January 19, 2021

@QCOMResearch

Qualcomm

Efficient Video Perception through AI

Qualcomm Technologies, Inc.

Agenda

- The role of video in our lives
- What is video perception & what makes it challenging
- Our research toward efficient video perception
- Forward looking video perception research

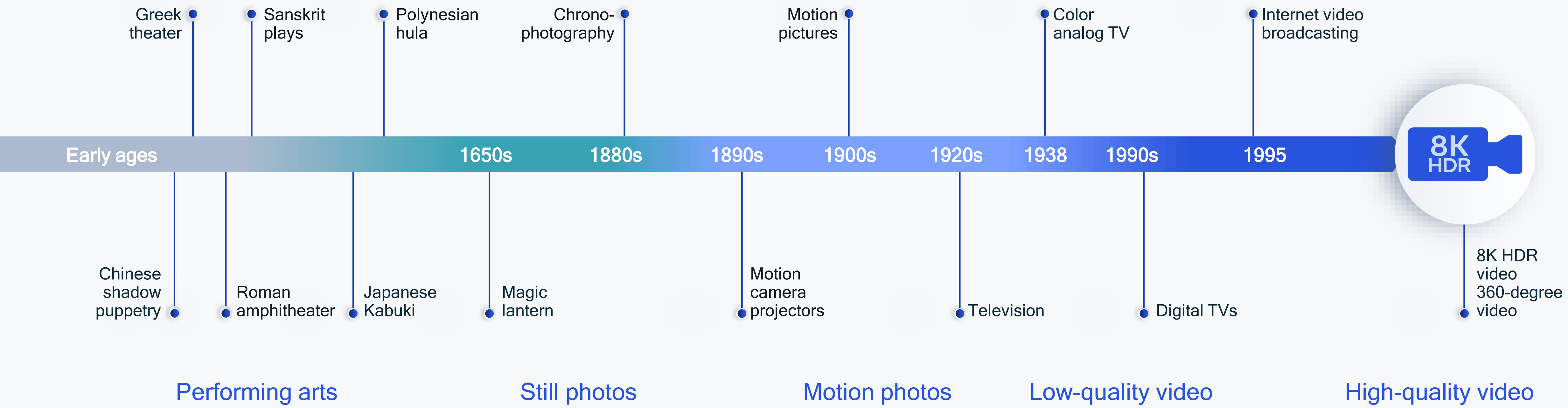
A picture is worth a thousand words

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

A minute of video has more than **1,000** pictures

How video came to be

From performing arts and still photos to high-quality video



The scale of video being created and consumed is massive

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

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

Data challenges

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

Video perception challenges

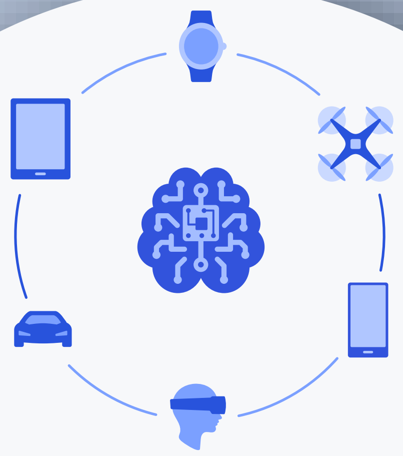
Implementation challenges

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

What makes video perception challenging?

The challenge of AI workloads

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



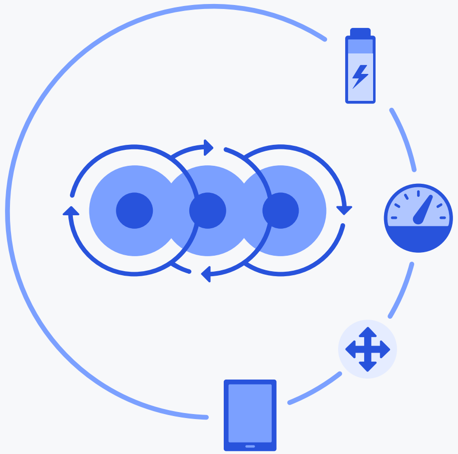
Power and thermal efficiency are essential for on-device video perception

Constrained mobile environment

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

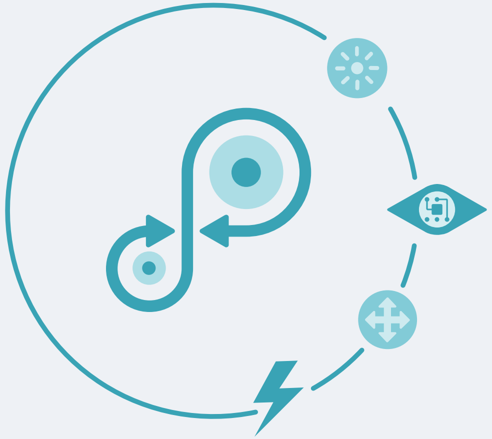
Making video perception ubiquitous

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



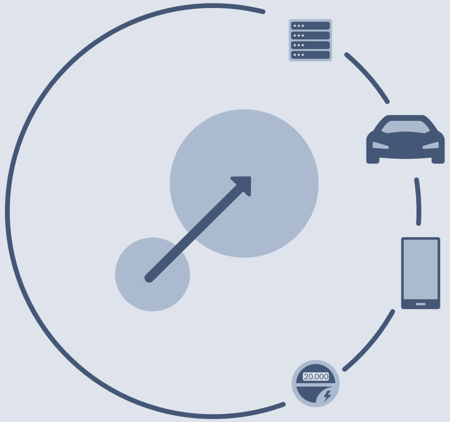
Robustness

Robust to data variations



Adaptability

Adaptable to different domains



Scalability

Scaling up and down, from IoT to the data center

Key concepts for efficient video perception

Leverage
Temporal redundancy

By reusing what is computed before

- Learning to skip regions
- Recycling features

Make
Early decisions

By dynamically changing the network architecture per input frame

- Early exiting
- Frame exiting

Efficiently running on-device video perception without sacrificing accuracy

Deep learning basics

Computable AI

Data driven learning

- Supervised, unsupervised, semi/self/weak supervised, adversarial

Neural network

- Convolutional neural networks
- Graph neural networks

Regression and classification tasks

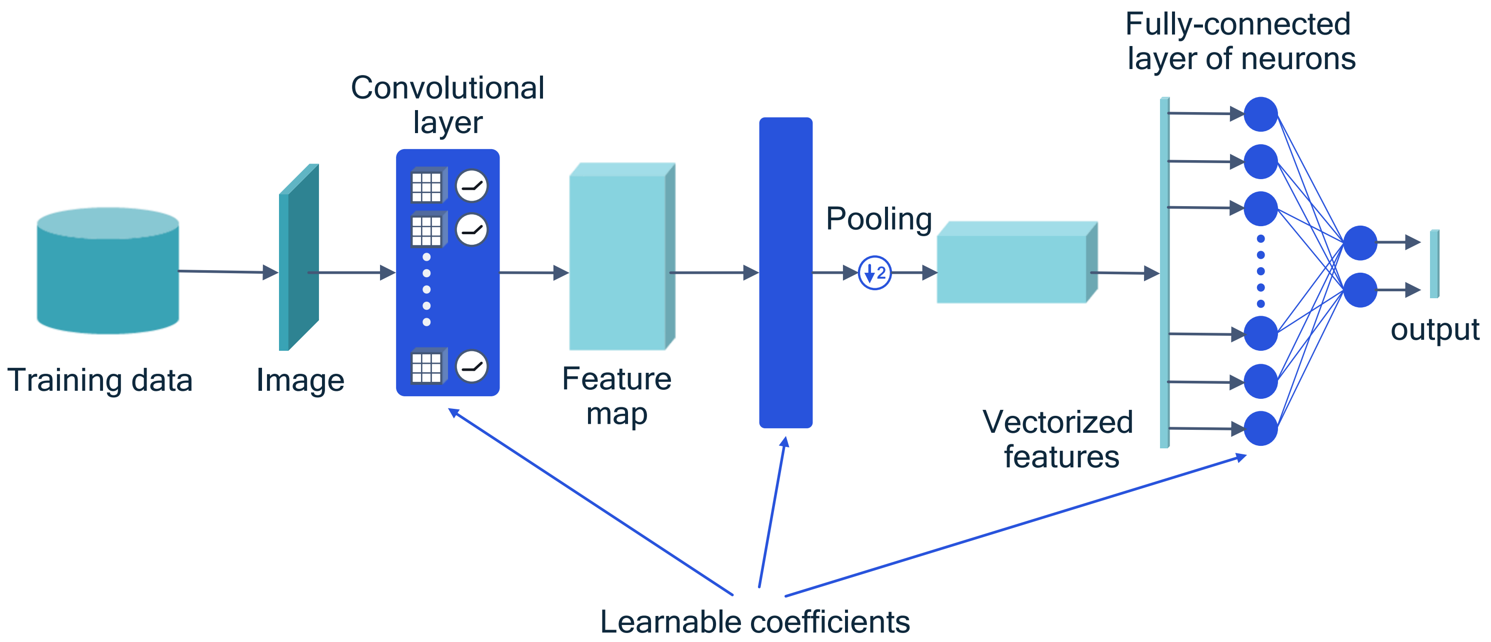
- Minimizing a loss function
- Back-propagation over differentiable layers

Convolutional layers (deep)

Non-linearity, pooling

Kernels, neurons

Feed-forward CNN



Inspired by the workings of the brain, drawing from data

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" (submitted 2021)

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

Skip-convolution

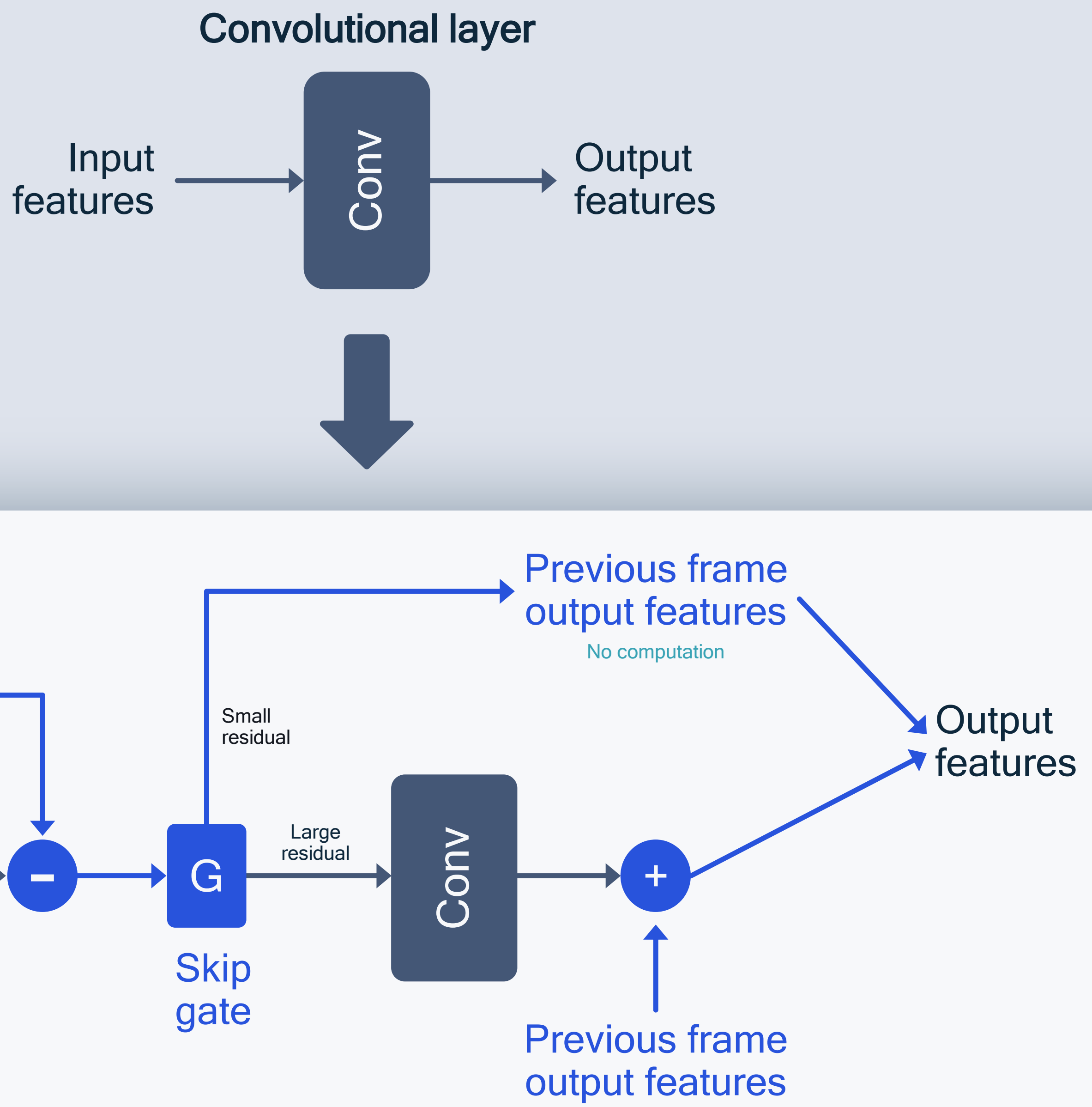
A convolutional layer with a **skip gate** that masks out negligible residuals

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" (submitted 2021)

Determining the gate for a skip convolution

Can we learn it?

Non-trainable

Based on thresholding the norm of the convolved residual

Trainable

Based on a tiny gating network that predicts gate probabilities

Approximate the norm of output without computing it

Apply a sigmoid function on the norm of the residual and the weight of the layer kernel

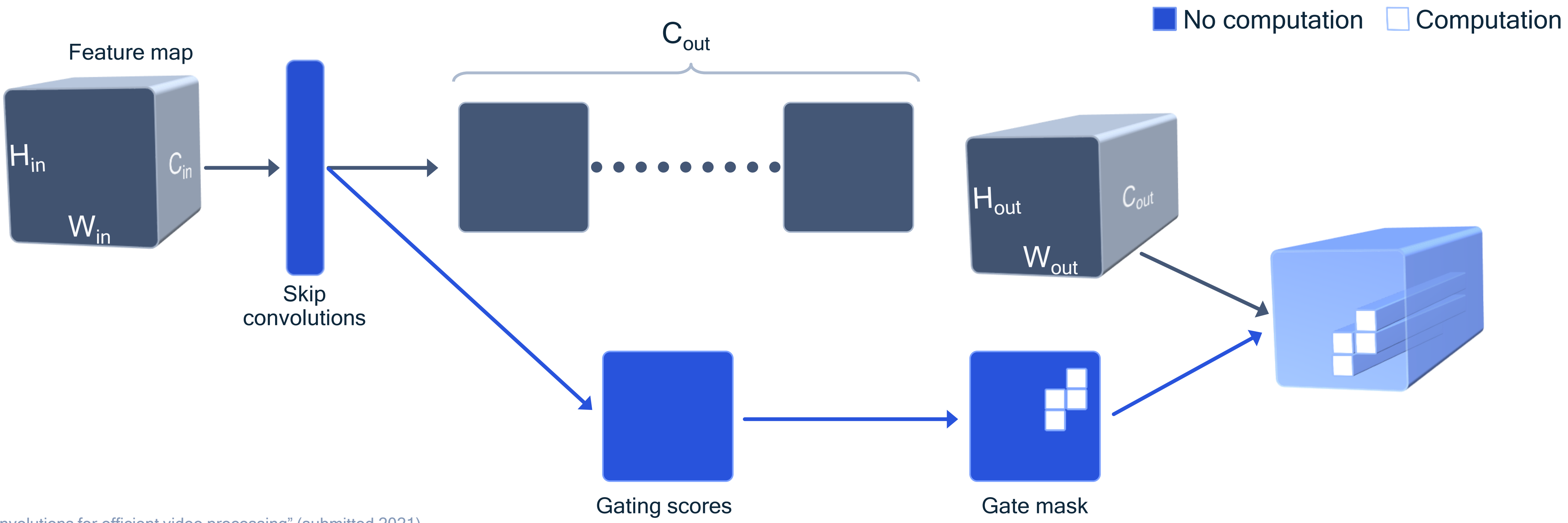
Implement as a convolution with a single output channel

Joint training to minimize the classification loss and average active gates

Apply a sigmoid function to gating network probabilities

Learned gate requires very low computational overhead

Gate network is an additional output channel



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

Hardware friendly implementation enforced by imposing structured sparsity into compute masks

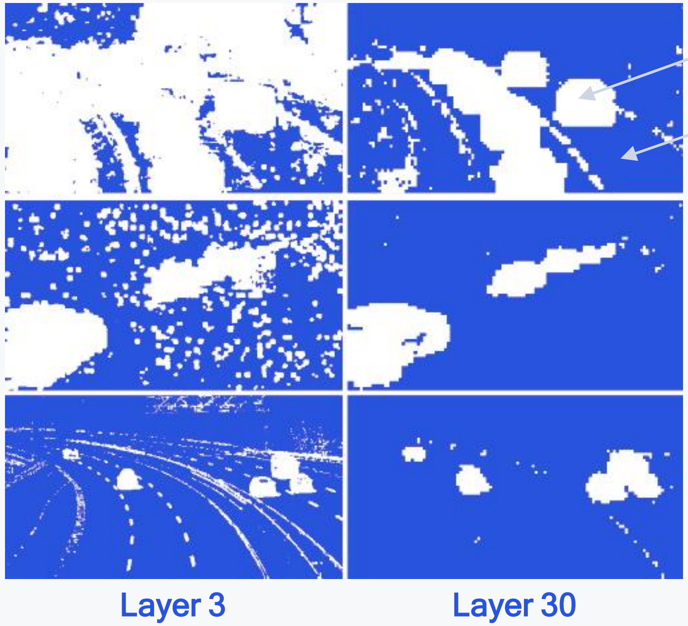
Block-wise structure by down/up sampling

Predicted gating masks reduce computations

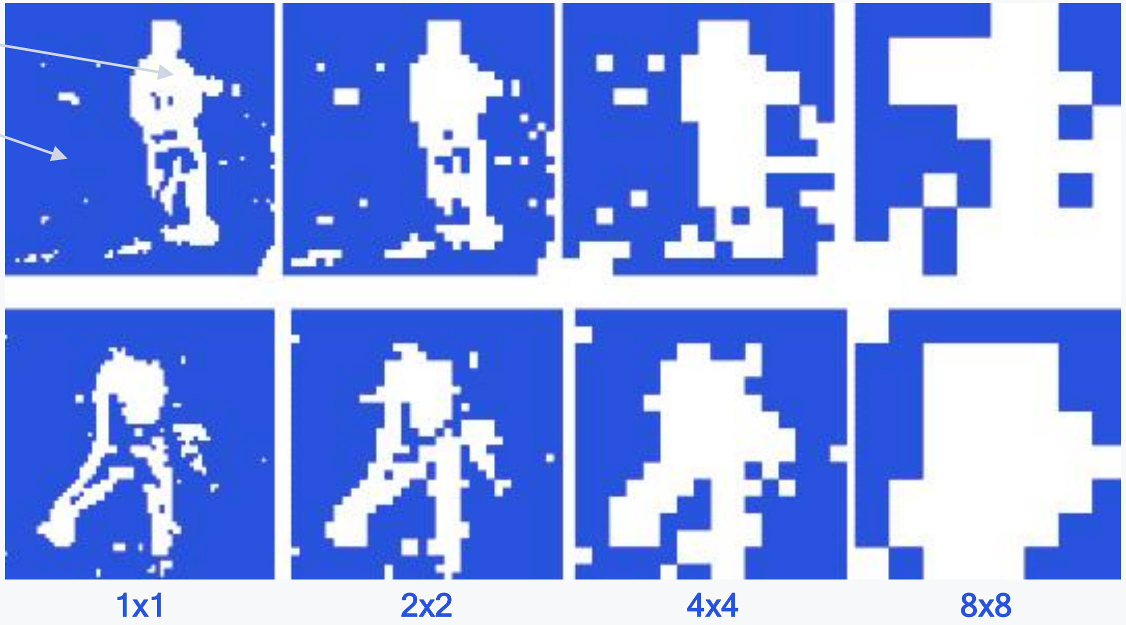
Example masks show where computations can be skipped

Gating masks become more selective at deeper layers, concentrating on task specific regions

Larger block-wise structured gates are more selective when training with different sizes



Gating masks for video object detection

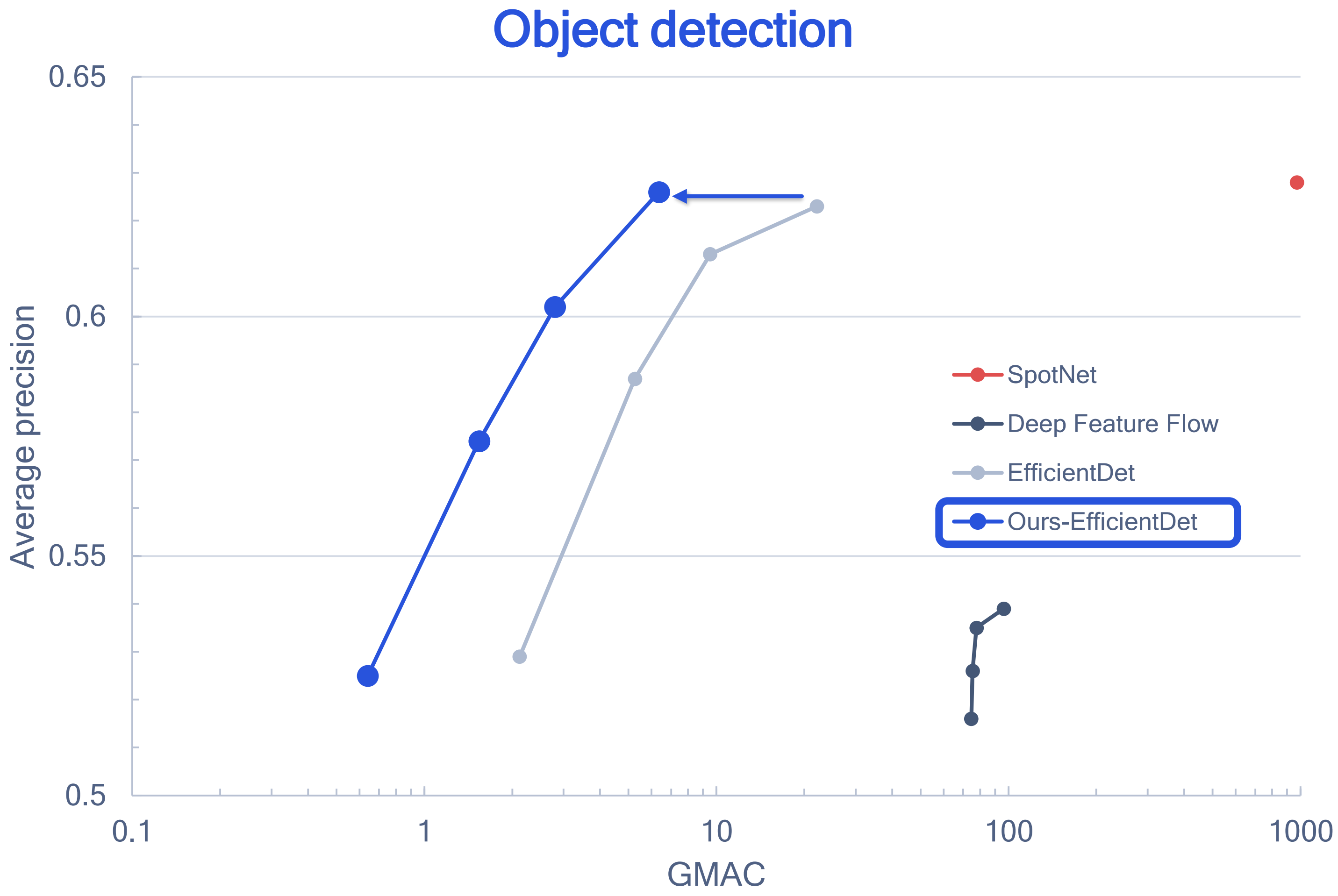


Gating masks for pose estimation

Learning to skip reduces compute and maintains accuracy

Results for object detection on video object detection dataset

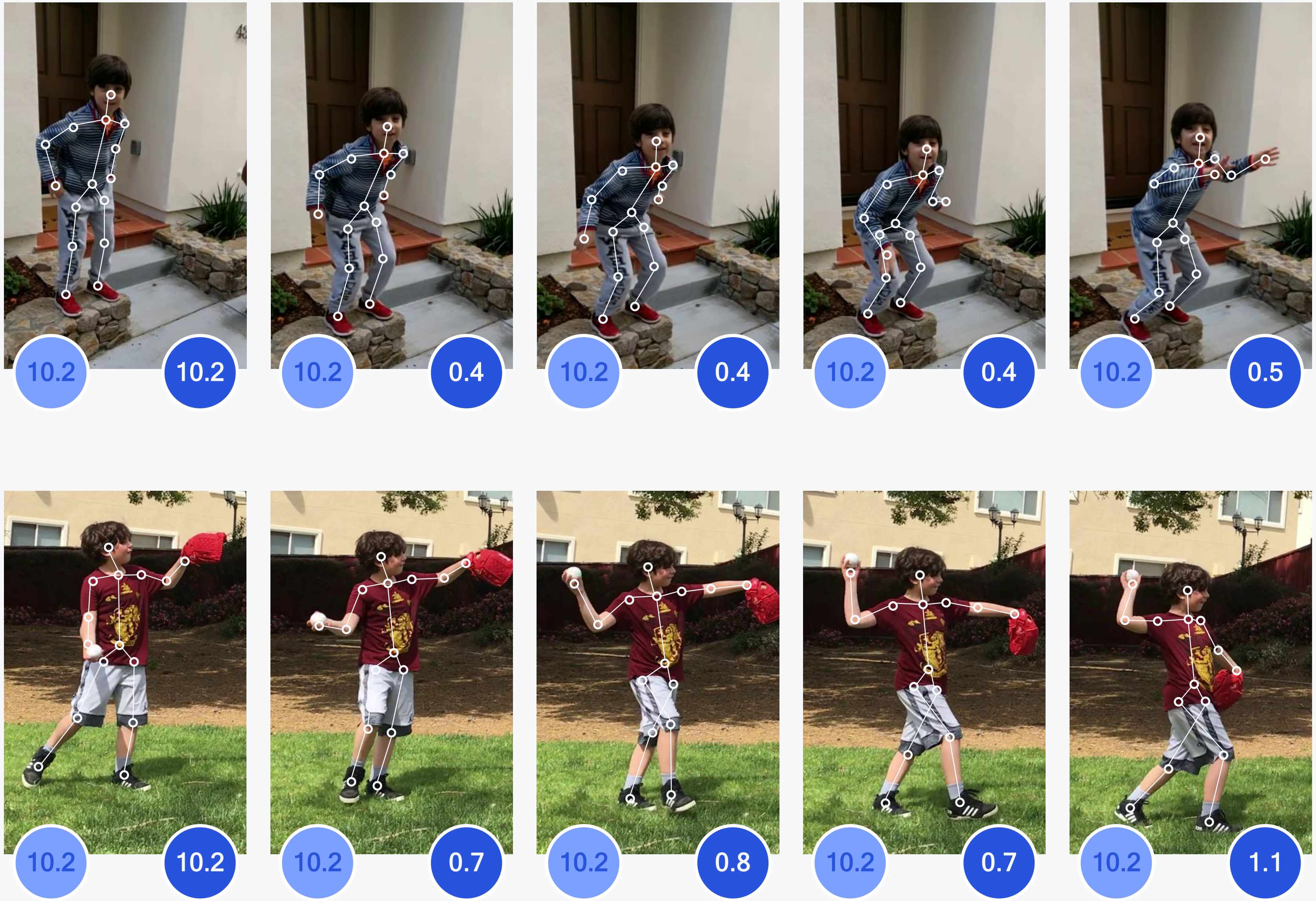
3x-5x speed-up over state-of-the-art



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

Learning to skip reduces compute for human pose estimation

Results for human pose estimation



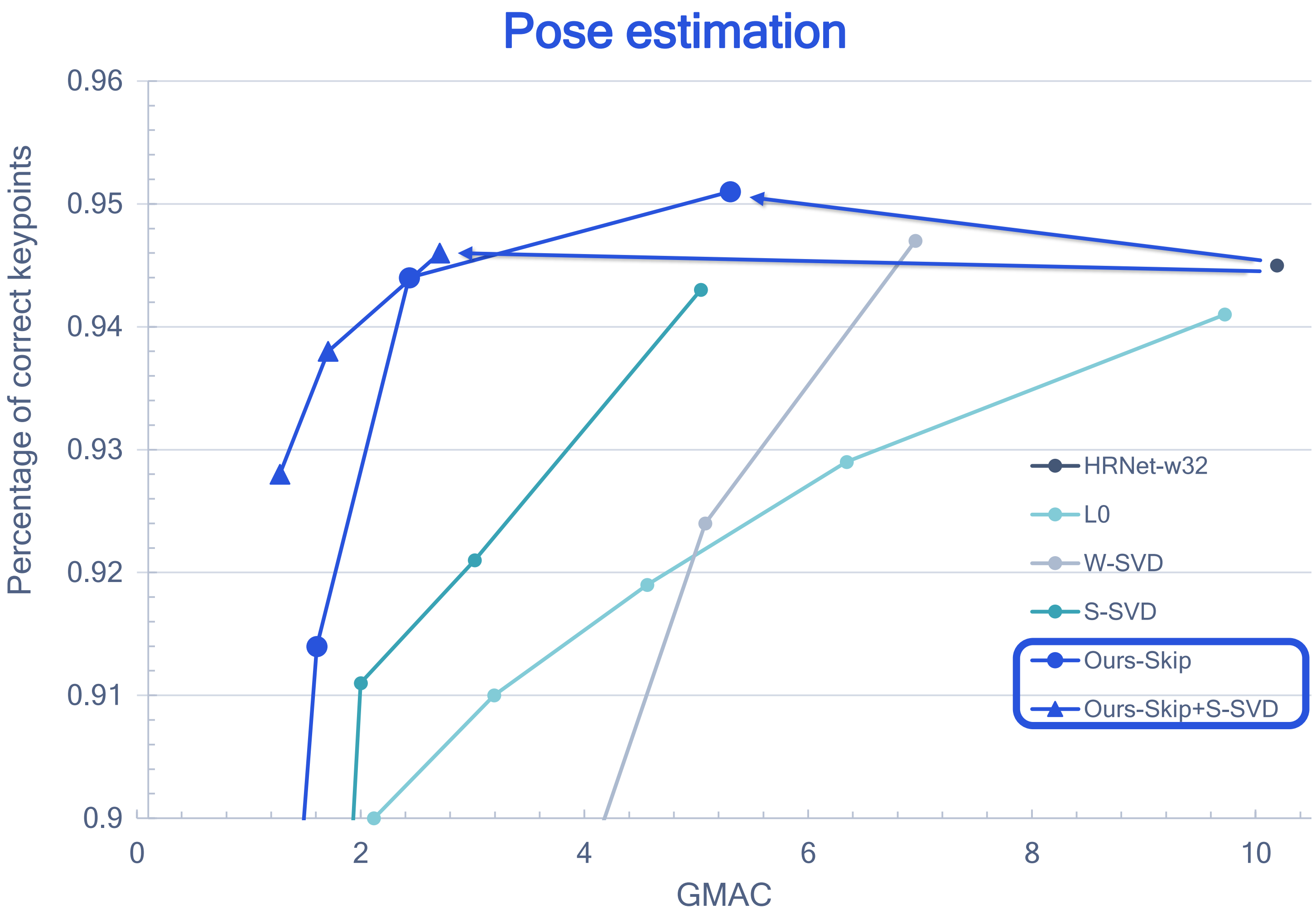
- GMACs **without** skip convolutions
- GMACs **with** skip-convolutions

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

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

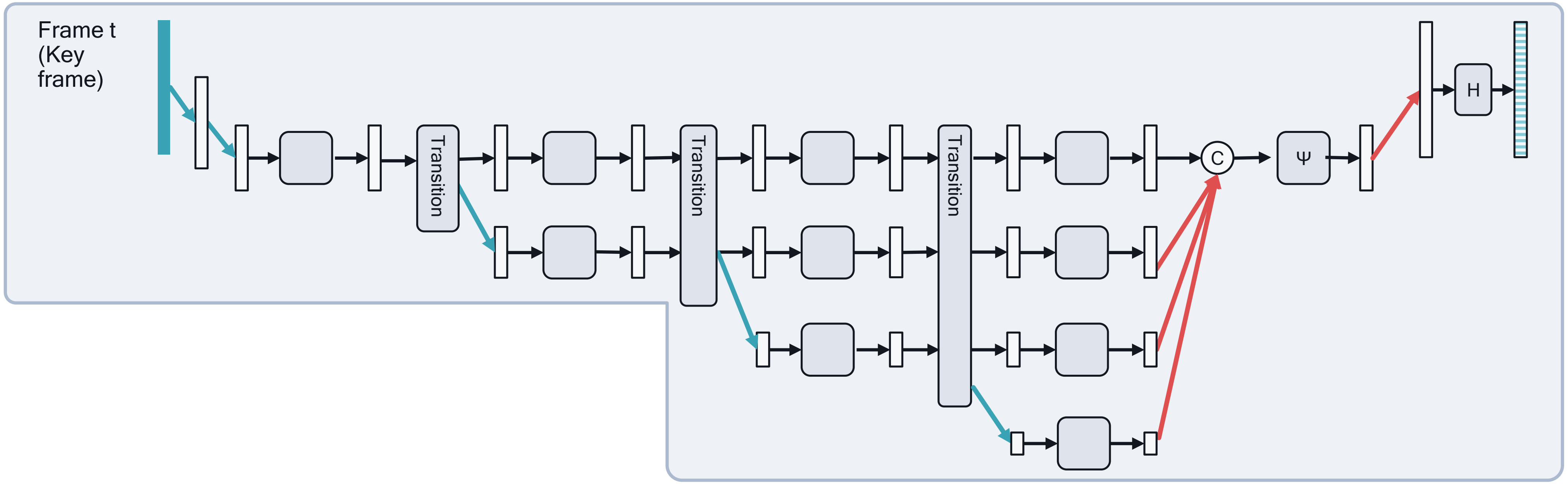
Compute shallow features for all frames

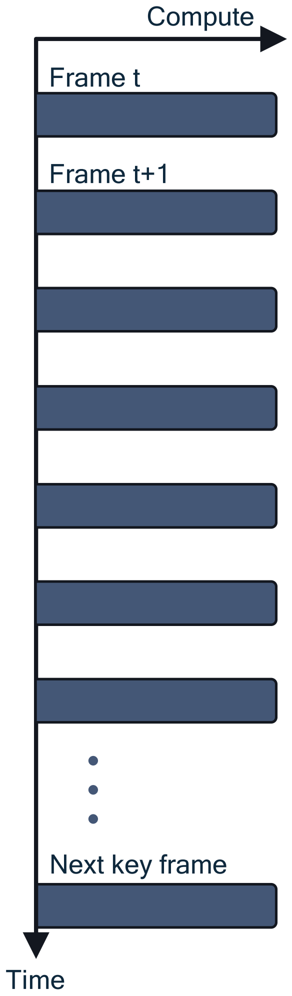
Applicable to any video neural network architectures including segmentation, optical flow, classification, and more

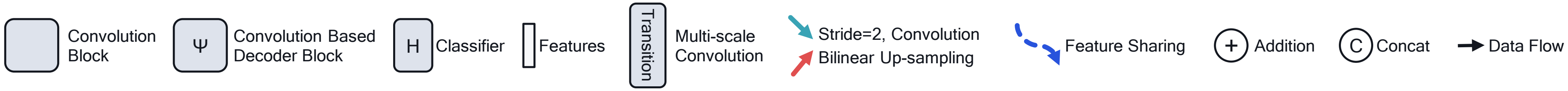
"Time-sharing networks for efficient semantic video segmentation" (submitted 2021)

Example of a feature recycling network

Making the best use of the deep features at hierarchical scales



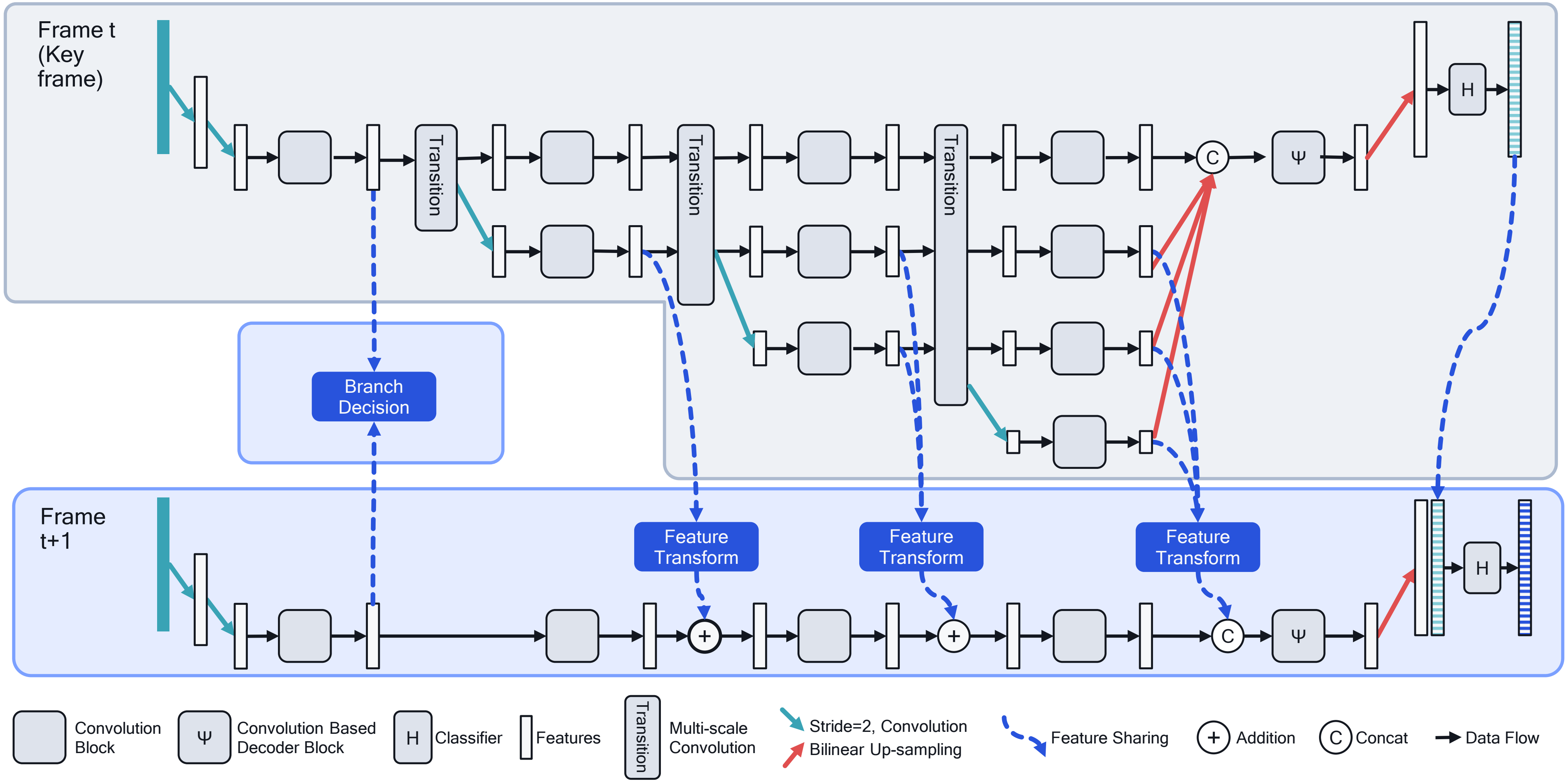


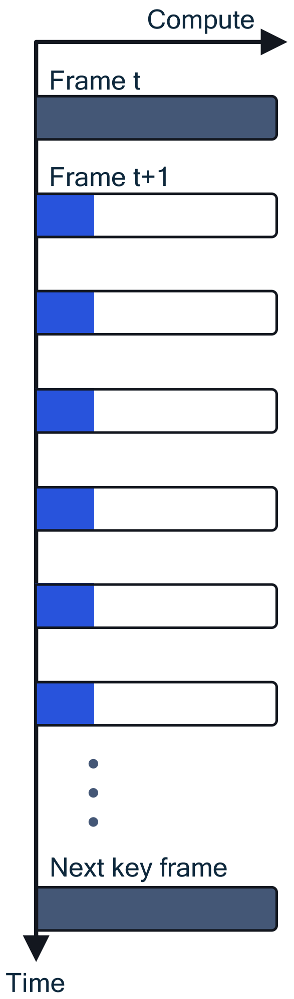


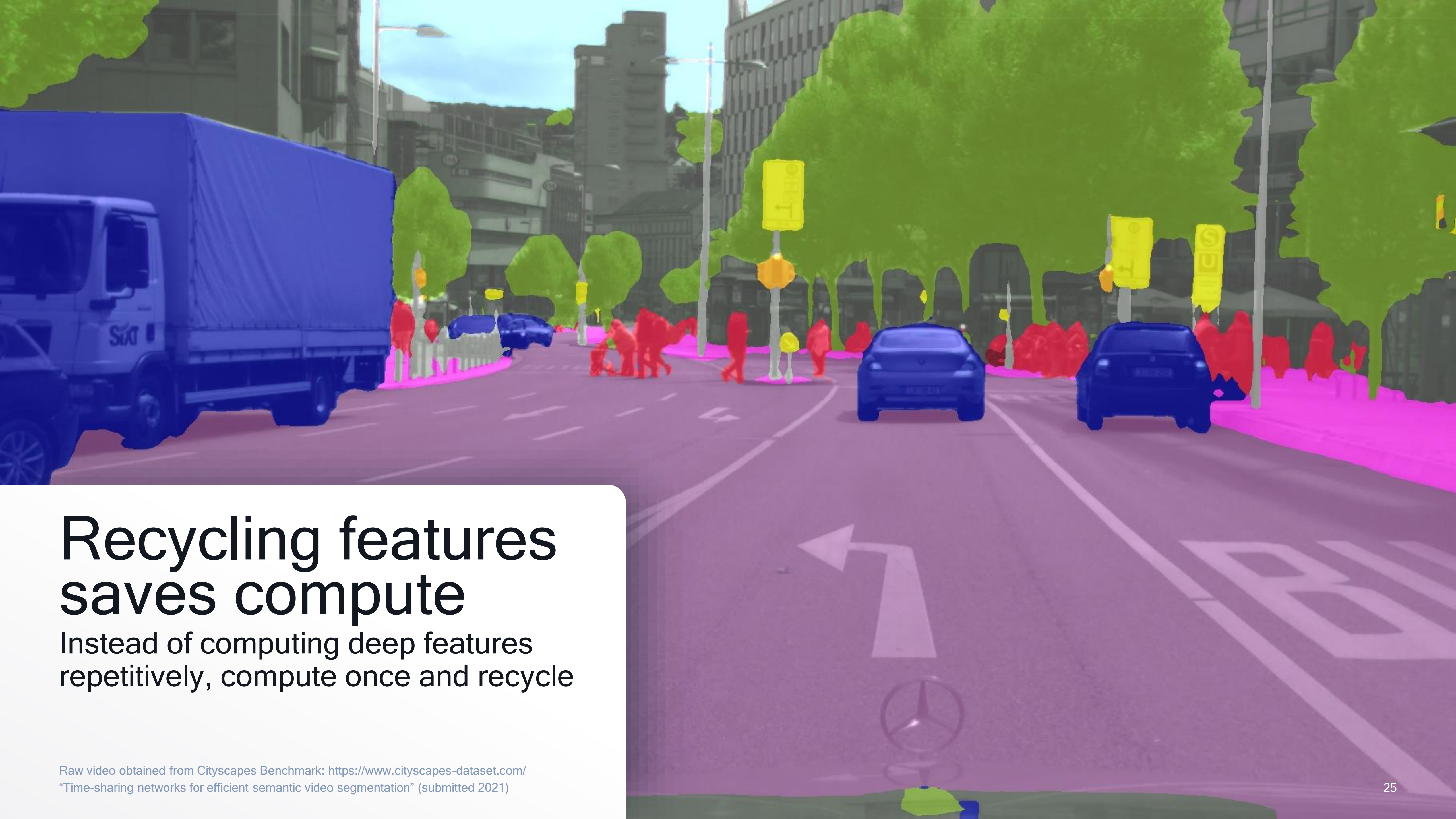
“Time-sharing networks for efficient semantic video segmentation” (submitted 2021)

Example of a feature recycling network

Most frames require a light branch; some frames require a heavy branch







Recycling features saves compute

Instead of computing deep features repetitively, compute once and recycle

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

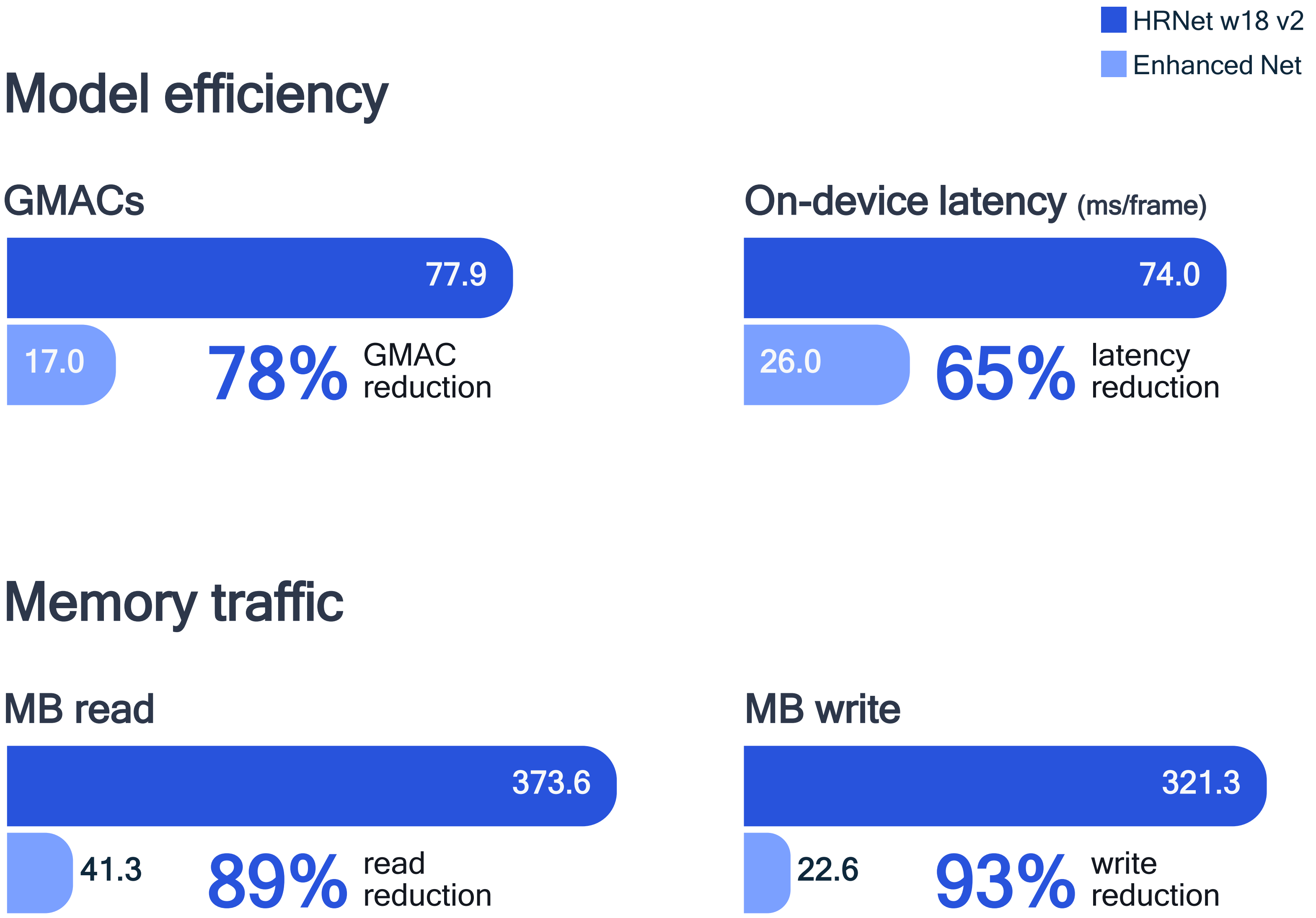
Feature recycling reduces compute and latency

Semantic segmentation example

Input:
2048x1024 RGB video

Output:
2048x1024,
19 object classes

Runs on:
Qualcomm® Snapdragon™ 888 Mobile Platform



“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

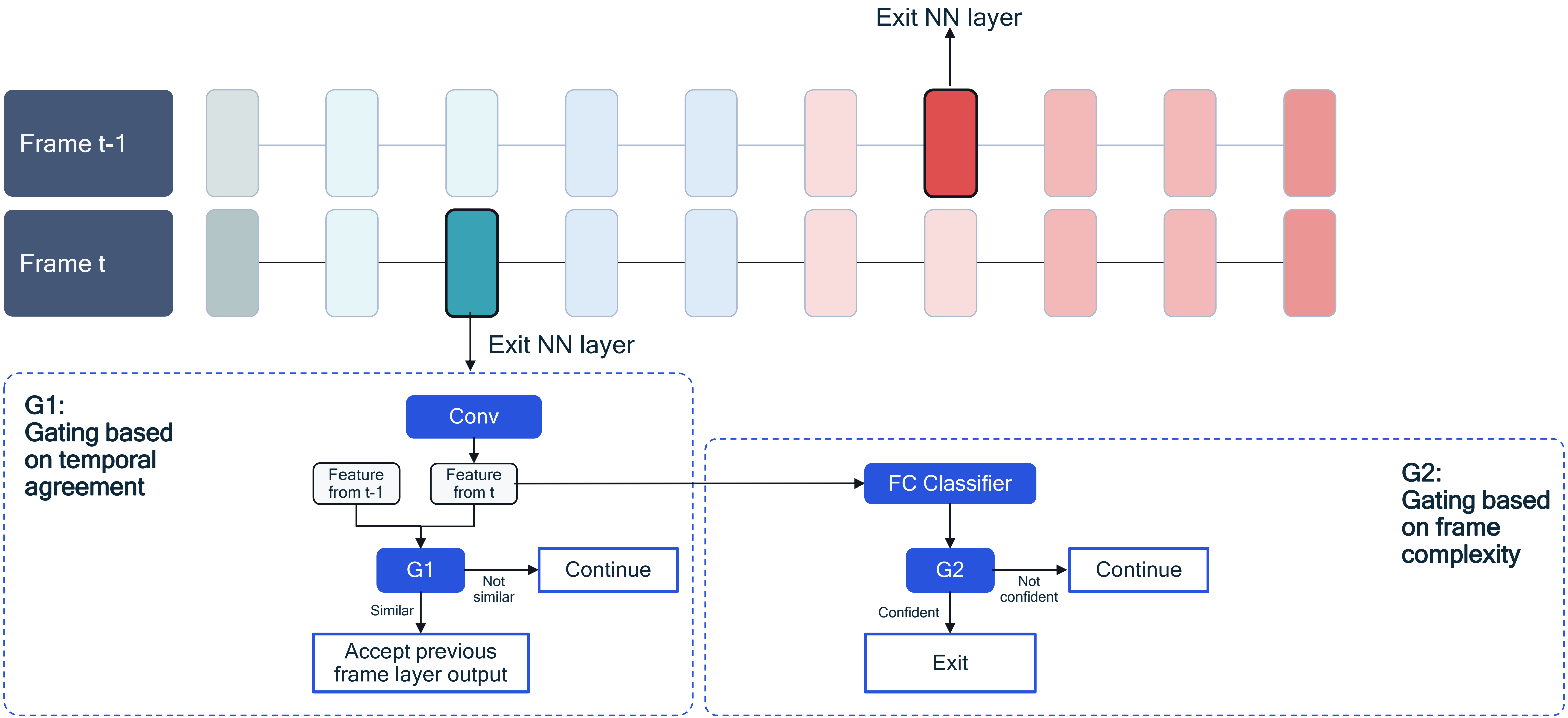
Simple examples

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

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

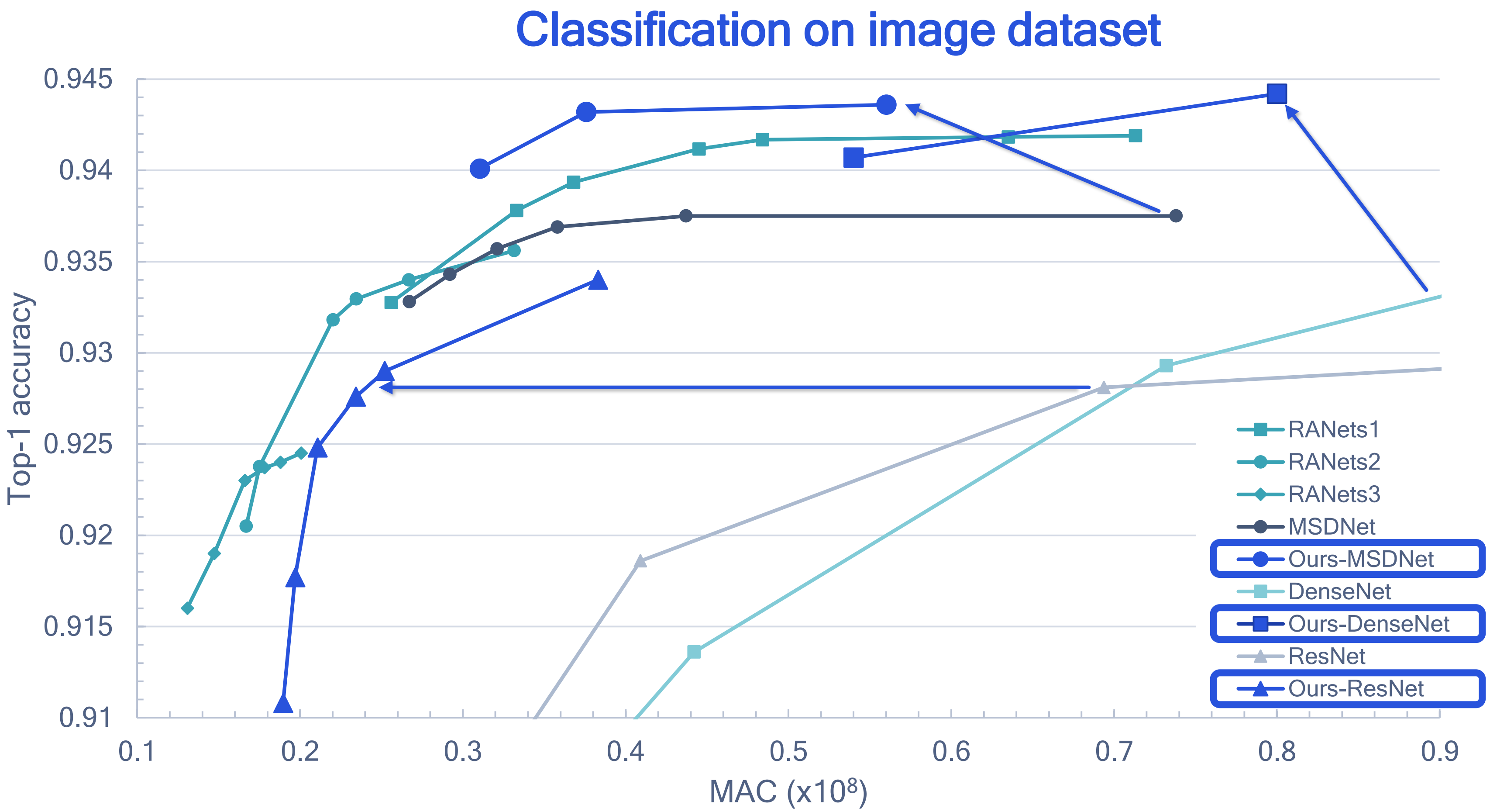
“FrameExit: Conditional early exiting for efficient video recognition” (submitted 2021)

Early exiting at the earliest possible NN layer for video



“FrameExit: Conditional early exiting for efficient video recognition” (submitted 2021)

Early exiting applies to most neural network backbones



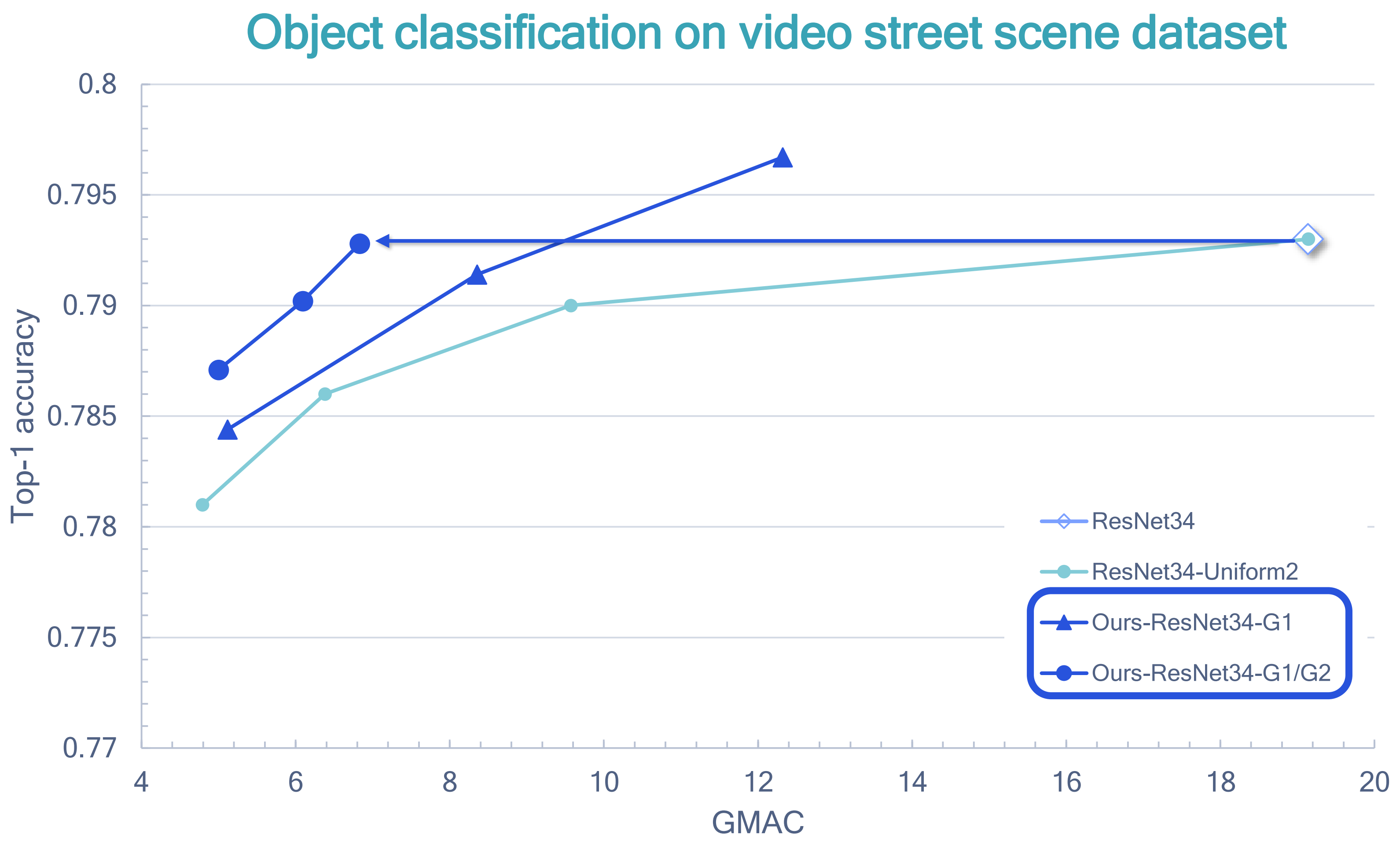
Early exiting reduces compute while maintaining accuracy

“FrameExit: Conditional early exiting for efficient video recognition” (submitted 2021)

Early exiting for object classification

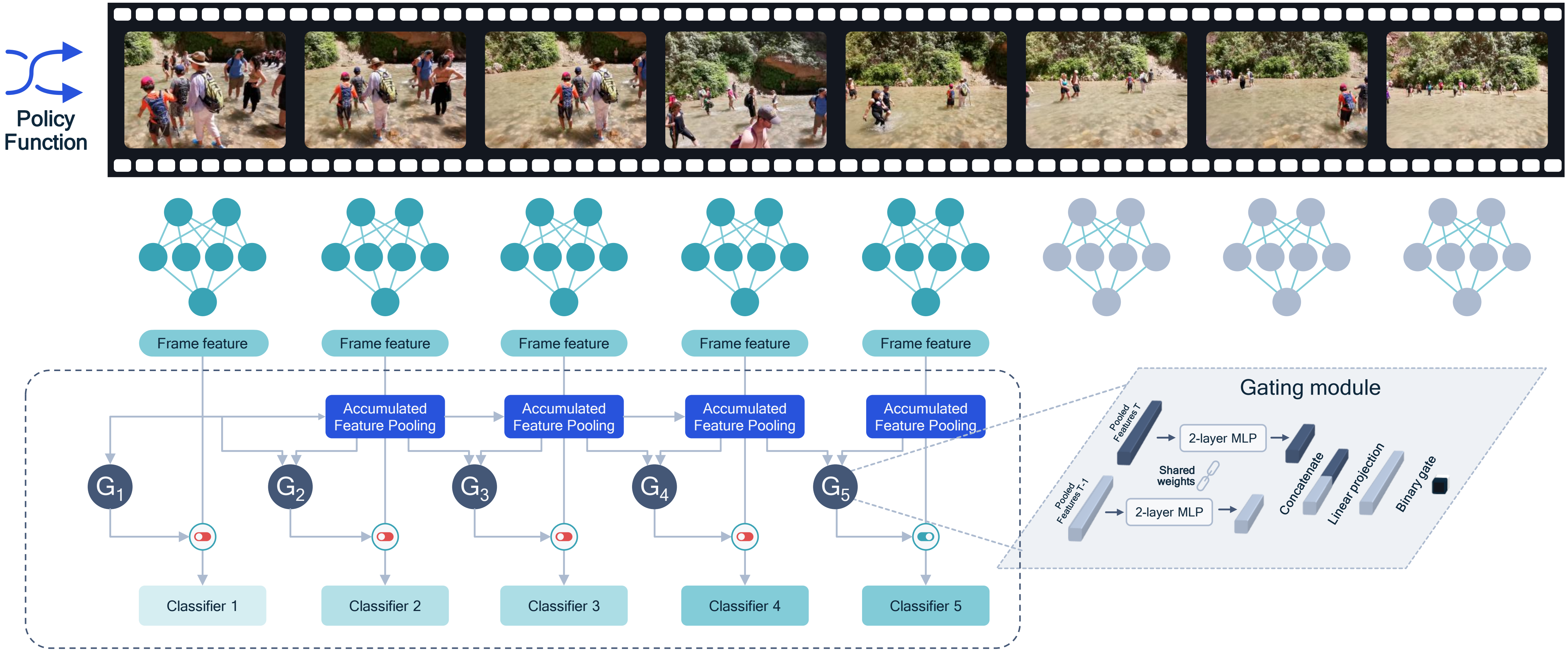
2.5x

less MACs while maintaining accuracy



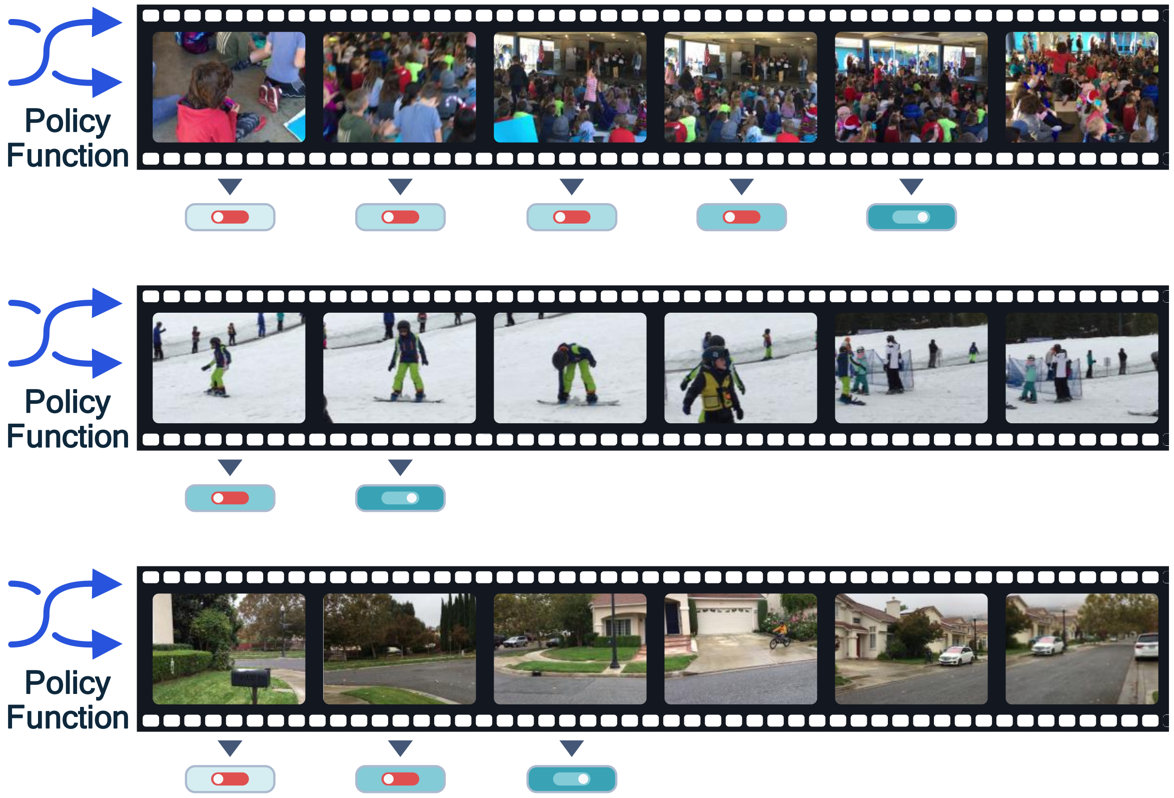
"FrameExit: Conditional early exiting for efficient video recognition" (submitted 2021)

Frame exiting also applies to action recognition tasks



“FrameExit: Conditional early exiting for efficient video recognition” (submitted 2021)

Frame exiting improves accuracy and reduces compute



Methods	Video activity dataset		Video action dataset	
	mAP (%)	GFLOPS	Top-1 (%)	GFLOPS
Resnet				
AdaFrame	71.5	79.0	–	–
LiteEval	72.7	95.1	61.0	99.0
ListenToLook	72.3	81.4	–	–
SCSampler	72.9	41.9	70.8	41.9
AR-Net	73.8	33.5	71.7	32.0
★ FrameExit	76.1	25.2	72.8	19.7
EfficientNet				
AR-Net	79.7	15.3	74.8	16.3
★ FrameExit	80.0	11.4	75.3	7.8

	mAP (%)	GFLOPS
2d/3d Resnet	Video human action dataset	
Uniform-10	44.7	41.2
Random-10	43.6	41.2
3D-ResNet18	35.4	38.6*
HATNet	39.6	41.8*
★ FrameExit (Efficient)	45.7	8.6
★ FrameExit (Accurate)	49.2	18.7

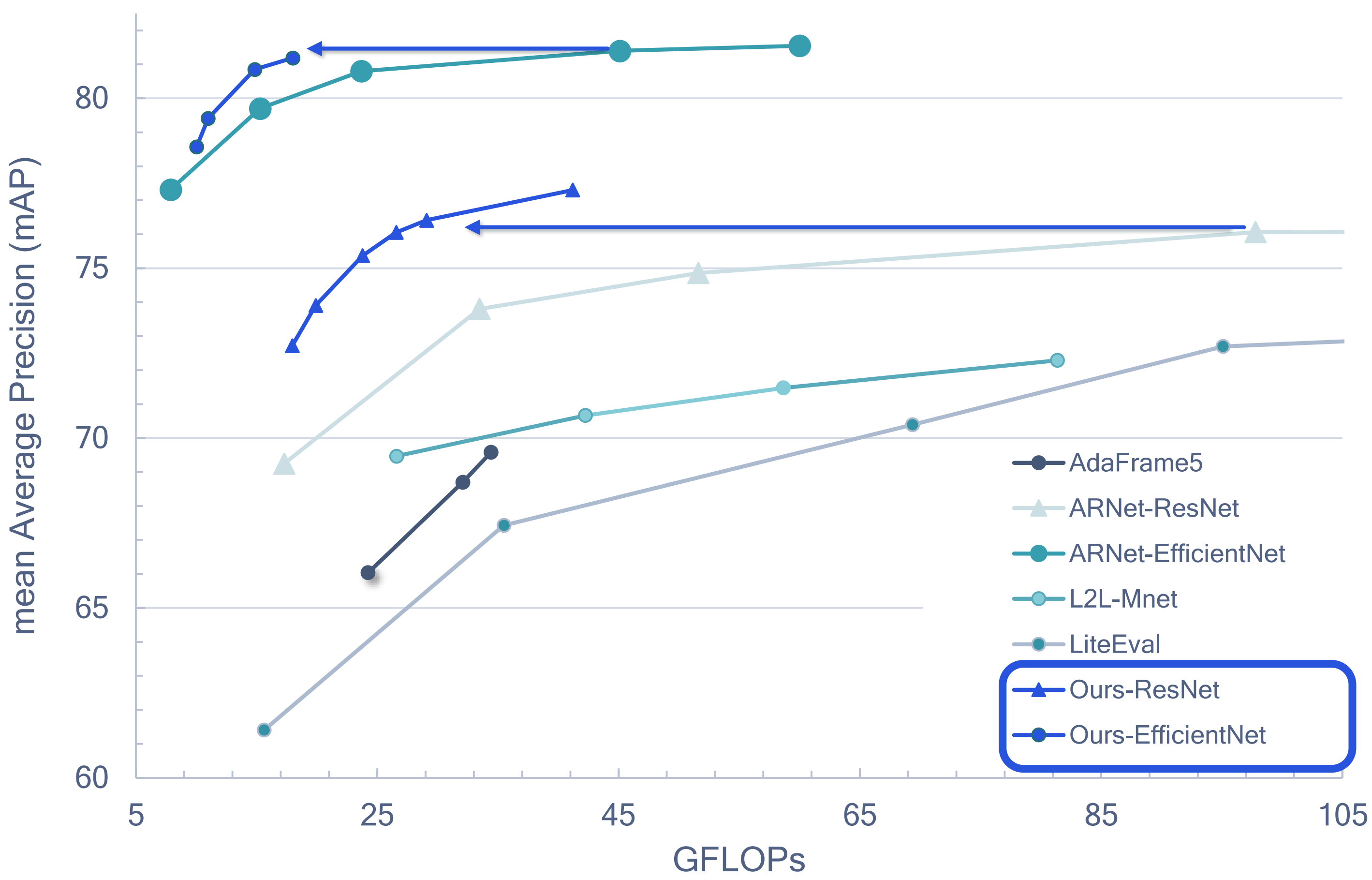
By adding gates to the NN architecture, deeper layers concentrate on the difficult decisions while earlier layers solve all the easy issues

Frame exiting for video classification

1.3x-5x

less GFLOPs while maintaining accuracy

Video classification on video activity dataset



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

Future work in video perception

Advance existing conditional compute techniques

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

Develop efficient video neural network solutions

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

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

Our perception research is much broader than video

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

Questions?

Connect with Us

www.qualcomm.com/ai

www.qualcomm.com/news/onq

@QCOMResearch

https://www.youtube.com/qualcomm?

http://www.slideshare.net/qualcommwirelessevolution

Thank you

Follow us on:

For more information, visit us at:

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