Efficient Video Perception Through AI

Fatih Porikli
Sr. Director, Technology
Qualcomm Technologies, Inc.

A picture is worth a thousand words

Out of all the five senses, **vision** is arguably the most important

A minute of video has more than 1000 pictures

The scale of video being created and consumed is massive

2021 embedded vision summit

1M
Minutes of video crossing the internet per second

82%
Of all consumer internet traffic is online video

76
Minutes per day watching video on digital devices by US adults

8B
Average daily video views on Facebook

300
Hours of video are uploaded every minute to YouTube

Cisco Visual Networking Index: Forecast and Trends, 2017–2022

Increasingly, video is all around us —

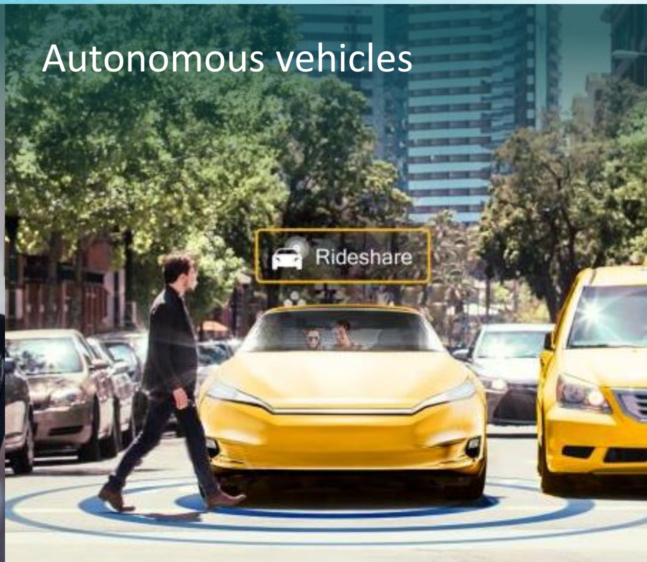
providing entertainment, enhancing collaboration, and transforming industries

2021 embedded vision summit

Smartphone

Sports

Video conferencing



Autonomous vehicles

Smart factories

Extended reality

Smart cities

Video monitoring

Qualcomm

What is video perception?

2021 embedded vision summit

Video perception

Making systems understand video content

Making

Developing mathematical representations, models, algorithms, rules, and frameworks

Systems

Any compute platform, including SoCs, CPUs, GPUs, TPUs, NPUs, and DSPs

Understand

Recognizing patterns, identities, objects, scenes, context, relations, compositions, changes, motions, actions, activities, events, 3D structures, surfaces, lightings, text, emotions, sentiments, sounds, and more

Qualcomm

What makes video perception challenging?

Data challenges

- Diversity in visual data
- Quality of data acquisition
- Availability of annotated datasets

Video perception challenges

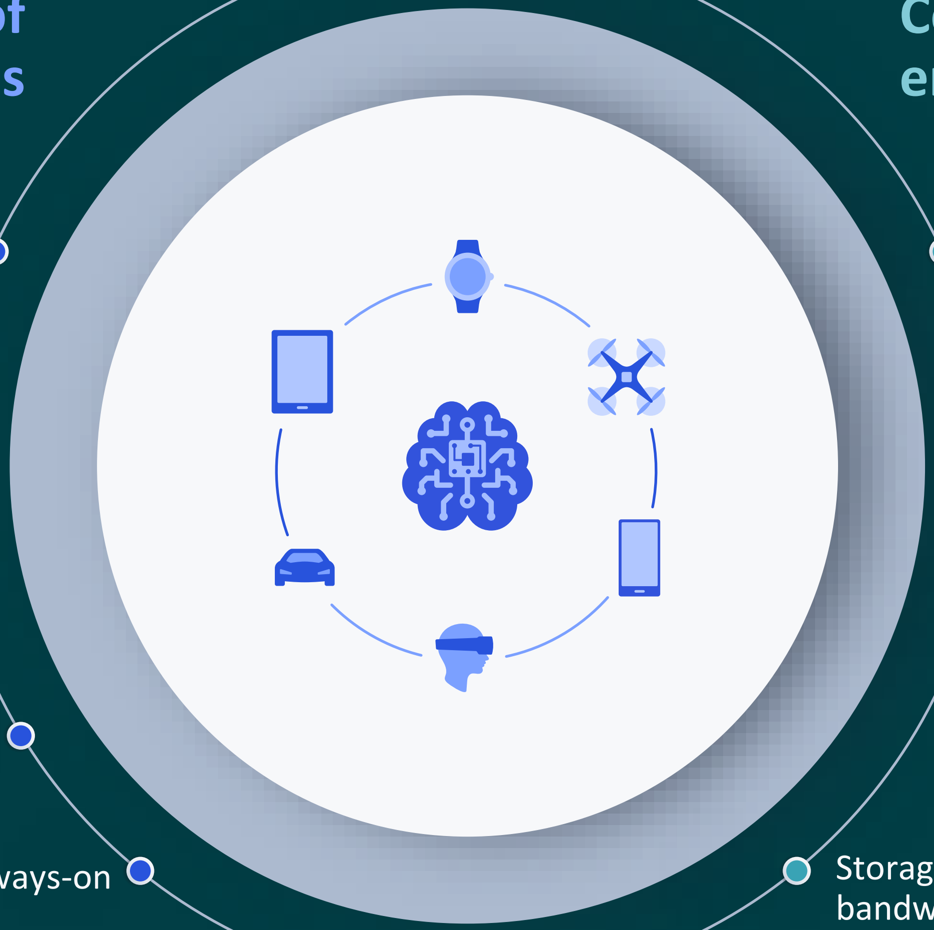
Implementation challenges

- Volume of video data (training/testing)
- Platform limitations
- Task diversity

Power and thermal efficiency are essential for on-device video perception

The challenge of AI workloads

- Very compute intensive
- Large, complicated neural network models
- Complex concurrencies
- Real-time
- Always-on

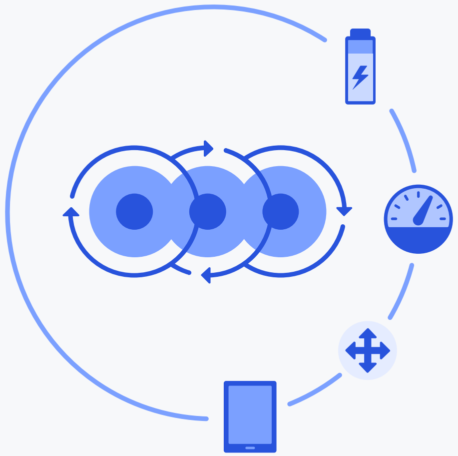


Constrained mobile & embedded environments

- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations

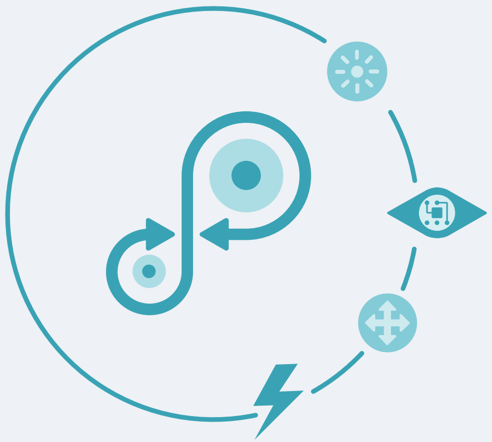
Making video perception ubiquitous

Solving additional key challenges to take video perception from the research lab to broad commercial deployment



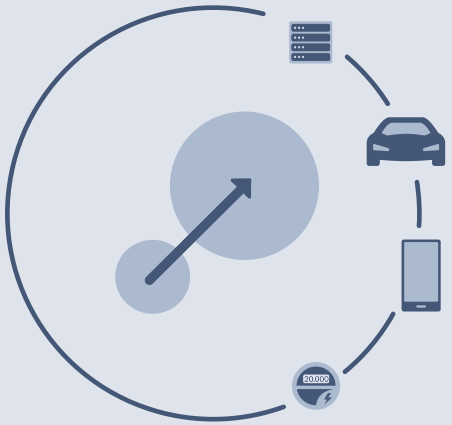
Robustness

Robust to data variations



Adaptability

Adaptable to different domains



Scalability

Scaling up and down, from IoT to the data center

Efficiently running on-device video perception without sacrificing accuracy

Leverage
Temporal redundancy

By reusing what is computed before

- **Learning to skip regions**
- **Recycling features**

Key concepts for efficient video perception

Make
Early decisions

By dynamically changing the network architecture per input frame

- **Early exiting**
- **Frame exiting**

Learning to skip redundant computations

Video frames are heavily correlated

frame t

−

frame t+10

=

residual

The residual frame, the difference between two consecutive frames, contains little information in most regions

“Skip-convolutions for efficient video processing” (CVPR 2021)

Limit the computation only to the regions where there are significant changes

Skip-convolution

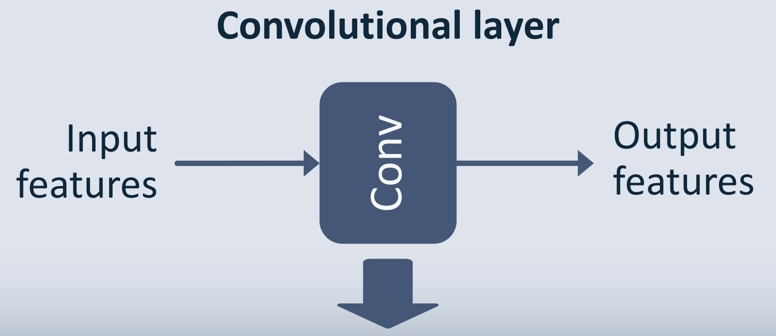
A convolutional layer with a skip gate that masks out negligible residuals

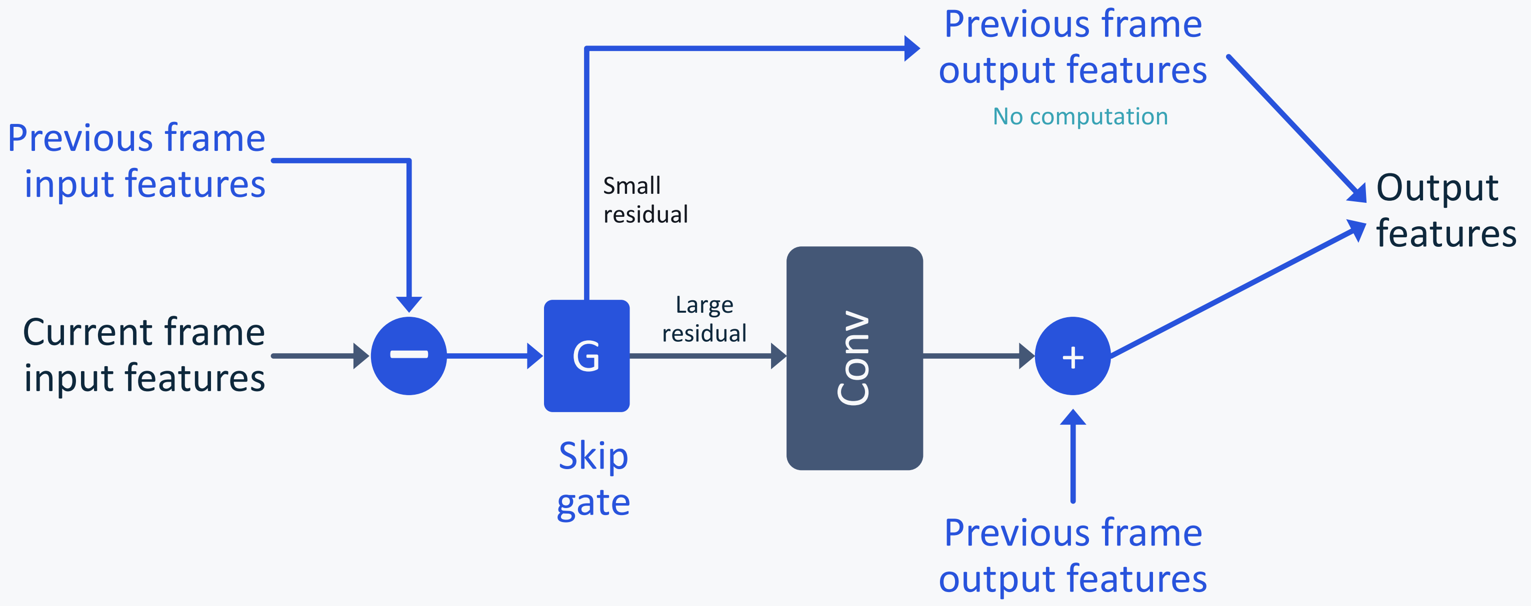
A convolution at a frame can be written as the previous frame's convolution plus the convolution of the residual

Computation is limited only to the regions where there are strong residuals

Reinforce residual's sparsity by removing negligible residuals

Can replace convolutional layers in any CNN with skip convolutions



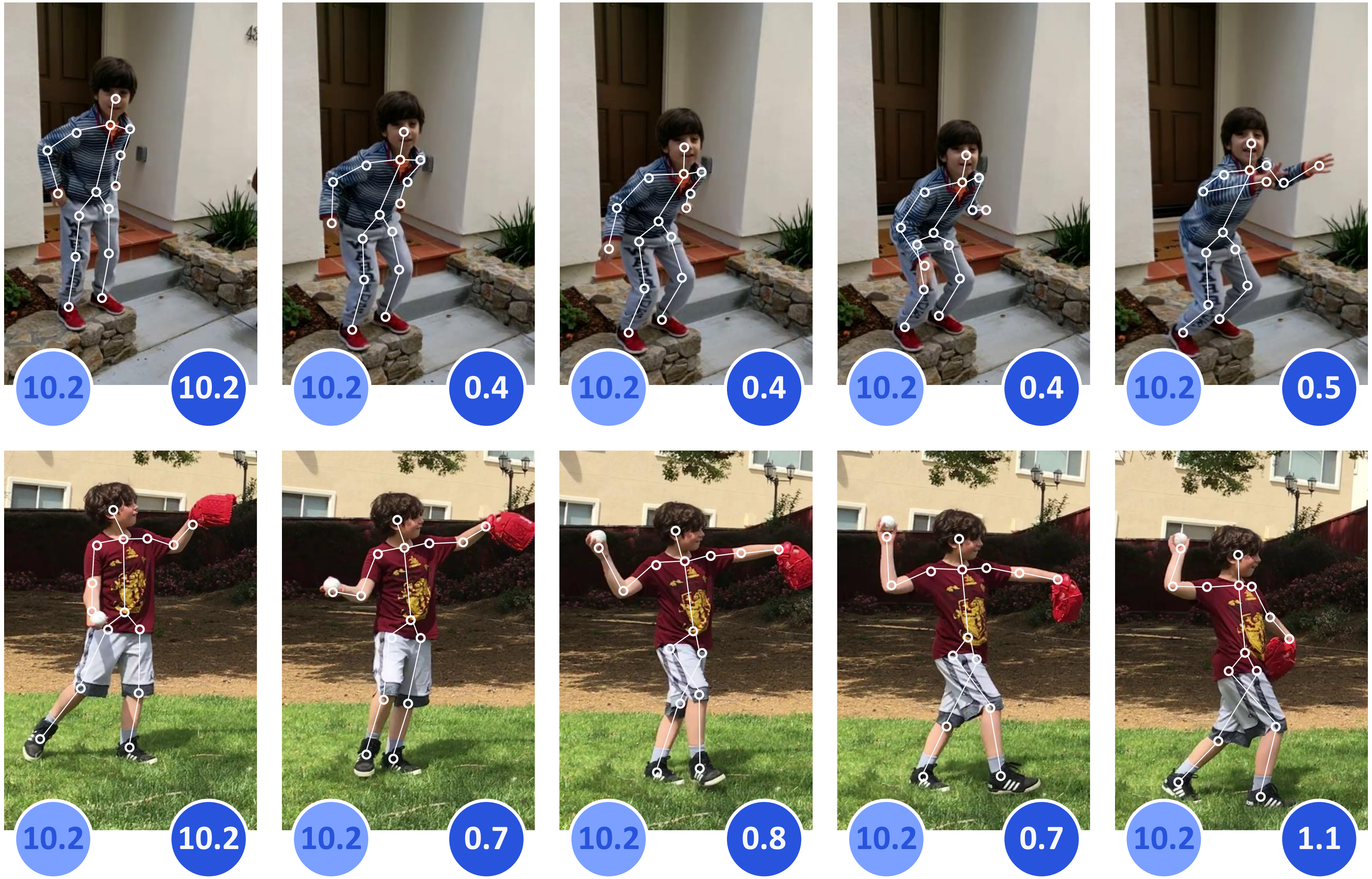


"Skip-convolutions for efficient video processing" (CVPR 2021)

Learning to skip reduces compute for human pose estimation

Results for human pose estimation

- GMACs **without** skip-convolutions
- GMACs **with** skip-convolutions



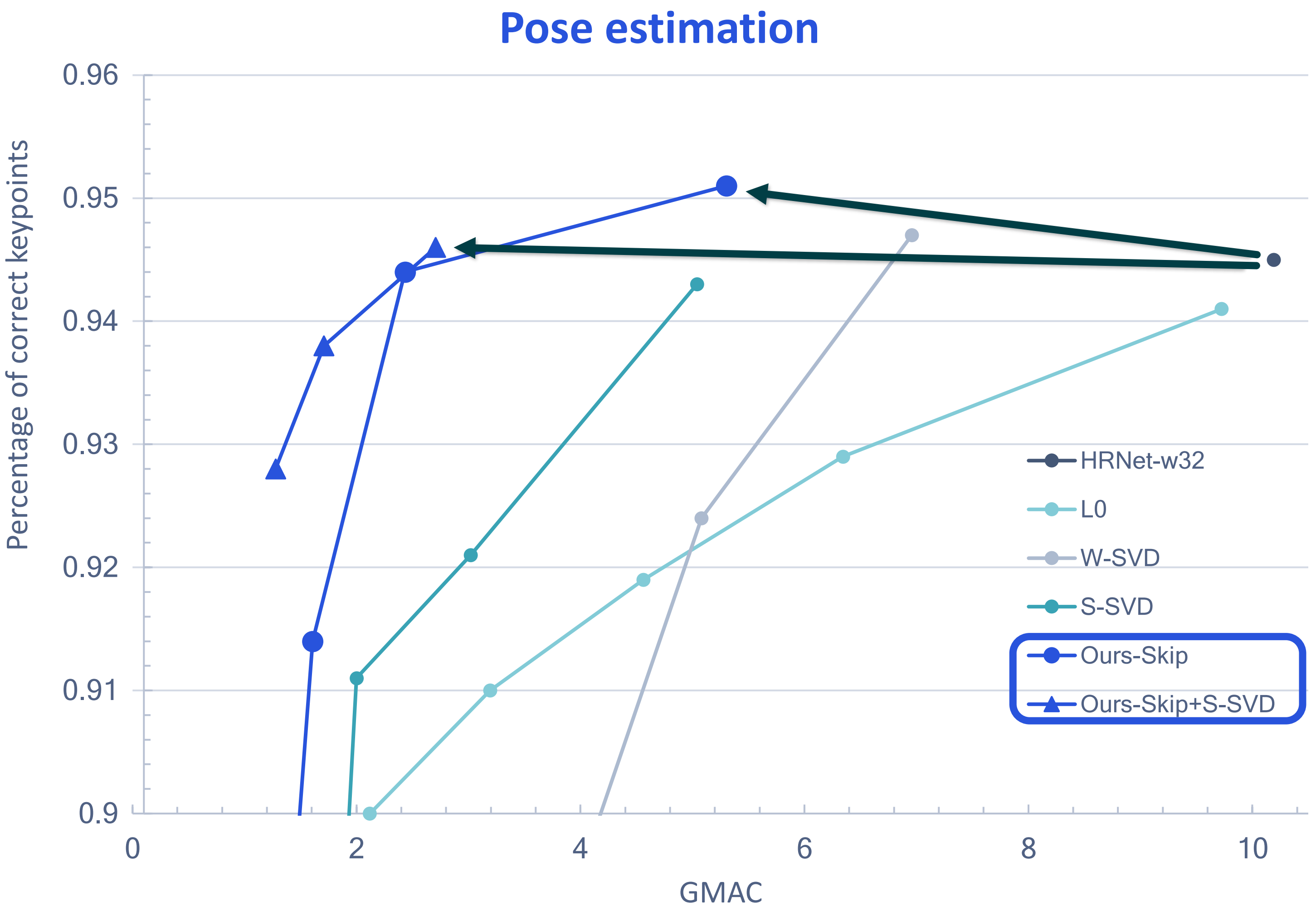
"Skip-convolutions for efficient video processing" (CVPR 2021)

Learning to skip is complementary to model compression

Results for human pose estimation on video human action dataset

2.5x- 8x
speed-up over HRNet

"Skip-convolutions for efficient video processing" (CVPR 2021)



Qualcomm

Recycling features saves compute

Instead of computing deep features repetitively, compute once and recycle

Deep features remain relatively stationary over time — they have lower spatial resolution

Compute deep features once and recycle — reuse from past frame

Shallow features are more responsive to smooth changes, encoding the temporally varying information

Compute shallow features for all frames

“Time-sharing networks for efficient semantic video segmentation” (submitted 2021)

Applicable to any video neural network architectures including segmentation, optical flow, classification, and more

Recycling features saves compute

Instead of computing deep features repetitively, compute once and recycle

Visual example of recycling features for a semantic segmentation task

Raw video obtained from Cityscapes Benchmark: https://www.cityscapes-dataset.com/
"Time-sharing networks for efficient semantic video segmentation" (submitted 2021)

Feature recycling reduces compute and latency

Semantic segmentation example

Input:
2048x1024 RGB video

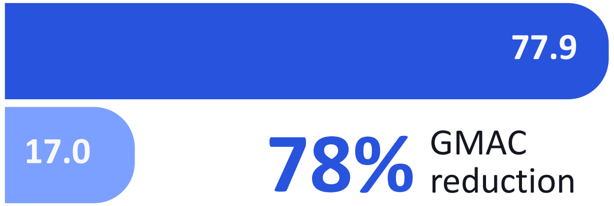
Output:
2048x1024,
19 object classes

Runs on:
Qualcomm® Snapdragon™ 888 Mobile Platform

Model efficiency

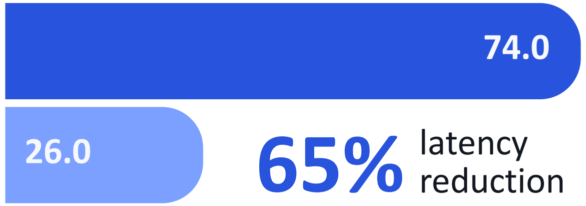
HRNet w18 v2
Enhanced Net

GMACs



78% GMAC reduction

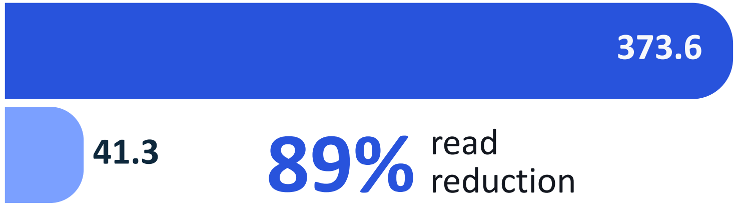
On-device latency (ms/frame)



65% latency reduction

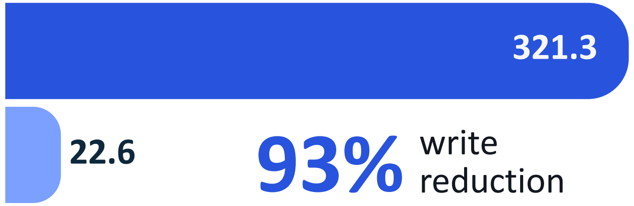
Memory traffic

MB read



89% read reduction

MB write



93% write reduction

"Time-sharing networks for efficient semantic video segmentation" (submitted 2021)

Early exiting a neural network saves compute

Exploit the fact that not all input examples require models of the same complexity

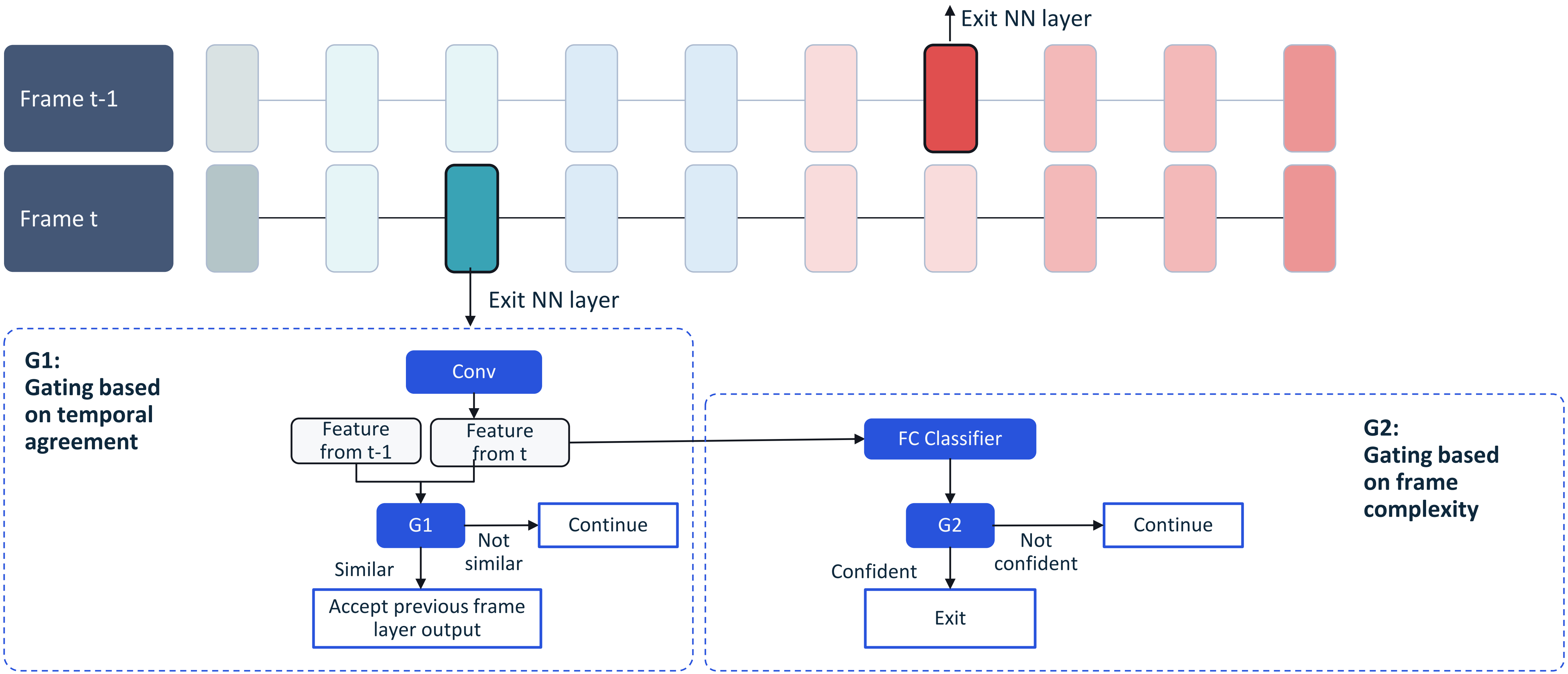
Complex examples — Very large, computationally intensive models are needed to correctly classify

Simple examples — Very small and compact models can achieve very high accuracies, but they fail for complex examples

"FrameExit: Conditional early exiting for efficient video recognition" (CVPR 2021)

Ideally, our system should be composed of a cascade of classifiers throughout the network

Early exiting at the earliest possible NN layer for video

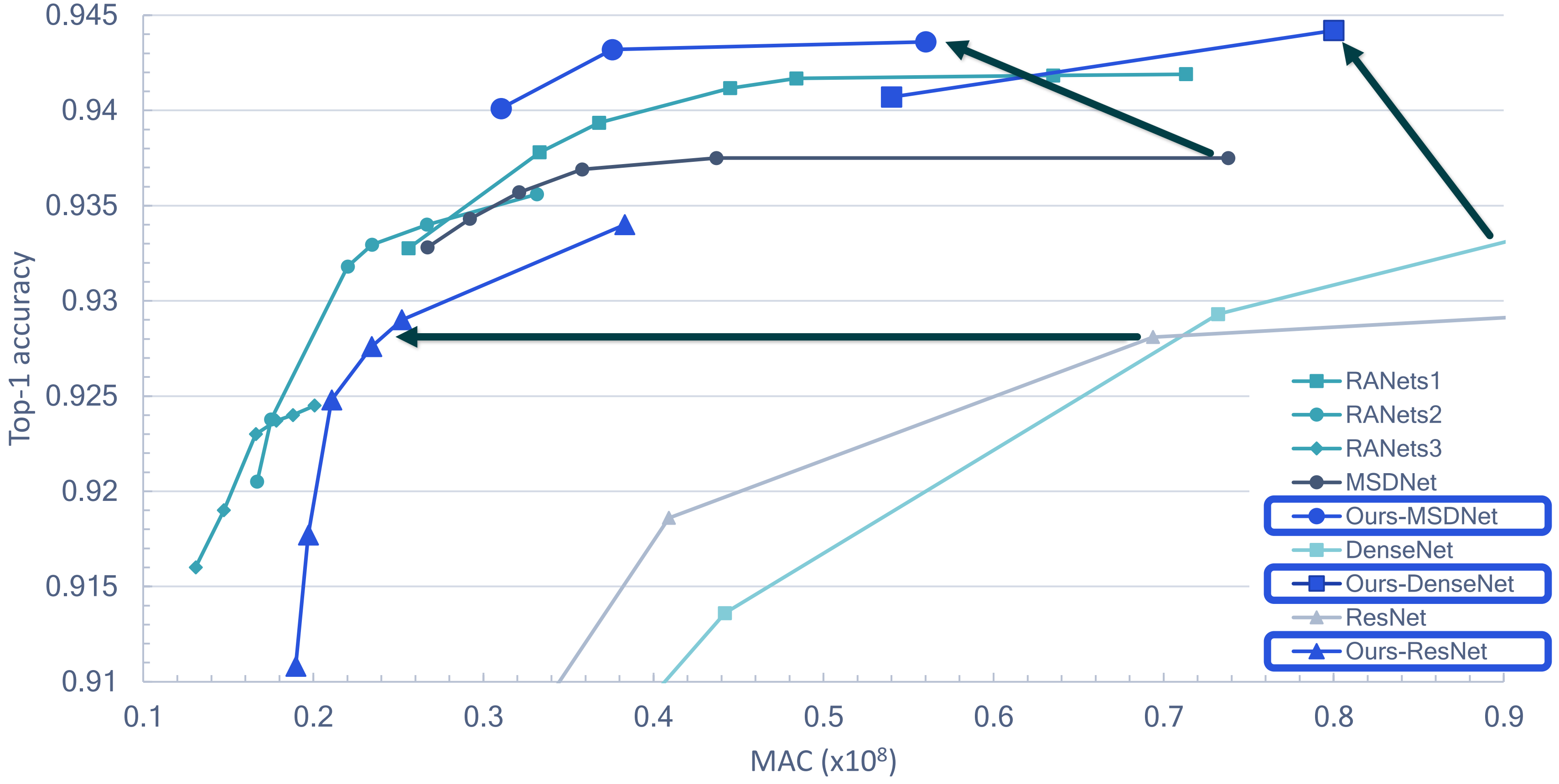


"FrameExit: Conditional early exiting for efficient video recognition" (CVPR 2021)

Early exiting reduces compute while maintaining accuracy

Early exiting applies to most neural network backbones

Classification on image dataset



"FrameExit: Conditional early exiting for efficient video recognition" (CVPR 2021)

What's next?

Advance existing conditional compute techniques

- Learning to skip regions
- Recycling features
- Early exiting
- Frame exiting

Future work in video perception

Develop efficient video neural network solutions

- Unsupervised / semi-supervised learning
- Efficient sparse convolutions
- Personalization
- Multi-task networks
- Quantization aware training
- Platform optimizations

Our perception research is much broader than video

Machine learning core

Research Areas

- Computer vision
- 3D Sensing
- RF Sensing
- Personalization
- Biometrics
- Data driven models

Leading high-impact research efforts in perception

Inventing technology enablers for important applications

Applications

- XR
- Camera
- Autonomous vehicles
- Wi-Fi
- Fingerprint ASP
- Mobile
- IOT
- Cloud

Takeaways

Qualcomm

Video perception is crucial for understanding the world and making devices smarter

We are conducting leading research and development in video perception

We are making power efficient video perception possible without sacrificing accuracy

Resource slide

- **Qualcomm AI page:**

https://www.qualcomm.com/invention/artificial-intelligence

- **Qualcomm AI Research page:**

https://www.qualcomm.com/invention/artificial-intelligence/ai-research

- **Qualcomm® Platform Solution Ecosystem:**

https://www.qualcomm.com/support/qan/platform-solutions-ecosystem

- **GitHub open-source projects:**

https://github.com/quic/aimet

https://github.com/quic/aimet-model-zoo/

- **Qualcomm Mobile AI page:**

https://www.qualcomm.com/products/smartphones/mobile-ai

- **Qualcomm Mobile AI blog:**

https://www.qualcomm.com/news/onq/2020/12/02/exploring-ai-capabilities-qualcomm-snapdragon-888-mobile-platform

- **Qualcomm® Cloud AI 100 blog:**

https://www.qualcomm.com/news/onq/2021/03/15/qualcomm-cloud-ai-100-amd-epyc-7003-series-processor-and-gigabyte-server

- **Qualcomm AI Research blog:**

https://www.qualcomm.com/news/onq/2020/09/01/pushing-boundaries-ai-research

Thank you

2021 embedded VISION summit

Qualcomm

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog