Information and Communications Technology (ICOIACT), 2022, pp. 355–360, doi: 10.1109/ICOIACT55506.2022.9971855.

[10] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 2641–2649, 2015, doi: 10.1109/ICCV.2015.303.

[11] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.

[12] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image Captioning Based on Deep Neural Networks," *MATEC Web Conf.*, vol. 232, pp. 1–7, 2018, doi: 10.1051/matecconf/201823201052.

[13] H. Shi, P. Li, B. Wang, and Z. Wang, "Image captioning based on deep reinforcement learning," *ACM Int. Conf. Proceeding Ser.*, vol. 01052, pp. 1–7, 2018, doi: 10.1145/3240876.3240900.

[14] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," *MM 2017 - Proc. 2017 ACM Multimed. Conf.*, pp. 1549–1557, 2017, doi: 10.1145/3123266.3123366.

[15] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," *ICMR 2016 - Proc. 2016 ACM Int. Conf. Multimed. Retr.*, pp. 271–275, 2016, doi: 10.1145/2911996.2912049.

[16] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 417–421, 2017, doi: 10.18653/v1/P17-2066.

[17] H. A. Al-muzaini, T. N. Al-yahya, and H. Benhidour, "Automatic Arabic image captioning using RNN-LSTM-based language model and CNN," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 67–73, 2018, doi: 10.14569/IJACSA.2018.090610.

[18] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, "Adaptive Attention Generation for Indonesian Image Captioning," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166244.

[19] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 3, pp. 2048–2057, 2015.

[20] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6298–6306, 2017, doi: 10.1109/CVPR.2017.667.

[21] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, observe and tell: Attribute-driven attention model for image captioning," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2018-July, pp. 606–612, 2018, doi: 10.24963/ijcai.2018/84.

[22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 4651–4659, 2016, doi: 10.1109/CVPR.2016.503.

[23] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, no. c, pp. 8927–8936, 2019, doi: 10.1109/ICCV.2019.00902.

[24] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 1–11, 2019.

[25] V. Atliha and D. Šešok, "Text augmentation using BERT for image captioning," *Appl. Sci.*, vol. 10, no. 17, 2020, doi: 10.3390/app10175978.

[26] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Appl. Sci.*, vol. 8, no. 5, 2018, doi: 10.3390/app8050739.

[27] W. Zhang, W. Nie, X. Li, and Y. Yu, "I m age Caption Generation With Adaptive Transfor m er," pp. 521–526, 2019.

[28] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image Captioning Through Image Transformer," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12625 LNCS, pp. 153–169, 2021, doi: 10.1007/978-3-030-69538-5_10.

[29] C. Cormier, "Bleu," *Landscapes*, vol. 7, no. 1, pp. 16–17, 2005, doi: 10.3917/chev.030.0107.

[30] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," *Intrinsic Extrinsic Eval. Meas. Mach. Transl. and/or Summ. Proc. Work. ACL 2005*, no. June, pp. 65–72, 2005.

[31] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.

[32] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4566–4575, 2015, doi: 10.1109/CVPR.2015.7299087.

[33] R. Staniute and D. Šešok, "A systematic literature review on image captioning," *Appl. Sci.*, vol. 9, no. 10, 2019, doi: 10.3390/app9102024.