Human-Centered Multi-Modal Spatiotemporal Modeling in The Era of Infrastructure

Pan He, University of Florida

My primary research interests are computer vision and deep learning. With the advent of cheaper and smarter sensors, we have seen concomitant skyrocketing interest but have yet to provide fundamental Artificial Intelligence (AI) techniques sufficient in infrastructure applications. Moreover, sensory data collected from infrastructure is multimodal, noisy, and often unreliable. It remains a significant challenge to effectively and efficiently process and understand sensory data and connect various perceived information intelligently. My ultimate research goal is to develop the essential building blocks of sensing, computing, and communication and seamlessly integrate them for creating or empowering real-time downstream spatiotemporal applications, thereby deserving an immediate investigation. Furthermore, as the next-generation infrastructure systems will become ubiquitous shortly, my research will impact various facets of our daily life, including but not limited to industrial and public spaces and academic and scientific fields. In particular, I advance methods in the following research areas:

- *Spatiotemporal Visual and Point Cloud Data Modeling* [3, 4, 5, 14]. I have developed deep learning frameworks to achieve effective distributed training, inference of sparse models, and a better understanding of motion and correspondence from dynamic point clouds.
- *Trustworthy and Collaborative Multi-Modal Perception* [2, 6, 7, 10, 11, 12, 15]. I have proposed a series of models for successful processing and aggregation of information from diverse and heterogeneous video streams for downstream infrastructure applications.
- *Data-driven Human-AI Collaboration* [1, 8, 13]. To build smart infrastructure, I have progressed in building digital replicas of infrastructures with the integration of heterogeneous streams, generating diverse safety-critical scenarios for behavior analysis, and achieving a theoretical development in imposing linear equality constraints for neural networks.

My research has resulted in more than 20 research papers in workshops, conferences, and journals in the fields of computer vision (IJCV, ICCV, ECCV, BMVC, TSAS), machine learning (AAAI, ICML, TNNLS), intelligent transportation (T-ITS, ITSC, IJITSR), multimedia (ACM-MM), security and privacy (BigSecurity), and others (HiPC, ORLR). The interdisciplinary perspective of my research will encourage close collaboration to bridge machine learning, intelligent transportation, internet of things (IoT), cloud computing (CC), high-performance computing (HPC), data analytics (DA), security and privacy (SP), and cyber-physical systems (CPS). Below I describe my research progress as initial steps in the research mentioned above and present my ongoing and future research for advancing fundamental AI to create impacts for social good.

1 Spatiotemporal Visual and Point Cloud Data Modeling

One initial step towards my research goal is efficiently and effectively processing sensory data and abstracting them with symbolic or numerical descriptions, thereby generalizing their concepts and benefiting higher-level perception and decision-making tasks. Due to the dynamic nature of our world, we continuously receive and interpret sensory information over time to understand the world. A fundamental problem for processing a spatiotemporal sequence from a dynamic environment is describing its complex motion patterns as a reflection of changes over time. This is particularly challenging due to 3D/4D spatiotemporal data characteristics, e.g., severe occlusions and sparsity on the unstructured data,

additional noise, and intensive computation/memory consumption for high-dimension data. Towards this end, my research has mainly revolved around asking fundamental research questions, developing conceptual frameworks, and designing practical algorithms to understand motion and correspondence (e.g., how points in one LIDAR scan correspond to points in another scan) from dynamic point cloud sequences with minimal human supervision. This benefits downstream applications requiring dynamic visual perception, such as those in autonomous driving, smart infrastructure, and geometry processing.

To achieve effective training and inference for large-scale point clouds collected from sparse sensors, e.g., LIDARs, In [14], I investigated existing deep-learning frameworks, such as TensorFlow and Pytorch, which are originally optimized for dense tensor processing using data-parallel pipelining. I identified their main limitations—their ineffective utilization of the underlying parallelization and computation resources and their incapability to process sparse point clouds. I developed *SparsePipe*, an efficient and asynchronous pipeline parallelism approach, the first work that addresses sparse computation frameworks for point clouds and incorporates a load balancing for exploiting differential GPU characteristics in multi-node training. Compared to data-parallel training frameworks, the *SparsePipe* framework is faster and memory-efficient while maintaining high accuracy.

A popular approach is to describe the 3D motion field of a scene via estimating scene flows from a pair of frames using self-supervised learning. Nevertheless, most existing self-supervised scene flow estimation (SFE) approaches operating on discrete point clouds adopt the de facto nearest-neighbor-based similarity metrics, i.e., Chamfer Distance (CD) or Earth Mover's Distance (EMD), for generating pseudo-ground-truth scene flow. In [5], I pointed out the main limitations of CD and EMD— sensitive to outliers or computationally intensive. Then, I presented a principled framework that represents discrete point clouds as continuous probability density functions (PDFs) using Gaussian mixture models to address these limitations. Specifically, I convert SFE into the problem of recovering motion from the alignment of PDFs, achieved by a closed-form expression of the classic Cauchy-Schwarz (CS) divergence. The developed framework generates a more robust and accurate scene flow in the presence of outliers and makes noticeable gains over CD and EMD in real-world environments.

Most SFE approaches have been limited to inferring the relative motion of a point cloud pair while under-exploring the spatiotemporal processing of point cloud video sequences in a dynamic scene. To fill the gap, in [4], I introduced sequential scene flow estimation (SSFE) for point cloud sequences, which is a novel extension of SFE and demonstrated its benefits in sequential point cloud forecasting (SPF). I developed the SPCM-Net architecture by introducing a new set-to-set cost volume layer for spatiotemporal modeling. To advance future research, I presented the first benchmark for point cloud sequences containing diverse backgrounds and multiple object motions in synthetic and realistic environments to standardize training and evaluation protocols. All show great promise in spatiotemporal perception for downstream applications.

Another fundamental aspect of my research is to find *similarity* and *correspondence* between geometric shapes, such as human and animal shapes. However, the major dissatisfaction is a need for supervised learning, where massive ground truth correspondences are scarce and difficult to obtain. Unfortunately, most unsupervised approaches, considered promising, suffer from non-trivial optimization and usually achieve limited generalization performance. I proposed the first LLE-inspired algorithm, LTENet [3], for unsupervised shape correspondence by representing maps between pairs of shapes as locally linear transformations and deploying the Cauchy-Schwarz reconstruction objective to optimize embeddings towards the same universal/canonical space. The proposed approach helps regularize embedding learning to find more reliable correspondences. Importantly, it highlights the application of manifold learning

and could be helpful in other matching problems of single- and cross-modality.

Ongoing and Future Directions Comprehensively understanding dynamic scenes involves great effort in annotating large-scale datasets and leveraging cues, e.g., semantics, motion, and occupancy grid mapping, and prior knowledge, e.g., high-definition maps used in autonomous driving and traffic modeling, for object and scene geometry inference. The perception of dynamic visual data in a spatiotemporal fashion can benefit AI downstream applications. For example, 2D and 3D object detection and tracking could potentially benefit from motion and correspondence cues, given the fact that object trajectories are high-level abstractions of object movement perceiving a series of individual moving elements, e.g., points or pixels, as a whole, according to common fate, Gestalt principles of perception [9]. I plan to develop new dynamic visual modeling techniques that enable a machine to perceive 3D scenes cost-effectively. This aims to reduce annotation efforts and solve perception tasks by exploiting appearance, geometry, and contextual information between tasks, with fewer annotations. Besides, I plan to generate realistic dynamic visual data via advanced simulation or generative modeling such that we create massive synthetic or augmented data suitable for boosting model performance. I have begun progressing in the aforementioned directions by developing a self-supervised learning algorithm leveraging spatiotemporal information for outdoor LIDAR segmentation using invariance, equivariance, and motion cues and exploring differentiable graduated assignments for graph matching to refine predictions in perception tasks.

2 Trustworthy and Collaborative Multi-Modal Perception

Currently, most intelligent systems rely on processing data streams of a single modality to build fundamental perception models such as detection, segmentation, tracking, and prediction. The developed intelligent systems have been limited by their sensor ranges, applicability, and algorithmic performance, resulting in an incomplete, unconfident, and nonrobust understanding of the environments. For example, a video camera deployed at a traffic intersection is prone to occlusions, limited field of view (FoV), and light and weather conditions, which inevitably leads to significant performance degradation and high risks in safety-critical scenarios, e.g., pedestrians and vehicles at traffic intersections. It is promising to transition from single-modal perception to multi-modal perception, given that different sensors provide complementary signals—cameras capture rich semantic information, LIDARs offer precise 3D geometric and spatial information, and radars give the accurate instant velocity of moving objects. Successful aggregation of information from various sensory modalities would eventually lead to a trustworthy and collaborative perception. It is trustworthy with accurate, safe, and robust modeling on reducing uncertainties of single-modal perception due to occlusion, FoV, light, and weather conditions. It is collaborative, with sensors simultaneously working in concert and joint modeling with shared knowledge.

My recent research has demonstrated its successful processing and aggregation of information from diverse and heterogeneous video streams, including pan-tilt and fisheye cameras. For highway scenarios, I developed deep learning and hybrid approaches [2, 11, 15] for truck and trailer classification via integrating deep learning models and geometric truck features. By further utilizing logo data on trucks that are predominantly text or images, I developed an integrated approach [6, 10] to combine predictions from the text- and image-based solutions [18, 19] to determine the commodity type potentially hauled by trucks. The developed approaches are human-understandable, enabling a successful collaboration and effective interaction with traffic agencies for making informed decisions. For intersections, I co-developed the first end-to-end trainable two-stream deep learning model [12] to perform real-time object

detection, tracking, and near accident detection of road users in traffic video data. For traffic corridors, I co-developed a real-time video processing system for multi-camera vehicle tracking and travel-time estimation [7] by introducing phase-based signature matching and utilizing vehicle arrival and departure information. My research outcomes have seen their deployment on real platforms funded by the Florida Department of Transportation and the National Science Foundation: I was able to integrate innovative point cloud and video processing models and real-time traffic datasets, thereby providing researchers with detailed real-time information on high-risk events such as near-misses and traffic violations.

Ongoing and Future Directions Despite the incredible progress [7, 12, 17], the multi-modal perception still faces some challenges caused by model designs in capturing collaborative interactions between heterogeneous sensor features, limited communications under allocated resources, lack of suitable benchmarks with standard training and evaluation protocols, robustness with the presence of noise and errors. Designing a unified representation that seamlessly integrates features in different views and modalities is promising. However, effective and efficient transformations between sensor features and the unified representation remain unclear. Moreover, perception tasks from multi-modal sensors tend to be imbalanced in sample sizes across tasks, diverse in classes of interest, and mixed in both sparse and dense representations. I plan to tackle the mentioned challenges in the future while providing potential directions for multi-modal perception. My exploratory research of building simulated datasets with multiple sensors of different modalities in advanced simulators such as CARLA and SUMO could be pivotal in investigating the direction.

3 Data-driven Human-AI Collaboration

In our era of smart infrastructure, we are witnessing increasing impacts of AI integrated into virtually all infrastructure systems to improve people's daily lives. However, the performance of AI techniques is strictly bounded by the limitations of existing algorithms and tools and by the availability and quality of multi-modal sensory data. To meet real-world requirements, it is crucial to bring active human-AI interaction such that humans keep observing how machines (mis-)behave and adapting their design, development, and operation, thereby improving machines' flexibility and productivity. Besides, humans should respond to the outcomes of machines by understanding the expected impact and potential biases and mitigating any risks of AI machines.

Ongoing and Future Directions I have begun investigating this direction, and my ongoing research includes but is not limited to building smart traffic infrastructures that improve road users' safety and reduce traffic congestion. I began to build digital replicas of infrastructures to integrate heterogeneous traffic streams and generate diverse safety-critical scenarios in which (virtual) pedestrians or vehicles are put in danger. The development directly provides preferential treatment to existing traffic modeling frameworks that are limited to a few scenarios where diverse sets of erratic maneuvers (e.g., U-turns, hard braking), complex interactions (e.g., yielding, merging), and accidents (e.g., near-miss, minor, injury, and fatal accidents) are not sufficiently captured in the real world datasets. Establishing a digital framework capable of mimicking real-world traffic behaviors is beneficial to validate and verify intelligent systems before deployment. Unfortunately, the current systems exist significant gaps between simulation and the real world. They have been limited to simulating traffic behaviors by replaying recorded data streams or relying on heuristic controllers, which fail to sufficiently reflect the real world. I have made progress in developing a new semi-annotation tool for a human-in-the-loop perception pipeline of detection and tracking based on LIDAR sequences, which generates real-world trajectories of road users for high-level behavior analysis [1]. I will work on accurately modeling road users' behaviors

by integrating perception from multi-modal sensor data, developing realistic and adversarial generation techniques [13, 16], and closing the sim-to-real gap while consistently bringing domain expert knowledge. Specifically, we have achieved a theoretical development in imposing linear equality constraints for neural networks [8], in which we obtain the surprising interpretation of Lagrange parameters as additional, penultimate layer hidden units with fixed weights stemming from the constraints. We envisage its wider application to generative models with constraints (a typical approach to generate counterfactual scenarios for behavior analysis) and model compression and acceleration shortly. My research strives to support the enactment of qualitative and salient distinctions (accident prediction and avoidance, the impact of sudden changes in light and weather conditions, etc.) germane to its users, e.g., traffic engineers and policymakers. Importantly, my research has potential in the applications of other fields, such as healthcare and robotics, given that AI technique remains emerging in health events (e.g., health monitoring via smartwatches) and new business models (e.g., self-service and robot-run supermarket).

Collaboration and Funding Opportunities

I have collaborated with many researchers from the industry, such as Alibaba, ByteDance, SenseTime, and some famous universities, institutions, and laboratories, such as the National Renewable Energy Lab (NREL), Florida Department of Transportation (FDOT), City of Gainesville (CoG), University of Oxford (UO), Chinese Academy of Sciences (CAS), University of Hong Kong (HKU), Chinese University of Hong Kong (CUHK). My future research direction is highly interdisciplinary and requires close collaboration with researchers and domain experts of different backgrounds to tackle fundamental research problems. There is significant interest in building smart infrastructure in various funding agencies. During my Ph.D. study, I was supported by National Science Foundation (NSF) and FDOT. Moving forward, I plan to actively pursue funds from NSF, particularly in Cyber-Physical Systems (CPS), CISE Community Research Infrastructure (CCRI), Computer and Network Systems (CNS), and Smart and Connected Communities (S&CC) programs, Department of Transportation (DoT), particularly in the University Transportation Centers program, Defense Advanced Research Projects Agency (DARPA) in programs such as Learning with Less Labeling (LwLL) and Assured Autonomy, Army Research Office (ARO), National Aeronautics and Space Administration (NASA), Department of Defense (DoD), Multidisciplinary Research Program of the University Research Initiative (MURI) and industries.

References

* indicates equal contributions. † indicates corresponding author(s) (if not the senior author)
Citations: 3226; H-index: 8. According to Google Scholar, by Feb. 6, 2022

[1] A. Wu, **P. He**†, X. Li, K. Chen, S. Ranka, and A. Rangarajan. "An Efficient Semi-Automated Scheme for Infrastructure LiDAR Annotation". In: *arXiv preprint arXiv:2301.10732* (2023). Under review for IEEE Transactions on Intelligent Transportation Systems (T-ITS).

[2] A. Almutairi*, **P. He***, A. Rangarajan, and S. Ranka. "Automated Truck Taxonomy Classification Using Deep Convolutional Neural Networks". In: *International Journal of Intelligent Transportation Systems Research* 20.2 (2022), pp. 483–494.

[3] **P. He**, P. Emami, S. Ranka, and A. Rangarajan. "Learning Canonical Embeddings for Unsupervised Shape Correspondence with Locally Linear Transformations". In: *arXiv preprint arXiv:2209.02152* (2022). Under review for IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

[4] **P. He**, P. Emami, S. Ranka, and A. Rangarajan. "Learning Scene Dynamics from Point Cloud Sequences". In: *International Journal of Computer Vision (IJCV)* (2022). **6 citations**, pp. 1–27.

[5] **P. He**, P. Emami, S. Ranka, and A. Rangarajan. "Self-Supervised Robust Scene Flow Estimation via the Alignment of Probability Density Functions". In: *AAAI Conference on Artificial Intelligence (AAAI)* 36.1 (2022). **3 citations**, pp. 861–869.

[6] **P. He**[*], A. Wu[*], X. Huang, A. Rangarajan, and S. Ranka. "Machine Learning-Based Highway Truck Commodity Classification Using Logo Data". In: *Applied Sciences* 12.4 (2022). **3 citations**, pp. 2075–2089.

[7] X. Huang, **P. He**, A. Rangarajan, and S. Ranka. "Machine-Learning-Based Real-Time Multi-Camera Vehicle Tracking and Travel-Time Estimation". In: *Journal of Imaging* 8.4 (2022). **3 citations**, pp. 101–119.

[8] A. Rangarajan, **P. He**, J. Lee, T. Banerjee, and S. Ranka. "Expressing Linear Equality Constraints in Feedforward Neural Networks". In: *arXiv preprint arXiv:2211.04395* (2022). **1 citation**.

[9] P. Emami, **P. He**, S. Ranka, and A. Rangarajan. "Efficient Iterative Amortized Inference for Learning Symmetric and Disentangled Multi-Object Representations". In: *International Conference on Machine Learning (ICML)*. Vol. 139. **26 citations**. PMLR, 2021, pp. 2970–2981.

[10] **P. He**[*], A. Wu[*], X. Huang, A. Rangarajan, and S. Ranka. "Video-based Machine Learning System for Commodity Classification". In: *International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS)*. **5 citations**. SCITEPRESS, 2020, pp. 229–236.

[11] **P. He**, A. Wu, X. Huang, J. Scott, A. Rangarajan, and S. Ranka. "Truck and Trailer Classification With Deep Learning Based Geometric Features". In: *IEEE Transactions on Intelligent Transportation Systems (T-ITS)* 22.12 (2020). **6 citations**, pp. 7782–7791.

[12] X. Huang, **P. He**, A. Rangarajan, and S. Ranka. "Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video". In: *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 6.2 (2020). **65 citations**, pp. 1–28.

[13] X. Yuan, **P. He**, X. Lit, and D. Wu. "Adaptive Adversarial Attack on Scene Text Recognition". In: *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. **17 citations**. IEEE. 2020, pp. 358–363.

[14] K. Zhai[*], **P. He**[*], T. Banerjee, A. Rangarajan, and S. Ranka. "SparsePipe: Parallel Deep Learning for 3D Point Clouds". In: *International Conference on High Performance Computing, Data, and Analytics (HiPC)*. **1 citation**. IEEE. 2020, pp. 51–61.

[15] **P. He**, A. Wu, X. Huang, J. Scott, A. Rangarajan, and S. Ranka. "Deep Learning based Geometric Features for Effective Truck Selection and Classification from Highway Videos". In: *Intelligent Transportation Systems Conference (ITSC)*. **10 citations**. IEEE. 2019, pp. 824–830.

[16] X. Yuan, **P. He**, Q. Zhu, and X. Li. "Adversarial Examples: Attacks and Defenses for Deep Learning". In: *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* 30.9 (2019). **1483 citations**, pp. 2805–2824.

[17] **P. He**, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li. "Single Shot Text Detector with Regional Attention". In: *International Conference on Computer Vision (ICCV)*. **318 citations**. 2017, pp. 3047–3055.

[18] **P. He**[*], W. Huang[*], Y. Qiao, C. C. Loy, and X. Tang. "Reading Scene Text in Deep Convolutional Sequences". In: *AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 30. 1. **306 citations**. 2016, pp. 3501–3508.

[19] Z. Tian, W. Huang, T. He, **P. He**, and Y. Qiao. "Detecting Text in Natural Image with Connectionist Text Proposal Network". In: *European Conference on Computer Vision (ECCV)*. **963 citations**. Springer. 2016, pp. 56–72.