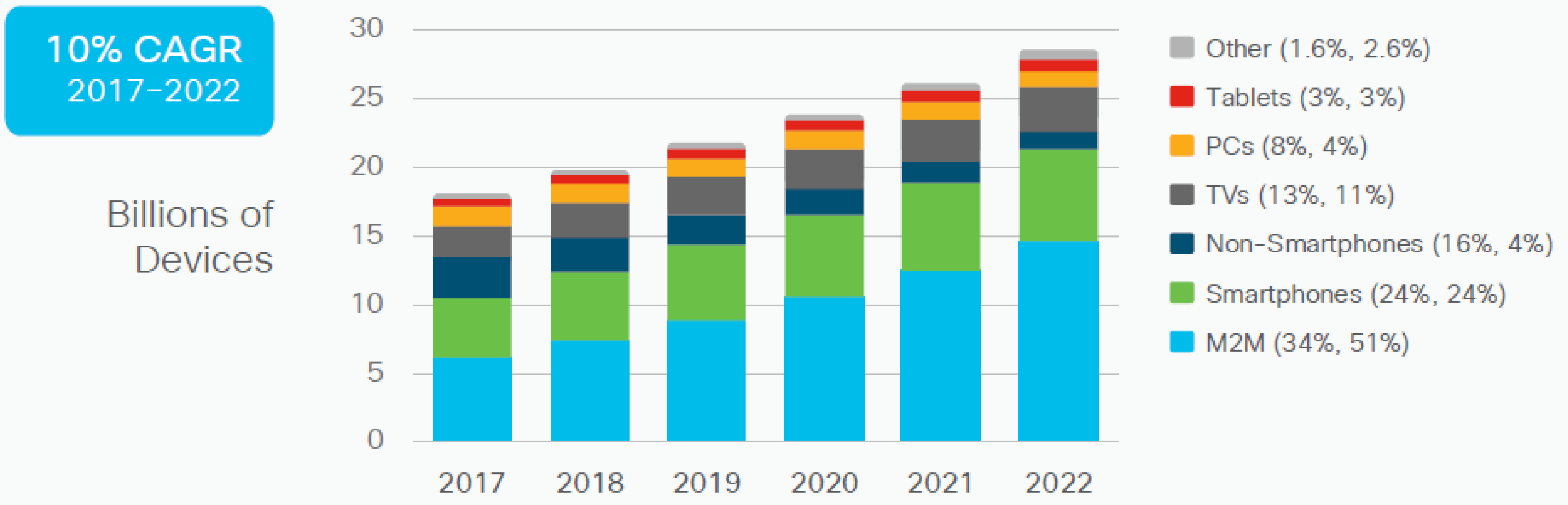
On Intra Video Coding and In-Loop Filtering for Neural Object Detection Networks

Kristian Fischer, Christian Herglotz, André Kaup
kristian.fischer@fau.de
Chair of Multimedia Communications
and Signal Processing

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Why do we need Video Coding for Machines (VCM)?

Global devices and connections growth



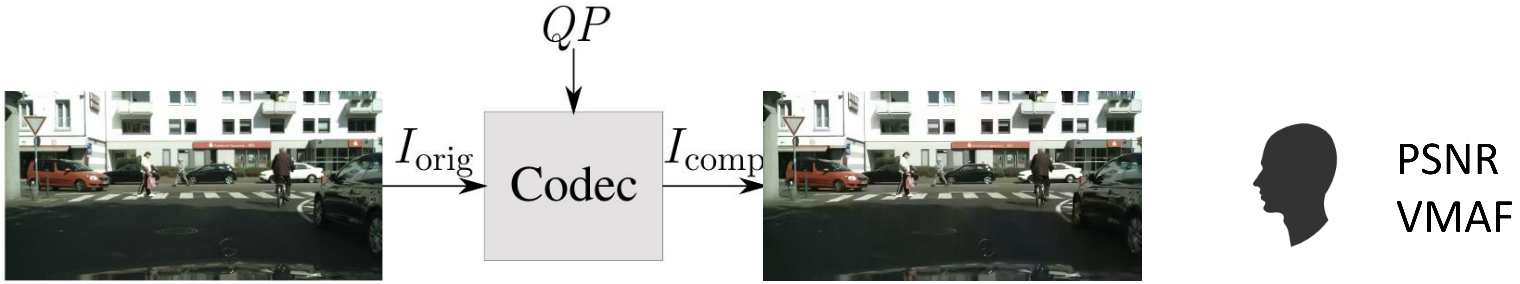
➔ Suitable compression for machine-to-machine (M2M) communication required

Image credit: Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022," Tech. Rep., Feb. 2019.

Outline

- General setup
  - Dataset
  - Used neural object detection networks
  - Evaluation metric
  - Coding framework
- Comparison between High Efficiency Video Coding (HEVC) and successor Versatile Video Coding (VVC) for VCM scenario
- Evaluating impact of VVC in-loop filtering for VCM scenario

Signal Flow Video Coding for Humans

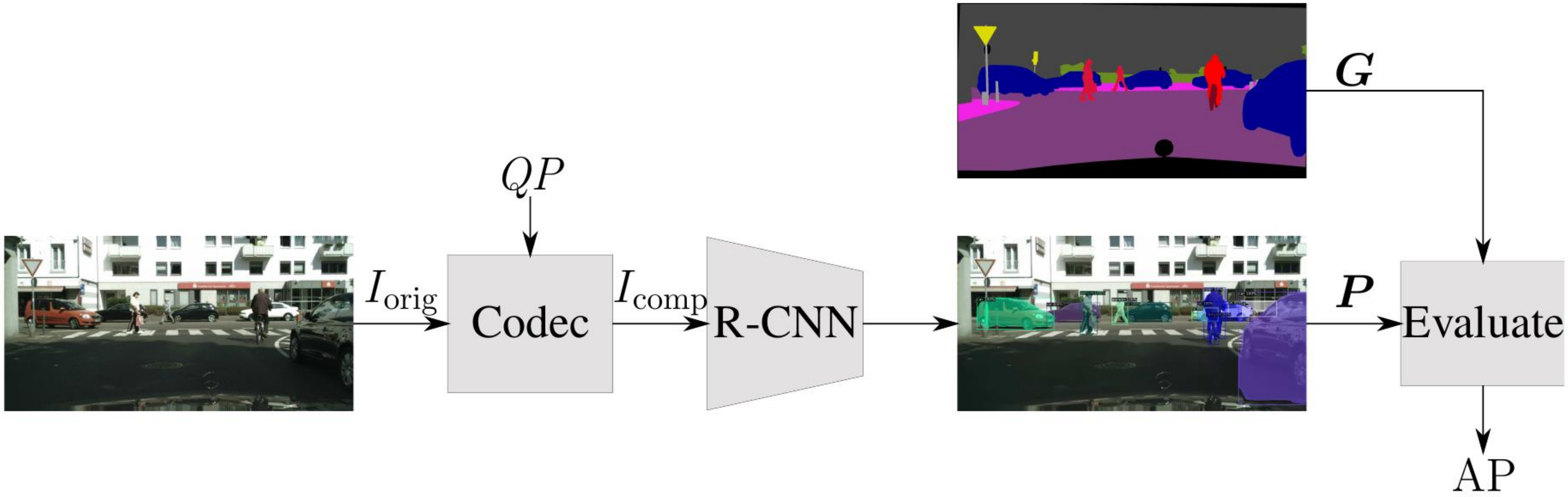


I_{orig}: Input image
I_{comp}: Compressed image
QP: Quantization parameter

PSNR: Peak-signal-to-noise ratio
VMAF: Video multi-method assesment fusion

VMAF: Netflix Inc., “VMAF – video multi-method assessment fusion,” https://github.com/Netflix/vmaf, 2016.

Signal Flow Video Coding for Machines



I_{orig}: Input image
I_{comp}: Compressed image
QP: Quantization parameter

R-CNN: Region-based convolutional neural network
P: Predicted objects
G: Ground-truth objects
AP: Average precision

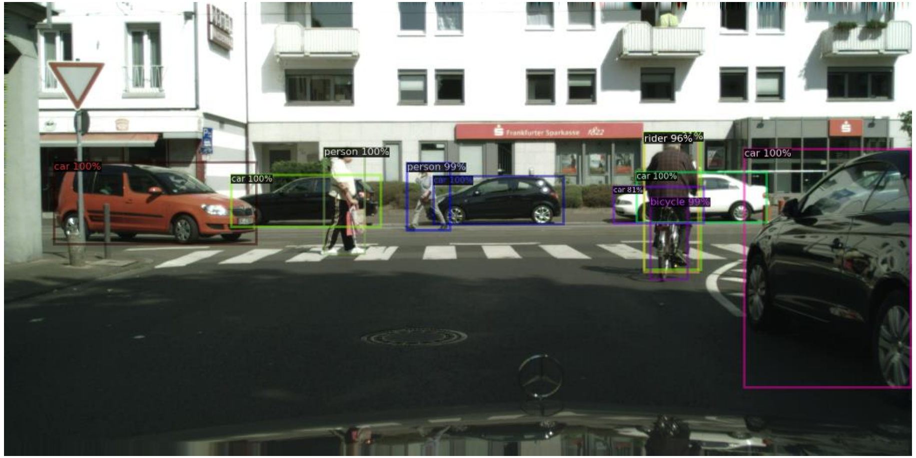
Cityscapes Dataset

- Stereo data observing urban scenes
- 5000 uncompressed images
- 2048x1024 pixels
- Pixel-wise labeled data for object detection and segmentation
- 8 object categories considered like car, person, truck, etc.

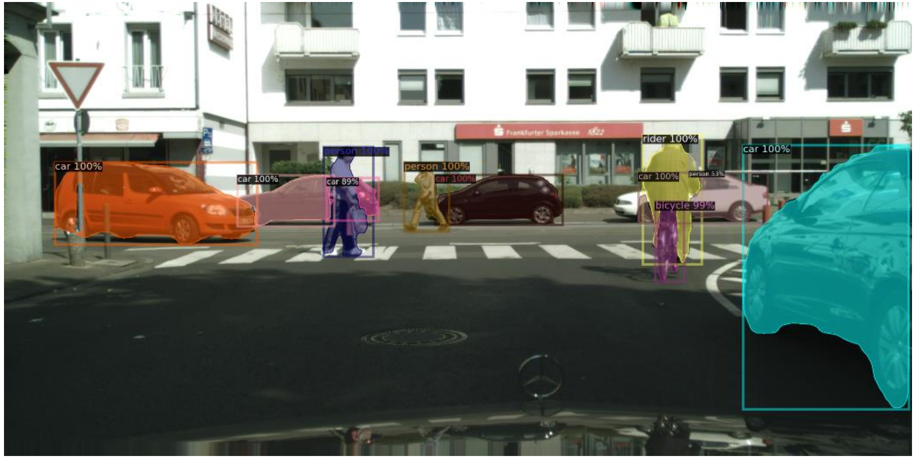
Cityscapes: Cordts et al., “The Cityscapes Dataset for Sementic Urban Scene Understanding,” *CVPR*, 2016.

Investigated Object Detection R-CNNs

Faster R-CNN



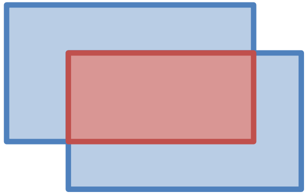
Mask R-CNN

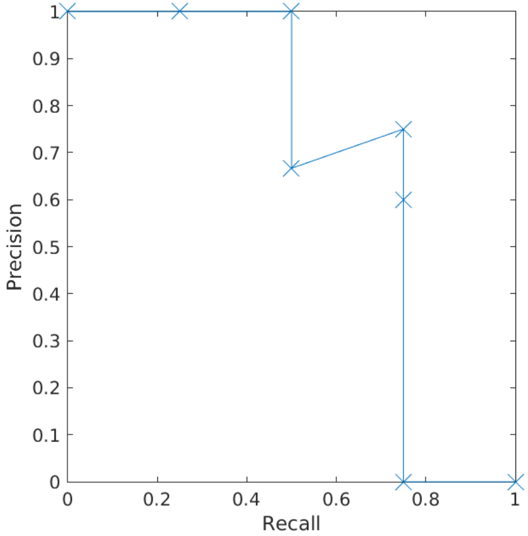


Faster R-CNN: Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *TPAMI*, 2017.
Mask R-CNN: He et al., "Mask R-CNN," *ICCV*, 2017.

Mean Average Precision (mAP)

- Metric to evaluate accuracy of object detection
- mAP used as proposed for Cityscapes challenge
- Considers precision and recall for certain intersection over union (IoU) thresholds
- mAP is the mean over the AP of each class
- Modification: mAP is calculated as weighted average depending on the number of instances of each class in the dataset





Cityscapes Scripts: Cordts et al., "The cityscapes dataset," https://github.com/mcordts/cityscapesScripts, 2017.

Coding Framework

- HEVC test model (HM-16.2)
- VVC test model (VTM-6.0)
- QP from 2 to 47 in steps of 5
- All-intra configuration
- Coded 500 images from Cityscapes validation set
- Transformation from RGB to YCbCr 4:2:0 and vice versa with Ffmpeg

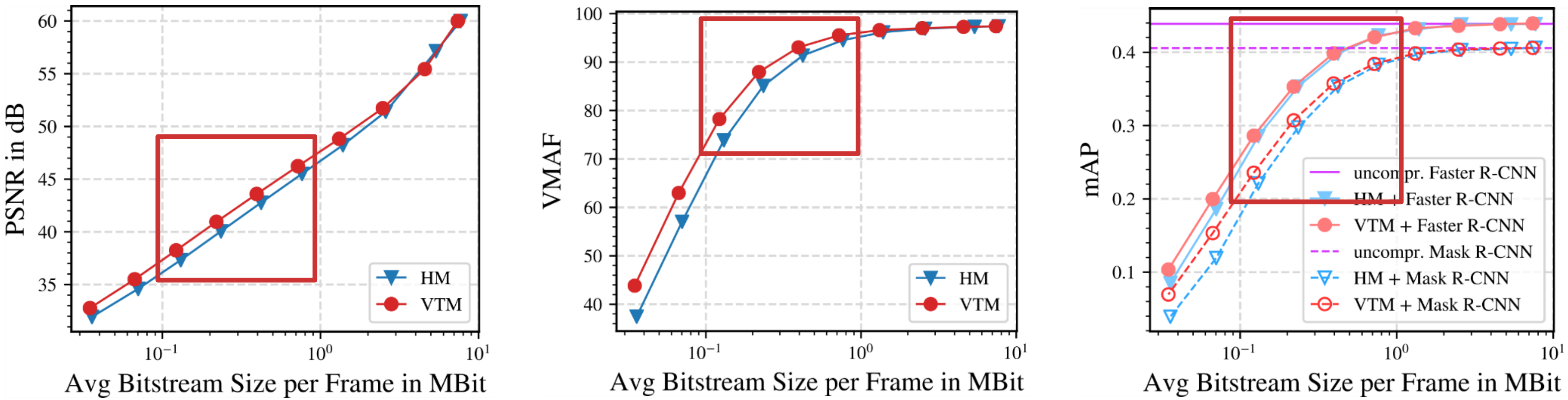
HEVC: Sullivan et al., “Overview of the high efficiency video coding (HEVC) standard,” *TCSVT*, 2012.
VVC: Chen et al., “JVET-O2002: Algorithm description for versatile video coding and test model 6 (VTM 6),” *Tech. Rep.*, 2019.

Used Faster and Mask R-CNN Models

- General Settings
  - PyTorch implementations from Detectron2 library
  - Residual net with 50 layers and feature pyramid network as backbone
- **Mask R-CNN**: Pre-trained model on Cityscapes training set from Detectron2 library
- **Faster R-CNN**: Pre-trained model on COCO dataset taken and further trained with Cityscapes training set for 42000 iterations

Detectron 2: Wu et al., “Detectron2,” https://github.com/facebookresearch/detectron2, 2019.

Comparison HEVC vs. VVC



Bjøntegaard Delta Rate (BDR) Savings in %	PSNR	VMAF	mAP
Faster R-CNN	-22.17	-25.55	-6.01
Mask R-CNN			-13.56

QP = {22, 27, 32, 37}
Anchor: HM

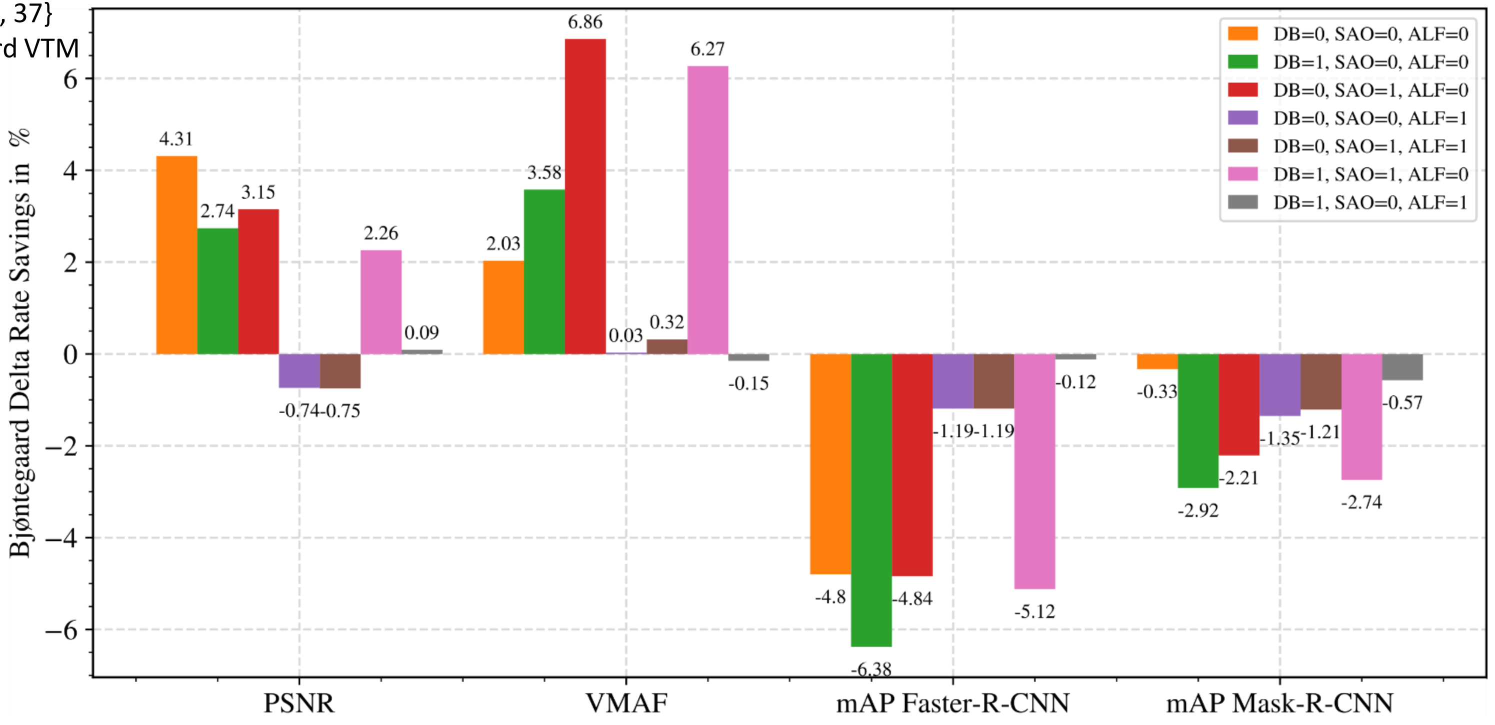
In-Loop Filters in VVC

- De-blocking filter (DB)
  - Minimizing block artifacts
- Sample adaptive offset filter (SAO)
  - Transmitting offset values depending on pixel category
- Adaptive loop filter (ALF)
  - Convolving output with suitable filter
- Standard VVC: All in-loop filters activated

DB: Norkin et al., “HEVC deblocking filter,” *TCSVT*, 2012.
SAO: Fu et al., “Sample adaptive offset in the HEVC standard,” *TCSVT*, 2012.
ALF: Tsai et al., “Adaptive loop filtering for video coding,” *JSTSP*, 2013.

Influence of In-loop Filters on R-CNNs

QP = {22, 27, 32, 37}
Anchor: Standard VTM



Losses compared to standard VTM

Savings compared to standard VTM

Conclusions

- Coding gains for VCM significantly smaller than for human visual system
  - Above 22 and 25% BDR savings of VVC over HEVC for PSNR and VMAF, respectively
  - Only 5 to 14% BDR savings for VCM use case with Faster and Mask R-CNN
- SAO and ALF not beneficial for VCM scenario
  - 6% BDR can be saved when deactivating SAO and ALF
  - Only DB filter improves the coding efficiency for VCM scenarios

➔ New VVC optimizations have to be found for VCM when coding for neural networks