Supplementary Materials
Neighbor Relations Matter in Video Scene Detection

Jiawei Tan, Hongxing Wang*, Jiaxin Li, Zhilong Ou, Zhangbin Qian
Key Laboratory of Dependable Service Computing in Cyber Physical Society (Chongqing University), Ministry of Education, China
School of Big Data and Software Engineering, Chongqing University, China
{jwtan, ihxwang, jiaxin_li}@cqu.edu.cn, zlou@stu.cqu.edu.cn, zbqian@cqu.edu.cn

Besides experiments reported on MovieNet [4], BBC [1], and OSVD [7] in our manuscript, we supplement more studies of our method on MovieNet.

1. Additional Experimental Results

The following results are obtained in a fully supervised manner, unless otherwise specified.

NeighborNet Architecture. Table S1 reports the performance of the two modules, *i.e*. relating neighbors in feature dimension (RNF) and relating neighbors in temporal dimension (RNT), operated in different orders in the proposed method. The results demonstrate that sequentially cascading RNF and RNT outperforms other configurations of these two modules. This superiority can be attributed to capability of RNF to distinguish similar shots in the same scene between those from different scenes, making RNT ease to enhance the similarity between shots, especially those dissimilar, within the same scene.

Edges of Temporal Graph. To justify the necessity of temporal edge selection when building temporal graph, we compare in Table S2 the temporal graph when using Eq. (8) selecting edge connections or full edge connections between shots. The results show that ours using Eq. (8) exhibits a substantial performance advantage in all metrics, underscoring the effectiveness of our customized temporal graph.

Operation for Aggregating Similarity Table S3 presents the impacts of different operation for aggregating similarity in Eq. 3. When a weighted sum operation in place of "max", it yields a decrease in AP from 64.0% to 62.1%.

Computation costs. Under identical hardware conditions, Table S4 below presents the training/inference throughput measured in samples per second (Sam./s), number of training parameters (Par.), and GPU memory costs per sample (GB/Sam.) for state-of-the-art methods. A sample corresponds to a time window with a length of 21 shots, equating to 21 nodes in our graph. As can be seen, our model requires fewer resources in terms of training/inference throughput, training params, and GPU memory costs compared to others utilizing the same backbone.

Architecture	AP	mIoU	F1
RNF ‖ RNT	60.4	59.3	55.4
RNT⊕RNF	62.7	59.8	56.4
RNF⊕RNT	64.0	61.2	57.8

Table S1. Impact of NeighborNet architecture. RNF denotes "relating neighbors in feature dimension" module, and RNT denotes "relating neighbors in temporal dimension" module. ‖ denotes a parallel operation, and ⊕ represents a series operation as per the given order.

Temporal Graph	AP	mIoU	F1
Fully Connected	49.8	54.1	49.2
Selectively Connected (Ours)	64.0	61.2	57.8

Table S2. Selectively *vs*. fully connected temporal Graph.

Similarity Aggregatation	AP
Weighted Sum	62.1
Max (Ours)	64.0

Table S3. Impacts of operations of aggregating similarity in Eq. 3.

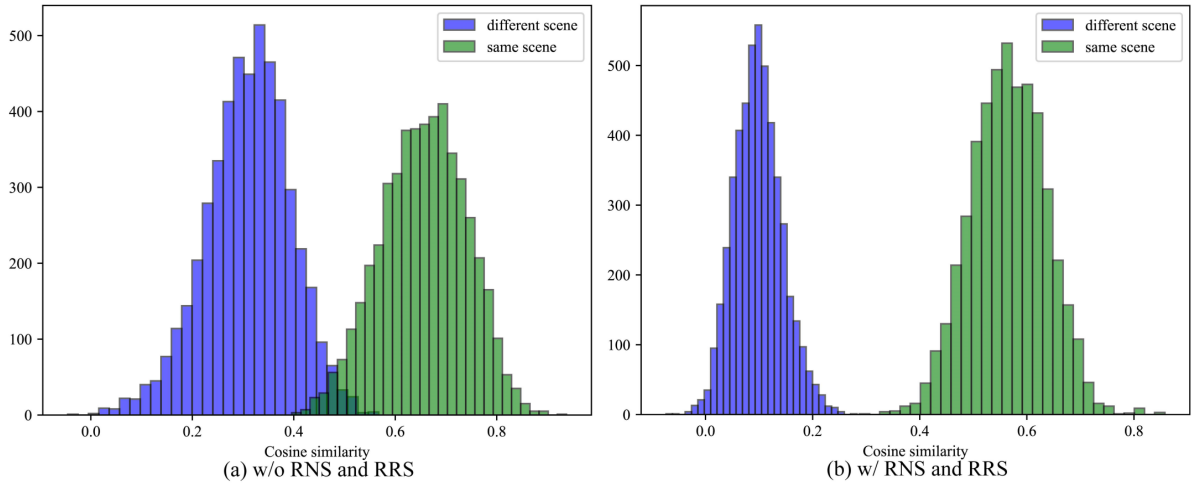
Method	Backbone	Train			Inference	
		Sam./s	Par.	GB/Sam.	Sam./s	GB/Sam.
BaSSL [6] (ACCV'22)	ResNet-50	1.0	43.8 M	34.3	2304.6	0.0013
Ours		1208.3	35.5 M	0.0058	3072.0	0.0012
TranS4mer [5] (CVPR'23)	ViT-S/16	2.0	32.0 M	5	32.0	0.36
Ours		2452.5	5.1 M	0.0015	4557.4	0.00057

Table S4. Computation cost compared with prior methods.

Metric for Reducing Similarity of Shots from Different Scenes. As depicted in the Fig. S1, the absence of the proposed RNS and RRS results in an overlap between the similarity distributions of shots from the same scene and different scenes. Conversely, once leveraging RNS and RRS, the two distributions are distinctly separated.

Dimensionality of Shot Features. Fig. S2 depicts the

*Corresponding author: Hongxing Wang.



(a) w/o RNS and RRS (b) w/ RNS and RRS

Figure S1. Cosine similarity distributions of shots from the same scene and different scenes in the test set of the MovieScenes dataset. Left: w/o RNS and RRS; Right: w/ RNS and RRS.

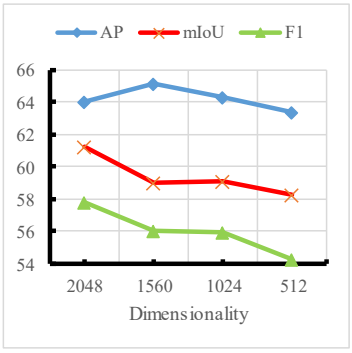
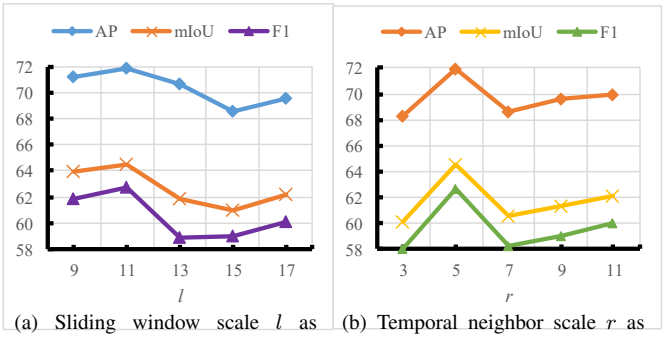


Figure S2. Impact of dimensionality of input shot features.



(a) Sliding window scale l as (b) Temporal neighbor scale r as

(c) Feature neighbor scale k as defined in Eq. (1). (d) Number of inputting shots N as defined in Sec. 3.2.

Figure S3. Impact of hyperparameters in the self-supervised transfer learning setting.

impact of different dimensions of shot feature defined in Sec. 3.2. The initial feature dimension is 2048 output by the backbone. For dimensions lower than the initial, we utilize an MLP to transform the original shot features to the target dimensions. As shown in Fig. S2, AP peaks at 1560 dimensions, while mIoU and F1 scores decline as the dimensions gradually decrease.

Impact of Hyper-parameters in the self-supervised transfer learning setting. Fig. S3 shows the impact of hyper-parameters in the self-supervised transfer leaning setting. The behaviours of hyper-parameters under self-supervised transfer learning is consistent with those (shown in Fig. 7 of our manuscript) under fully supervised learning.

2. Additional Visualizations

The qualitative results presented below are obtained from the model trained with self-supervised transfer learning.

Detection Results. Fig. S4 shows additional samples of video scene detection. The results highlight that BaSSL [6] is susceptible to background changes and shot transitions in adjacent frames, leading to over-segmentation of scenes. In contrast, our method could address these challenges without causing scene over-segmentation, which thanks to the ability of our TCS to enhance the relations between dissimilar shots in the same scene.

Feature Similarity Maps. Fig. S5 illustrates the correlations among shot features in consecutive scenes. Shot features of being input are extracted using the ResNet shot en-

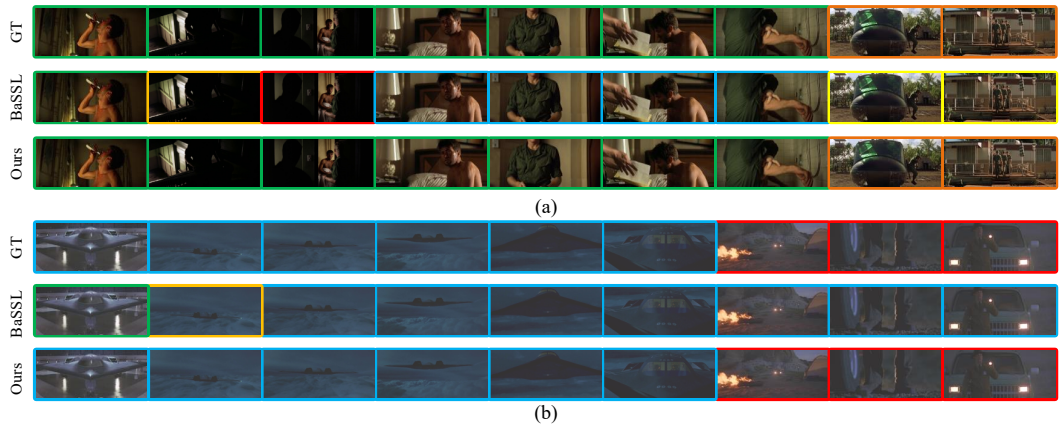


Figure S4. Visualized comparisons of the proposed method with the previous BaSSL [6]. GT denotes ground-truth scene boundaries for reference. The border of the same color represents frames from the same scene.

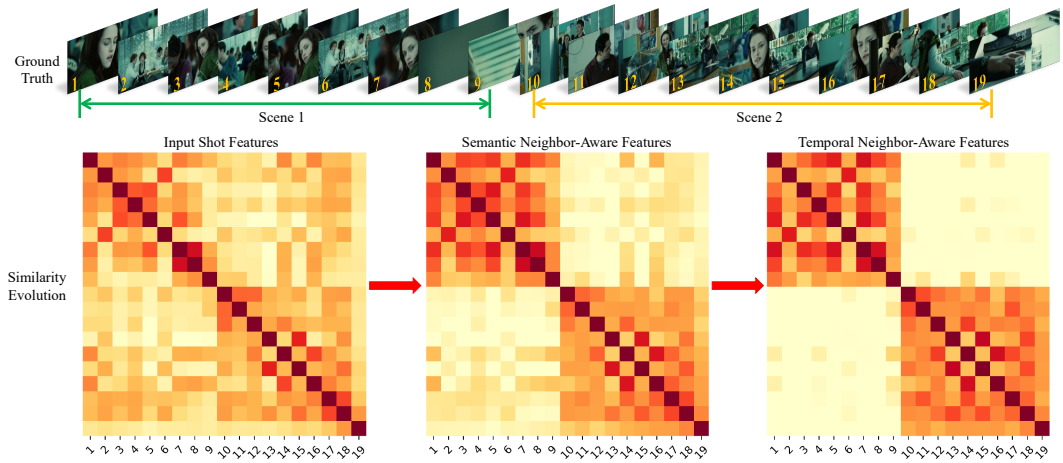


Figure S5. Visualization of feature similarity graphs at different stages. Shot features are initially extracted using the backbone. As detailed in Sec. 3.2, semantic neighbor-aware features are produced by the RNF module, while temporal neighbor-aware features are generated by the RNT module.

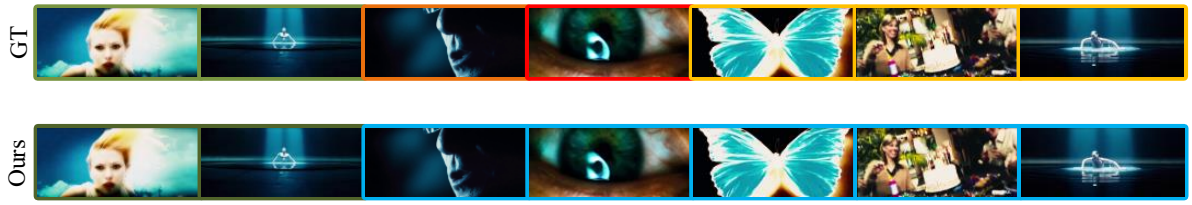


Figure S6. Visualization of failure scene detection. Our model occasionally cannot detect video scenes due to lacking consideration of audio or artistic information.

coder. Semantic neighbor-aware features defined in Eq. (5) are produced by the RNF module, and temporal neighbor-aware features defined in Eq. (9) are output by the RNT module. As shown in Fig. S5, distinguishing two scenes solely from original shot features can be challenging due to the prevalence of similar shots in both scenes. In contrast, the semantic neighbor-aware features allows for distinguishing between scenes as they weaken the relationships between similar shots from different scenes while enhancing those between similar shots within the same scene. More than that, the similarity map derived from temporal neighbor-aware features exhibits a clearer boundary be-

tween two scenes because our NeighborNet further goes beyond semantic neighbor-awareness and strengthens the connections between shots within the same scene.

3. Broader Impact and Limitation

Broader Impact. We propose the NeighborNet that compares multiple neighboring shots across various dimensions to establish relations between shots. This model can serve as a valuable reference for tasks that necessitate contextual comparisons, such as video summarization [3], video highlight detection [2], and movie trailer generation [8].

Limitation. Although our proposed method excels at detecting video scenes, as shown in Fig. S6, it cannot detect some video cases because it leaves more factors such as audio or artistic techniques (montage) unconsidered. This fact inspires us to explore future research directions that consider multiple modalities and incorporate artistic techniques.

References

[1] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. A deep siamese network for scene detection in broadcast videos. In *ACM Int. Conf. Multimedia*, pages 1199–1202, 2015. 1

[2] Bei Gan, Xiujun Shu, Ruizhi Qiao, Haoqian Wu, Keyu Chen, Hanjun Li, and Bo Ren. Collaborative noisy label cleaner: Learning scene-aware trailers for multi-modal highlight detection in movies. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18898–18907, 2023. 4

[3] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14867–14878, 2023. 4

[4] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Eur. Conf. Comput. Vis.*, pages 709–727, 2020. 1

[5] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18749–18758, 2023. 1

[6] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Bassl: Boundary-aware self-supervised learning for video scene segmentation. In *ACCV*, pages 485–501, 2022. 1, 2, 3

[7] Daniel Rotman, Dror Porat, and Gal Ashour. Robust and efficient video scene detection using optimal sequential grouping. In *ISM*, pages 275–280, 2016. 1

[8] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N. Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *Eur. Conf. Comput. Vis.*, pages 300–316, 2020. 4