REFERENCES

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," *arXiv:2101.06072 [cs]*, Jan. 2021, Accessed: Jul. 25, 2021. [Online]. Available: http://arxiv.org/abs/2101.06072

[2] M. Barbieri, L. Agnihotri, and N. Dimitrova, "Video summarization: methods and landscape," Orlando, FL, Nov. 2003, pp. 1–13. doi: 10.1117/12.515733.

[3] K. Zhou, Y. Qiao, and T. Xiang, "Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward," *arXiv:1801.00054 [cs]*, Feb. 2018, Accessed: Jul. 19, 2021. [Online]. Available: http://arxiv.org/abs/1801.00054

[4] G. Liang, Y. Lv, S. Li, S. Zhang, and Y. Zhang, "Unsupervised Video Summarization with a Convolutional Attentive Adversarial Network," *arXiv:2105.11131 [cs]*, May 2021, Accessed: Jul. 19, 2021. [Online]. Available: http://arxiv.org/abs/2105.11131

[5] M. A. Samsuden, N. M. Diah, and N. A. Rahman, "A Review Paper on Implementing Reinforcement Learning Technique in Optimising Games Performance," in *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, Shah Alam, Malaysia, Oct. 2019, pp. 258–263. doi: 10.1109/ICSEngT.2019.8906400.

[6] Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, "Deep Reinforcement Learning for Query-Conditioned Video Summarization," *Applied Sciences*, vol. 9, no. 4, p. 750, Feb. 2019, doi: 10.3390/app9040750.

[7] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video Captioning via Hierarchical Reinforcement Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 4213–4222. doi: 10.1109/CVPR.2018.00443.

[8] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He, "Hierarchically Structured Reinforcement Learning for Topically Coherent Visual Story Generation," *arXiv:1805.08191 [cs]*, Jan. 2019, Accessed: Jul. 25, 2021. [Online]. Available: http://arxiv.org/abs/1805.08191

[9] J.-H. Huang and M. Worring, "Query-controllable Video Summarization," *arXiv:2004.03661 [cs]*, Apr. 2020, Accessed: Jul. 19, 2021. [Online]. Available: http://arxiv.org/abs/2004.03661

[10] H. Wei, B. Ni, Y. Yan, H. Yu, and X. Yang, "Video Summarization via Semantic Attended Networks," p. 8.

[11] K. Zhang, K. Grauman, and F. Sha, "Retrospective Encoders for Video Summarization," in *Computer Vision – ECCV 2018*, vol. 11212, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 391–408. doi: 10.1007/978-3-030-01237-3_24.

[12] A. Kanehira, L. Van Gool, Y. Ushiku, and T. Harada, "Viewpoint-Aware Video Summarization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 7435–7444. doi: 10.1109/CVPR.2018.00776.

[13] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, "Stacked Memory Network for Video Summarization," in *Proceedings of the 27th ACM*

International Conference on Multimedia, Nice France, Oct. 2019, pp. 836–844. doi: 10.1145/3343031.3350992.

[14] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video Summarization With Attention-Based Encoder–Decoder Networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, Jun. 2020, doi: 10.1109/TCSVT.2019.2904996.

[15] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos," *AAAI*, vol. 33, pp. 8393–8400, Jul. 2019, doi: 10.1609/aaai.v33i01.33018393.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, Accessed: Feb. 27, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," p. 12.

[18] X. Shang, Z. Yuan, A. Wang, and C. Wang, "Multimodal Video Summarization via Time-Aware Transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China, Oct. 2021, pp. 1756–1765. doi: 10.1145/3474085.3475321.

[19] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 2017, pp. 1092–1102. doi: 10.18653/v1/D17-1114.

[20] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?," *arXiv:2102.05095 [cs]*, Jun. 2021, Accessed: Jan. 06, 2022. [Online]. Available: http://arxiv.org/abs/2102.05095

[21] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *arXiv:1910.10683 [cs, stat]*, Jul. 2020, Accessed: Feb. 26, 2022. [Online]. Available: http://arxiv.org/abs/1910.10683

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, vol. 25. Accessed: Feb. 26, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[24] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," *arXiv:1406.2661 [cs, stat]*, Jun. 2014, Accessed: Mar. 11, 2022. [Online]. Available: http://arxiv.org/abs/1406.2661

[25] Yong Rui, T. S. Huang, and S. Mehrotra, "Exploring video structure beyond the shots," in *Proceedings. IEEE International Conference on Multimedia Computing and Systems (Cat. No.98TB100241)*, Austin, TX, USA, 1998, pp. 237–240. doi: 10.1109/MMCS.1998.693648.

[26] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video Summarization with Long Short-term Memory," *arXiv:1605.08110 [cs]*, Jul. 2016, Accessed: Feb. 25, 2022. [Online]. Available: http://arxiv.org/abs/1605.08110

[27] J. Lee and S. Abu-El-Haija, “Large-Scale Content-Only Video Recommendation,” in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Oct. 2017, pp. 987–995. doi: 10.1109/ICCVW.2017.121.

[28] C. C. Park and G. Kim, “Expressing an Image Stream with a Sequence of Natural Sentences,” in *Advances in Neural Information Processing Systems*, 2015, vol. 28. Accessed: Feb. 25, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/17e62166fc8586dfa4d1bc0e1742c08b-Abstract.html

[29] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, “Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation,” in *Advances in Neural Information Processing Systems*, 2016, vol. 29. Accessed: Feb. 25, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/f442d33fa06832082290ad8544a8da27-Abstract.html

[30] Z. Li and L. Yang, “Weakly Supervised Deep Reinforcement Learning for Video Summarization With Semantically Meaningful Reward,” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, Jan. 2021, pp. 3238–3246. doi: 10.1109/WACV48630.2021.00328.

[31] X. He *et al.*, “Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks,” in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice France, Oct. 2019, pp. 2296–2304. doi: 10.1145/3343031.3351056.

[32] J. Gao, X. Yang, Y. Zhang, and C. Xu, “Unsupervised Video Summarization via Relation-aware Assignment Learning,” *IEEE Trans. Multimedia*, pp. 1–1, 2020, doi: 10.1109/TMM.2020.3021980.

[33] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, “Discriminative Feature Learning for Unsupervised Video Summarization,” *arXiv:1811.09791 [cs]*, Nov. 2018, Accessed: Jul. 19, 2021. [Online]. Available: http://arxiv.org/abs/1811.09791

[34] K. Zhou, T. Xiang, and A. Cavallaro, “Video Summarisation by Classification with Deep Reinforcement Learning,” *arXiv:1807.03089 [cs]*, Sep. 2018, Accessed: Jul. 25, 2021. [Online]. Available: http://arxiv.org/abs/1807.03089

[35] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, “Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video,” *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 996–1009, May 2019, doi: 10.1109/TKDE.2018.2848260.

[36] A. Vaswani *et al.*, “Attention Is All You Need,” *arXiv:1706.03762 [cs]*, Dec. 2017, Accessed: Feb. 26, 2022. [Online]. Available: http://arxiv.org/abs/1706.03762

[37] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, “Cycle-SUM: Cycle-consistent Adversarial LSTM Networks for Unsupervised Video Summarization,” *arXiv:1904.08265 [cs]*, Apr. 2019, Accessed: Jul. 19, 2021. [Online]. Available: http://arxiv.org/abs/1904.08265

[38] G. Jocher *et al., ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations*. Zenodo, 2021. doi: 10.5281/zenodo.4679653.

[39] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” *arXiv:1405.0312 [cs]*, Feb. 2015, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/1405.0312
[40] R. Chandra *et al.*, “METEOR: A Massive Dense & Heterogeneous Behavior Dataset for Autonomous Driving,” *arXiv:2109.07648 [cs]*, Sep. 2021, Accessed: Jan. 06, 2022. [Online]. Available: http://arxiv.org/abs/2109.07648
[41] “Documentation - WebNLG Challenges.” https://webnlg-challenge.loria.fr/docs/ (accessed Mar. 02, 2022).
[42] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.
[43] Z. Liu *et al.*, “Swin Transformer V2: Scaling Up Capacity and Resolution,” *arXiv:2111.09883 [cs]*, Nov. 2021, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/2111.09883
[44] L. Yuan *et al.*, “Florence: A New Foundation Model for Computer Vision,” *arXiv:2111.11432 [cs]*, Nov. 2021, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/2111.11432
[45] M. Xu *et al.*, “End-to-End Semi-Supervised Object Detection with Soft Teacher,” *arXiv:2106.09018 [cs]*, Aug. 2021, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/2106.09018
[46] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Scaled-YOLOv4: Scaling Cross Stage Partial Network,” *arXiv:2011.08036 [cs]*, Feb. 2021, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/2011.08036
[47] D. Tran, H. Wang, M. Feiszli, and L. Torresani, “Video Classification With Channel-Separated Convolutional Networks,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 5551–5560. doi: 10.1109/ICCV.2019.00565.
[48] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, “Temporal Pyramid Network for Action Recognition,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 588–597. doi: 10.1109/CVPR42600.2020.00067.
[49] J. Clive, K. Cao, and M. Rei, “Control Prefixes for Text Generation,” *arXiv:2110.08329 [cs]*, Oct. 2021, Accessed: Mar. 04, 2022. [Online]. Available: http://arxiv.org/abs/2110.08329