Research Statement: Human-Centric Architectures for Visual Computing

Yuhao Zhu

Over the past two decades, computer architecture research has moved from *general-purpose* computing toward *domain-specific* accelerator architectures. I believe the field is on the cusp of a new revolution — *human-centric* architectures — driven by emerging platforms such as Augmented Reality (AR), Virtual Reality (VR), and autonomous machines that all intimately interact with humans: they continuously capture and interpret visual data *from* humans and generate visual data *for* humans to consume. These computing platforms must be designed, from the ground up, with principled considerations of human cognition.

My research over the past five years is centered around building human-centric visual computing systems, both to obtain unprecedented efficiency guided by human cognition and to augment human cognition through computing technologies. The **approach** I take is to bridge the conventional computing domain with *sensing* and *human perception*, the two fundamental components that connect computing with humans. The key **tenet** of my work is that a computing problem that seems challenging may become significantly easier when one considers how computing interacts with sensing and human perception in an end-to-end system.

Our work on human-centric computer systems have brought broad and positive societal impact in domains such as AR/VR, autonomous machines, and cultural heritage. For instance, many of our computing systems make their way to self-driving cars built by PerceptIn and deployed in US, Japan, China, and Switzerland. Our rendering algorithms and accelerators are used in a variety of high-profile cultural heritage projects, including digitally reconstructing in VR, for the first time, Elmina Castle in Ghana, a UNESCO World Heritage site known as the first European slave trading base in Sub-Saharan Africa. The virtual reconstruction allows tens of thousands of African Americans who could not physically visit the castle to seek insights into their ancestors' experiences of enslavement.

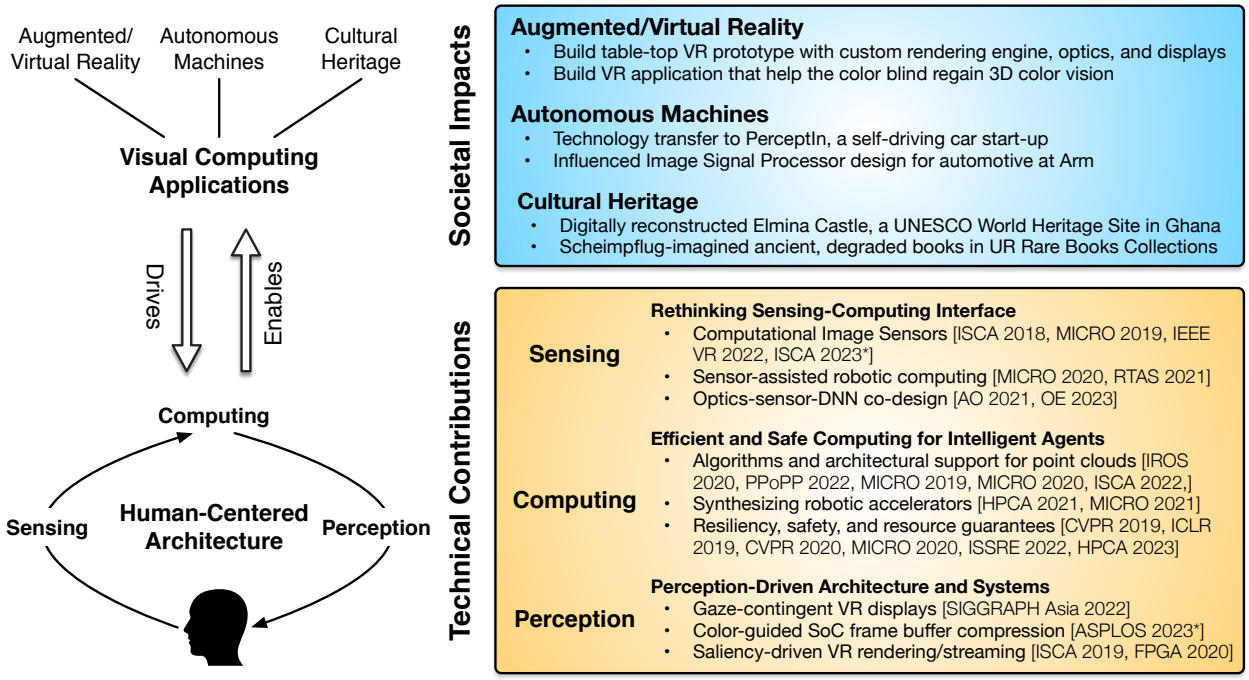

Fig. 1: My human-centered architecture research program.

Rethinking Sensing-Computing Interface

Sensing and computing, which acquire and interpret visual data, respectively, are traditionally designed in isolation and simply stitched together in a system, resulting in a sub-optimal whole. My research rethinks the sensing-computing interface and co-designs the two to 1) unlock new application capabilities and to 2) deliver orders of magnitude efficiency gains.

Our main contributions are in *Computational Image Sensors* (CIS), where image sensors are equipped with compute capabilities. Moving computations inside sensors unlocks new application capabilities (e.g., per-pixel exposure control to widen the dynamic range) and, more importantly, offers significant processing efficiency gains. The efficiency gains come from consuming large volumes of pixel data *in-situ*, thereby reducing the data communication overhead and keeping the power-hungry application processors idle.

Unleashing the power of CIS, however, requires making a myriad of interlocked design decisions. For instance, in-sensor computations are inefficient because sensors tend to be fabricated using older process nodes compared to standard CMOS nodes (limited by the photon sensing sensitivity), which offsets the gains from reducing the communication cost; 3D-stacking the compute logic with the pixel array offers more room for integration but increases thermal-induced noises that might, in turn, increase downstream processing power. We build CAMJ [1], the first CIS modeling framework that empowers designers, at an early stage, to explore architectural design trade-offs such as computing inside vs. off a sensor, 2D vs. 3D stacked design, and analog vs. digital computation inside the sensor.

We demonstrate concrete architectural augmentations to computational image sensors. Specifically, we show in our EUPHRATES [40] and ASV [12] work that judiciously extracting and sharing motion metadata, a natural byproduct during sensing, with the Systems-on-a-Chip (SoC) leads to an order of magnitude energy reduction by enabling incremental computing in downstream vision algorithms. Depth sensing based on this idea is in PerceptIn's self-driving vehicles. The EUPHRATES [40] work is selected as an Honorable Mention in IEEE Micro Top Picks of Computer Architecture, 2018. In collaboration with Meta, we propose in-sensor hardware augmentations for gaze tracking in AR/VR, which extracts Region of Interest inside an image sensor and reduces gaze tracking energy consumption by half [8]. The paper won the Best Paper Honorable Mention at IEEE VR 2022.

Pushing the sensing-computing co-design to its extreme, we expand the scope of sensing to consider the optics (e.g., lenses) before the sensor. From an information-theoretical perspective, certain information (e.g., wavelength, phase, polarization, and light field) is forever lost once light is transduced from the optical domain to the electrical domain. Our idea is that clever manipulations of those information in the optical domain could unlock new application capabilities otherwise unobtainable purely in the electrical domain.

Specifically, we design a table-top camera prototype [5, 4], in which optics, sensors parameters, and DNN weights are *jointly learned*. The end system, without losing any task quality, 1) reduces overall computation and 2) protects user privacy by ensuring no human-recognizable image is ever captured. The key is that the jointly-designed optics use the phase and polarization information inherent in scene lights to obfuscate private information while extracting features in the scene for the downstream task.

Finally, our work expands to modalities other than images. We build a series of algorithms and systems where intensive computations are replaced with inexpensive sensing. Examples include using GPS for correcting vehicle drifting in localization and using Radar for object tracking [31, 21]. Both systems are deployed in PerceptIn vehicles.

Perception-Guided Architecture and Systems

Today's computer systems waste a substantial amount of data transmission and computations for work that is *imperceptible to humans*. My work bridges the gap between human visual perception and computing sys-

tems. Our approach is to develop computational models that quantitatively capture human visual perception, and design techniques across the systems stack that leverage the perceptual models for efficient processing.

We build a VR system [7] that reduces the display power by 20% by exploiting *peripheral color confusion*, where our visual acuity is extremely bad in visual peripheries (beyond 10° eccentricity). Through over 8,000 (IRB-approved) trials of psychophysical measurements on real participants, we build a computational model that predicts, for a given reference color at a given eccentricity, the set of colors that are perceptually no different from the reference color. Leveraging the perception model, we design a VR rendering system that modulates pixel colors to minimize display power (dictated by colors) without affecting human color perception. Building on the principle of peripheral color confusion, we propose a compression algorithm that encodes perceptually similar colors together [2]. This algorithm, efficiently implemented in hardware, reduces the memory traffic in a VR SoC by up to 70%.

Complementary to peripheral vision, we exploit human visual *saliency* in our EVR work [25, 20] to reduce the rendering and streaming cost of VR (360°) videos. EVR is a cloud-client collaborative system. The cloud service, deployed on Amazon EC2, extracts trajectories of salient objects in a video (i.e., stimuli that most likely attract user attention), pre-render them, and store them as much smaller "videolets". At rendering time given the real-time visual field of a user, only the best matching videolets are transmitted. EVR reduces the data transmission cost and avoid expensive on-device rendering, amounting to 58% overall energy reduction. This work is selected as one of the IEEE Micro Top Picks in Computer Architecture, 2019.

Efficient and Safe Computing for Intelligent Agents

My group has also made contributions to improving the efficiency and safety of the computing stack for intelligent agents that operate on visual data, e.g., AR/VR devices and autonomous machines.

Architectural Support for Point Clouds. Our group is among the first to systematically investigate architectural support for point cloud processing [27, 11, 9] before it becomes prevalent in today's intelligent agents. The key challenges of point cloud processing are its input-dependent, irregular computation and memory access patterns. Our idea to tame the irregularities is to build algorithms and accelerators with *structured* compute and memory access patterns to begin with and mitigate the resulting accuracy loss in the training process. Some structures that are proven effective include avoiding backtracking in neighbor search [27] and eliding on-chip bank conflicts [9].

Coupled with architectural support, we also investigate foundational point cloud algorithms, including compression and neighbor search. We propose an algorithm that compresses LiDAR-generated point clouds by up to 90 times [10] and is deployed in PerceptIn's infrastructure-assisted autonomous vehicles [22]. We repurpose ray tracing hardware in modern GPUs for neighbor search in point clouds and achieve two orders of magnitude speedups [32].

Synthesizing Robotics Accelerators. In a departure from the manual design of robotics accelerators today, we build a framework to *automatically synthesize* robotics accelerators [14, 23, 19]. Given the complexity of the robotics software stack, the key challenge is to raise the level of abstractions. Taking inspiration from the block variant of classic dataflow architectures, our approach is to compile an algorithm to a macro data-flow graph (M-DFG), where each node is usually equivalent to billions of instructions when compiled to a conventional CPU ISA. The core contribution of our framework is to 1) use constraint optimization to identify the optimal mapping from the M-DFG to a prescribed hardware template that respects latency/power requirements, and 2) dynamically reconfigures the mapping to adapt to the complexity of the operating environment (e.g., number of pedestrians, traveling straight vs. making turns).

Safety. Our group has made significant stride in improving the safety of intelligent agents, which I initially started thinking about when I was concluding my Ph.D., when I explained my vision in an article invited by *IEEE Micro* [38]. Two themes underline our efforts. First, *quantitative resource guarantees*

lead to predictable system behaviors with less variability. We develop algorithms to train DNNs that, by design, are guaranteed to meet prescribed resource specifications on both GPUs [30, 28] and specialized accelerators [29] while maximizing accuracy.

Second, cost-effective safety mechanisms must be situational. Our PTOLEMY work [24, 15] proposes a DNN architecture that dynamically reconfigures itself based on inputs to actively detect and defend against adversarial attacks. Our BRAUM work [16] observes that algorithms used in autonomous machines inherently possess different forms of fault masking mechanisms. We propose a dynamic protection system that selectively elides and/or relaxes protection of certain algorithms in software. We present an RTL-free method for estimating Architectural Vulnerability Factors of DNN accelerators [26], which can be used for selectively protecting different structures in an accelerators.

Ph.D. Work: Human-Centric Mobile Web Computing

My Ph.D. thesis research was one of the first to address the human-perceivable efficiency of mobile (Web) computing. Existing mobile and server systems are fundamentally built to optimize for system-level metrics such as energy-delay products while being largely agnostic to user Quality-of-Experience. My work develops new ways to improve the human-perceivable efficiency of mobile Web by understanding when and how to make a calculated trade-off between performance and energy consumption in a user-driven manner. To that end, I have developed hardware accelerators [36, 39], runtime mechanisms [6, 13, 35, 33], and programming language extensions [37].

The analysis and workload characterization framework that I built in my dissertation research is now part of Chrome's main code base and is shipped with binary on all platforms. As part of the research, my collaborators and I perform the first large-scale performance analysis of mobile Web applications [34], conduct the first crowdsourcing characterization of mobile CPU designs on user satisfaction [18], and develop the first deterministic record and replay tool for mobile applications [17]. The *ACM Queue* magazine and *Communications of the ACM* invited me to write an article to introduce the cutting-edge research in Web computing to industry practitioners [3].

References

[1] "CamJ: Enabling Early-Stage Energy Modeling and Architectural Exploration for In-Sensor Visual Computing," in *Under Submission*.

[2] "Perceptual Framebuffer Compression for Visual Computing SoCs," in *Under Submission*.

[3] P. Bailis, J. Yang, V. J. Reddi, and Y. Zhu, "Research for practice: web security and mobile web computing," *Communications of the ACM*, vol. 60, no. 1, pp. 50–53, 2016.

[4] C. M. V. Burgos, P. Xiong, L. Qiu, Y. Zhu, and A. N. Vamivakas, "Co-designed metaoptoelectronic deep learning," *Optics Express*, 2023.

[5] C. M. V. Burgos, T. Yang, Y. Zhu, and A. N. Vamivakas, "Design framework for metasurface optics-based convolutional neural networks," *Applied Optics*, vol. 60, no. 15, pp. 4356–4365, 2021.

[6] W. Cui, D. Richins, Y. Zhu, and V. J. Reddi, "Tail latency in node. js: energy efficient turbo boosting for long latency requests in event-driven web services," in *Proceedings of the 15th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments*, 2019, pp. 152–164.

[7] B. Duinkharjav, K. Chen, A. Tyagi, J. He, Y. Zhu, and Q. Sun, "Color-perception-guided display power reduction for virtual reality," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–16, 2022.

[8] Y. Feng, N. Goulding-Hotta, A. Khan, H. Reyserhove, and Y. Zhu, "Real-time gaze tracking with event-driven eye segmentation," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2022, pp. 399–408.

[9] Y. Feng, G. Hammonds, Y. Gan, and Y. Zhu, "Crescent: taming memory irregularities for accelerating deep point cloud analytics," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, 2022, pp. 962–977.

[10] Y. Feng, S. Liu, and Y. Zhu, "Real-time spatio-temporal lidar point cloud compression," in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 10 766–10 773.

[11] Y. Feng, B. Tian, T. Xu, P. Whatmough, and Y. Zhu, "Mesorasi: Architecture support for point cloud analytics via delayed-aggregation," in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020, pp. 1037–1050.

[12] Y. Feng, P. Whatmough, and Y. Zhu, "Asv: Accelerated stereo vision system," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 643–656.

[13] Y. Feng and Y. Zhu, "Pes: Proactive event scheduling for responsive and energy-efficient mobile web computing," in *Proceedings of the 46th International Symposium on Computer Architecture*, 2019, pp. 66–78.

[14] Y. Gan, Y. Bo, B. Tian, L. Xu, W. Hu, S. Liu, Q. Liu, Y. Zhang, J. Tang, and Y. Zhu, "Eudoxus: Characterizing and accelerating localization in autonomous machines industry track paper," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 827–840.

[15] Y. Gan, Y. Qiu, J. Leng, M. Guo, and Y. Zhu, "Ptolemy: Architecture support for robust deep learning," in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020, pp. 241–255.

[16] Y. Gan, P. Whatmough, J. Leng, B. Yu, S. Liu, and Y. Zhu, "Braum: Analyzing and protecting autonomous machine software stack," in *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2022, pp. 85–96.

[17] M. Halpern, Y. Zhu, R. Peri, and V. J. Reddi, "Mosaic: cross-platform user-interaction record and replay for the fragmented android ecosystem," in *2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2015, pp. 215–224.

[18] M. Halpern, Y. Zhu, and V. J. Reddi, "Mobile cpu's rise to power: Quantifying the impact of generational mobile cpu design trends on performance, energy, and user satisfaction," in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2016, pp. 64–76.

[19] Y. Hao, B. Yu, Q. Liu, S. Liu, and Y. Zhu, "Factor graph accelerator for lidar-inertial odometry," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–7.

[20] Y. Leng, C.-C. Chen, Q. Sun, J. Huang, and Y. Zhu, "Energy-efficient video processing for virtual reality," in *Proceedings of the 46th International Symposium on Computer Architecture*, 2019, pp. 91–103.

[21] S. Liu, B. Yu, Y. Liu, K. Zhang, Y. Qiao, T. Y. Li, J. Tang, and Y. Zhu, "Brief industry paper: The matter of time—a general and efficient system for precise sensor synchronization in robotic computing," in *2021 IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 2021, pp. 413–416.

[22] S. Liu, B. Yu, J. Tang, Y. Zhu, and X. Liu, "Communication challenges in infrastructure-vehicle cooperative autonomous driving: A field deployment perspective," *IEEE Wireless Communications*, 2022.

[23] W. Liu, B. Yu, Y. Gan, Q. Liu, J. Tang, S. Liu, and Y. Zhu, "Archytas: A framework for synthesizing and dynamically optimizing accelerators for robotic localization," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 2021, pp. 479–493.

[24] Y. Qiu, J. Leng, C. Guo, Q. Chen, C. Li, M. Guo, and Y. Zhu, "Adversarial defense through network profiling based path extraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4777–4786.

[25] Q. Sun, A. Taherin, Y. Siatitse, and Y. Zhu, "Energy-efficient 360-degree video rendering on fpga via algorithm-architecture co-design," in *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2020, pp. 97–103.

[26] A. Tyagi, Y. Gan, S. Liu, B. Yu, P. Whatmough, and Y. Zhu, "Thales: Formulating and estimating architectural vulnerability factors for dnn accelerators," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023.

[27] T. Xu, B. Tian, and Y. Zhu, "Tigris: Architecture and algorithms for 3d perception in point clouds," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 629–642.

[28] H. Yang, S. Gui, Y. Zhu, and J. Liu, "Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2178–2188.

[29] H. Yang, Y. Zhu, and J. Liu, "Energy-constrained compression for deep neural networks via weighted sparse projection and layer input masking," in *International Conference on Learning Representations*, 2018.

[30] H. Yang, Y. Zhu, and J. Liu, "Ecc: Platform-independent energy-constrained deep neural network compression via a bilinear regression model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 206–11 215.

[31] B. Yu, W. Hu, L. Xu, J. Tang, S. Liu, and Y. Zhu, "Building the computing system for autonomous micromobility vehicles: Design constraints and architectural optimizations," in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020, pp. 1067–1081.

[32] Y. Zhu, "Rtnn: accelerating neighbor search using hardware ray tracing," in *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 2022, pp. 76–89.

[33] Y. Zhu, M. Halpern, and V. J. Reddi, "Event-based scheduling for energy-efficient qos (eqos) in mobile web applications," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2015, pp. 137–149.

[34] Y. Zhu, M. Halpern, and V. J. Reddi, "The Role of the CPU in Energy-Efficient Mobile Web Browsing," *IEEE Micro*, vol. 35, no. 1, pp. 26–33, 2015.

[35] Y. Zhu and V. J. Reddi, "High-performance and energy-efficient mobile web browsing on big/little systems," in *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2013, pp. 13–24.

[36] Y. Zhu and V. J. Reddi, "Webcore: Architectural support for mobile web browsing," in *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*. IEEE Computer Society, 2014, pp. 541–552.

[37] Y. Zhu and V. J. Reddi, "Greenweb: language extensions for energy-efficient mobile web computing," in *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2016, pp. 145–160.

[38] Y. Zhu and V. J. Reddi, "Cognitive computing safety: The new horizon for reliability," *IEEE Micro*, vol. 37, pp. 15–21, 2017.

[39] Y. Zhu, D. Richins, M. Halpern, and V. J. Reddi, "Microarchitectural implications of event-driven server-side web applications," in *Proceedings of the 48th International Symposium on Microarchitecture*, 2015, pp. 762–774.

[40] Y. Zhu, A. Samajdar, M. Mattina, and P. Whatmough, "Euphrates: algorithm-soc co-design for low-power mobile continuous vision," in *Proceedings of the 45th Annual International Symposium on Computer Architecture*, 2018, pp. 547–560.