Research Statement

My research goal is to bridge **physical principles** with **computer vision**. I pursue this goal through the knowledge from **computational photography**, leveraging physical image formation models to enhance visual perception and reconstruction under diverse real-world conditions. In recent years, with the rapid advancement of foundation models and large-scale datasets, researchers have provided remarkable answers to the question of ***"What can images do?"***. Alongside the development of various 2D and 3D computer vision tasks, these advances have profoundly transformed our daily lives (e.g., autonomous driving, scene surveillance, smart healthcare). Meanwhile, the classical yet profound concept of computational photography has been increasingly focused on another fundamental question: ***"Where do images come from?"***

The answer is light. Computational photography encompasses the entire process by which light interacts with a scene (e.g., reflection, refraction, scattering), passes through the camera imaging system (e.g., lens, point spread function (PSF), image signal processor (ISP)), and ultimately forms the final output image. In this area, mature camera and smartphone manufacturers have developed complete processing pipelines to achieve superior generated image quality, thus satisfying human visual perception. **In my view**, computational photography not only enhances image quality but also provides **physically consistent cues**, which are crucial for reliable performance in downstream vision tasks such as object detection, semantic segmentation, and 3D reconstruction. This is especially useful under diverse real-world conditions, including varying illumination, weather, and data degradations (e.g., noise, blur). In other words, leveraging the knowledge of computational photography can improve the **robustness of computer vision**.

In this direction, I have made extensive efforts and explorations, which include: **(i).** developing physics-based data augmentation to enhance computer vision robustness [1; 2; 3; 4], **(ii).** leveraging camera RAW data—the original sensor response prior to RGB—for enhanced visual performance [5; 6; 7], and **(iii).** toward an understanding of 3D physical world with the 3D vision techniques (e.g., multi-view generation and synthesis) [8; 9; 10]. **In the future**, I envision building interdisciplinary collaborations with researchers in fields such as robotics, electrical engineering, and optics, as well as with scholars in the social sciences, including sociology, economics, and anthropology. By bridging hardware design with physics-based vision, I aim to establish a research agenda that not only advances core scientific understanding but also creates real-world impact with long-term benefits for society.

Research Progress

Physics-based Data Augmentation for Robust Vision

My research on leveraging computational photography for real-world downstream tasks began in mid-2020, at a time when most studies were focused on designing better low-level image restoration models to enhance human visual perception (e.g., super-resolution, denoising, low-light image enhancement). However, these works often overlooked whether such low-level algorithms could improve performance on real-world high-level vision tasks, such as object detection.

In this way, my first study [1] focuses on low-light conditions object detection, as challenging lighting environments can significantly degrade high-level visual perception. I discovered that even the state-of-the-art (SOTA) low-light image enhancement algorithms at the time failed to improve object detection performance in nighttime scenes; in many cases, they even proved counterproductive (see Fig. 1).

After that, my subsequent research [2] revealed a parallel limitation: conventional image restoration techniques, including super-resolution, deblurring and denoising, typically do not lead to performance gains in high-level computer vision tasks. We identified that this limitation stems primarily from the misalignment between **human** perception-based metrics and **machine** vision performance criteria [1; 2; 3].

To address these limitations, I proposed physics-based data augmentation that simulates diverse degradations during training to enhance model robustness. In MAET [1], we followed the camera ISP pipeline to synthesize realistic low-light images from large-scale datasets, while incorporating self-supervised learning to improve low-light condition object detection. In AERIS [2], we extended this idea by analyzing resolution, noise, and blur, and designing a composite augmentation strategy with a joint super-resolution and detection framework. Together, these approaches yielded detectors with significantly improved robustness under real-world corruptions.

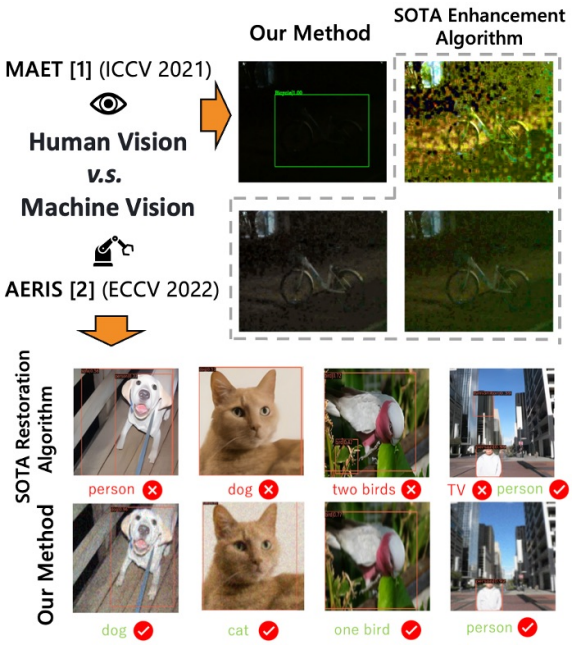

Figure 1: Our studies [1; 2] demonstrate that state-of-the-art (SOTA) image enhancement and restoration methods generally cannot improve performance on high-level vision tasks; in some cases, they adversely affect object detection accuracy.

Edge-device Image Enhancement & Restoration Models

Beyond high-level vision robustness, I also explore light-weight model design for real cameras and edge devices. Capturing high-quality images under diverse lighting remains difficult due to natural factors (e.g., low light) and camera exposure. In [3], I proposed the Illumination Adaptive Transformer (IAT), a 90K-parameter model combining global and local branches for low-light enhancement, retouching, and exposure correction—later the most cited paper of BMVC 2022. More recently, I introduced Image-Adaptive Cartesian coordinates (IAC) [4], which unify curve- and 3D LUT-based methods by learning image-specific coordinates, enabling faster and more efficient enhancement than IAT.

Beyond RGB, Seeing Further with Camera RAW Data

Building on the understanding that vision models need physically accurate inputs, I shifted my focus to the source: **camera RAW data**. As the original sensor readout before any non-linear camera ISP processing (see Fig. 2), RAW images preserve abundant details for a more faithful scene representation. Compared with normally used 8-bit sRGB data, they also feature a higher bit depth (e.g., 14- or 16-bit), a wider color gamut, and structured noise, maintaining a linear relationship with scene irradiance. A key drawback is their larger storage requirement. As a result, RAW datasets (which often contain only thousands of images) are typically much smaller than sRGB datasets.

To tackle this issue, RAW-Adapter [5] employs adapter tuning to leverage the knowledge-rich pre-trained sRGB models while using the information-rich camera RAW data as input. Specifically, it adopts a dual-adapter design, consisting of an input-level adapter (comprising learnable ISP stages) and a model-level adapter. In [5], we demonstrate the potential of camera RAW data under various lighting conditions. In the extension work [6], we further propose RAW-Bench, a new benchmark that includes 17 types of degradation conditions—such as those related to lightness, weather, and

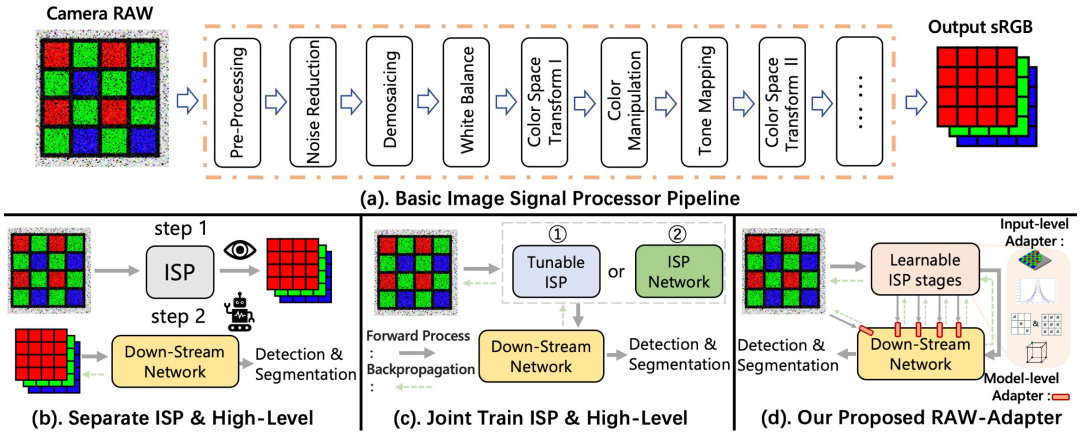

Figure 2: Beyond RGB, RAW-Adapter [5] directly uses camera RAW data for visual perception.

imaging—to further evaluate the capabilities of RAW data in real-world computer vision.

Additionally, our recent work [7] further explores the potential of camera RAW data across multiple vision tasks, including object detection, semantic segmentation, instance segmentation, and pose estimation. We propose a more efficient tuning strategy that achieves superior performance to RAW-Adapter by training only 10%–20% of the parameters.

Toward an Understanding of the 3D Physical World

Beyond my work on 2D vision, I have increasingly focused on 3D vision as a means to model natural phenomena in the physical world. Recent advances such as neural radiance fields (NeRF) and 3D Gaussian Splatting (3DGS) have greatly facilitated 3D representation and reconstruction, making them more accessible for practical applications, opening broad opportunities in applications like virtual/augmented reality and autonomous driving.

In Aleth-NeRF [8], we draw an analogy between NeRF's rendering process and the ancient Greek **emission theory**, introducing the concept of a "concealing field" to explain darkness in 3D space and releasing the first multi-view dataset for low-light and overexposure scenarios. Building on this, I2-NeRF [10] proposes a novel framework that enhances isometric and isotropic metric perception under media degradation, we further present a general radiative formulation that unifies emission, absorption, and scattering under the **Beer–Lambert law**, extending naturally to environments such as underwater, haze, and low-light scenes. Most recently, Luminance-GS [9] adopts a curve-adjustment strategy, using view-adaptive curves to map multi-view images under varying illumination into physically consistent 3D reconstructions.

Ongoing and Future Directions

Computational Photography for Embodied AI and Robotics

Building on my research in physics-based vision and robustness, I plan to extend my work toward bridging computational photography with **embodied AI** and **robotic perception**, an area that is progressing rapidly and poised for long-term impact. Similar to computer vision, camera-based robotic perception also suffers from lighting and weather degradations, and in even more extreme scenarios such as **polar environments** or **planetary exploration** (e.g., Mars).

To this end, the key directions may include: ① Lighting-Adaptive Visual Policies: Integrate lightweight image processing models into robotic control loops for real-time adaptation to extreme photometric variations (e.g., glare, overexposure) during navigation and manipulation. ② Robot-

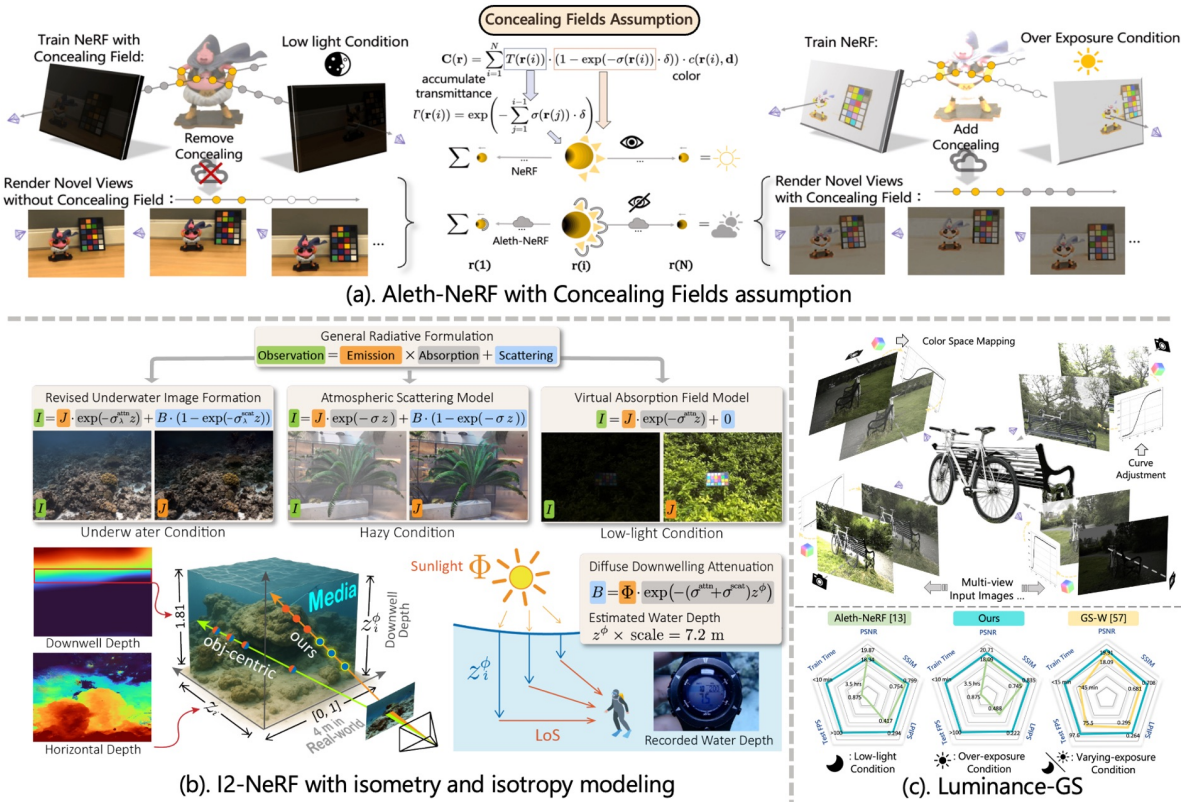

Figure 3: My research forms a series of works aimed at understanding the 3D physical world. Aleth-NeRF [8] models changes in lightness in 3D by introducing the concept of a "concealing field." I2-NeRF [10] integrates two physical principles—isometry and isotropy—into the neural radiance field. And Luminance-GS [9] focuses on fundamental curve adjustments to handle a wide range of real-world challenging lighting conditions.

Perception with camera RAW Data: Exploit the high dynamic range and linearity of camera RAW data to achieve superior robustness for high-speed robotic tasks. ③ Physically Faithful Sim2Real Transfer: Leverage our 3D physical world models to create high-fidelity simulators that enable seamless and reliable policy transfer from simulation to real robots.

Next Generation Camera Hardware Design

I will pioneer task-specific camera designs through cross-disciplinary collaboration with optics and electronics experts. The ECCV 2024 Best Paper, "Minimalist Vision with Freeform Pixels," [11] presents a promising pathway from camera design to vision algorithms. I hope to develop programmable ISPs optimized for high-level vision tasks and create lightweight, low-power imaging systems for vision and robots, establishing a new paradigm for hardware-software co-design.

Toward a Better Understanding of the 3D Physical World

Building on my work in neural rendering (Aleth-NeRF [8], Luminance-GS [9], I2-NeRF [10]), I will develop more interpretable and physically grounded **3D world models**. This includes designing differentiable rendering frameworks that unify light transport phenomena (e.g., scattering, absorption) and embedding explicit material and lighting properties into neural representations. The goal is to enable robust applications in autonomous driving, virtual reality, and environmental sensing under complex real-world conditions.

Publications

[1] **Ziteng Cui**, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2553–2562, October 2021.

[2] **Ziteng Cui**, Yingying Zhu, Lin Gu, Guo-Jun Qi, Xiaoxiao Li, Renrui Zhang, Zenghui Zhang, and Tatsuya Harada. Exploring resolution and degradation clues as self-supervised signal for low quality object detection. In *European Conference on Computer Vision (ECCV)*, pages 473–491. Springer, 2022.

[3] **Ziteng Cui**, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, ZhengKai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.

[4] **Ziteng Cui**, Lin Gu, and Tatsuya Harada. Discovering an image-adaptive coordinate system for photography processing. In *35rd British Machine Vision Conference 2022, BMVC 2024*. BMVA Press.

[5] **Ziteng Cui** and Tatsuya Harada. Raw-adapter: Adapting pretrained visual model to camera raw images. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

[6] **Ziteng Cui**, Jianfei Yang, and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images and a benchmark. *arXiv preprint arXiv:2503.17027*, 2025.

[7] Wenjun Huang*, **Ziteng Cui*** (co first), Yinqiang Zheng, Yihui He, Tatsuya Harada, and Mohsen Imani. Dr. RAW: Towards general high-level vision from RAW with efficient task conditioning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[8] **Ziteng Cui**, Lin Gu, Xiao Sun, Xianzheng Ma, Yu Qiao, and Tatsuya Harada. Aleth-nerf: Illumination adaptive nerf with concealing field assumption. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[9] **Ziteng Cui**, Xuangeng Chu, and Tatsuya Harada. Luminance-gs: Adapting 3d gaussian splatting to challenging lighting conditions with view-adaptive curve adjustment. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[10] Shuhong Liu, Lin Gu, **Ziteng Cui**, Xuangeng Chu, and Tatsuya Harada. I2-nerf: Learning neural radiance fields under physically-grounded media interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[11] Jeremy Klotz and Shree K. Nayar. Minimalist vision with freeform pixels. In *European Conference on Computer Vision (ECCV)*, 2024.