Salient Object Detection in 360 videos

#	Title	Year	Venue	Description
1	Salient Object Detection: A Survey [8]	2014	Computational Visual Media	A review of SOD methods before 2014.
2	A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception [46]	2015	Comput.Graph.Forum	A report introduces exhaustive concepts of human eye; a review of eye-gaze-related researches edifies future CV development.
3	Saliency Prediction in the Deep Learning Era: Successes and Limitations [7]	2019	TPAMI	A review of deep learning-based models and datasets/metrics for 2D saleincy prediction.
4	Salient Object Detection in the Deep Learning Era: An In-Depth Survey [54]	2020	arXiv	A review of SOD methods before 2020.
5	VR content creation and exploration with deep learning: a survey [52]	2020	Computational Visual Media	A review of deep learning-based methods for VR images/videos processing.
6	On the Synergies between Machine Learning and Stereo: a survey [41]	2020	arXiv CVPR2019(tutorial)	A review of deep learning-based models for monocular depth estimation in panoramas.
7	Deep Audio-Visual Learning: A Survey [74]	2020	arXiv	A review of audio-visual learning methods before 2019.

Table 1: Summary of previous reviews.

No.	Model	Year	Pub.	Task	SL.	Base	Label	Loss	Metric	Training Set	# training	code
1	360-spatialization [37]	2018	NIPS	SS	Sel.	STFT/UNet	Non	STFT-distance	STFT/ENV/EMD	REC-STREET YT-ALL	123k 0.1s samples 3976k samples	py-o
2	AVE/AVOL-Net [4]	2018	ECCV	cmR/L	Sel.	CNN/FC	Non	AVC [3]	nDCG/heatmap [38]	AudioSet [22]	263K 10s clips	py-n/o
3	DMRFE/AVDLN [50]	2018	ECCV	EL/cmL	S/W	CNN/LSTM	Tem.	MCE/L_c	heatmap/accuracy	AVE [50]	4K(T.) $\geq$2s clips	py-o
4	PixelPlayer [71]	2018	ECCV	SS/L	Sel.	ResNet/STFT	Non	SCE/L1	NSDR/SIR/SAR	MUSIC [71]	500 videos	py-o
5	SoundLoc [47]	2018	CVPR	L	W	CNN/FC	bbox	SSL [47]	cIoU [47]	Flickr-SoundNet [5]	144K frames	Non
6	**A/V-CoSeg** [45]	2019	ICASSP	Seg/SS	Sel.	UNet/ResNet	pol.	BCE	IoU/SDR/SIR	AVE [50]	3,339 $\geq$2s clips	Non
7	VehicleTrack [17]	2019	ICCV	Track	Sel.	YOLOv2	bbox	Rank [6]/OD [44]	AP	AudioVideoTrack [17]	227K 1s clips	Non
8	DDT [70]	2019	ICCV	L/SS	Sel.	I3D [9]	Non	BCE(on spec.)	SDR/SIR/SAR/HE	MUSIC-21 [70, 71]	1,065 videos	Non
9	CO-SEPARATION [20]	2019	ICCV	SS	Sel.	UNet/STFT	Non	CE [20]/L1	SDR/SIR/SAR	[22, 71]/AVBench [18]	122K 10s clips	py
10	MONO2BINAURAL [19]	2019	CVPR	m2b/SS	Sel.	UNet/STFT	Non	L2/L1	STFT/ENV-Dis. SDR/SIR/SAR	FAIR-Play [19]	1,497 10s clip	py-o
11	VSLNet [67]	2020	ACL	NLVL	S	CNN	mom.	CE	IoU	[67]	60K moments	py-o
12	IMGAUD2VID [21] IMGAUD-SKIMMING [21]	2020	CVPR	AR	U	LSTM/R21D distillation	cls	L1/KL	mAP	Kinetics [24]	300K 10s clips	py-o

Table 2: **Summary of recently proposed models for audio-visual learning.** cmR = cross-modal retrieval. L = (sound source) localization. EL = event localization. cmL = cross-modal localization. SL. = supervision level. S = supervised. W = weakly supervised. U = un-supervised. Sel. = self supervised. T = traditional method. CNN = convolutional neural network. FC = fully connected layer. py = python. n/o = non official. Tem. = temporally labeled segments (viusal/audio). MCE = multi-class cross-entropy loss. L_c = contrastive loss function. SS = sound separation. STFT = Short- Time Fourier Transform. NSDR = Normalized Signal-to-Distortion Ratio. SIR = Signal-to-Interference Ratio. SAR = Signal-to-Artifact Ratio. SCE = sigmoid cross entropy. SSL = semi-supervised loss. frm = frames. BCE = binary cross entrophy. pol. = polygon. OD = object detection. HE = human evaluation. CE = cross entrophy. m2b = mono to binaural. AR = action recognition. cls = class. NLVL = natural language video localization. mom. = moments annotations.

No.	Dim.	Model	Year	Pub.	Task	Base	Training Set	Label	# training	F_β	F_β^ω	M	S_α	E_ξ	code
1	360-RGB	DDS [31]	2020	JSTSP	SOD	CNNs	360SOD [31]	o-pw	400 images	.650	.652	.023	-	-	py-o
2	2D-RGBD	UCNet [68]	2020	CVPR	SOD	CVAE [48]	AugedGT [68]	o-pw	-	.855∼.919	-	.019∼.066	.864∼.934	.901∼.967	-
3	2D-RGBD	JLDCF [15]	2020	CVPR	SOD	VGG16 ResNet101	NLPR NJU2K	o-pw	2,2K images	.862∼.919	-	.022∼.078	.854∼.929	.893∼.968	-
4	2D-RGBD	SSF [69]	2020	CVPR	SOD	CIM CAU/BSU	DUT-RGBD NLPR	o-pw	1,485 images 700 images	.867∼.915	-	.025∼.044	.859∼.915	-	-
5	2D-RGB	F^3Net [57]	2020	AAAI	SOD	CFD/CFM MLS	DUTS-TR	o-pw	10,533 images	.766∼.925	-	.028∼.062	.838∼.924	.859∼.953	-
6	2D-RGB	DFI [33]	2020	arXiv	SOD ed./sk.	CNNs PPM [72]	DUTS-TR [51]	o-pw	10,533 images	.829∼.945	-	.031∼.100	.802∼.921	-	-
7	2D-RGB	SISO [26]	2019	WACV	SOD	3D FCN [25]	SESIV [26]	i-pw	58 videos (3,944 frames)	-	-	-	-	-	m.-o
8	2D-RGB	SVSNet [56]	2019	ACM MM	r-SOD	FCN	RVSOD [56]	o-pw	242 videos (7140 frames) DAVIS [40]/DUT [65]	.816 .745(DAVIS)	-	.089 .047(DAVIS)	-	-	py-o
9	2D-RGB	RSDNet [2]	2018	CVPR	r-SOD	ResNet101	Pascal-S [32]	o-pw	425 images	.880	-	.090	-	-	ca-o

Table 3: **Summary of recently proposed models for salient object detection.** SOD = salient object detection. F_β = F-measure [1]. F_β^ω = weighted F-measure [34]. M = mean absolute error [39]. S_α = S-measure [13]. E_ξ = E-measure [14]. (n/)o = (non) official. o(i)-pw = object(instance)-level pixel-wise annotations. m. = matlab. ca = caffe. py = python. ed. = edge detection. sk. = skeleton detection. r-SOD = ranking SOD.

No.	Dim.	Model	Year	Pub.	Base	Training Set	# Training	Label	Code	Key Issue
1	360	MT-DNN [42]	2020	TMM	CNNs/convLSTM	[64]	65 videos (3,501 viewports)	SalMap	py-o	viewports influence fixations
2	2D	UVA-Net [16]	2020	AAAI	knowledge distillation	AVS1K...	1K aerial videos	SalMap	-	accelerating SP
3	360	DHP [64]	2019	TPAMI	DRL [36]	PVS-HM [64]	61 videos	HM map	py-o	-
4	2D	DAVE [49]	2019	arXiv	3D ResNet log mel-spectrogram	AVE [49]	150 videos	SalMap	py-o	visual-audio SP
5	2D	SKD-DVA [30]	2019	TIP	knowledge distilation	-	-	SalMap	-	accelerating SP
6	2D	TASED-Net [35]	2019	ICCV	3D FCN (S3D [63])	DHF1K [55]	700 videos	Fixations	-	3D-FCN for video SP
7	360	E/H-SalPredict [75]	2019	TMM	EMP, HMP	Salient360! [43]	85 images	SalMap	-	-

Table 4: **Summary of recently proposed models for saliency prediction.** SP = saliency prediction. HM = head movement.

No.	Task	Method	Year	Pub.	Components	Training Set	#Training	Label
1	Classification Semantic Segmentation	tangent-360 [12]	2020	CVPR	-	-	-	-
2	D-epth Estimation	(waiting for paper...) [27]	2020	CVPR	-	-	-	-
3	Depth Estimation	OmniMVS [58]	2020	TPAMI	uncertainty prior	Weather/House/Thing	700/2048/9216 images	-
4	Depth Estimation	360SD-Net [53]	2020	ICRA	ASPP [10]	MP3D/SF3D	1602/800 images	-
5	Classification	SGCN [66]	2020	CVPR	GConv, HPool	ModelNet40 [61]	9843 samples	cls
6	VP	MDN [59]	2020	AAAI	s2cnn [59]	PanoUCF101 [59]	35 users records	cls
7	OD Evaluation	Rep R-CNN [73]	2020	AAAI	SphBB, SphIoU	ImageNet	-	bbox
8	VQA	V-CNN [29]	2019	CVPR	VP/VQ-Net	VQA-ODV [28]	432 impired videos	HM info.
9	OD/IS/SS	Pano-BlitzNet [23]	2019	arXiv	BlitzNet [11]	SUN360 [62]	400 images	i-pw
10	VQA	FAST-VQA [60]	2019	TMM	spatial quality degradation	-	-	tra.

Table 5: **Summary of recent methods for 360 processing.** OD = object detection. IS = instance segmentation. SS= semantic segmentation. o(i)-pw = object(instance)-level pixel-wise annotations. cls = class. VP = viewport prediction. VQA = video quality assessment. tra. = salient trajectories.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009.

[2] Md Amirul Islam, Mahmoud Kalash, and Neil D. B. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[7] A. Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[8] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 1411, 11 2014.

[9] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[11] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. BlitzNet: A real-time deep network for scene understanding. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. *CVPR*, 2019.

[13] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE ICCV*, pages 4548–4557, 2017.

[14] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *IJCAI*, pages 698–704, 2018.

[15] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Kui Fu, Jia Li, Yafei Song, Yu Zhang, Shiming Ge, and Yonghong Tian. Ultrafast video attention prediction with coupled knowledge distillation. *arXiv preprint arXiv:1904.04449*, 2019.

[17] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[18] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.

[19] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019.

[20] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019.

[21] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.

[22] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[23] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J Guerrero. What's in my room? object recognition on indoor panoramic images. *arXiv preprint arXiv:1910.06138*, 2019.

[24] W Kay, J Carreira, K Simonyan, B Zhang, C Hillier, S Vijayanarasimhan, F Viola, T Green, T Back, P Natsev, et al. The kinetics human action video dataset. arxiv 2017. *arXiv preprint arXiv:1705.06950*.

[25] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 38.1–38.13. BMVA Press, September 2017.

[26] Trung-Nghia Le and Akihiro Sugimoto. Semantic instance meets salient object: Study on video semantic salient instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1779–1788. IEEE, 2019.

[27] Jia Zheng Junfei Zhang Rui Tang Shugong Xu Jingyi Yu Shenghua Gao Lei Jin, Yanyu Xu. Geometric structure based and regularized depth estimation from 360° indoor imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[28] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 932–940, New York, NY, USA, 2018. Association for Computing Machinery.

[29] Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, and Xiaoming Tao. Viewport proposal cnn for 360deg video quality assessment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] J. Li, K. Fu, S. Zhao, and S. Ge. Spatiotemporal knowledge distillation for efficient estimation of aerial video saliency.

IEEE Transactions on Image Processing, 29:1902–1914, 2020.

[31] J. Li, J. Su, C. Xia, and Y. Tian. Distortion-adaptive salient object detection in 360° omnidirectional images. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):38–48, 2020.

[32] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.

[33] Jiang-Jiang Liu, Qibin Hou, and Ming-Ming Cheng. Dynamic feature integration for simultaneous detection of salient object, edge and skeleton. *arXiv preprint arXiv:2004.08595*, 2020.

[34] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *IEEE CVPR*, pages 248–255, 2014.

[35] Kyle Min and Jason J. Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[36] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[37] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, pages 362–372, 2018.

[38] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[39] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, pages 733–740, 2012.

[40] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[41] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and stereo: a survey. *arXiv preprint arXiv:2004.08566*, 2020.

[42] M. Qiao, M. Xu, Z. Wang, and A. Borji. Viewport-dependent saliency prediction in 360° video. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[43] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. A dataset of head and eye movements for 360 degree images. In *MMSys*, pages 205–210. ACM, 2017.

[44] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[45] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361, 2019.

[46] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. *Comput. Graph. Forum*, 34(6):299–326, Sept. 2015.

[47] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[48] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3483–3491, Cambridge, MA, USA, 2015. MIT Press.

[49] Hamed R Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019.

[50] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[51] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, 2017.

[52] Miao Wang, Xu-Quan Lyu, Yufang Li, and Fang-Lue Zhang. Vr content creation and exploration with deep learning: A survey. *Computational Visual Media*, 6:28 – 3, 2020.

[53] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sd-net: 360° stereo depth estimation with learnable cost volume. *ICRA 2020*, 2019.

[54] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.

[55] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[56] Zheng Wang, Xinyu Yan, Yahong Han, and Meijun Sun. Ranking video salient object detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 873–881, New York, NY, USA, 2019. Association for Computing Machinery.

[57] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. *AAAI*, 2020.

[58] C. Won, J. Ryu, and J. Lim. End-to-end learning for omnidirectional stereo matching with uncertainty prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[59] Chenglei Wu, Ruixiao Zhang, Zhi Wang, and Lifeng Sun. A spherical convolution approach for learning long term viewport prediction in 360 immersive video. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI '20, New York, NY, USA, 2020.

[60] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin. Quality assessment for video with degradation along salient trajectories. *IEEE Transactions on Multimedia*, 21(11):2738–2749, 2019.

[61] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[62] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702, 2012.

[63] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[64] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2693–2708, 2019.

[65] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE CVPR*, pages 3166–3173, 2013.

[66] Qin Yang, Chenglin Li, Wenrui Dai, Junni Zou, GuoJun Qi, and Hongkai Xiong. Rotation equivariant graph convolutional network for spherical image classification. In *IEEE CVPR*, 2020.

[67] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *ACL 2020*, 2020.

[68] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

[69] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *CVPR*, 2020.

[70] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[71] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[72] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[73] Pengyu Zhao, Ansheng You, Yuanxing Zhang, Jiaying Liu, Kaigui Bian, and Yunhai Tong. Spherical criteria for fast and accurate 360 object detection. In *AAAI*, 2020.

[74] Hao Zhu, Mandi Luo, Rui Wang, Aihua Zheng, and Ran He. Deep audio-visual learning: A survey. *arXiv preprint arXiv:2001.04758*, 2020.

[75] Y. Zhu, G. Zhai, X. Min, and J. Zhou. The prediction of saliency map for head and eye movements in 360 degree images. *IEEE Transactions on Multimedia*, pages 1–1, 2019.