Towards Accurate, Adaptive, and Real-time Machine Perception on Resource-constrained Platforms

Jeho Lee
Yonsei University
Seoul, Republic of Korea
jeholee@yonsei.ac.kr

Abstract

Accurate, real-time machine perception is a key enabler of emerging mobile applications such as augmented reality and autonomous driving. However, running complex vision models within the tight latency budgets of resource-limited platforms remains challenging. We address two root causes: (i) the growing computational demands of state-of-the-art vision models and (ii) the variability of compute resource availability in on-device AI deployments. In this extended abstract, we introduce two adaptive perception systems that leverage AI-system co-design. Deployed on commercial devices and evaluated on representative perception workloads, our systems demonstrate high-performance perception under practical latency and resource constraints.

CCS Concepts

• **Computer systems organization → Embedded systems**; • **Computing methodologies → Computer vision**.

Keywords

On-device AI, Mobile and Edge Computing, AI-system Co-design, Machine Perception

ACM Reference Format:
Jeho Lee. 2025. Towards Accurate, Adaptive, and Real-time Machine Perception on Resource-constrained Platforms. In *The 23rd Annual International Conference on Mobile Systems, Applications and Services (MobiSys '25), June 23–27, 2025, Anaheim, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3711875.3736677

1 Motivation and Challenges

Alongside advances in deep neural networks (DNNs), machine perception has become a core component of modern mobile applications. Autonomous driving and augmented reality (AR) are two key examples requiring continuous, accurate, and real-time understanding of surrounding spaces. In autonomous systems, omnidirectional 3D perception enables safe navigation through dynamic environments, while AR applications demand pixel-level understanding for realistic content rendering and context-aware interaction. These tasks are traditionally offloaded to a powerful cloud infrastructure. However, there is a growing need to perform such computations directly *on-device*—e.g., on smartphones or mobile robots—to preserve privacy and enable operation in bandwidth-limited environments. Accordingly, contemporary System-on-Chip (SoC) architectures integrate heterogeneous AI accelerators—such as GPUs and neural processing units (NPUs)—to support the intense computation of perception models on mobile and edge devices.

A key requirement is to maximize perception accuracy while meeting strict real-time constraints—e.g., 30 frames per second (FPS) for smooth AR rendering. Achieving this on resource-constrained devices introduces major system design challenges. First, modern vision models are increasingly complex in both architecture and compute demands [1, 3]. To ensure robust perception, state-of-the-art models often adopt sophisticated architectures like vision transformers (ViTs), which substantially increase parameter counts and computation depth. These models also rely on high computational fidelity—e.g., high-resolution and floating-point operations—to preserve prediction quality. Moreover, emerging applications like autonomous driving require multi-view perception, demanding simultaneous processing of multiple camera streams. These trends dramatically increase inference workloads, making it harder to meet latency constraints on limited platforms.

Second, mobile and edge devices exhibit significant variability in available compute resources, both statically across devices and dynamically over time. Inter-device heterogeneity causes variation in compute and memory budgets across deployment targets [3]. At the same time, intra-device heterogeneity—such as the coexistence of GPUs and NPUs within a single SoC—necessitates workload partitioning strategies that align with each processor's strengths [1]. Compounding this, OS-level decisions such as thermal throttling and dynamic voltage and frequency scaling (DVFS) cause runtime fluctuations in processing capability [1, 4]. These variations require inference systems to adapt execution strategies to both device architecture and current operating conditions.

2 Proposed Systems

To address these challenges, we have developed adaptive and efficient machine perception systems for various mobile and edge platforms, as illustrated in Figure 1. Rather than relying on static inference pipelines, we explore AI-system co-design approaches that dynamically adapt model execution to the underlying hardware capabilities and environmental context. We formulate the goal of perception systems as a constrained optimization problem—achieving high task accuracy while satisfying strict latency constraints imposed by real-time applications:

$$\text{maximize} \quad \mathrm{Acc}(\mathcal{M}) \quad \text{subject to} \quad \mathrm{Lat}(\mathcal{M}) \leq \tau,$$

where $\mathcal{M}$ denotes the perception model, $\mathrm{Acc}(\mathcal{M})$ its task accuracy, and $\mathrm{Lat}(\mathcal{M})$ its inference latency on a given device. The constraint τ reflects the application-specific latency requirement. We summarize how this objective can be achieved in practice through the design

MobiSys '25, June 23–27, 2025, Anaheim, CA, USA

ACM ISBN 979-8-4007-1453-5/2025/06
https://doi.org/10.1145/3711875.3736677

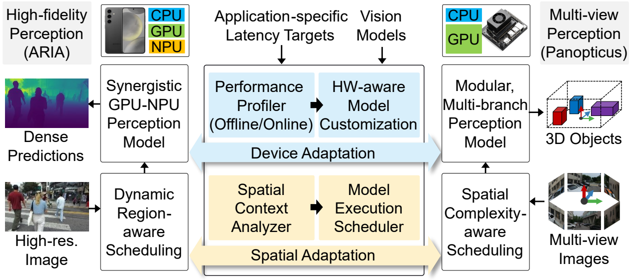
High-fidelity Perception (ARIA)
CPU
GPU
NPU
Application-specific Latency Targets
Vision Models
CPU
GPU
Multi-view Perception (Panopticus)
Dense Predictions
Synergistic GPU-NPU Perception Model
Performance Profiler (Offline/Online)
HW-aware Model Customization
Device Adaptation
Modular, Multi-branch Perception Model
3D Objects
High-res. Image
Dynamic Region-aware Scheduling
Spatial Context Analyzer
Model Execution Scheduler
Spatial Adaptation
Spatial Complexity-aware Scheduling
Multi-view Images

Figure 1: Overview of our adaptive machine perception systems for representative real-time applications: augmented reality (left) and autonomous driving (right).

and implementation of the two systems that target distinct but representative perception workloads.

High-fidelity Perception on Mobile SoCs. Recent advances in ViTs have led to the emergence of vision foundation models (VFMs) [2, 6] with strong generalization to diverse environmental conditions. These models are especially promising for mobile AR scenarios, where robustness to unseen environments is critical. However, realizing their full potential requires high computational fidelity—high-resolution and floating-point operations—which poses challenges on resource-limited mobile devices.

To understand how VFMs behave under different hardware configurations, we analyzed the characteristics of mobile GPUs and NPUs. GPUs are capable of executing high-fidelity inference due to their high parallelism capabilities, but still struggle to meet real-time latency targets due to the computational demands of ViT-based architectures. Lowering computational fidelity enables real-time inference on NPUs, which are highly optimized for small-sized data processing with specialized hardware, but at the expense of a noticeable accuracy degradation.

Based on these observations, we developed ARIA [1], a system designed to enable high-quality, real-time VFM inference on mobile devices. ARIA leverages the intra-device heterogeneity of mobile SoCs through a parallel and selective execution scheme: the GPU periodically performs full-frame, high-resolution inference, while the NPU concurrently processes low-latency updates for dynamic regions within the scene. To identify these regions efficiently, ARIA analyzes temporal variations in the intermediate patch-level features produced by the VFM encoder, enabling fine-grained detection of scene changes with minimal overhead. It further prioritizes the dynamic regions based on their estimated motion saliency, ensuring that fast-moving regions are updated promptly to maintain visual consistency and responsiveness. To adapt to runtime variability—such as GPU performance degradation from thermal throttling or abrupt scene changes due to camera motion—ARIA adjusts the frequency and resolution of GPU executions based on camera motion magnitude and thermal feedback.

Multi-view Perception on Edge GPUs. Many autonomous systems rely on omnidirectional visual perception to operate safely in dynamic environments. To support such capabilities, multi-camera setups are often used to capture 360° views, which are then processed to detect and localize objects in 3D [5]. A naïve approach that applies a full, high-capacity model to all camera views fails to meet the strict latency targets on edge GPUs. Conversely, using lightweight models for every view leads to degradation in detection accuracy, particularly in complex or high-speed scenarios.

To better understand this trade-off, we analyzed how perception workloads vary across individual views in multi-camera setups. We observed that the spatial complexity—shaped by factors such as object density, motion, and scene geometry—can differ significantly both temporally and across camera views. For example, some camera views often encounter high-speed moving objects and dense traffic, while others may remain sparse and static. These variations result in heterogeneous computational demands across views.

Based on these insights, we developed Panopticus [3], a system for real-time, accurate multi-view 3D perception on edge GPUs. To effectively adapt to the spatial diversity, Panopticus introduces a multi-branch detection model that enables each camera view to be processed with a tailored inference configuration. The model comprises modular branches with varying combinations of backbones, depth estimation networks, and temporal fusion strategies, each offering distinct accuracy-latency trade-offs. At deployment, the model is pruned to respect the latency and memory constraints of the target device. At runtime, Panopticus performs spatial-adaptive execution by analyzing the current object distributions and selecting the most suitable branch for each view. This allows the system to allocate heavier configurations to views with high spatial complexity (e.g., distant or fast-moving objects) while preserving efficiency by assigning lightweight paths to simpler views.

Deployment and Results. We implemented and evaluated ARIA and Panopticus on commercial mobile and edge platforms. ARIA was deployed on two smartphone SoCs (e.g., Snapdragon 8 Gen 3), operating on high-resolution video streams captured via a custom-built AR camera rig. ARIA achieved a 99.9% deadline success rate under 30 and 60 FPS latency constraints, and improved prediction accuracy by 27.6–72.3% over its baselines across monocular depth estimation and semantic segmentation tasks. Panopticus was evaluated on three NVIDIA Jetson platforms using a 360° camera rig designed to emulate real-world autonomous perception. Under a strict 33 ms/frame latency budget, Panopticus achieved up to 62% improvement in 3D object detection accuracy over static single-model baselines, while reducing overall inference latency by 2.1×. These results demonstrate the effectiveness of our spatially- and device-adaptive inference strategies in supporting high-quality, real-time perception on resource-constrained platforms.

3 Future Direction

A key direction for future work is to support *multi-task* and *multi-view* perception in a unified inference framework. Real-world applications like AR increasingly demand the simultaneous execution of diverse perception tasks across multiple camera inputs, which exacerbates system-level complexity. Supporting such scenarios requires efficient architectural designs that minimize redundant computation, e.g., by sharing feature encoders across tasks or views, and runtime schedulers that can dynamically prioritize tasks (and views) based on contextual importance. Furthermore, managing compute resource contention (e.g., heterogeneous compute units, memory) across multiple tasks remains an open challenge. Addressing these issues is critical to realizing fully-featured, scalable perception systems for next-generation AI applications.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00344323).

References

[1] Chanyoung Jung*, Jeho Lee*, Gunjoong Kim, Jiwon Kim, Seonghoon Park, and Hojung Cha. 2025. ARIA: Optimizing Vision Foundation Model Inference on Heterogeneous Mobile Processors for Augmented Reality. In *Proceedings of the 23rd ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)* (*Co-primary author).

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision (CVPR).*

[3] Jeho Lee, Chanyoung Jung, Jiwon Kim, and Hojung Cha. 2024. Panopticus: Omnidirectional 3D Object Detection on Resource-constrained Edge Devices. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom).*

[4] Seonghoon Park, Yeonwoo Cho, Hyungchol Jun, Jeho Lee, and Hojung Cha. 2023. OmniLive: Super-Resolution Enhanced 360° Video Live Streaming for Mobile Devices. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys).*

[5] Jonah Philion and Sanja Fidler. 2020. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *European Conference on Computer Vision (ECCV).*

[6] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Short Bio

Jeho Lee is currently pursuing his PhD in Computer Science at Yonsei University, Seoul, Korea. He is working under the supervision of Prof. Hojung Cha. His research focuses on efficient on-device AI systems for resource-constrained computing platforms, paving the way for next-generation mobile AI applications.