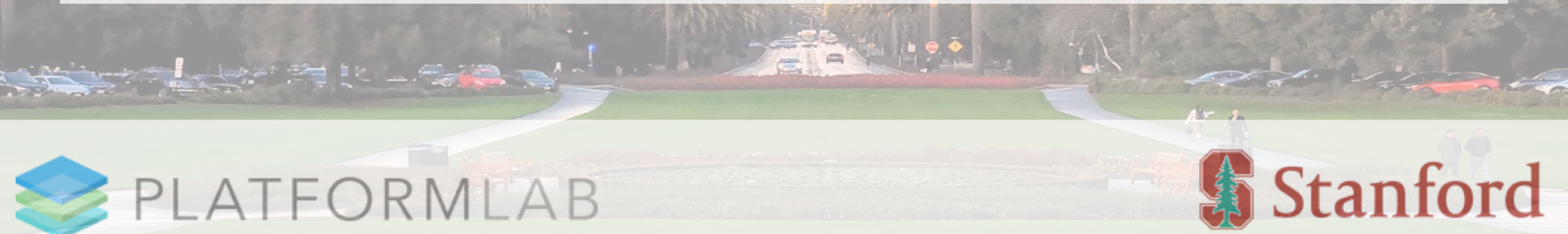
Starfish: Resilient Image Compression for AIoT Cameras

Pan Hu

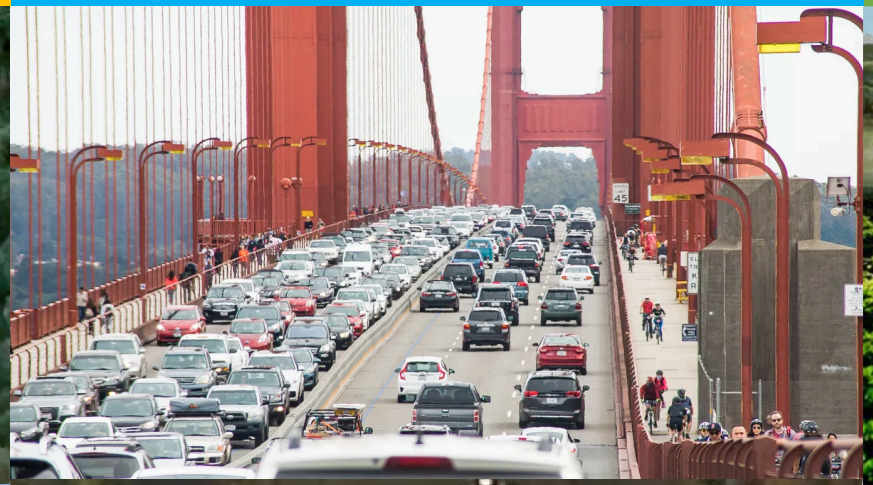


PLATFORMLAB

Stanford

Wide-area Camera Analytics Enables Future Applications

Smart Cities	*Intelligent Transportation*	*AI-enabled Farms and Factories*

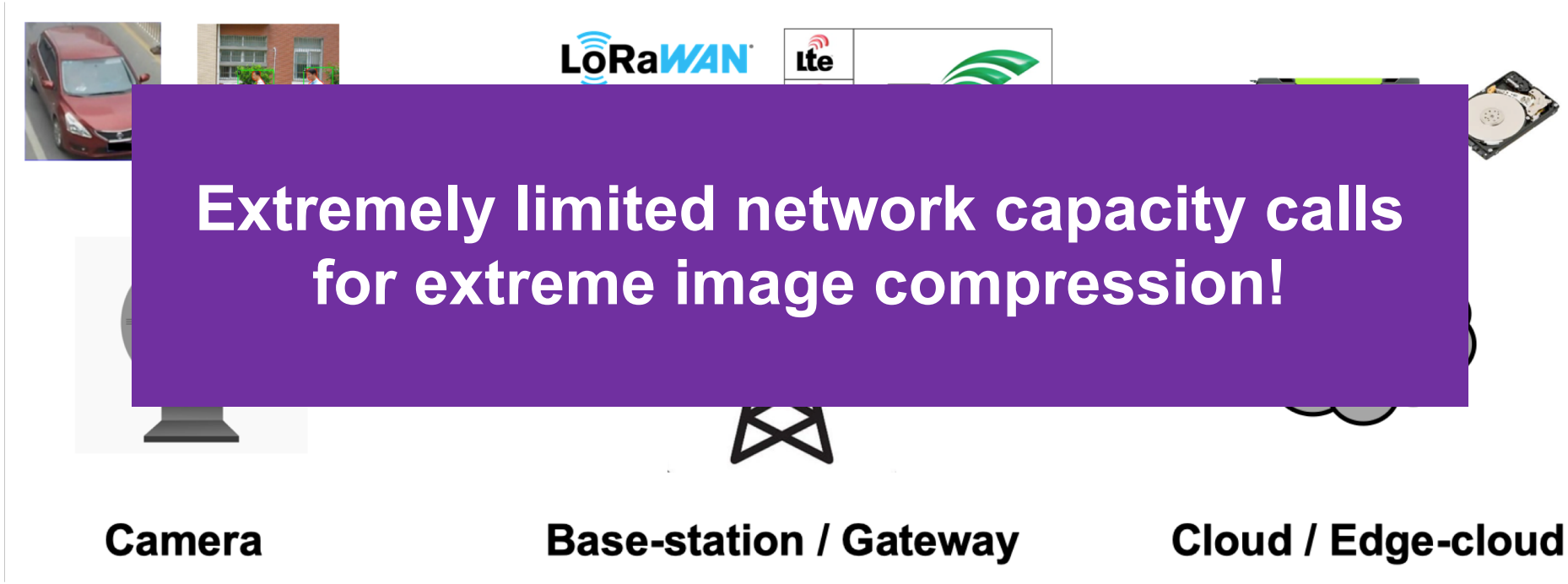


AI for IoT (AIoT) Cameras needs Efficient Compression

Images needs to be uploaded to Cloud
- Computation-heavy
- Aggregate from multiple cameras
- Cloud storage

⟷

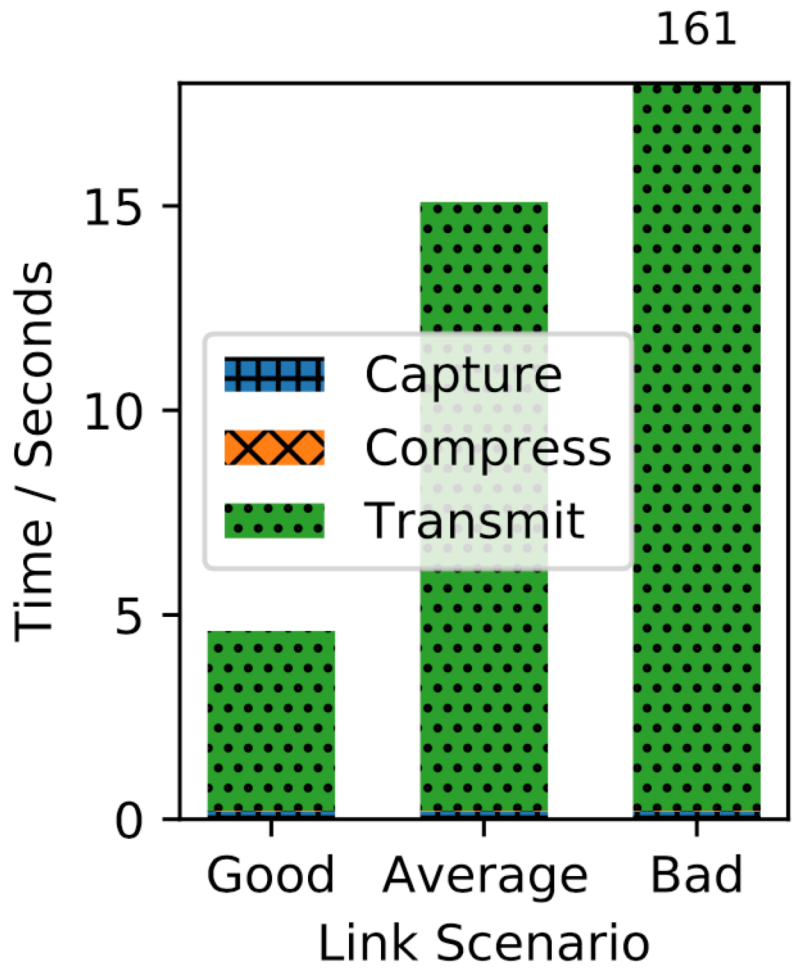
Capacity of each gateway:
- LoRaWAN: < 100kbps
- Sigfox: < 10kbps

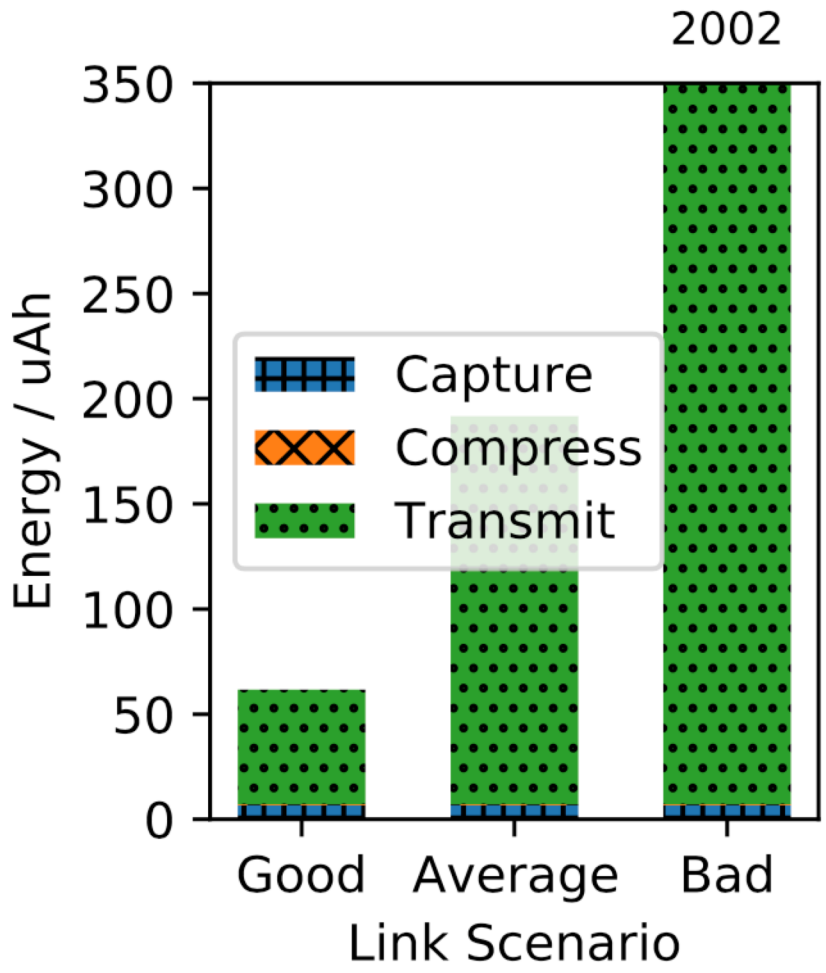


Ghena, Branden, et al. "Challenge: Unlicensed LPWANs Are Not Yet the Path to Ubiquitous Connectivity." *The 25th Annual International Conference on Mobile Computing and Networking*. 2019.

Imbalance in Computation vs. Communication

- **Good:** 5.47kbps
- **Average:** 1.76kbps
- **Bad:** 245 bps





System Limitations AIoT Device

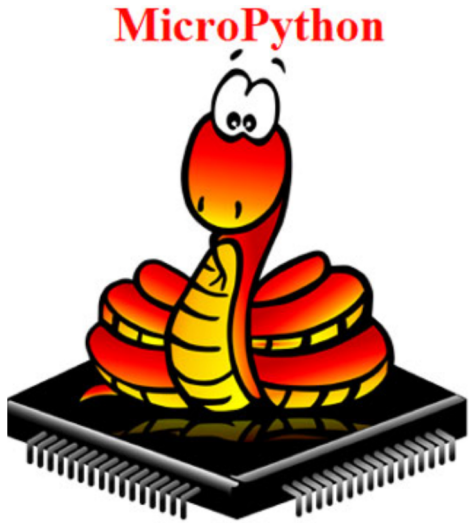
Hardware

JPEG ⟷ webp

bpg
Better Portable Graphics

	GAP8	GAP9	K210
Frequency /MHz	250	400	400
RAM/KB	512	1536	8192 (2048 for AI)

Software



- High-level abstraction
- Basic APIs:

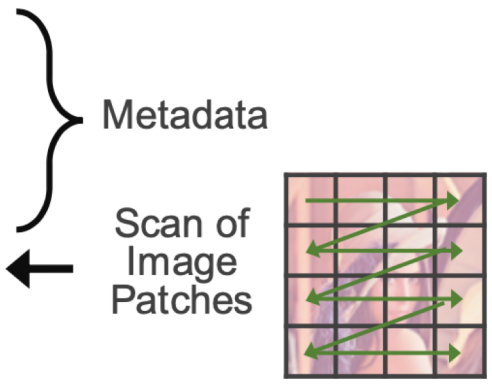
image.save(img, quality=10)

Current method vs. Proposed method on AIoT Camera

Current method (JPEG)

- Susceptible to loss
- Task-agnostic
- Content-agnostic

Marker	Segment
0xFFD8	SOI(Start Of Image)
0xFFC0	SOF0(Start Of Frame 0)
0xFFDB	DQT(Define Quantization Table)
0xFFC4	DHT(Define Huffman Table)
0xFFDA	SOS(Start Of Scan)
0xFFD9	EOI(End Of Image)



Current method vs. Proposed method on AIoT Camera

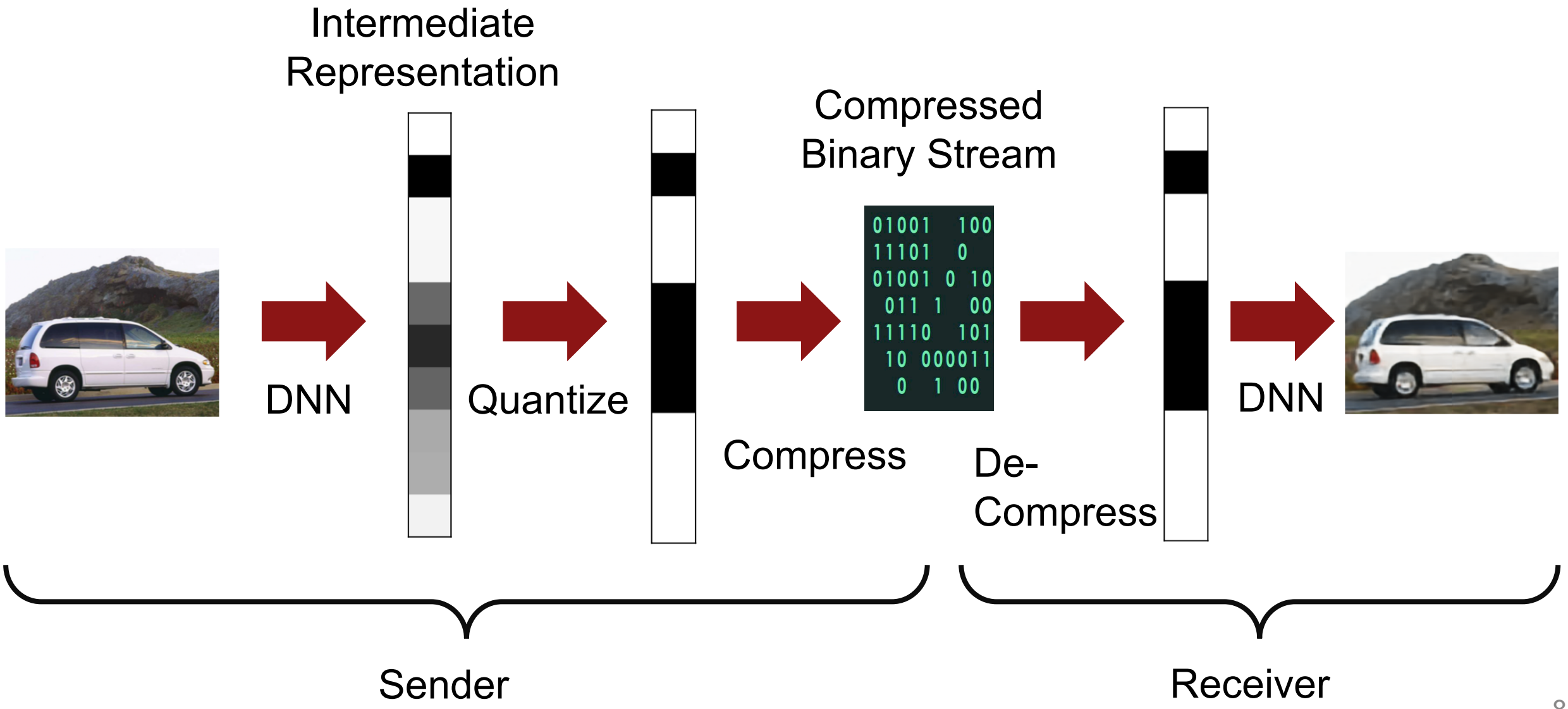
Current method (JPEG)

- Susceptible to loss
- Task-agnostic
- Content-agnostic

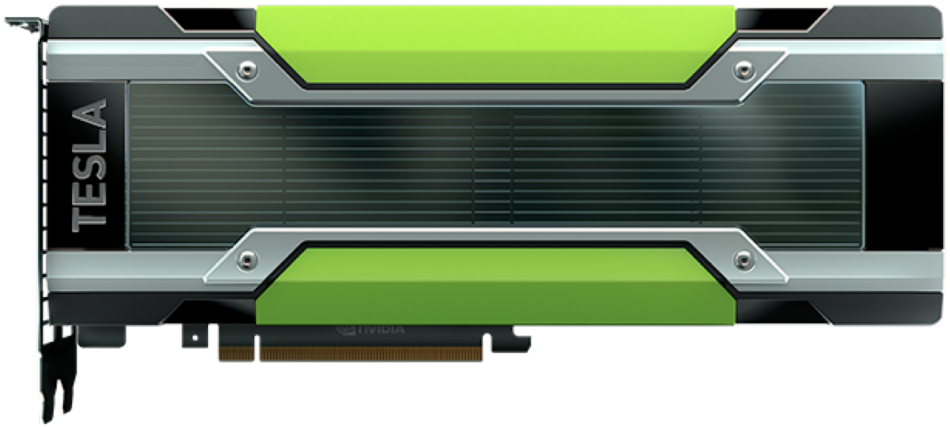
StarFish (DNN-based)

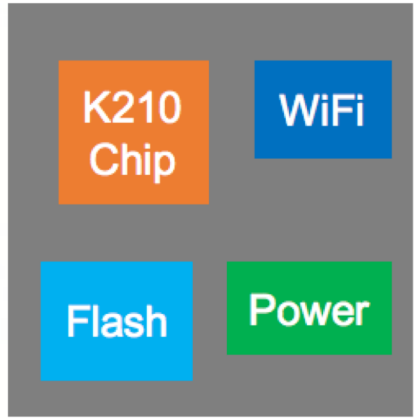
- Loss-resilient
- Task-aware
- Specialize to applications

Workflow of DNN-based Image Compression



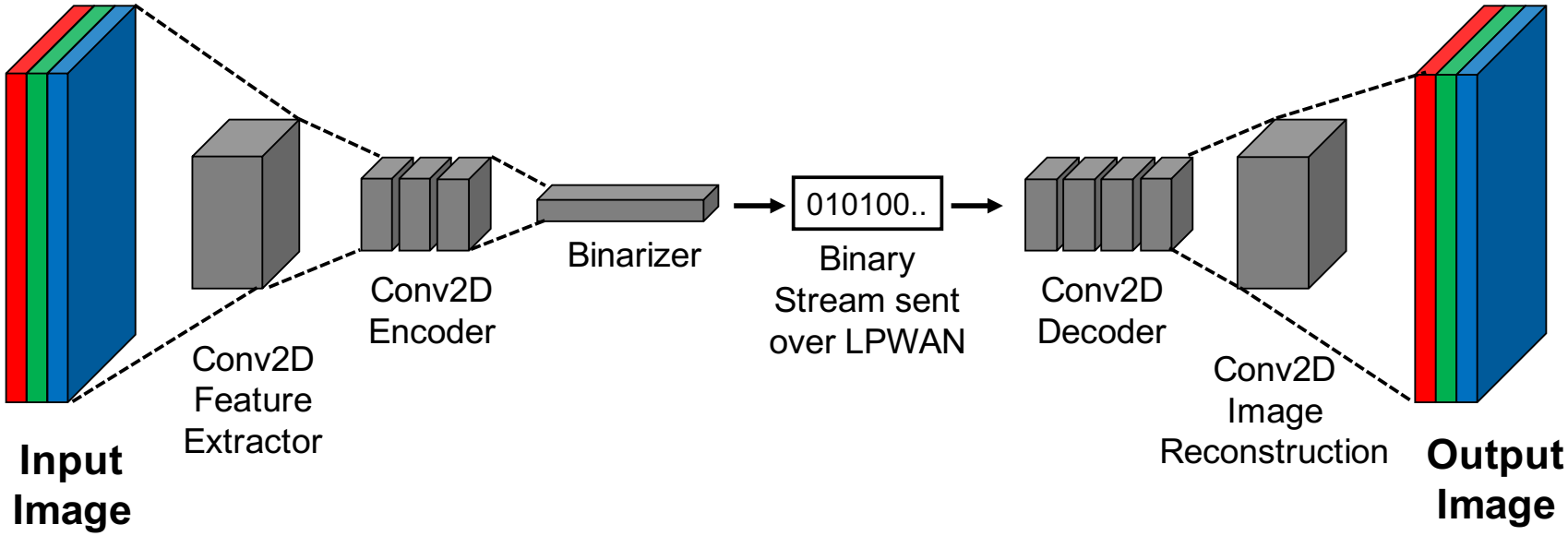
Extreme Resource Limitation on AIoT Devices

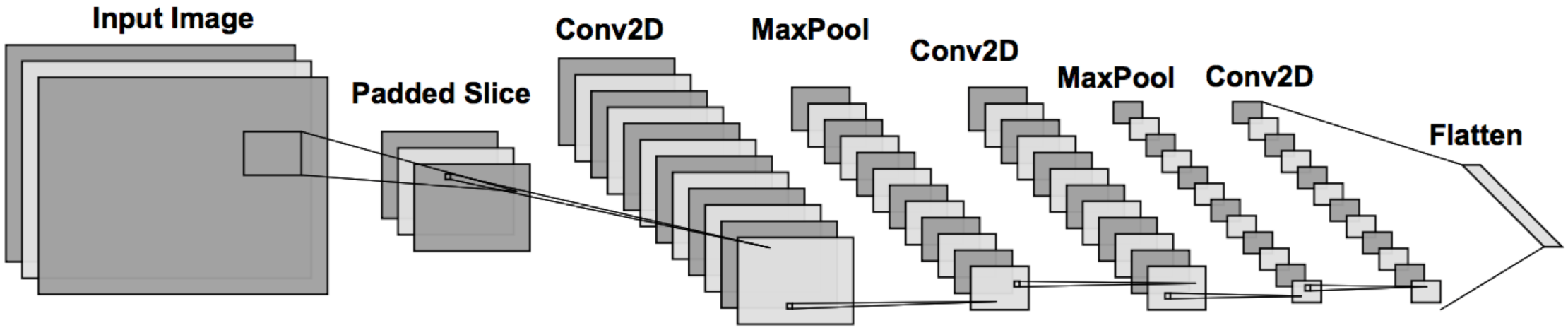




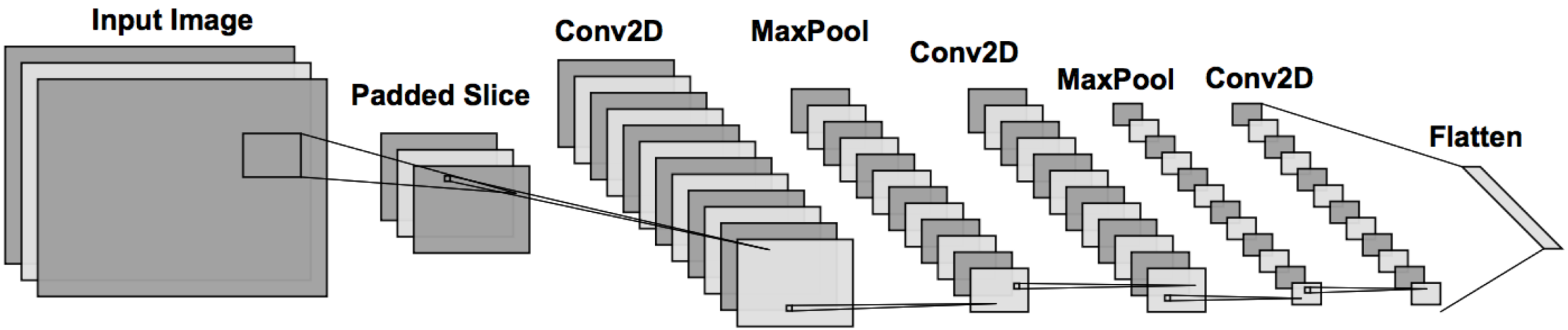
	Desktop GPU (Nvidia K80)	AIoT Accelerator (K210)
Cost	$900 (GPU only)	$8 ($3 chip only)
Power	300W, AC powered	300mW, battery-powered
Memory	24GB	2MB
Speed	13.45TFLOPS FP32	230GOPS INT8

Compression / De-Compression DNN Architecture

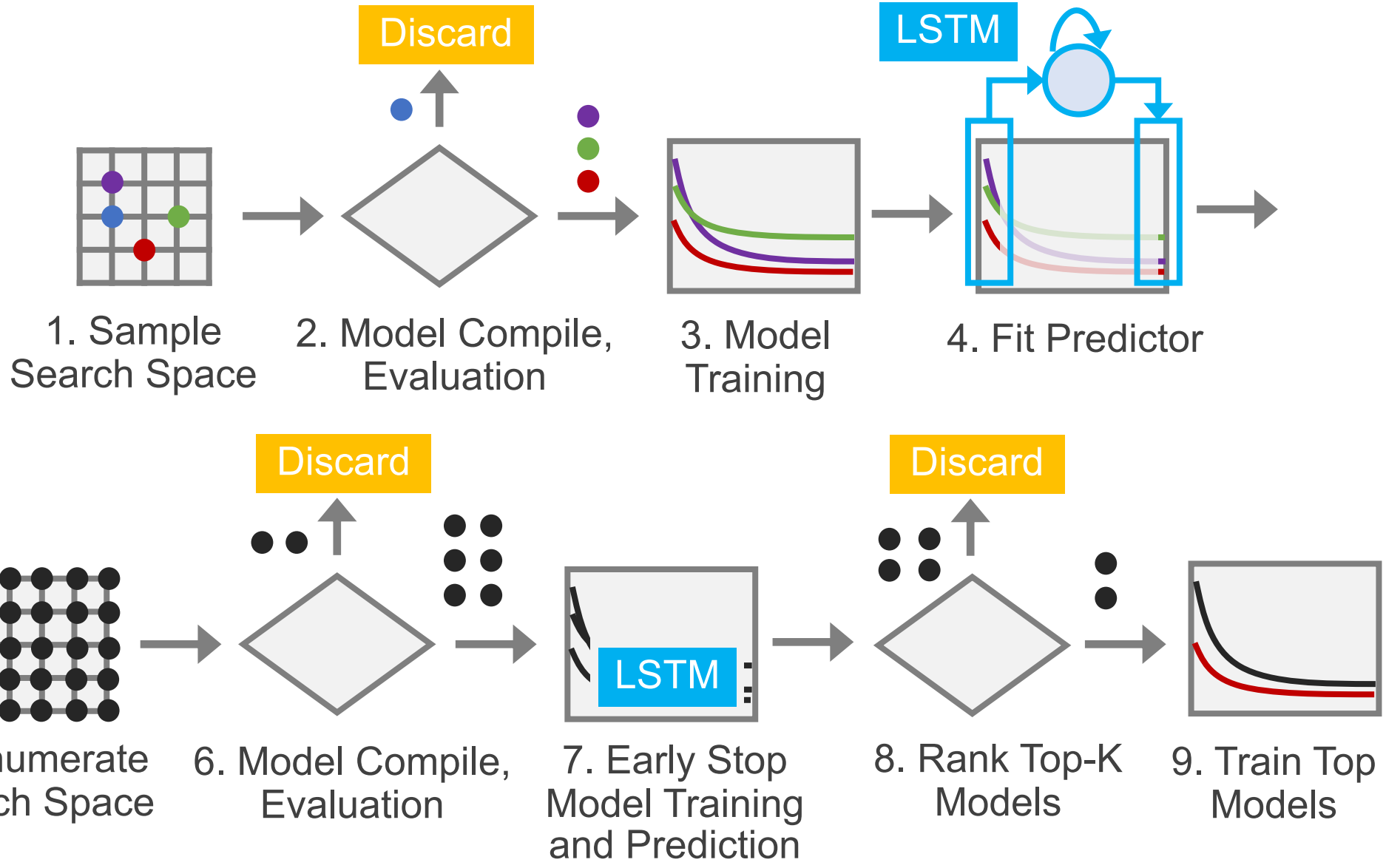




Search for Efficient DNN on AIoT Devices



Search for Efficient DNN on AIoT Devices



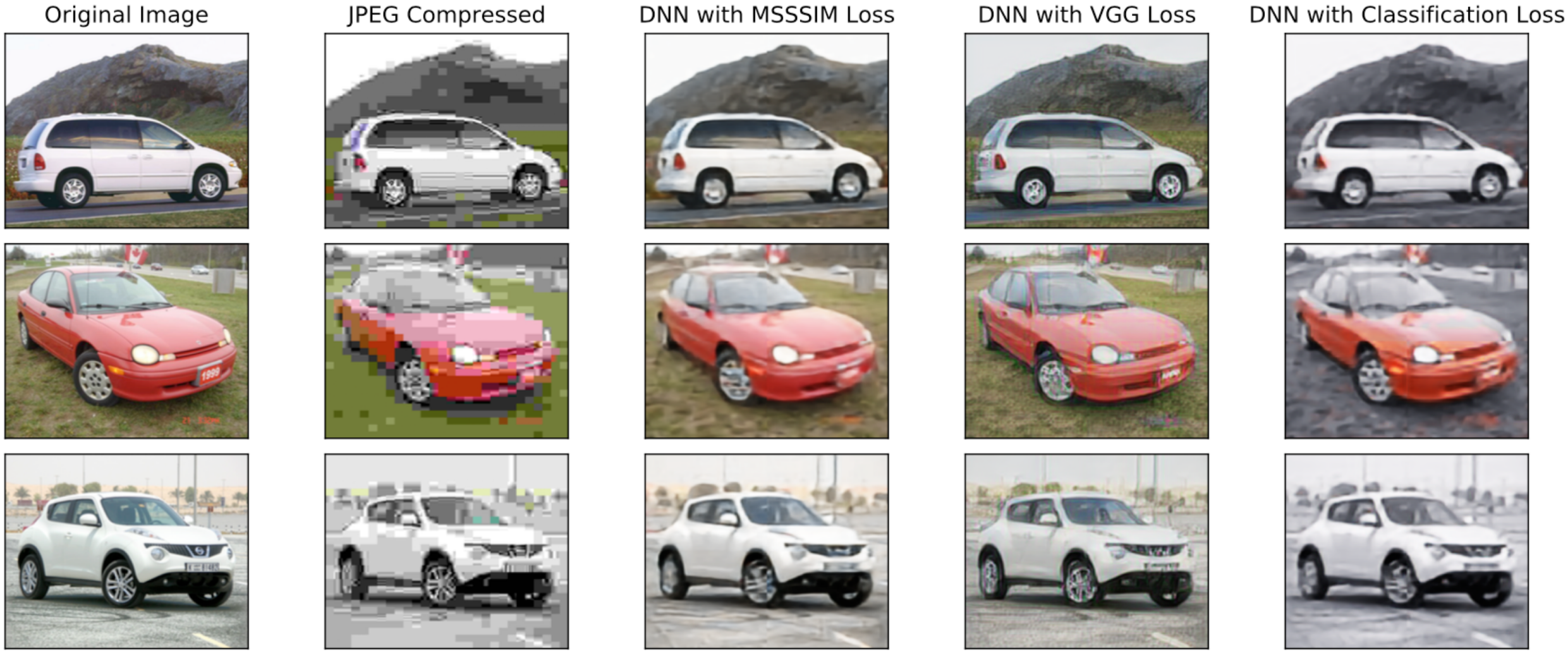
Implementation

- **Four widely used datasets**
- **>500 GPU hours** for training and evaluation




Dataset	Labeled Images	Classes
Stanford Cars [36]	8144	196
Caltech Birds 2011 [66]	11788	200
TensorFlow Flowers [62]	3670	5
Caltech 101 [26]	9144	102

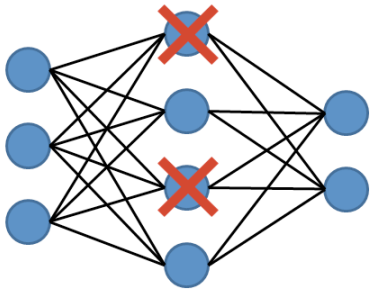
Directly Optimized for Task/Application

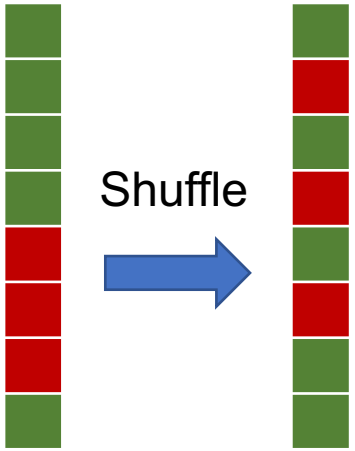


*All images for each compression method averages ~2.5kB in size.

Train DNN for Resiliency

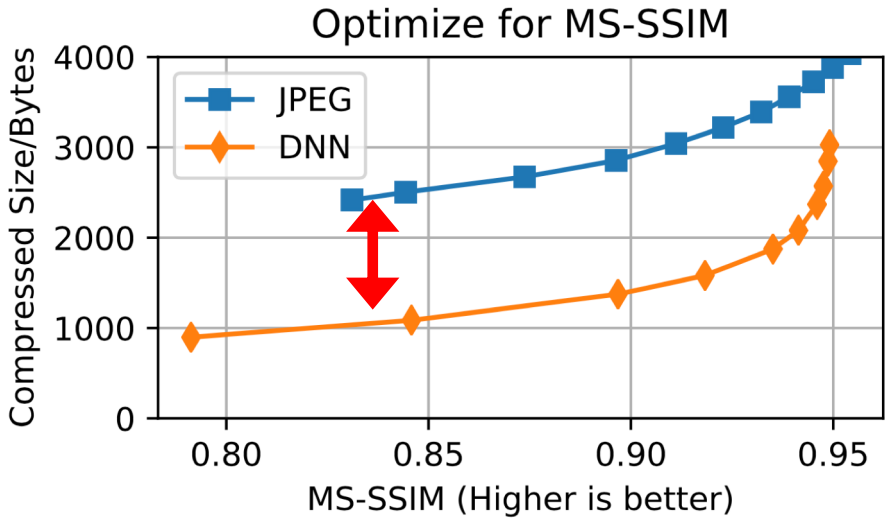
- **Dropout:** originally used to avoid overfitting, repurposed to simulate packet loss
- **Shuffle:** spread data loss over the entire image

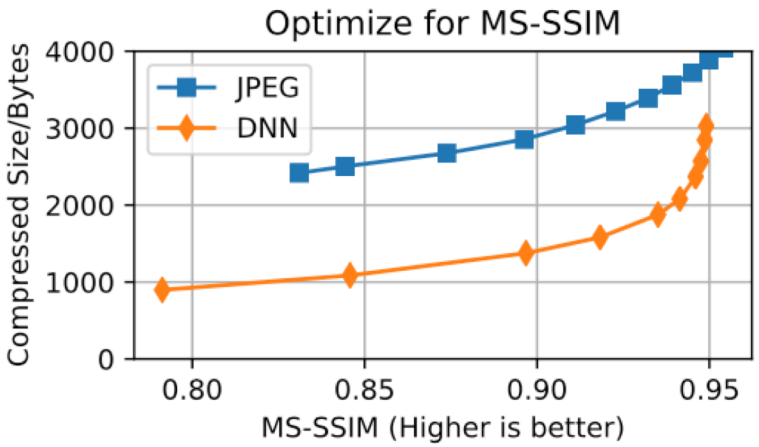


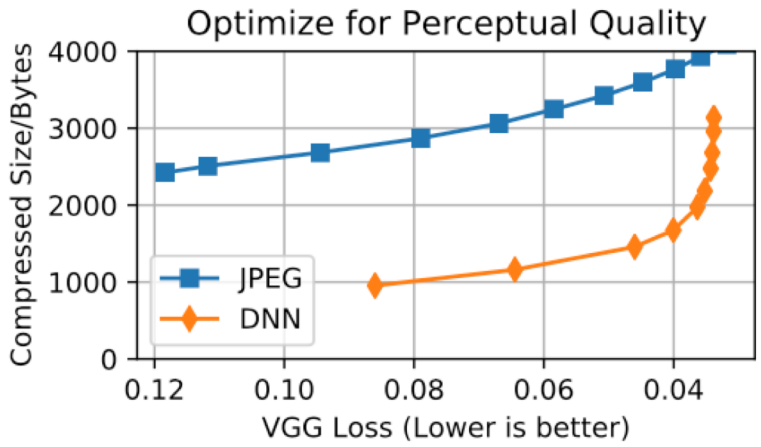


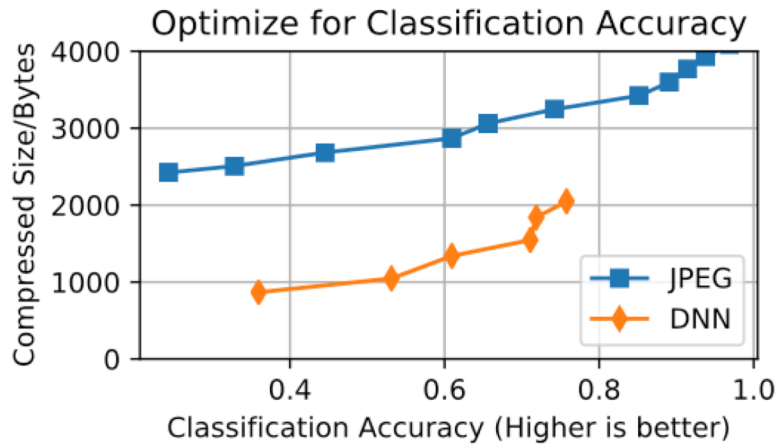
Compression Efficiency Benchmark

2.5x as efficient for SSIM



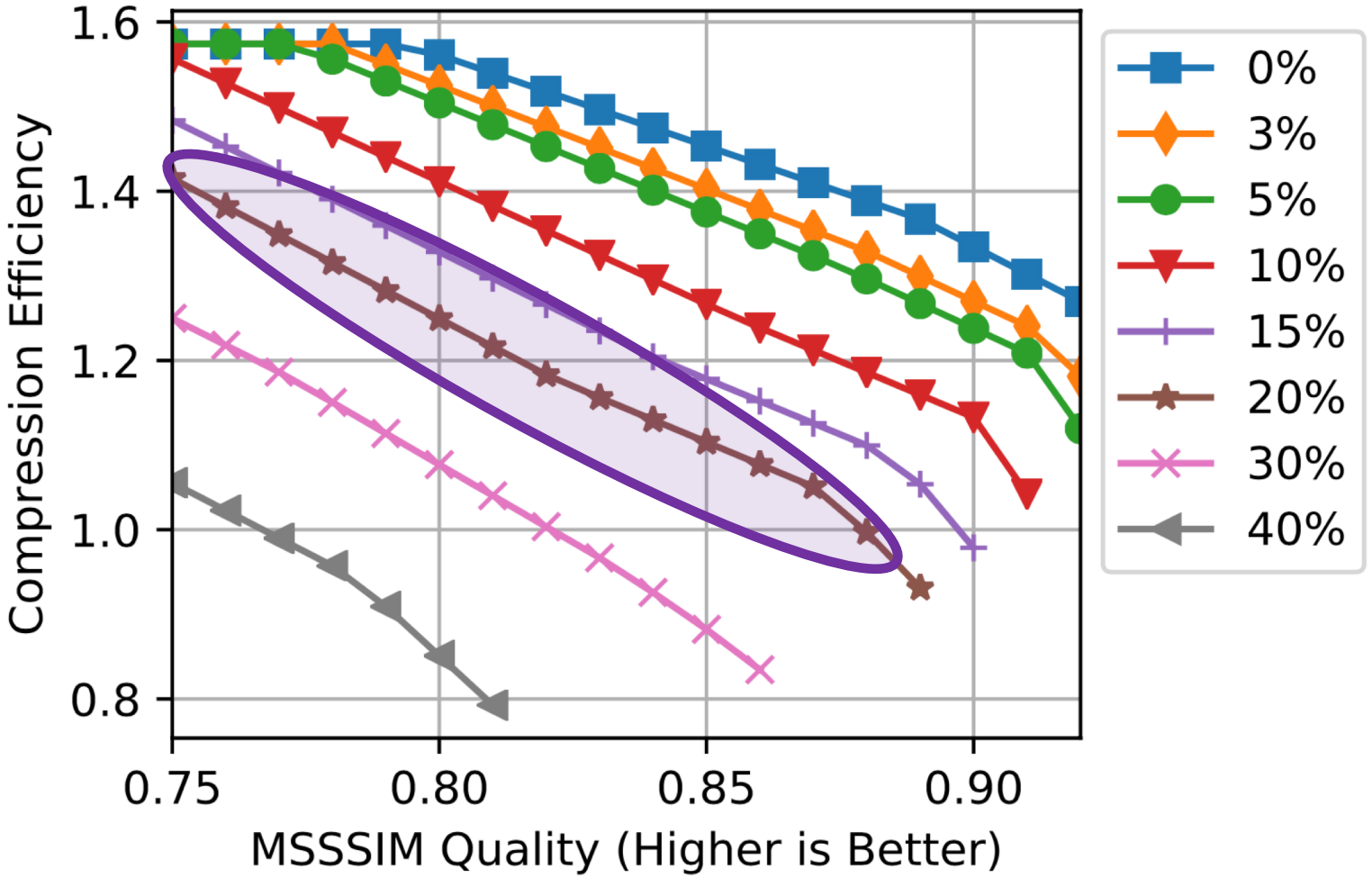






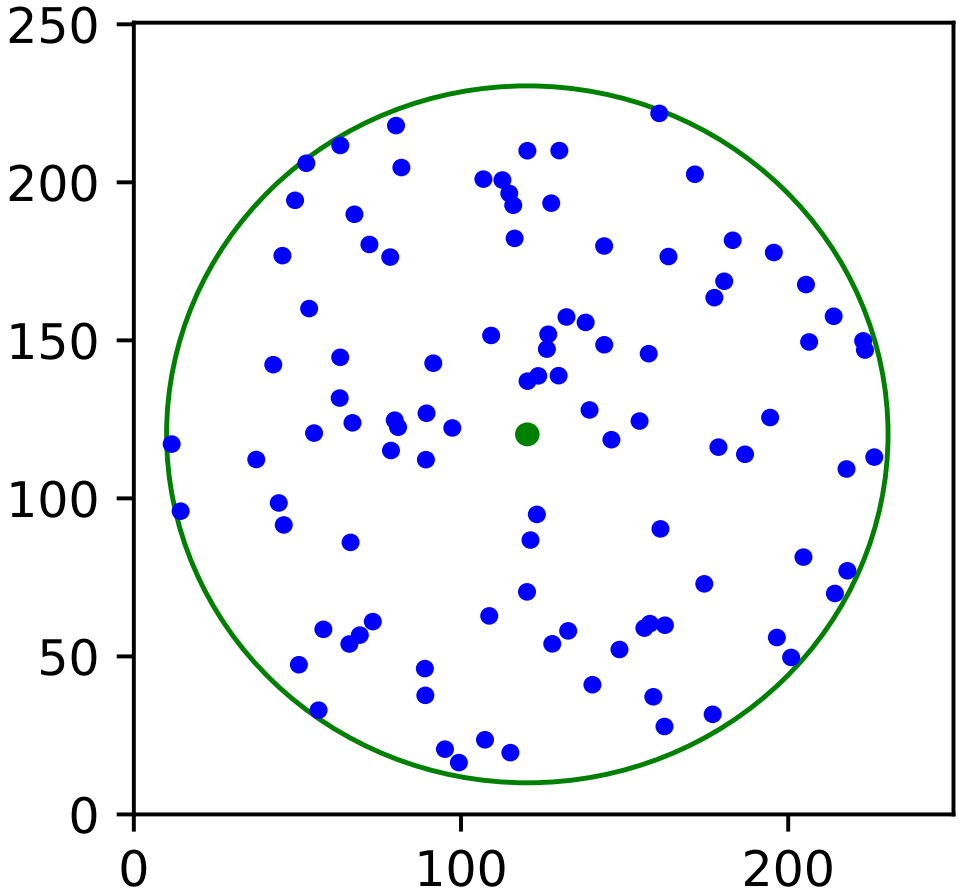
2.5~3x as efficient for various quality metrics

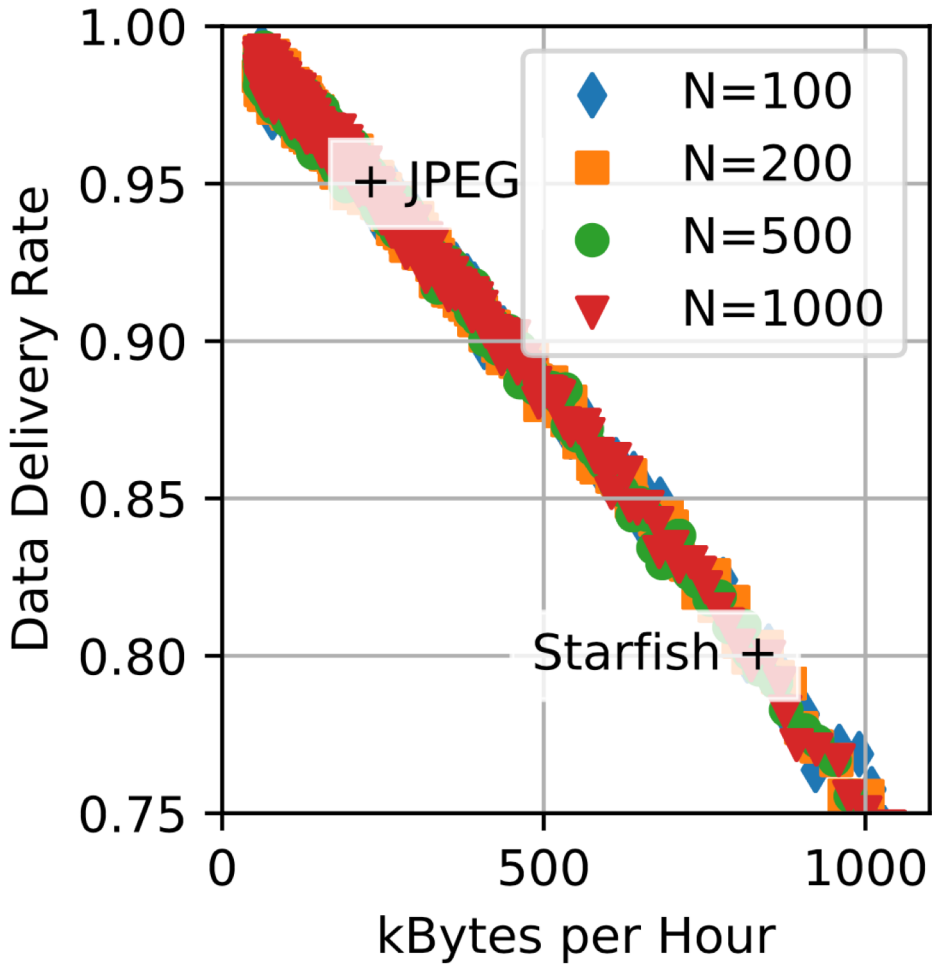
StarFish Loss Resiliency



StarFish with heavy loss is better than JPEG for given range

Large-scale simulation

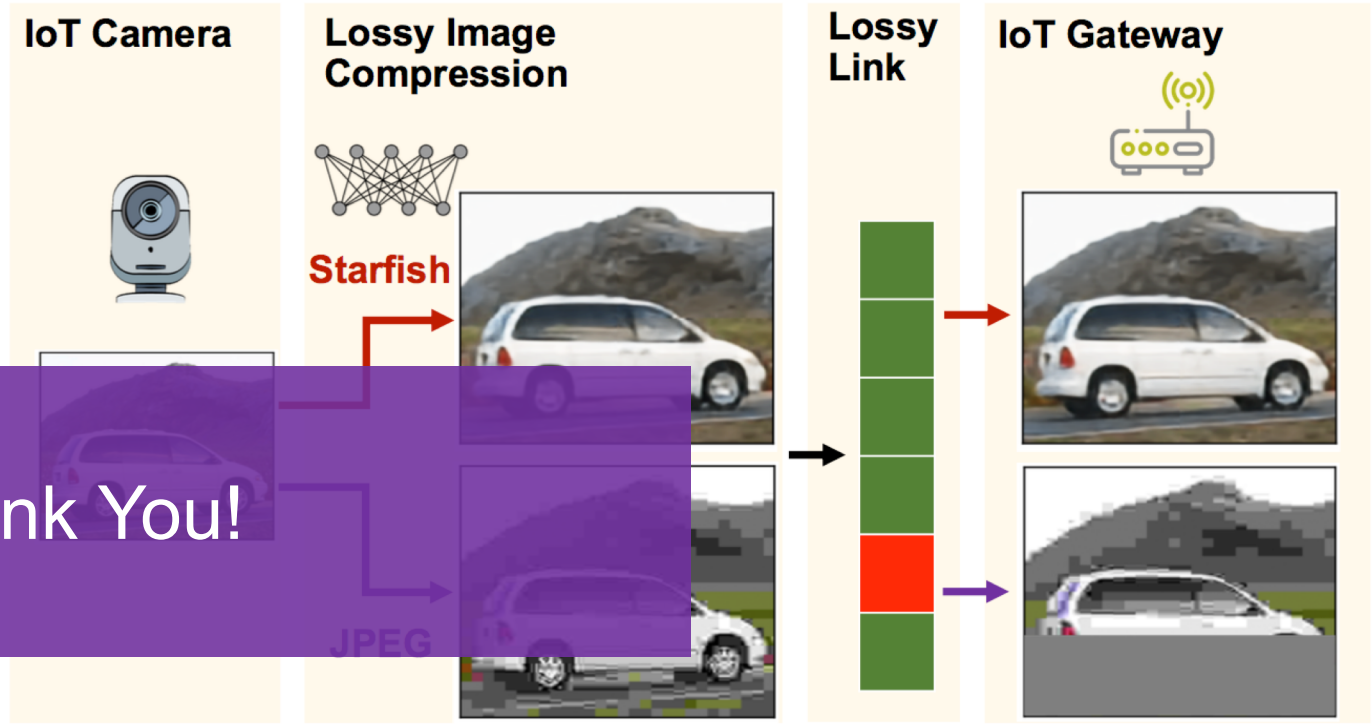




StarFish leads to much higher throughput by tolerating loss

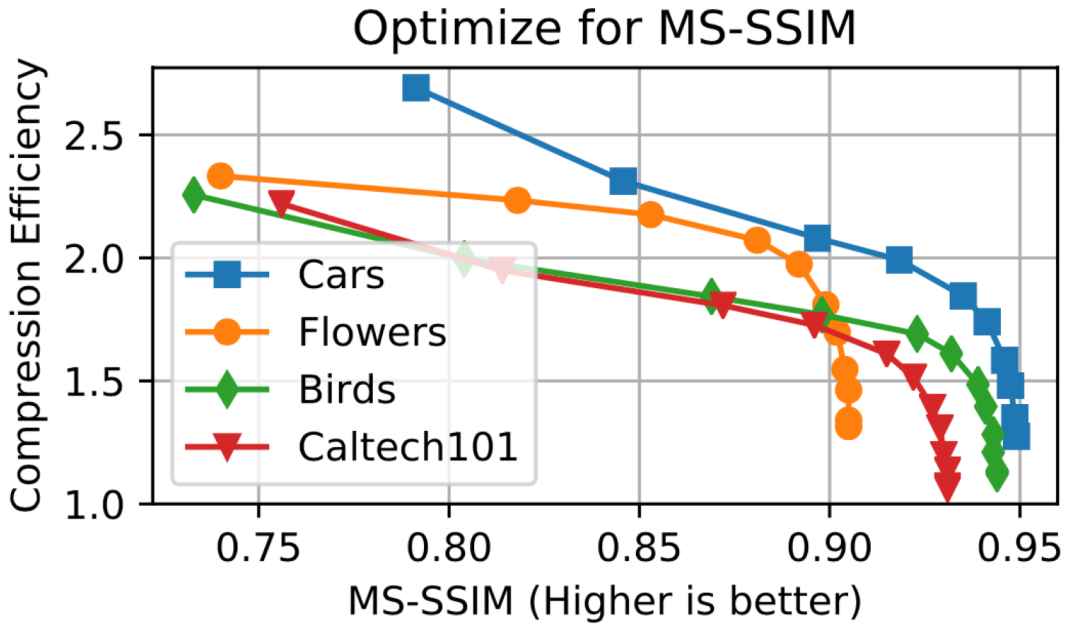
Summary of StarFish

- **Resilient** image compression framework designed for LPWAN that process all the information loss in the application layer
- **First DNN-based compression** runs efficiently on low-cost AIoT devices
- **Flexible, convenient, and universal** solution, more efficient than basic JPEG, especially in lossy scenarios



Thank You!

Compression Efficiency Benchmark



>2x as efficient for MS-SSIM