# MUHAMMAD MAAZ

muhammad.maaz@mbzuai.ac.ae, +971-52-5326156
Website: muhammadmaaz.com, GitHub: mmaaz60 LinkedIn: mmaaz60

## PERSONAL PROFILE

Computer Vision researcher with comprehensive hands-on expertise spanning the entire cycle of Deep Learning-enabled Computer Vision solutions, from initial design to final deployment. Current focus lies in harnessing multi-modal understanding derived from vision and text to augment machines' common-sense reasoning capabilities. Specifically, engaged in the application of LLMs to build robust, scalable vision systems.

## EDUCATION

**Mohamed bin Zayed University of Artificial Intelligence, UAE** *Jan 2023 - Continue*
PhD in Computer Vision

**Mohamed bin Zayed University of Artificial Intelligence, UAE** *Dec 2020 - Dec 2022*
Research Based Masters in Computer Vision
CGPA: 4.0/4.0

**University of Engineering and Technology, Pakistan** *Sep 2014 - Aug 2018*
B.Sc. Electrical Engineering
CGPA: 3.7/4.0 (First class with honors)

## RESEARCH

**GLaMM: Pixel Grounding Large Multimodal Model** *(Nov-2023, Under Review)*
*Hanoona Rasheed, **Muhammad Maaz**, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, Fahad Khan*
This work introduces Video-ChatGPT, a conversational model adept at producing meaningful conversations about videos. It combines the capabilities of LLMs with a pretrained visual encoder adapted for spatiotemporal video representation. A new dataset of 100,000 video-instruction pairs is used to train Video-ChatGPT acquired via manual and semi-automated pipeline that is easily scalable and robust to label noise.

**Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models** *(Jun-2023, Under Review)*
***Muhammad Maaz**, Hanoona Rasheed, Salman Khan, Fahad Khan*
This work introduces Video-ChatGPT, a conversational model adept at producing meaningful conversations about videos. It combines the capabilities of LLMs with a pretrained visual encoder adapted for spatiotemporal video representation. A new dataset of 100,000 video-instruction pairs is used to train Video-ChatGPT acquired via manual and semi-automated pipeline that is easily scalable and robust to label noise.

**MaPLe: Multi-modal Prompt Learning** *(CVPR-2023)*
*Muhammad Uzair Khattak, Hanoona Rasheed, **Muhammad Maaz**, Salman Khan, Fahad Khan*
This work proposes to learn prompts in both vision and language branches of pretrained CLIP to adapt it to various downstream tasks. Multi-modal Prompt Learning (MaPLe) improves the alignment between the vision and language representations, encouraging mutual synergy.

**Fine-tuned CLIP Models are Efficient Video Learners** *(CVPR-2023)*
*Hanoona Rasheed, Muhammad Uzair Khattak, **Muhammad Maaz**, Salman Khan, Fahad Khan*
This work demonstrates that a simple Video Fine-tuned CLIP (ViFi-CLIP) baseline is generally sufficient to bridge the domain gap from images to videos. The analysis indicates that frame-level processing from the CLIP image-encoder, feature pooling, and similarity matching help model the temporal cues within ViFi-CLIP.

**Class-agnostic Object Detection with Multi-modal Transformer** *(ECCV-2022)*

***Muhammad Maaz****, Hanoona Rasheed, Salman Khan, Fahad Khan, Rao M. Anwer, Ming-Hsuan Yang*

This work explores the potential of Multi-modal Vision Transformers (MViTs) for class-agnostic object detection. Through extensive experiments across various domains and unique objects, the efficacy of MViTs in localizing generic objects in images was established.

**Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection** *(NeurIPS-2022)*

*Hanoona Rasheed,* ***Muhammad Maaz****, Muhammad Uzair Khattak, Salman Khan, Fahad Khan*

This work proposes to solve the Open-vocabulary detection (OVD) problem using a pretrained CLIP model, adapting it for object-centric local regions. A region-conditioned weight transfer method is introduced to leverage complementary benefits from both region-based distillation and image-level supervision.

## WORK EXPERIENCE

**Hazen.ai** *Jul 2020 - Dec 2020*

*Computer Vision Engineer*

Developed a traffic light phase detection solution for road safety applications. I trained a network to learn embeddings for traffic light phases (red, yellow, green and black) using triplet loss. The network was robust enough to handle different road scenarios including day and night scenes. The product was deployed on the NVIDIA Jetson devices using TensorRT.

**Confiz Limited** *Jun 2018 - Jul 2020*

*Computer Vision Engineer*

Led Shopper Value - Computer Vision Team where I was responsible for technological evolution and scalability of Computer Vision Products; Visitor Tracking and Visitor Profile.

- **Visitor Tracking:** A Person Detection and Tracking solution to identify the engaged and ignored areas of a retail store. Our utmost challenge was to process 7 to 10 video streams on an i5 CPU or NVIDIA Jetson device with fair enough accuracy. We experimented with Yolov3 and pruned it to get the desired speed and accuracy balance. We used Network Distillation to train camera specific small neural networks. We also focused to effectively utilize the CPU cores and use optimized inference frameworks like Intel's OpenVino and TensorRT for edge deployment.
- **Visitor Profile:** A face recognition solution capable of generating visitor's and buyer's demographics and visit frequency data for the brick and mortar retail stores. FaceNet like architecture was being used to prepare face embeddings.

## CERTIFICATES

| | |
|---|---|
| **Computer Vision Nano Degree** | *Udacity* |
| **Deep Learning Specialization by deeplearning.ai** | *Coursera* |
| **Machine Learning with TensorFlow on Google Cloud Platform Specialization** | *Coursera* |
| **Advance Machine Learning with TensorFlow on Google Cloud Platform Specialization** | *Coursera* |

## TECHNICAL STRENGTHS

| | |
|---|---|
| **Computer Sciences** | Computer Vision, Deep Learning, Machine Learning |
| **Programming Languages & Tools** | Python, C, Java, PyTorch |

## EXTRA-CURRICULAR AND ACHIEVEMENTS

- Acted as a reviewer for Transformer Models in Vision, Special Issue in IEEE TPAMI, ECCVW-22, ACCV-22, NeurIPS-22, CVPR-23 and ICCV-23.
- Former Secretary of Graduate Student Council at MBZUAI, and VP Operations at IET UET Chapter.
- Enjoy travelling, cricket and table tennis.

## REFERENCES

Dr. Salman Khan
Academic Advisor
Associate Professor at MBZUAI
✉ salman.khan@mbzuai.ac.ae

Prof. Fahad Khan
Academic Advisor
Professor at MBZUAI
✉ fahad.khan@mbzuai.ac.ae

Mr. Hashim Ali
Chief Operating Officer
Confiz Limited
✉ hashim.ali@confiz.com