**REFERENCES**

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2017**). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering**. *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.1707.07998>

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015). **Microsoft COCO Captions: Data Collection and Evaluation Server.** *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.1504.00325>

Hochreiter, S., & Schmidhuber, J. (1997). **Long Short-Term Memory.** *Neural Computation, 9*(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., & Girshick, R. (2023). **Segment anything**. *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* <https://doi.org/10.1109/iccv51070.2023.00371>

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). **ImageNet classification with deep convolutional neural networks.** *Communications of the ACM, 60(6), 84–90.* <https://doi.org/10.1145/3065386>

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). **BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation**. *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.2201.12086>

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). **PyTorch: An Imperative Style, High-Performance Deep Learning Library.** *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.1912.01703>

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). **Learning Transferable Visual Models From Natural Language Supervision**. *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.2103.00020>

Ren, S., He, K., Girshick, R., & Sun, J. (2015). **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.** *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.1506.01497>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). **Attention Is All You Need**. *arXiv (Cornell University).* <https://doi.org/10.48550/arxiv.1706.03762>