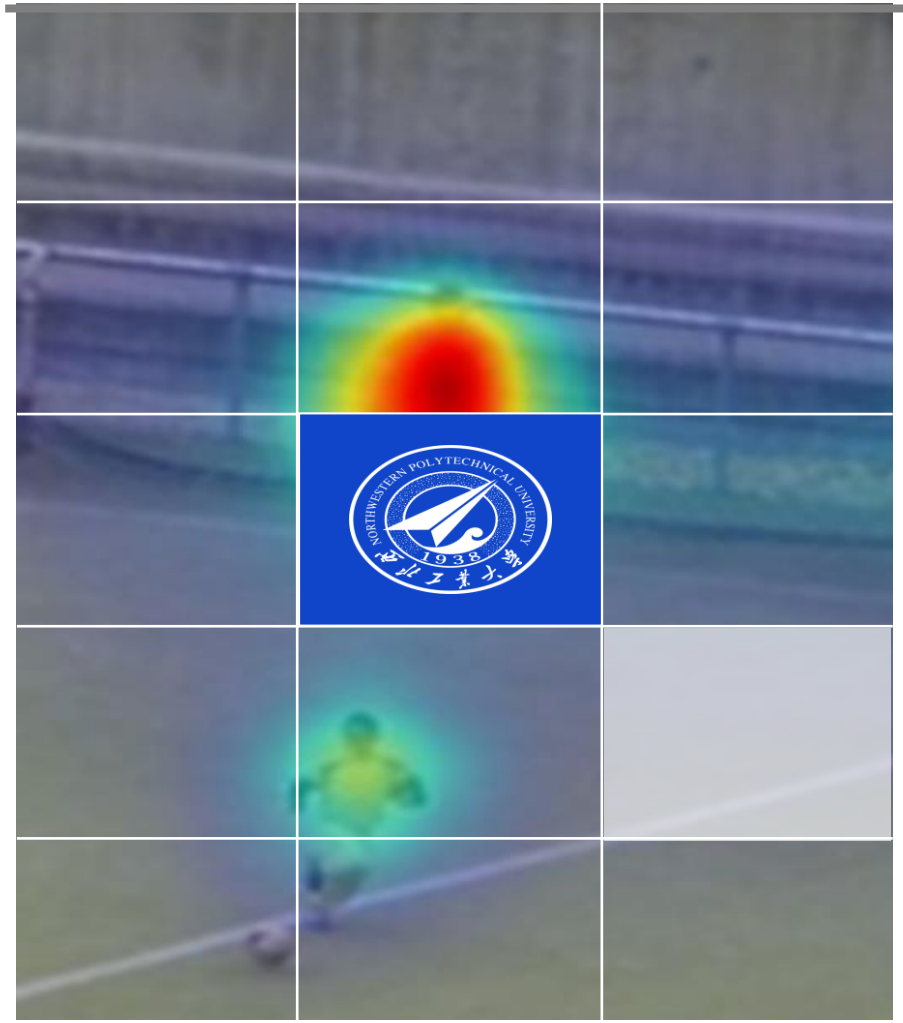
TSMSAN: A Three-Stream Multi-Scale Attentive Network for Video Saliency Detection

Jingwen Yang1, Guanwen Zhang1*, Jiaming Yan1， Wei Zhou1
1 School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China

January 10，2021

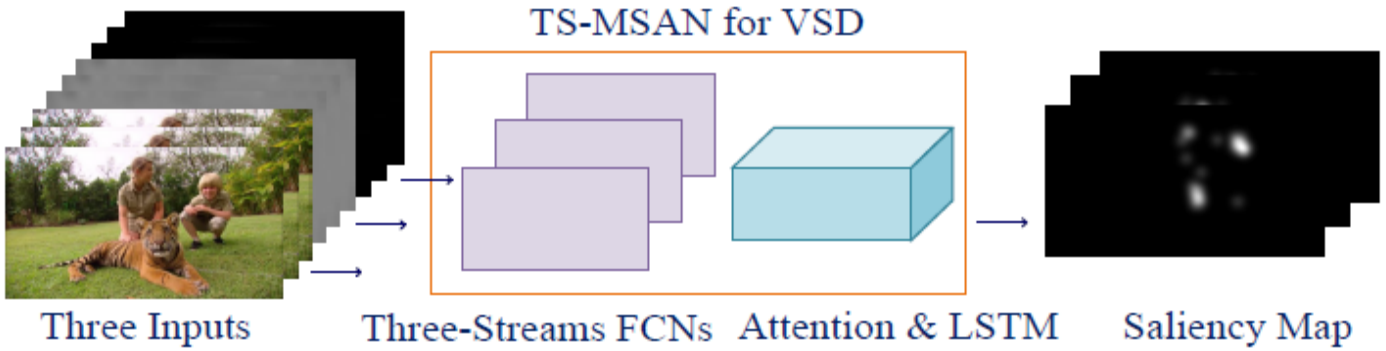
CONTENTS



Introduction

We proposed a three-stream multi-scale attentive network (TSMSAN) for saliency detection in dynamic scenes.

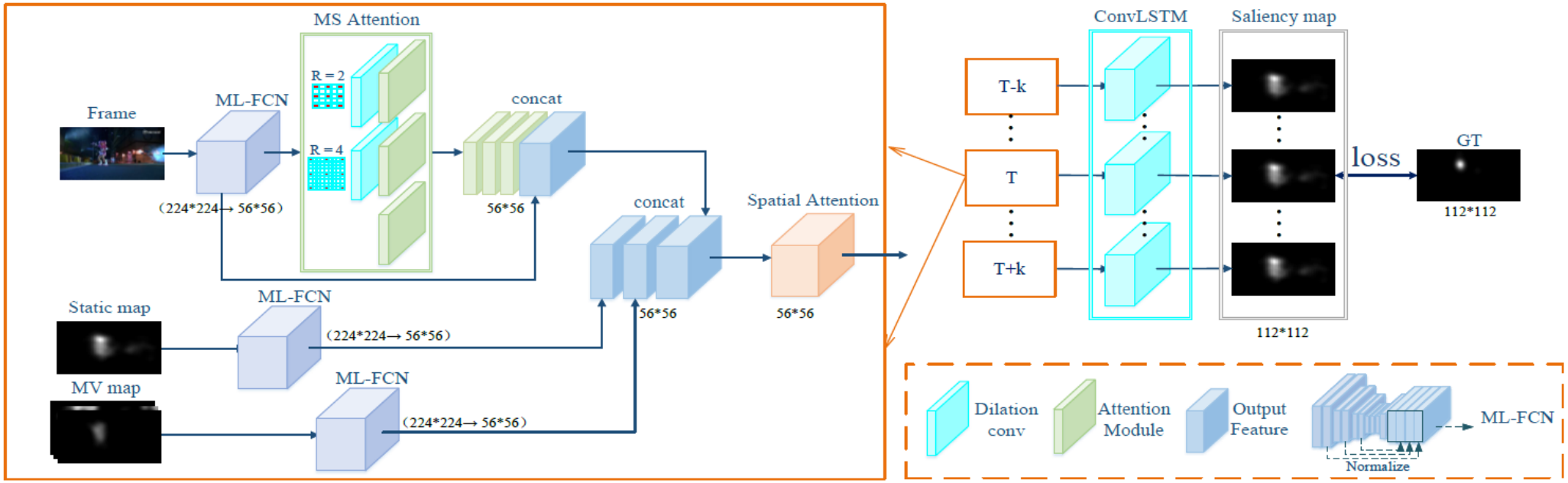
TSMSAN integrates motion vector (MV) representation, static saliency map, and RGB information in multi-scales together into one framework on the basis of Fully Convolutional Network (FCN) and spatial attention mechanism.



a) On the one hand, the respective motion features, spatial features, as well as the scene features can provide abundant information for video saliency detection.

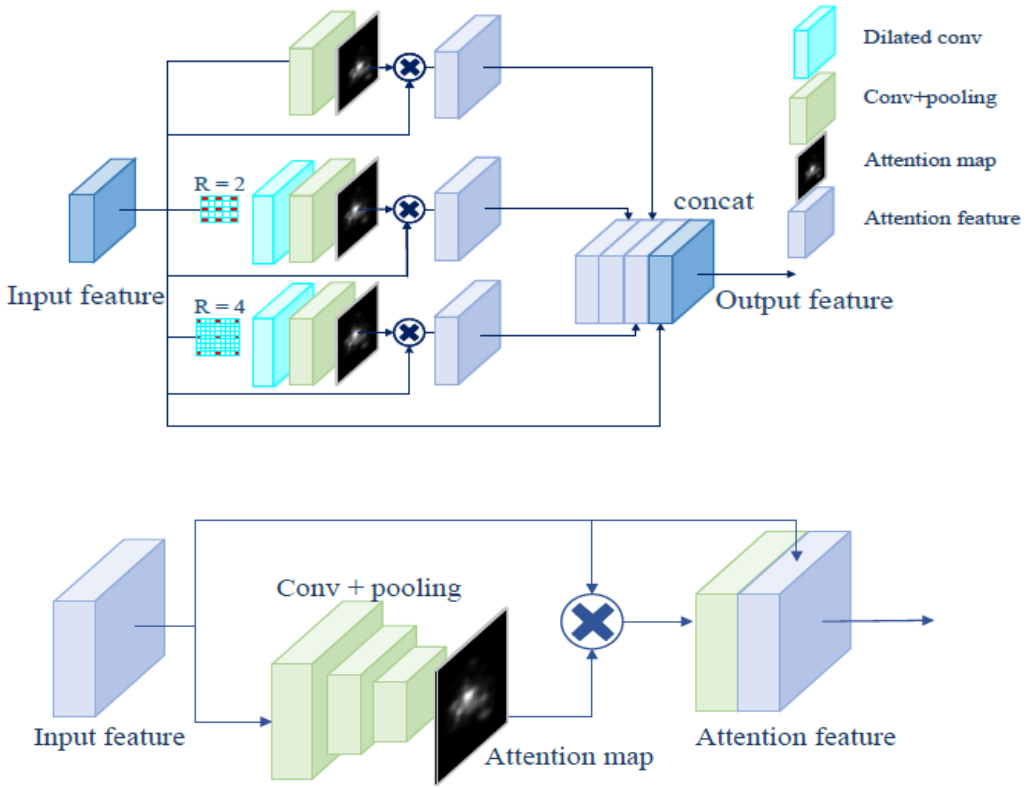
b) On the other hand, spatial attention mechanism can combine features with multi-scales to focus on key information in dynamic scenes.

Methods--Architecture



Architecture of the proposed TSMSAN. Three Multi-Level FCNs extract the features from three inputs. A Multi-Scale Attention module with dilated convolution is implemented in the frame stream. A spatial attention module and a convLSTM follow the output features from the three streams to further encode the spatiotemporal features.

Method--Attention Modules



a) A Multi-Scale Attention module based on spatial attention mechanism and dilated convolution combining features with multi-scales is adopted in the stream that takes RGB frames as input.

b) The spatial attention module follows the concatenated feature from the three streams. It outputs an attention map after further spatial feature extraction. Afterwards, the attention feature and the input feature are concatenated as the final output feature to retain the less important information.

Results of three inputs

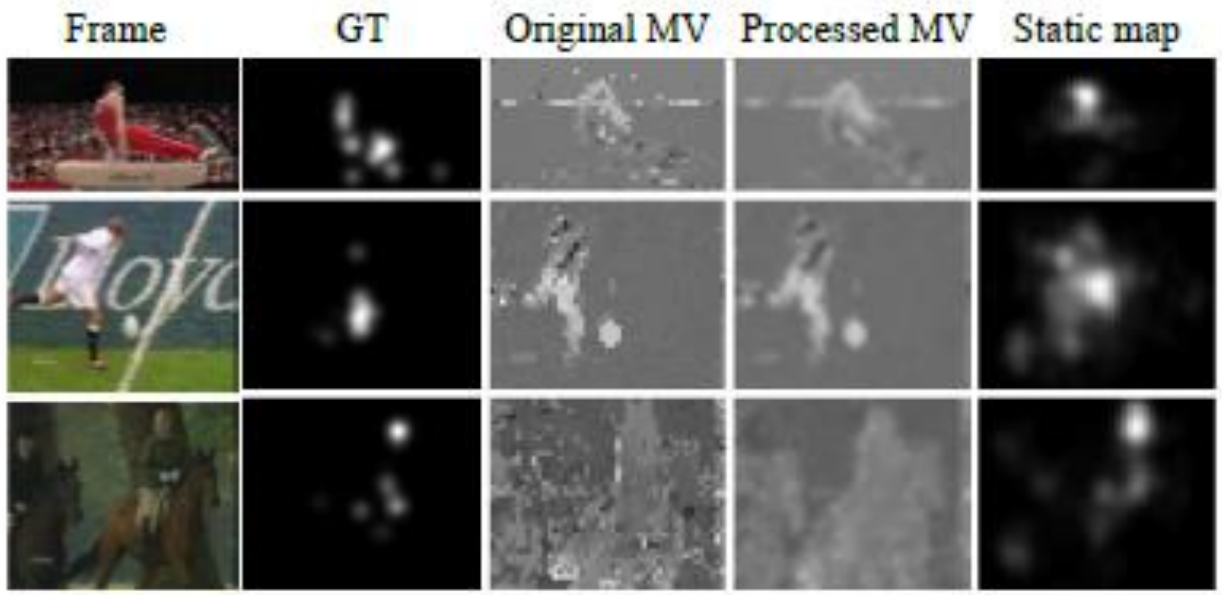


Fig. 5. Samples of three inputs in TSMSAN

The preprocessed MV representation shows more explicit motion information. The static saliency map obtained can roughly grasp the object in the scene, but it has little reflection of motion information.

Results on saliency metrics

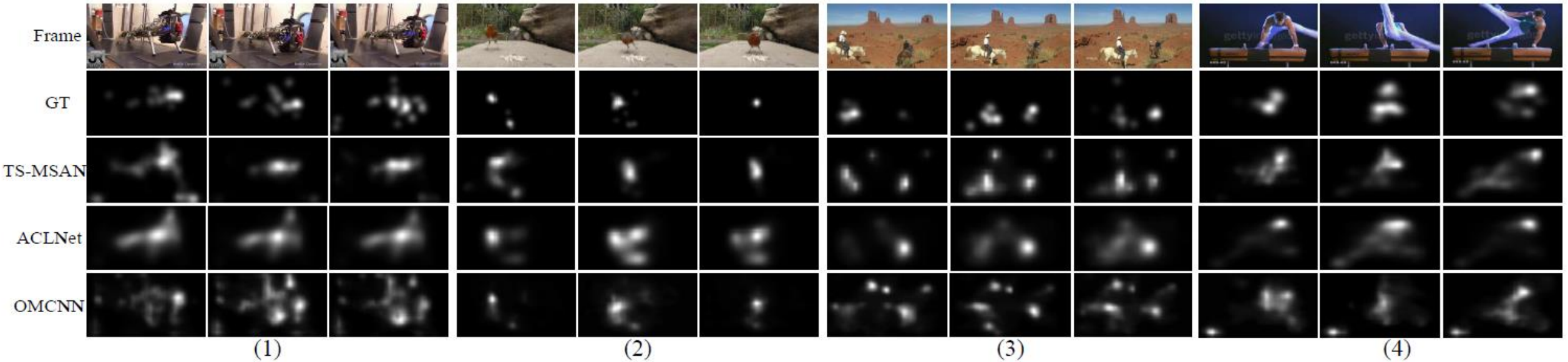
TABLE I
COMPARISON WITH STATE-OF-THE-ARTS

Testing set	Method	NSS↑	CC↑	SIM↑
UCF-sports	OMCNN [20]	2.089	0.405	0.321
	Two-stream [17]	1.753	0.343	0.264
	ACLNet [23]	3.200	0.603	0.496
	TSMSAN	3.589	0.616	0.490
Hollywood-2	OMCNN [20]	2.313	0.446	0.356
	Two-stream [17]	1.748	0.382	0.276
	ACLNet [23]	3.049	0.609	0.519
	TSMSAN	3.150	0.584	0.502

TABLE II
RESULTS ON UCF-SPORTS USING DIFFERENT TRAINING SETS

Training set	Method	NSS↑	CC↑	SIM↑
LEDOV	OMCNN [20]	2.089	0.405	0.321
	TSMSAN	2.347	0.454	0.334
DHF1K	ACLNet [23]	2.559	0.517	0.403
	TSMSAN	2.660	0.480	0.410
UCF-sports	ACLNet [23]	3.200	0.603	0.496
	TSMSAN	3.589	0.616	0.490
Hollywood-2	ACLNet [23]	2.186	0.452	0.364
	TSMSAN	2.577	0.465	0.395

Performance comparison



Reference

I. M. Carrasco, “Visual attention: The past 25 years,” Vision research, vol. 51, no. 13, pp. 1484–1525, 2011.

II. X. Huang, C. Shen, X. Boix, and Q. Zhao, “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 262–270.

III. C. Bak, A. Kocak, E. Erdem, and A. Erdem, “Spatio-temporal saliency networks for dynamic saliency prediction,” IEEE Transactions on Multimedia, vol. 20, no. 7, pp. 1688–1698, 2017.

IV. L. Jiang, M. Xu, and Z. Wang, “Predicting video saliency with objectto-motion cnn and two-layer convolutional lstm,” arXiv preprint arXiv: 1709.06316, 2017.

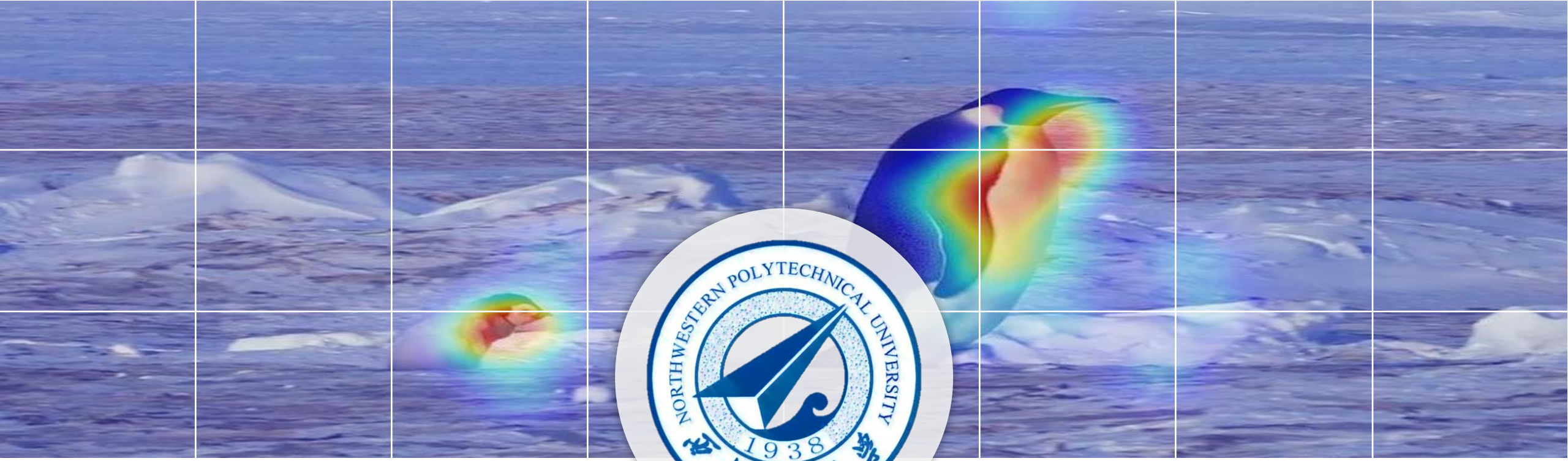
V. W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, “Revisiting video saliency: A large-scale benchmark and a new model,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4894–4903. II-B, IV-A, I, II, IV-C1, IV-C2, IV-C3

VI. P. Linardos, E. Mohedano, J. J. Nieto, N. E. O’Connor, X. Giro-I-Nieto, and K. Mcguinness, “Simple vs complex temporal recurrences for video saliency prediction,” 2019.

VII. K. Min and J. J. Corso, “Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp.2394–2403.

VIII. S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in Advances in neural information processing systems, 2015, pp. 802–810.

IX. S. Mathe and C. Sminchisescu, “Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 7, pp.1408–1424, 2014.



THANK YOU FOR ATTENTION