Tiny Overhead Person Detection On Synaptics Low-Power AI SoC

Omar Oreifej, Karthikeyan Shanmuga Vadivel, Patrick Worfolk
Synaptics Incorporated

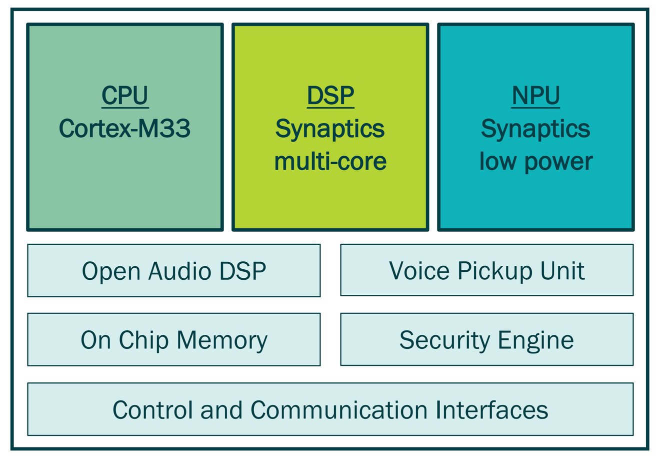
Target Application


Overhead person detection system

- Counts people in offices/conference rooms
- Wakes up on motion based on PIR detector
- Captures images at 5 fps
  - 3-5 images of a person entering or exiting the room at normal walking speeds
- Uses on-device NN processing of images to detect people
- Connects wirelessly to periodically send reports
- Runs for multiple years on batteries

Katana: Synaptics Low-Power AI SoC

Example applications
- Person and object detection
- Inventory tracking
- Keyword spotting/audio event detection
- Environmental sensing
- Emerging battery-powered audio and vision IoT products



Challenges

- Limited compute and memory for AI at the edge
- Low image resolution: 160x320 grayscale
- Data
  - No publicly available databases of overhead person
  - Requires collection and annotation of large amounts of data in order to train a robust model


Example overhead images of people captured by our system

Person Detection Model Architecture

We developed a person detection model based on RetinaNet architecture.

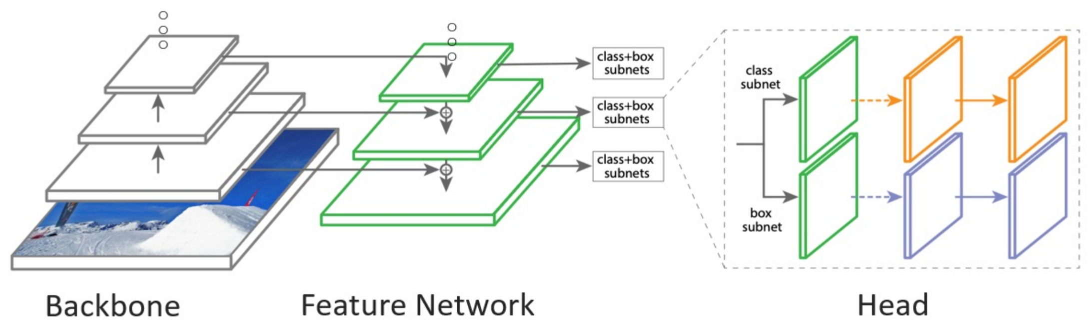


Figure from Lin et al. Focal loss for dense object detection. CVPR 2017

We reduced RetinaNet architecture significantly:
- Substituted ResNet backbone with MobileNetV2
- Used a single scale for regression and classification heads
- Reduced the backbone filters ($\alpha = 0.02$)
- Reduced the head filters by a factor of 8

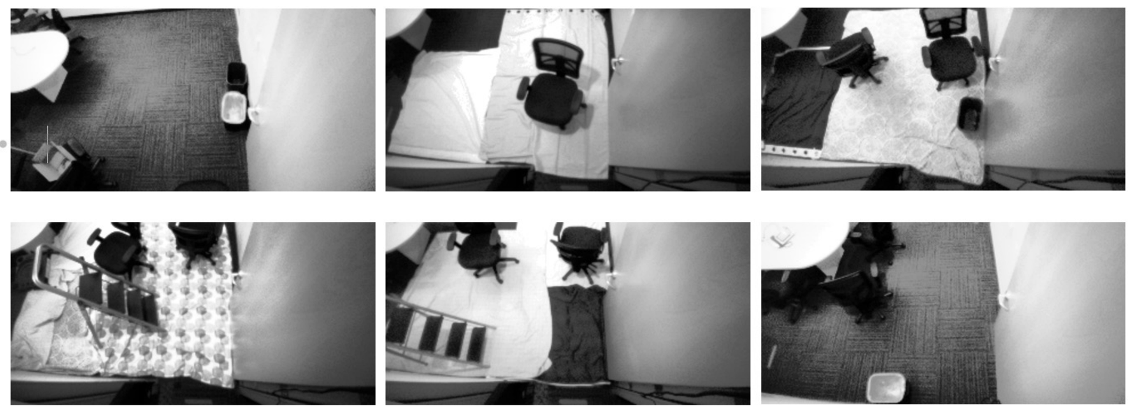
We optimized the architecture for Katana NPU:
- Combined Pointwise + Depthwise Conv → Full 2D Conv

	MobileNetV2 RetinaNet	+ Single scale head	+ Reduced backbone and head filters	+ Full 2D Conv (KatanaNet)
MACs	148 M	59 M	22 M	41 M
Weights	744 KB	313 KB	113 KB	150 KB
Total Memory	1967 KB	1536 KB	517 KB	381 KB

Dataset

Publicly available databases have limited representation of overhead person images.

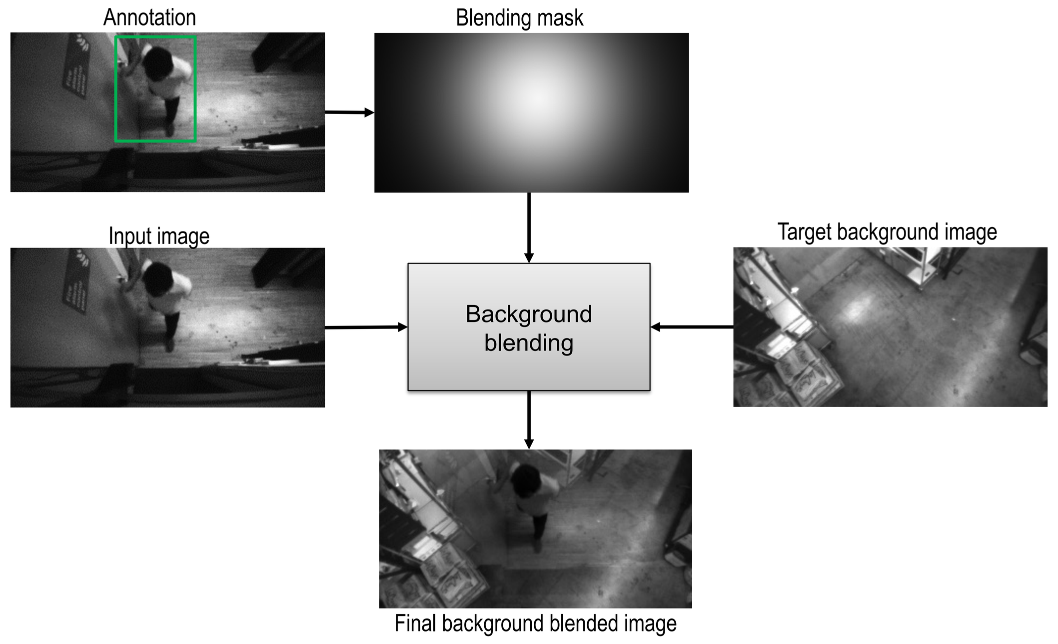
We collected our own dataset
- 18 rooms from Synaptics offices, with variations in lighting, furniture, and person appearance:
  - 13 rooms for training
    - 42K frames: 29K positive, 13K negative
  - 5 rooms for validation and testing
    - 8K frames: 5K positive, 3K negative
  - Total of 39 people
    - 12 women, 27 men
- Annotated each frame with ground truth bounding boxes tightly containing all parts of the person
- As the office environment has similar flooring, we increased diversity in flooring by using sheets


Example room from our data collection, in a variety of configurations

Data Augmentation

We use standard data augmentation approaches to increase the data diversity. This includes image rotation, scaling, mirroring, contrast and brightness adjustments.

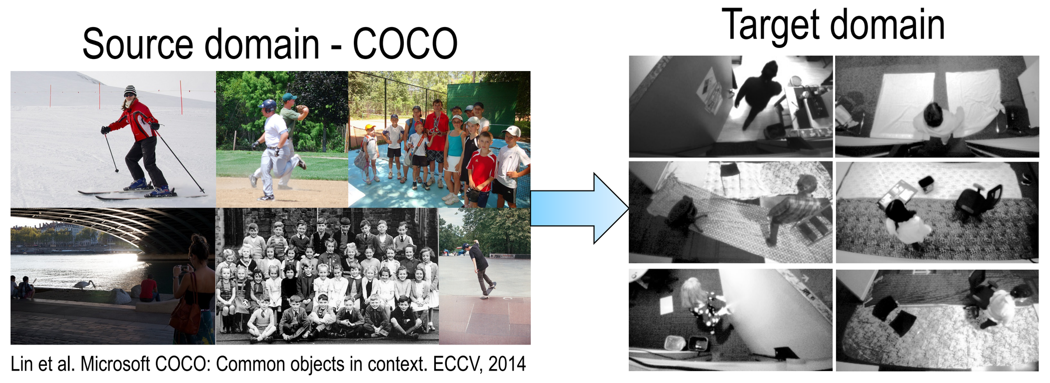
The primary shortcoming in our data collection is the limited diversity of the background. We address this issue by a novel background augmentation approach.



Our background augmentation approach

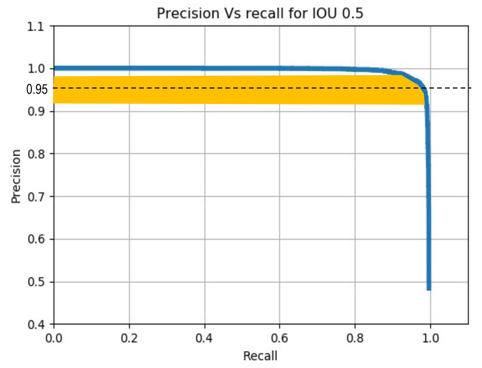
Training and Evaluation

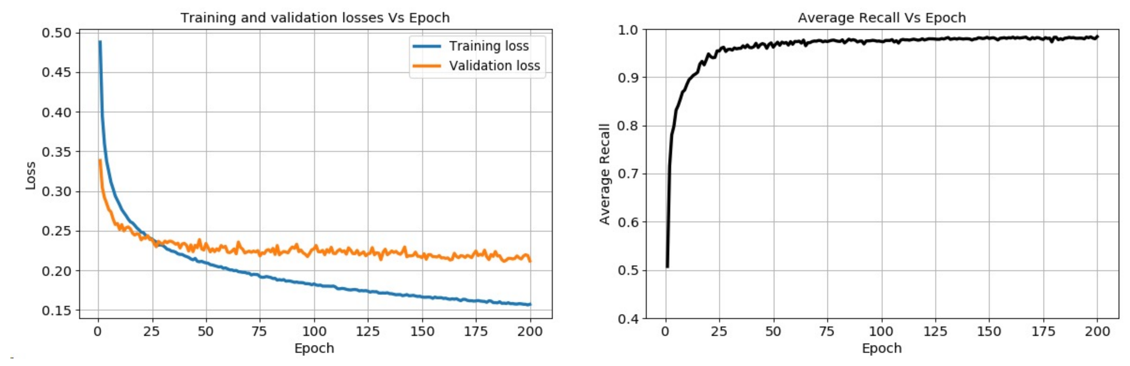
- We use domain adaptation for training our model
- We first pre-train the network on COCO 2017 person detection dataset, then finetune the model further on our dataset



Lin et al. Microsoft COCO: Common objects in context. ECCV, 2014

For evaluation, we use the standard precision recall metrics. However, since we expect the person counting application to operate only at high precision (~95%), we report the average recall for the precision region between 0.92 and 0.98. We refer to this metric as Average Recall.

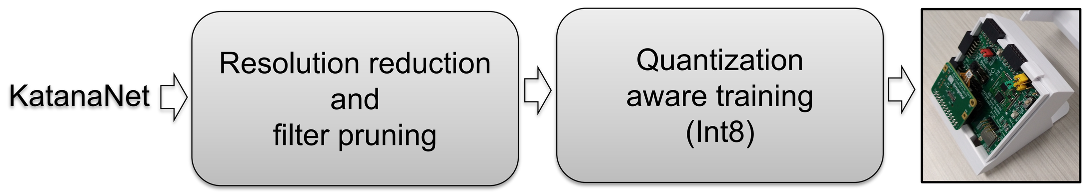




Model	Average Recall (AR)
KatanaNet	0.983
KatanaNet without background augmentation	0.958 (-2.5%)
KatanaNet without background blending	0.968 (-1.5%)
KatanaNet without domain adaptation	0.804 (-18.2%)

Ablation study: The contribution of the training elements in the performance

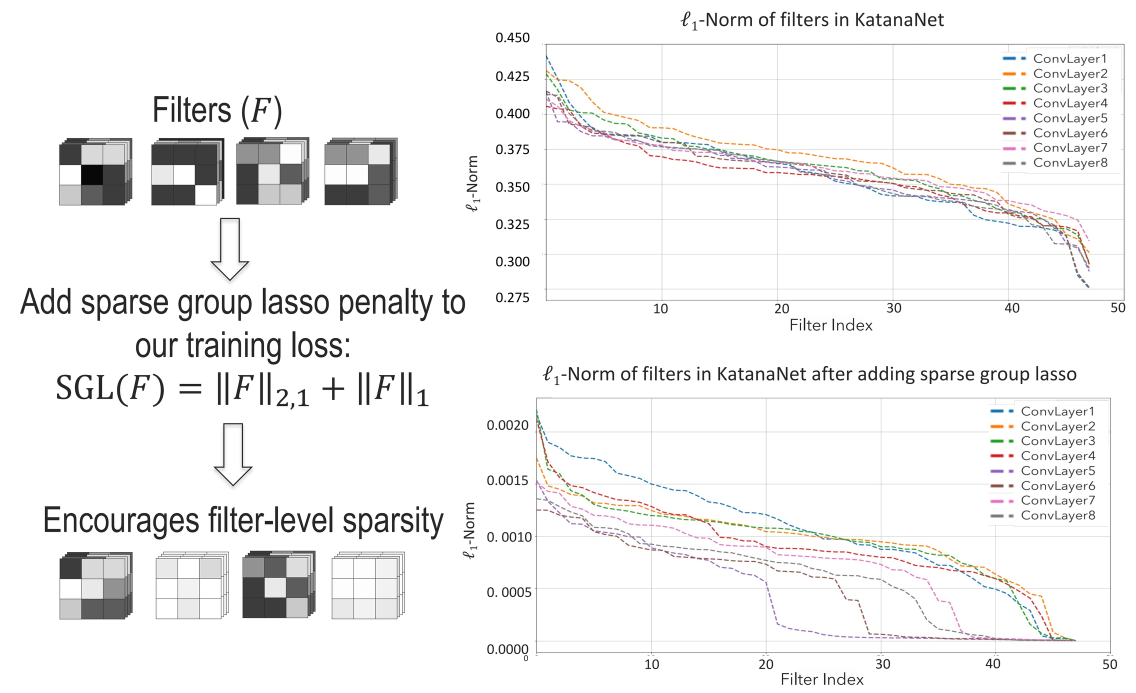
Deployment On Katana



- Resolution reduction

Input Image Size	MACs	Weights	Average Recall	Total Memory
320x160	41 M	150 KB	0.98	381 KB
160x80	14 M	134 KB	0.92	209 KB

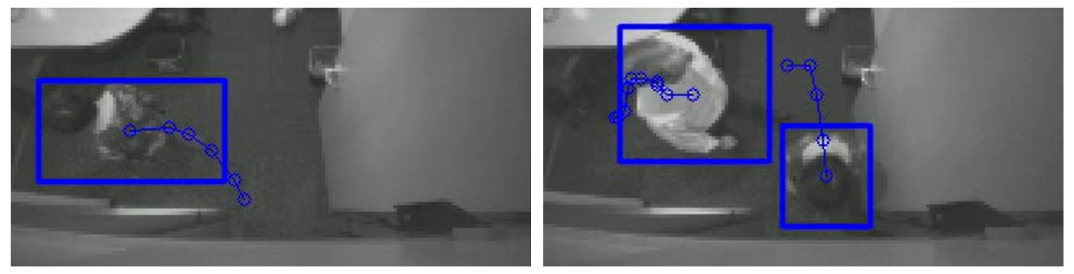
- Filter pruning:
  - Modified the loss function to identify the filters not contributing to the performance
  - Removed filters with small ℓ_1-Norm



Pruning	MACs	Average Recall		Total Memory
		Before Quantization	After Quantization	
Before Pruning	14 M	0.92	0.91	209 KB
After Pruning	10 M	0.91	0.91	161 KB

People Counting

- We developed a simple motion-based tracking algorithm to count the number of people entering and exiting the scene
- The tracker associates the detections from KatanaNet across the frames in order to create tracks



Conclusions

- We designed a compact overhead people detection neural network for the Synaptics Katana SoC
- Currently available to our customers as part of the Katana EVK

Acknowledgments: Thanks to Maryam Jaberi, Sohum Shenoy, and Ted Montreevat for their help in this work