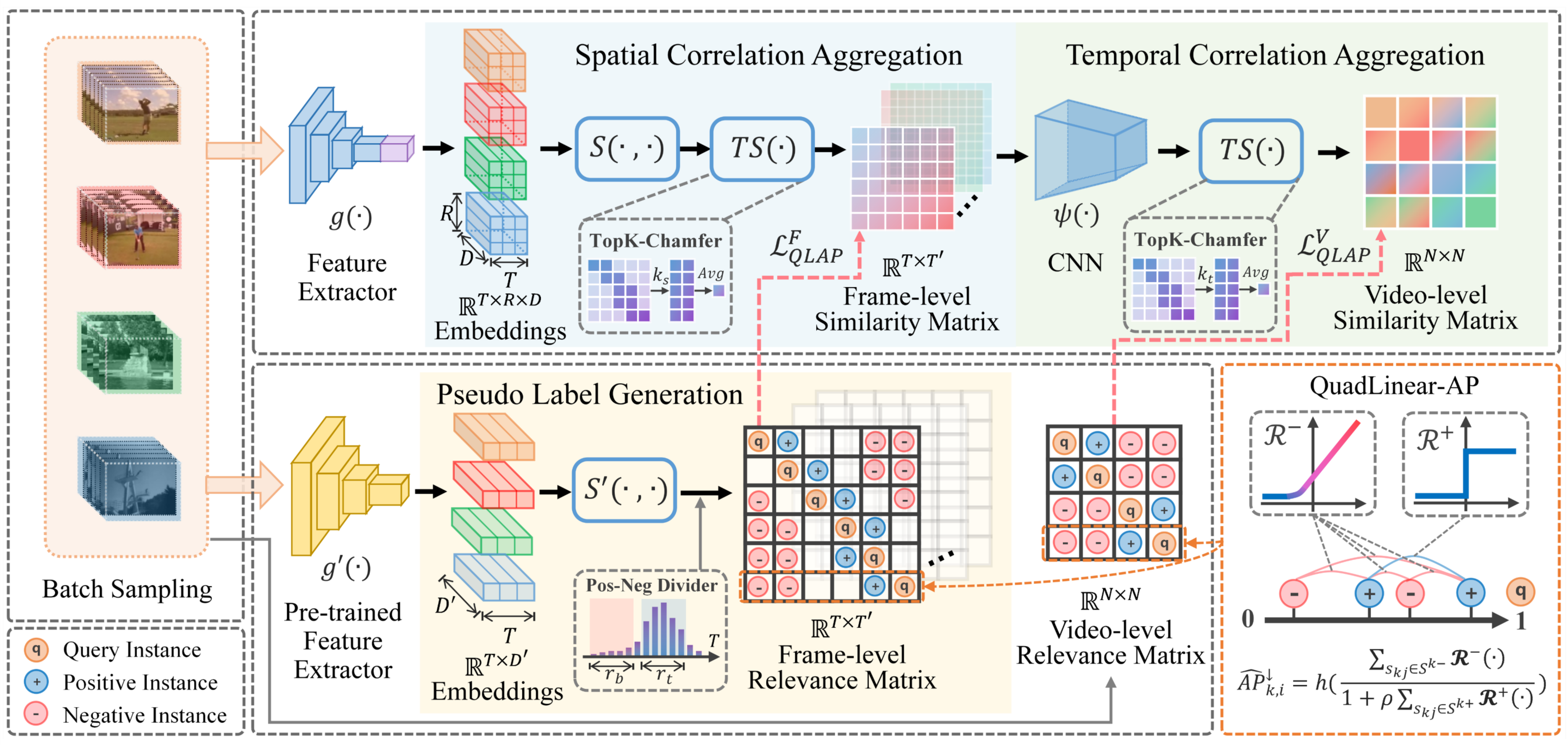
Two Challenges for Video-Oriented AP Optimization

- Current AP losses are suboptimal for **video-level** retrieval
- Noisy **frame-level** matching leads to a biased AP estimation

Hierarchical Learning Framework

Overall Framework: HAP-VR

Hierarchical learning for Average-Precision-oriented Video Retrieval



Video-Oriented AP Optimization Algorithm

- **Step1**: Bottom-up video similarity measure
- **Step2**: Pseudo-label generation
- **Step3**: Hierarchical AP optimization

$$\mathcal{L} = \underbrace{\lambda_f \mathcal{L}^F_{QLAP}}_{frame-level} + \underbrace{\lambda_v \mathcal{L}^V_{QLAP}}_{video-level} + \mathcal{L}_{base}$$

Gradient-Enhanced AP Optimization

Optimization Problem:

- Maximizing the AP score

$$\max_f \quad AP(f) = \frac{1}{N}\sum_{k=1}^{N} AP_k(f)$$

$$AP_k(f) = \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} \frac{\mathcal{R}(s_{ki}, \boldsymbol{S}^{k+})}{\mathcal{R}(s_{ki}, \boldsymbol{S}^{k+}\cup\boldsymbol{S}^{k-})} = \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} \frac{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{H}(d^k_{ji})}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}\cup\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})}$$

Objective Reformation:

- Minimizing the AP risk

$$\min_f \quad AP^{\downarrow}(f) = \frac{1}{N}\sum_{k=1}^{N} AP^{\downarrow}_k(f)$$

$$\boxed{\mathcal{R}(s, \boldsymbol{S}) = 1 + \sum_{s'\in\boldsymbol{S}}\mathcal{H}(s'-s) \qquad d^k_{ji} = s_{kj} - s_{ki} \quad h(x) = \frac{x}{1+x}}$$

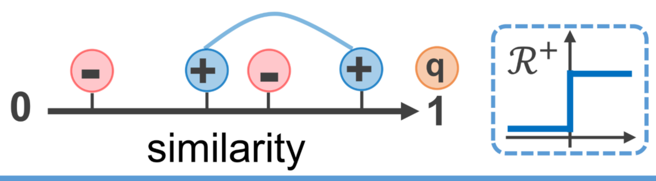
$$AP^{\downarrow}_k(f) = 1 - \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} \frac{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{H}(d^k_{ji})}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}\cup\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})} = \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} \frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}\cup\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})} = \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} h\left(\boxed{\frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{H}(d^k_{ji})}}\right)$$

Rethinking the Components of AP Risk: QuadLinear-AP

$$AP^{\downarrow}_k(f) = \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} h\left(\frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{H}(d^k_{ji})}\right) \Rightarrow \widehat{AP}^{\downarrow}_k(f) = \frac{1}{|\boldsymbol{S}^{k+}|}\sum_{s_{ki}\in\boldsymbol{S}^{k+}} h\left(\frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{R}^-(d^k_{ji};\delta)}{1+\rho\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{R}^+(d^k_{ji})}\right)$$

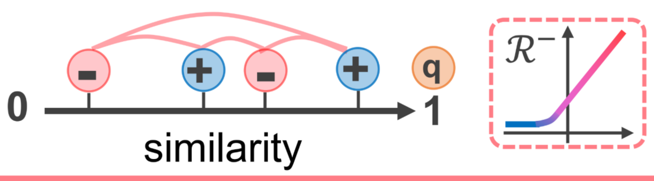
Positive-positive Pairs

- Serve as weights
- No need for optimization
- $\mathcal{R}^+(x) = \mathcal{H}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x \le 0 \end{cases}$



Positive-negative Pairs

- Rankings should be corrected
- Need proper gradients for optimization
- $\mathcal{R}^-(x;\delta) = \begin{cases} \mathcal{H}(-x)\cdot\frac{1}{\delta^2}x^2 + \frac{2}{\delta}x + 1, & \text{if } x \ge -\delta \\ 0, & \text{if } x < -\delta \end{cases}$



Original AP

$$\frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{H}(d^k_{ji})}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{H}(d^k_{ji})}$$

Non-differentiability

Smooth-AP

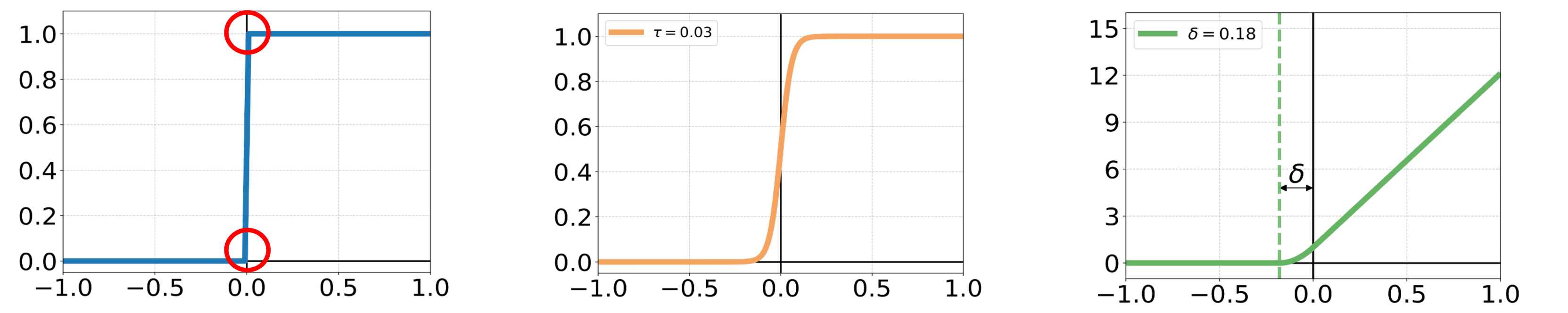
$$\frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{G}(d^k_{ji};\tau)}{1+\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{G}(d^k_{ji};\tau)}$$

Gradient Vanishing

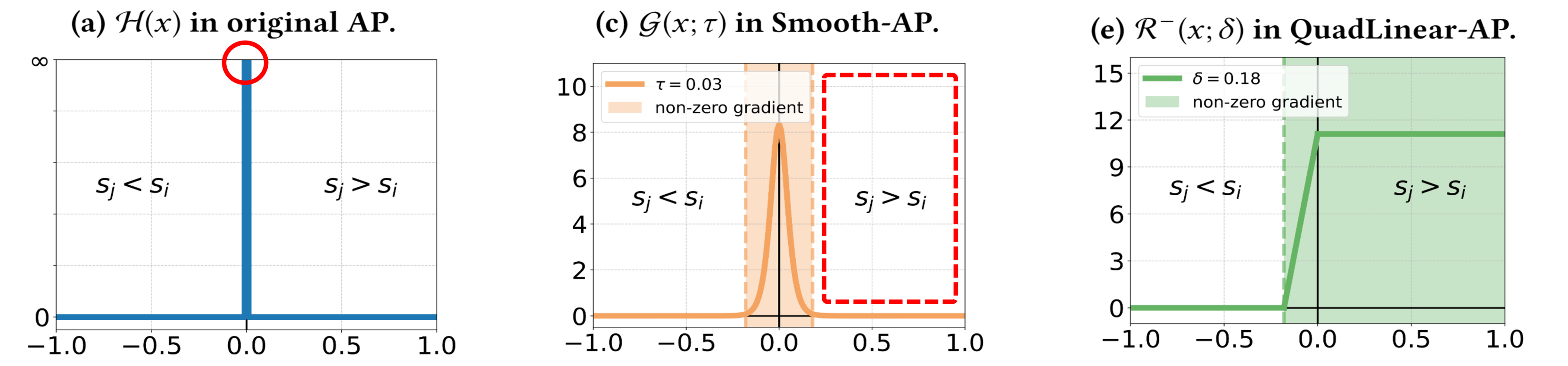
QuadLinear-AP (Ours)

$$\frac{\sum_{s_{kj}\in\boldsymbol{S}^{k-}}\mathcal{R}^-(d^k_{ji};\delta)}{1+\rho\sum_{s_{kj}\in\boldsymbol{S}^{k+}}\mathcal{R}^+(d^k_{ji})}$$

Favorable properties



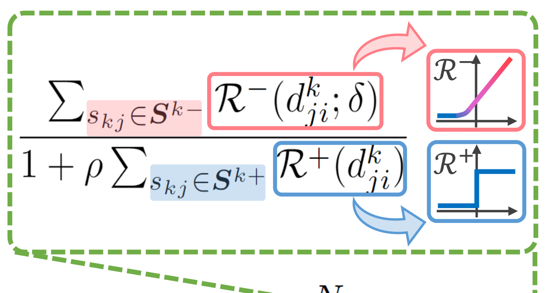
(a) $\mathcal{H}(x)$ in original AP. (c) $\mathcal{G}(x;\tau)$ in Smooth-AP. (e) $\mathcal{R}^-(x;\delta)$ in QuadLinear-AP.



(b) Derivative of $\mathcal{H}(x)$. (d) Derivative of $\mathcal{G}(x;\tau)$. (f) Derivative of $\mathcal{R}^-(x;\delta)$.

Favorable Properties of QuadLinear-AP

- Differentiable AP optimization
- Suitable gradients for low AP area
- Continuous, Smooth, and Convex
- Monotonically increasing (non-strictly)
- Upper bound of Heaviside function



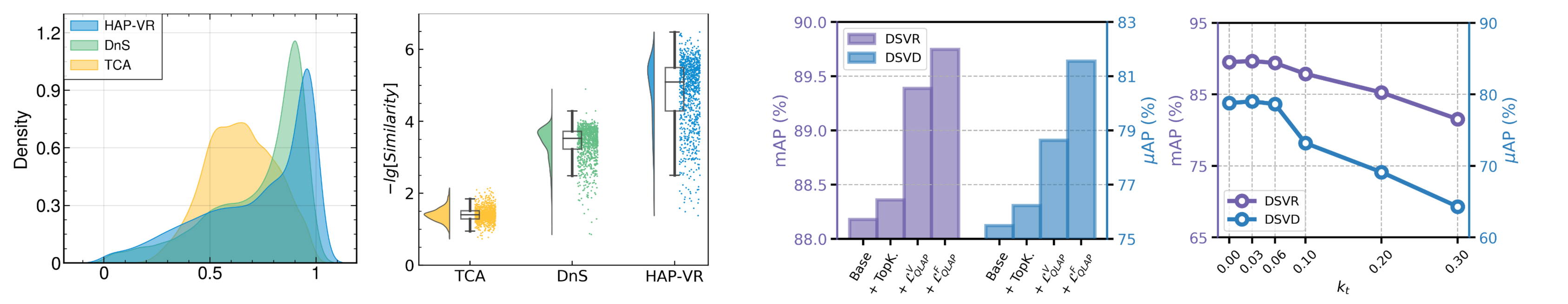
$$\mathcal{L}_{QLAP} = \frac{1}{N}\sum_{k=1}^{N}\widehat{AP}^{\downarrow}_k(f)$$

Evaluation Results

Overall Performance of Our Proposed HAP-VR

- HAP-VR improves mAP and μAP effectively on multiple tasks

Method	Label	Trainset	Retrieval (mAP)					Detection (μAP)				
			EVVE	SVD	FIVR-200K DSVR	FIVR-200K CSVR	FIVR-200K ISVR	EVVE	SVD	FIVR-200K DSVD	FIVR-200K CSVD	FIVR-200K ISVD
DML† [32]	✓	VCDB ($\mathcal{C}$&$\mathcal{D}$)	61.10	85.00	52.80	51.40	44.00	75.50	/	39.00	36.50	30.00
TMK† [46]	✓	internal	61.80	86.30	52.40	50.70	42.50	/	/	/	/	/
TCA [53]	✓	VCDB ($\mathcal{C}$&$\mathcal{D}$)	63.08	89.82	86.81	82.31	69.61	76.90	56.93	69.09	62.28	49.24
ViSiL† [30]	✓	VCDB ($\mathcal{C}$&$\mathcal{D}$)	65.80	88.10	89.90	85.40	72.30	79.10	/	75.80	69.00	53.00
DnS (S_a) [34]	✓	DnS-100K	65.33	90.20	92.09	87.54	74.08	74.56	72.24	79.66	69.51	54.20
DnS (S_b) [34]	✓	DnS-100K	64.41	89.12	90.89	86.28	72.87	75.80	66.53	78.05	68.52	53.48
LAMV† [2]	✗	YFCC100M	62.00	88.00	61.90	58.70	47.90	80.60	/	55.40	50.00	38.80
VRL† [24]	✗	internal	/	/	90.00	85.80	70.90	/	/	/	/	/
ViSiL$_f$† [30]	✗	ImageNet	62.70	/	89.00	84.80	72.10	74.60	/	66.90	59.50	45.90
S^{2}VS [33]	✗	VCDB ($\mathcal{D}$)	67.17	88.40	92.53	87.73	74.51	80.72	65.04	86.12	77.41	63.26
HAP-VR (Ours)	✗	VCDB ($\mathcal{D}$)	69.15	89.00	92.83	88.21	74.72	82.88	67.87	88.41	79.85	64.79

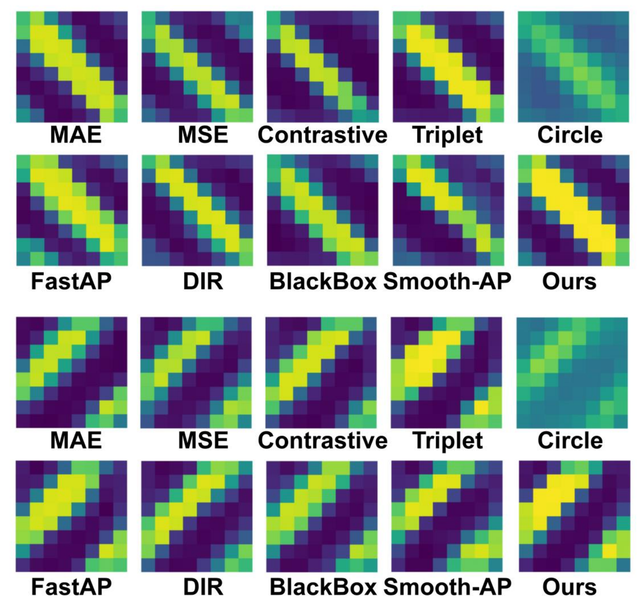


(a) Relevant pair distribution. (b) Irrelevant pair distribution. (c) Effect of components. (d) Impact of k_γ.

Effectiveness of Our Proposed QuadLinear-AP

- QuadLinear-AP outperforms previous pair-wise/AP-based losses

Losses	Retrieval (mAP)			Detection (μAP)		
	DSVR	CSVR	ISVR	DSVD	CSVD	ISVD
MAE	89.07	88.03	80.86	78.08	75.69	65.26
MSE	89.22	88.26	80.80	78.66	76.07	65.44
Contrastive [18]	88.67	88.09	80.97	75.12	74.23	67.41
Triplet [50]	88.11	87.77	81.21	72.94	73.18	69.23
Circle [55]	87.53	86.11	78.77	73.26	71.15	59.33
FastAP [6]	89.30	88.42	81.16	78.83	77.51	69.95
DIR [47]	89.65	88.57	80.64	78.50	76.22	65.42
BlackBox [45]	89.70	88.55	80.53	80.07	77.37	66.00
Smooth-AP [4]	89.36	88.33	80.73	79.85	77.75	68.42
QuadLinear-AP (**Ours**)	90.80	89.68	81.31	82.92	80.03	71.45



Conclusions

- **Methodologically**: Propose a self-supervised framework (**HAP-VR**) for video retrieval to **bridge the gap** of the objective and metric.
- **Analytically**: Introduce a gradient-enhanced AP surrogate loss (**QuadLinear-AP**) and design a **hierarchical learning strategy** for **AP optimization** on both video and frame levels.
- **Empirically**: Experimental results demonstrate the effectiveness of our proposed framework on various video retrieval tasks.

Paper Code WeChat