T-DEED: Temporal-Discriminability Enhancer Encoder-Decoder for Precise Event Spotting in Sports Videos

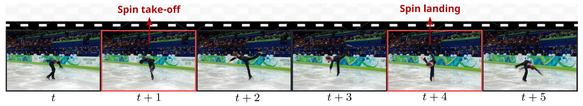
Artur Xarles[1,2] Sergio Escalera[1,2,3] Thomas B. Moeslund[3] Albert Clapés[1,2]

[1]Universitat de Barcelona [2]Computer Vision Center [3]Aalborg University

CVsports @ CVPR
JUNE 17-21, 2024
SEATTLE, WA

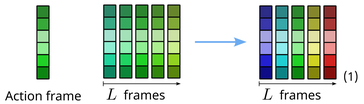
Problem Definition

Precise Event Spotting (PES): accurate localization of events represented by single keyframes within untrimmed videos.
Requires a small margin of error, with tolerances of 1 or 2 frames.



3 main challenges:

(1)- Need for discriminative per-frame representations to differenciate between adjacent frames with high spatial similarity.
(2)- Necessity of high output temporal resolution to avoid losing prediction precision.
(3)- Variability in the amount of temporal context required for different events, depending on event dynamics.

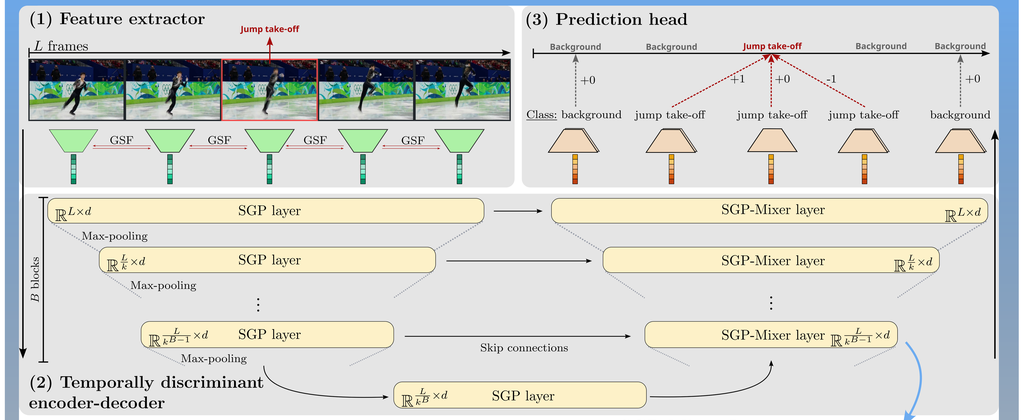


Contributions

(1)- We incorporate residual connections into the SGP[3] layer, enabling the integration of features from multiple temporal scales and resulting in the SGP-Mixer layer.
(2)- We introduce the SGP-Mixer module within the SGP-Mixer layer, which adapts the SGP module to aggregate information from different temporal scales while enhancing token discriminability.
(3)- We conduct extensive ablations of T-DEED components and achieve SOTA performance on FigureSkating[1] and FineDiving[5] datasets.

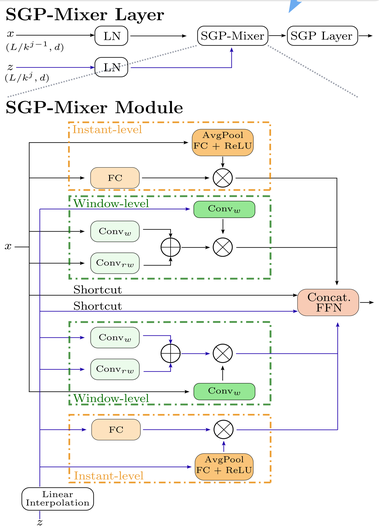
[1]- Hong et al. Video pose distillation for few-shot, fine-grained sports action recognition. (ICCV'21)
[2]- Hong et al. Spotting temporally precise, fine-grained events in video. (ECCV'22)
[3]- Shi et al. Tridet: Temporal action detection with relative boundary modeling. (CVPR'23)
[4]- Sudhakaran et al. Gate-shift-fuse for video action recognition. (PAMI)
[5]- Xu et al. Finediving: A fine-grained dataset for procedure-aware action quality assessment. (CVPR'22)

T-DEED Architecture



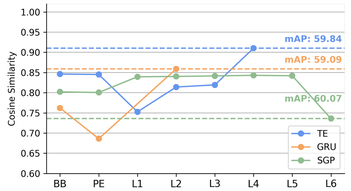
- **Feature extractor** comprising a 2D backbone with Gate-Shift-Fuse (GSF)[4] modules to generate per-frame representations.
- **Temporally discriminant encoder-decoder** with SGP-based layers to enhance discriminability across tokens of the same sequence.
  - Encoder blocks: SGP layer and max-pooling operation to reduce the temporal dimension.
  - Decoder blocks: SGP-Mixer layer and upsampling to increase the temporal dimension.
- **Prediction head** consisting of a classification component to differentiate between actions, and a displacement component to pinpoint the exact temporal position.

- **SGP-Mixer module** comprising two instant-level branches and two window-level branches.
  - Instant-level branches: to increase token discriminability.
  - Window-level branches: to aggregate information from both inputs while capturing different temporal contexts.

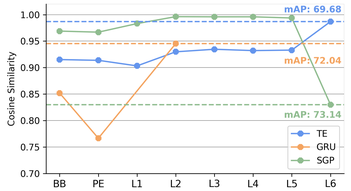


Ablations

- Rank-loss problem in Transformers → token similarity increases across layers, losing discriminability.
- SGP layers have the lowest similarity → effectiveness in enhancing token discriminability.
- Improved discriminability in SGP layers → best performance among the different temporal modules.



(a) FineDiving



(b) FigureSkating (FS-Comp)

- Importance of using multiple temporal scales.
- Skip connections through SGP-Mixer lead to best results.

Experiments	FS-Comp mAP ($\delta = 1$)	FineDiving mAP ($\delta = 1$)
(b) Skip connection		
w/o	74.29	63.01
sum	73.27	60.14
concat	76.78	61.90
SGP-Mixer (sum)	76.88	61.34
SGP-Mixer	**77.96**	**63.67**

Comparison to SOTA

- In the FineDiving dataset, T-DEED outperforms the SOTA for both tolerances, with improvements of +4.87 on the tight metric and +3.58 on the loose one.
- In the FigureSkating dataset, T-DEED outperforms the SOTA on the tight metric, while performance on the loose one is comparable or slightly worse.

Model	Size	Input	FS-Comp		FS-Perf		FineDiving	
			$\delta = 1$	2	1	2	1	2
E2E-Spot [2]	200MF	RGB	81.00	93.50	85.10	95.70	68.40	85.30
E2E-Spot [2]	800MF	RGB	84.00	-	83.60	-	64.60	-
E2E-Spot [2]	800MF	RGB + OF	83.40	**94.90**	83.30	96.00	66.40	84.80
T-DEED	200MF	RGB	**85.15**	91.70	86.79	**96.05**	71.48	87.62
T-DEED	800MF	RGB	84.77	92.86	**88.17**	95.87	**73.23**	**88.88**