REFERENCES

[1] Haoye Cai, Chunyan Bai, Yuwing Tai, and Chikeung Tang. 2018. Deep Video Generation, Prediction and Completion of Human Action Sequences. In *European Conference on Computer Vision*. 374–390.

[2] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 190–200.

[3] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine learning*. 160–167.

[4] Alexis Conneau, Douwe Kiela, Holger Schwenk, LoÃŕc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly Supervised Dense Event Captioning in Videos. In *Advances in Neural Information Processing Systems*. 3059–3069.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[8] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In *Proceedings of the 32th International Conference on Machine Learning*. 1462–1471.

[9] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. 2018. Probabilistic video generation using holistic attribute control. In *European Conference on Computer Vision*. 452–467.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[11] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4193–4202.

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 221–231.

[13] Shuiwang Ji, Ming Yang, Kai Yu, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 221–231.

[14] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. (2013). arXiv:1312.6114.

[15] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[16] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. 2017. Attentive Semantic Video Generation Using Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1435–1443.

[17] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which Training Methods for GANs Do Actually Converge? (2018). arXiv:1801.04406.

[18] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. 2017. Sync-Draw: Automatic Video Generation using Deep Recurrent Attentive Architectures. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1096–1104.

[19] Katsunori Ohnishi. 2018. Hierarchical Video Generation from Orthogonal Information : Optical Flow and Texture. (2018). arXiv:1711.09618v2.

[20] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To Create What You Tell: Generating Videos from Captions. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1789–1798.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[22] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning Text-to-image Generation by Redescription. (2019). arXiv:1903.05854.

[23] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proceedings of International Conference on Learning Representations*.

[24] Scott Reed, Zeynep Akata, Xinchen Yan, and Lajanugen Logeswaran. 2016. Generative Adversarial Text to Image Synthesis. In *Proceedings of the 33th International Conference on Machine Learning*. 1060–1069.

[25] Tim Salimans, Ian Goodfellow, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*. 1–10.

[26] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing Human Actions: A Local SVM Approach. In *Proceedings of International Conference on Pattern Recognition*. 32–36.

[27] Guangyao Shen, Wenbing Huang, Chuang Gan, Mingkui Tan, Junzhou Huang, Wenwu Zhu, and Boqing Gong. 2019. Facial Image-to-Video Translation by a Hidden Affine Transformation. In *Proceedings of the 27th ACM international conference on Multimedia*.

[28] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of International Conference on Learning Representations*.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*. 3104–3112.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3d Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4489–4497.

[32] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. (2014). arXiv:1412.4729.

[33] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*. 613–621.

[34] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. 2017. The Pose Knows: Video Forecasting by Generating Pose Futures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3352–3361.

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. 2015. Towards Good Practices for Very Deep Two-Stream ConvNets. (2015). arXiv:1507.02159.

[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained Text to Image Generation with Attentional Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1316–1324.

[37] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2868–2876.

[38] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. 2010. Deconvolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[39] Han Zhang, Tao Xu, and Hongsheng Li. 2017. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5908–5916.

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 1–1.

[41] Junyan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2242–2251.