Domain Adaptation for Fair and Robust Computer Vision.

Tarun Kalluri (sskallur@ucsd.edu)

Overview In recent years, the field of computer vision and deep learning has been propelled by the emergence of large-scale foundational models, which unlock remarkable capabilities in various tasks - from scene understanding and robot navigation to text-to-image synthesis and nuanced multimodal dialogue. However, their reliance on uncurated, web-sourced data presents new challenges, where the biases in training data can lead to unfair outcomes for under-represented subgroups and their lack of robustness outside their training domain limits their universal adoption. Thus, my doctoral research focuses on improving the generalizability of vision models across under-represented domains in the real world. My past research proposed successful scalable solutions for domain adaptation Kalluri et al. (2022), highlighted the severe limitation of current models in showcasing geographical robustness through a novel large-scale dataset Kalluri et al. (2023b) and devised efficient solutions to bridge such challenging domain gaps through natural language guidance Kalluri et al. (2024). My future research goals lies in identifying potential harms of emerging multimodal foundational models towards sub-groups that are under-represented in training datasets and proposing effective mitigation strategies, thus enabling AI models to generate images of living rooms in Columbia as good as living rooms in Europe, and equip them to answer questions about festivals in India with the same efficiency as they would about president of the USA. My PhD research achieved prestigious recognitions, including a WACV'23 best paper finalist and the IPE PhD fellowship in 2021.

Research Highlights: Large-Scale Datasets and Multimodal Solutions for Domain Adaptation
A major impediment to research progress in geographical fairness is the lack of suitable benchmarks informing the geographical sensitivity of existing methods. To address this limitation, I led the efforts in creating a large scale dataset and evaluation benchmark called GeoNet Kalluri et al. (2023b) dedicated to study geographical disparities on standard vision tasks. We analyze several salient properties of geographic adaptation, and highlight the limitations of several modern algorithms in bridging geographic domain gaps. We posit that GeoNet would not only provide researchers the opportunity to assess the suitability of state-of-the-art algorithms towards universal deployment, but also inspire design of robust AI models that can efficiently handle dynamic changes in geographies while maintaining superior performance. I also led the organization of a successful workshop and challenge based on the dataset at ICCV 2023.

A key potential lies in leveraging natural language guidance in improving domain transfer across challenging shifts for visual perception tasks. Our recent work LaGTran Kalluri et al. (2024) devises a simple yet highly efficient framework to incorporate readily available or easily acquired text supervision in conferring additional generalization capabilities for both images and videos. Motivated by our observation that semantically richer text modality has more favorable transfer properties, we devise a transfer mechanism to use a source-trained text-classifier to generate predictions on the target text descriptions, and utilize these predictions as supervision for the corresponding images. Through this emphasis on cost-effective or easily producible text supervision, we open new possibilities for advancing robustness in scenarios with limited manual supervision.

Alongside these fundamental innovations, my past research also proposed practical solutions for challenging applications like slomo video generation Kalluri et al. (2023a), where our novel architecture inspired by 3D U-Net and training recipe based on large-scale, unlabeled video data yielded unprecedented generation quality while being upto 6x faster than all existing methods.

Future Vision: Trustworthy Foundational AI Models
I plan to focus my research on two pivotal thrusts in the future. Firstly, I aim to explore the fairness properties of emerging generative AI applications, particularly Large Language Models (LLMs) and text-to-image (T2I) models. Notably, these models often face challenges in delivering optimal performance across low-resource domains like low and mid-income societies, presenting an open-challenge in extending their applicability to diverse populations. Secondly, I aspire to harness the recent advancements in foundational models towards enhancing the robustness capabilities in diverse applications. These models trained on multimodal, web-scale datasets showcase strong zero-shot and emergent intelligence capabilities which can drive progress in out-of-distribution generalization on several downstream tasks. This dual-pronged approach seeks to offer novel insights into the inclusivity challenges posed by the rapid progress in generative AI while providing practical solutions for the widespread adoption of these models in the future.

In summary, my research focuses on developing data-efficient algorithms to enhance test-time robustness and fairness of computer vision models, and my longer term goal is to make emerging AI technology universally deployable and more widely accessible.

References

Kalluri, T., Majumder, B. P., and Chandraker, M. (2024). Tell, don't show: Language guidance eases transfer across domains in images and videos.

Kalluri, T., Pathak, D., Chandraker, M., and Tran, D. (2023a). Flavr: Flow-agnostic video representations for fast frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2071–2082.

Kalluri, T., Sharma, A., and Chandraker, M. (2022). Memsac: Memory augmented sample consistency for large scale domain adaptation. In *European Conference on Computer Vision*, pages 550–568. Springer.

Kalluri, T., Xu, W., and Chandraker, M. (2023b). Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15368–15379.