Cagri Gungor

☎ (+1) 412 925 8651 ✉ m.cagrigungor@gmail.com
⌂ cagrigungor.github.io 🎓 Google Scholar in mcagrigungor

Education

University of Pittsburgh *Pittsburgh, PA*
Ph.D. *in Intelligent Systems Program* Aug 2021 - January 2026 (Expected)
M.Sc. *in Intelligent Systems Program* Aug 2021 - April 2025 (Completed)

- Advisor: Adriana Kovashka

Bilkent University *Ankara, Turkey*
B.Sc. *in Computer Science* Sep 2016 - Jun 2021 (Completed)

- Received full merit scholarship given to the students having exceptional success in the university entrance exam.

Interests

Computer Vision Generative AI Multimodal Learning Foundational Models (MLLMs, VLMs, LLMs)

Publications

- *Towards Generalization of Tactile Generation: Reference-Free Evaluation in a Leakage-Free Setting*, Under Submission [Paper]
- *Integrating Audio Narrations to Strengthen Domain Generalization in Action Recognition*, Accepted to 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing **ICASSP'25** [Paper]
- *Enhancing Weakly-Supervised Object Detection on Static Images through (Hallucinated) Motion*, To appear, the 3rd Workshop on Large Language and Vision Models for Autonomous Driving at **WACV'25** [Paper]
- *Boosting Weakly Supervised Object Detection using Fusion and Priors from Hallucinated Depth*, To appear, 2024 IEEE/CVF Winter Conference on Applications of Computer Vision **WACV'24** [Paper]
- *Complementary Cues from Audio Help Combat Noise in Weakly-supervised Object Detection*, To appear, 2023 IEEE/CVF Winter Conference on Applications of Computer Vision **WACV'23** [Paper]

Experience

Amazon *Santa Cruz, CA*
Incoming Applied Scientist Intern Aug 2025 - December 2025 (Expected)

- Will be joining the Amazon Last Mile – New Initiatives team for a stealth-mode project focused on multimodal computer vision.

University of Pittsburgh *Pittsburgh, PA*
Graduate Research Assistant Aug 2021 - Present

- Researching **multimodal computer vision**, integrating diverse sensory inputs—**audio, language, depth, motion,** and **touch**—to investigate the **complementary role** of different modalities in advancing tasks such as **object detection, video action recognition, domain generalization, image generation,** and the development of **evaluation metrics** for image generation.

Dolby Laboratories *San Francisco, CA*
Research Intern Summer 2023

- Conducted cutting-edge **audio-visual research** on **temporal activity localization** and **video summarization**, contributing to next-generation **video content understanding**.
- Developed a multi-modal deep learning pipeline using **temporal and cross-modal attention** to prioritize key frames by fusing **audio-visual cues**, improving temporal activity detection mAP by $>$ **15%**.
- Built an demo showcasing the research to a broad audience, including technical and non-technical stakeholders, leading to positive feedback from senior leadership.

Lenovo Research *Chicago, IL*
Research Intern Summer 2022

- Conducted research on **low-light image enhancement** and **image deblurring**, developing deep learning models to improve image quality for Motorola (a Lenovo company) smartphones enhancing visibility in low-light conditions.

- Spearheaded a data collection initiative independently collecting nighttime outdoor samples, **tripling the dataset size** and significantly improving model generalization to real-world low-light scenarios.
- Designed and optimized a deep learning-based **image de-blurring algorithm**, integrating a **multi-scale CNN with attention-based** feature refinement to enhance motion-degraded images, achieving $>$ **10**% improvement.

3DUniversum *Amsterdam, Netherlands*
Research Intern Summer 2020

- Conducted research on **visual emotion manipulation** in videos, developing a **GAN-based framework** for 3D facial expression synthesis and seamless emotion transfer.
- Integrated **audio-driven facial animation** into the GAN pipeline to address **lip-sync issues**, improving temporal consistency and speech synchronization in video synthesis.

Selected Projects

Enhancing Generalization in Tactile Image Generation *Under Submission*
Generative AI, Stable Diffusion, Vision-to-Touch, Reference-Free Metrics, Multimodal Learning

- Developed a **vision-to-touch** synthesis framework using **latent diffusion** to convert visual images into tactile images, leveraging **text descriptions** for more accurate, **material-specific texture** details.
- Identified and resolved up to **90% data leakage** in common tactile datasets, introducing a leak-free evaluation protocol and new **reference-free metrics** to ensure **robust** generalization and **reliable** performance.

Domain Generalization in Multimodal Action Recognition *ICASSP'25*
Domain Generalization, Multimodal Fusion, Audio Narrations, Multimodal LLMs (MLLMs)

- Discovered that **audio** and **motion** modalities exhibit **greater resilience** to domain shifts than **appearance**, with performance drops of only **32.7**% and **25.8**%, respectively, versus **54.8**% for appearance—underscoring their **pivotal role** in **domain generalization** across unseen scenarios and locations.
- Proposed a novel framework that aligns **audio narrations** with audio features to reinforce action representations and leverages **consistency ratings** between audio and visual narrations to optimize **audio's role**, achieving **4.8**% higher average accuracy on ARGO1M.

Boosting Object Detection with Hallucinated Depth *WACV'24*
Weakly Supervised Object Detection (WSOD), Depth Fusion, Contrastive Learning, Depth-Language

- Proposed an **amplifier** method for WSOD that incorporates **hallucinated depth** information through a **Siamese network** and **contrastive learning**, enhancing **representation learning** and enabling effective **fusion** of depth and RGB features.
- Introduced a novel mechanism that leverages language captions to compute **context-aware depth priors** per object class, which **re-weight** pseudo ground-truth boxes, improving object detection mAP by $>$ **14**%.

Complementary Audio to Enhanced Object Detection *WACV'23*
WSOD, Audio-Visual Learning, Sound localization, Multimodal Attention

- Proposed a framework that integrates audio with visual data, enhancing object detection by leveraging **audio cues** in a weakly supervised setting.
- Introduced mechanisms such as **indirect path** linking visual features to predictions via audio and an **attention path** prioritizing key visual regions leveraging audio, achieving **7**% higher mAP on AudioSet.

Technical Skills

Programming Languages	Python, Java, C/C++, MATLAB, R, SQL, Linux
ML & Deep Learning	PyTorch, TensorFlow, OpenCV, SciKit-Learn, SpaCy, NLTK, Pandas, NumPy
Database	MySQL, MongoDB, SQLite
Software Engineering	Git, Jupyter Notebook, Agile, Scrum

Professional Services

Conference Reviewer:	Conference on Computer Vision and Pattern Recognition (CVPR), 2024-2025
	Winter Conference on Applications of Computer Vision (WACV), 2025
	Association for the Advancement of Artificial Intelligence (AAAI), 2024-2025
	IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2025