# DAFTAR PUSTAKA

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile. IEEE.

Bai, L., Islam, M., and Ren, H. (2023). CAT-ViL: Co-Attention Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery. arXiv:2307.05182 [cs].

Bazi, Y., Rahhal, M. M. A., Bashmal, L., and Zuair, M. (2023). Vision–Language Model for Visual Question Answering in Medical Imagery. *Bioengineering*, 10(3):380. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Beddiar, D.-R., Oussalah, M., and Seppänen, T. (2023). Automatic captioning for medical imaging (MIC): a rapid review of literature. *Artificial Intelligence Review*, 56(5):4019–4076.

Bercovich, E. and Javitt, M. C. (2018). Medical Imaging: From Roentgen to the Digital Revolution, and Beyond. *Rambam Maimonides Medical Journal*, 9(4):e0034. Publisher: Rambam Health Corporation.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, Virtual Only Conference.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334. ISSN: 1063-6919.

Han, D., Shi, J., Zhao, J., Wu, H., Zhou, Y., Li, L.-H., Khan, M. K., and Li, K.-C. (2025). LRCN: Layer-residual Co-Attention Networks for visual question answering. *Expert Systems with Applications*, 263:125658.

Karaca, Z. and Aydin, I. (2025). Med-VQA: Performance Analysis of Question-Answering Systems on Medical Images. In *2025 29th International Conference on Information Technology (IT)*, pages 1–4. ISSN: 2836-3744.

Kim, B. S., Kim, J., Lee, D., and Jang, B. (2025). Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges. *ACM Comput. Surv.*, 57(10):249:1–249:35.

Lameesa, A., Silpasuwanchai, C., and Alam, M. S. B. (2025). VG-CALF: A vision-guided cross-attention and late-fusion network for radiology images in Medical Visual Question Answering. *Neurocomputing*, 613:128730.

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251. Publisher: Nature Publishing Group.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., and Ge, Z. (2023). Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. (2021). Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. ISSN: 1945-8452.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sharma, D., Purushotham, S., and Reddy, C. K. (2021). MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826. Publisher: Nature Publishing Group.

Vasuki, P., Kanimozhi, J., and Devi, M. B. (2017). A survey on image preprocessing techniques for diverse fields of medical imagery. In *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, pages 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, , Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System:

Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs].

Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, Honolulu, HI. IEEE.

Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6274–6283.