Video Coding for Deep Learning-Based Machine-to-Machine Communication in an Autonomous Driving Scenario

Kristian Fischer
kristian.fischer@fau.de
Chair of Multimedia Communications
and Signal Processing

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Agenda

1. Motivation Video Coding for Machines
2. Evaluation Framework
3. Classic Coding Performance vs. Video Coding for Machines
4. Feature-based Rate-Distortion Optimization
5. Conclusions

Common Video Coding Scenario

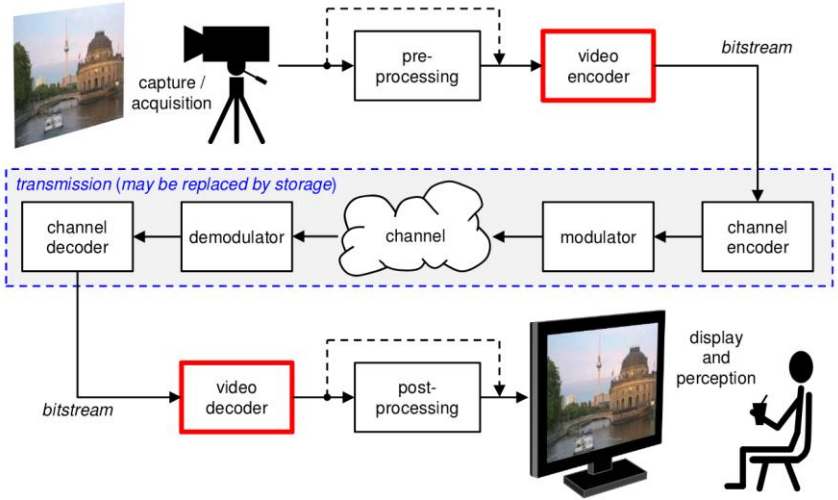
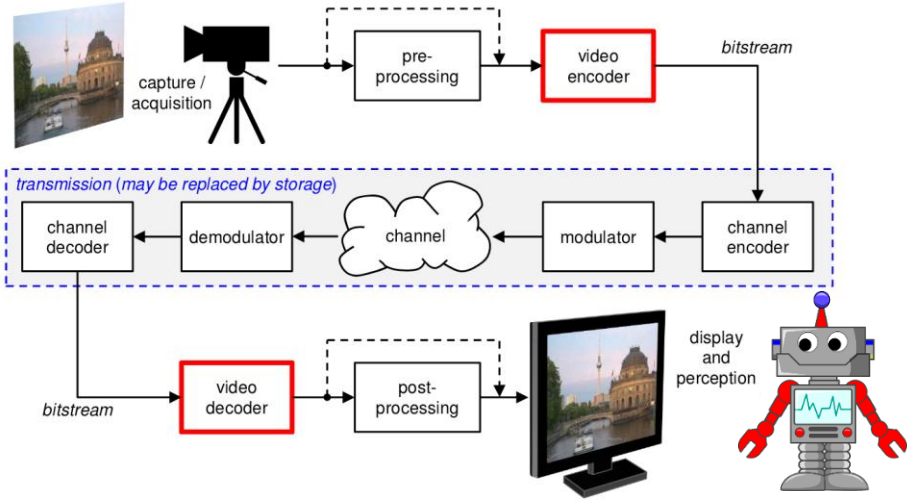


Image Credit: Schwarz and Wiegand: „Video Coding: Part II of Fundamentals of Source and Video Coding“, 2016

Machine-to-Machine (M2M) Communication



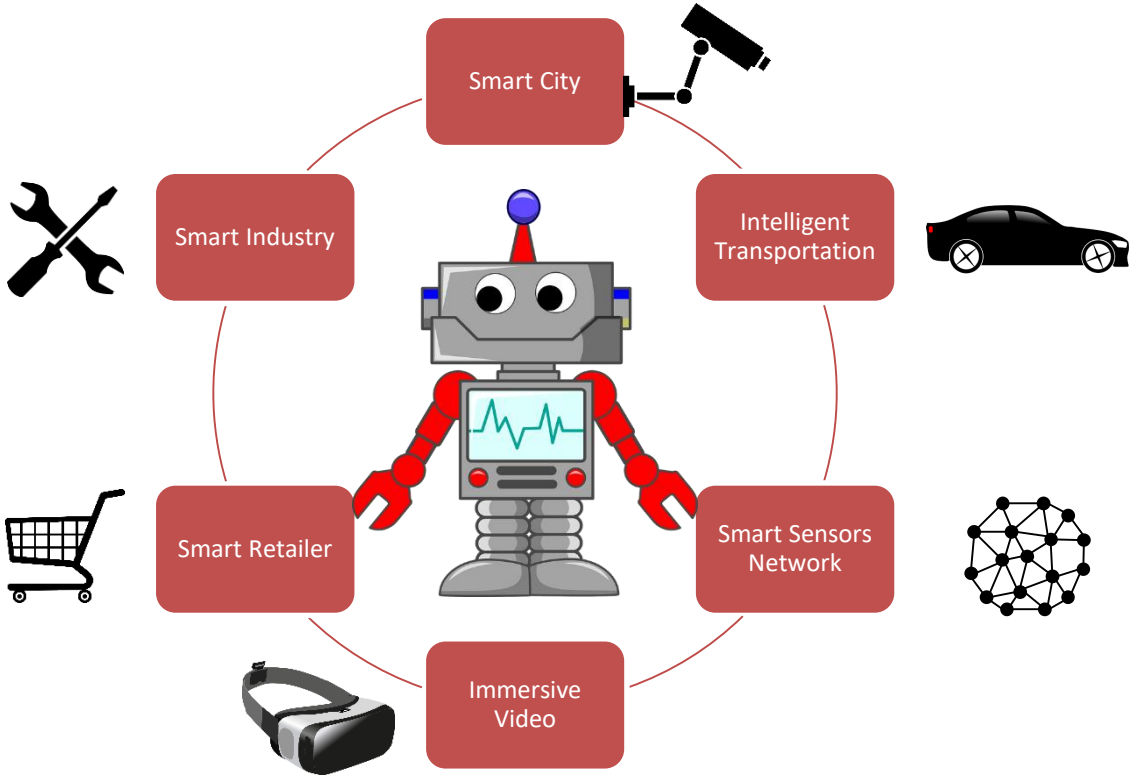
- Compressed image is analyzed by machine instead of human
- Target: Bitrate reduction at high detection rates
- Problem: Traditional codecs optimized for human observer

→ New coding methods have to be found

→ Targeted by MPEG "Video Coding for Machines (VCM)" ad-hoc group

VCM: Zhang et al., "MPEG-M49944: Report of the AhG on VCM," *MPEG*, 2020.
Image Credit: Schwarz and Wiegand: „Video Coding: Part II of Fundamentals of Source and Video Coding", 2016 (adapted)

VCM Use Cases



Agenda

1. Motivation Video Coding for Machines
2. Evaluation Framework
3. Classic Coding Performance vs. Video Coding for Machines
4. Feature-based Rate-Distortion Optimization
5. Conclusions

Signal Flow Video Coding for Machines

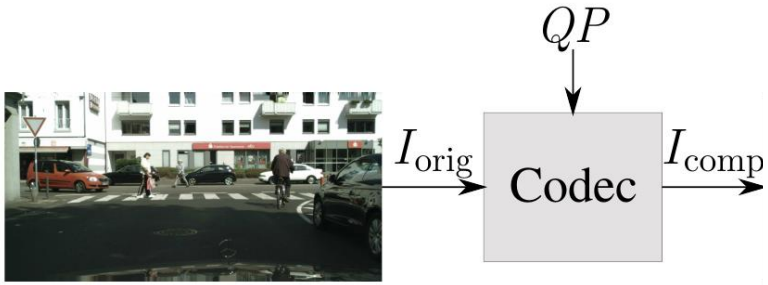
I_{orig}: Input image

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.

Signal Flow Video Coding for Machines



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

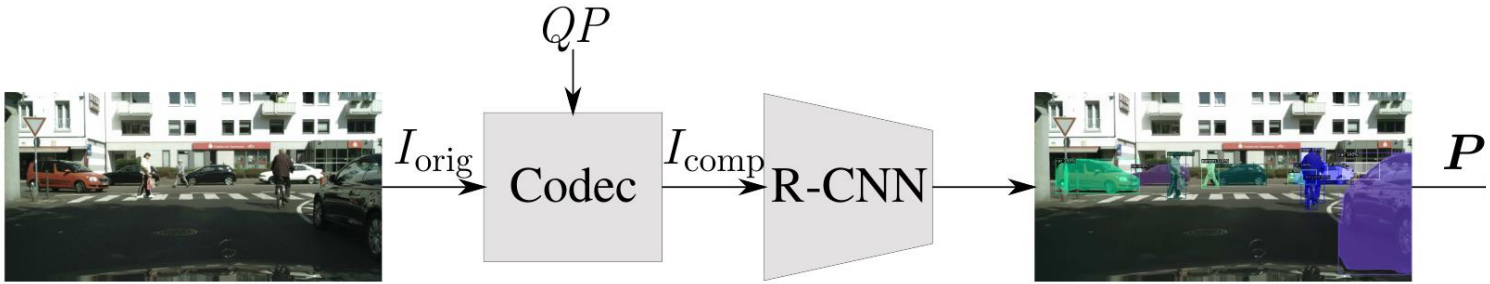
R-CNN: Region-based convolutional neural network
P: Predicted objects
G: Ground-truth objects
AP: Average precision

Coding Framework

- HEVC test model (HM-16.18)
- VVC test model (VTM-6.0)
- All-intra configuration
- QP from 2 to 47 in steps of 5
- 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.

Signal Flow Video Coding for Machines

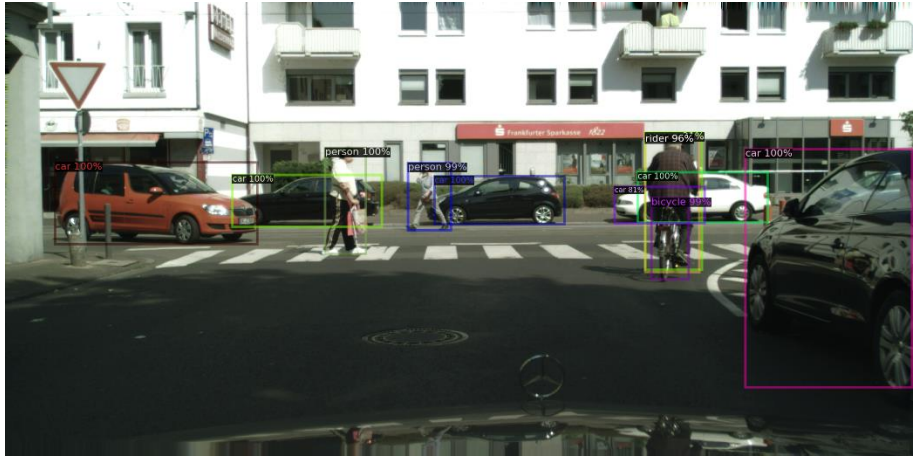


I_{orig}: Input image
I_{comp}: Compressed image
QP: Quantization parameter

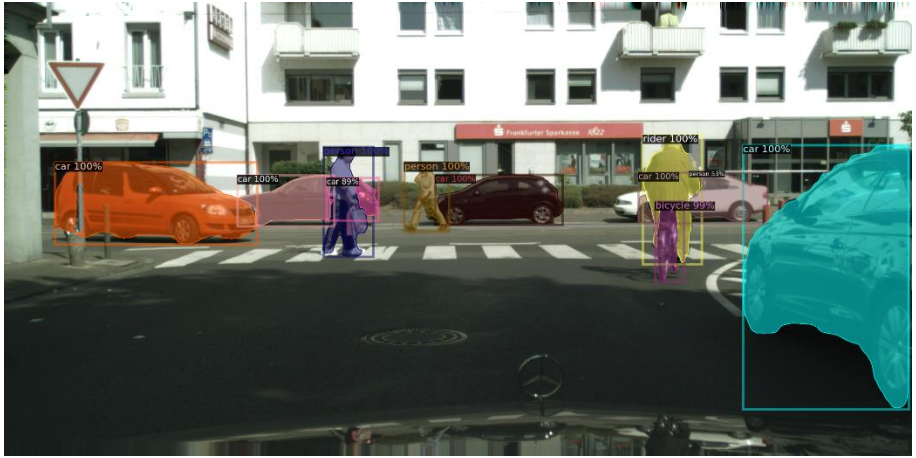
R-CNN: Region-based convolutional neural network
P: Predicted objects

Investigated Object Detection R-CNNs

Faster R-CNN

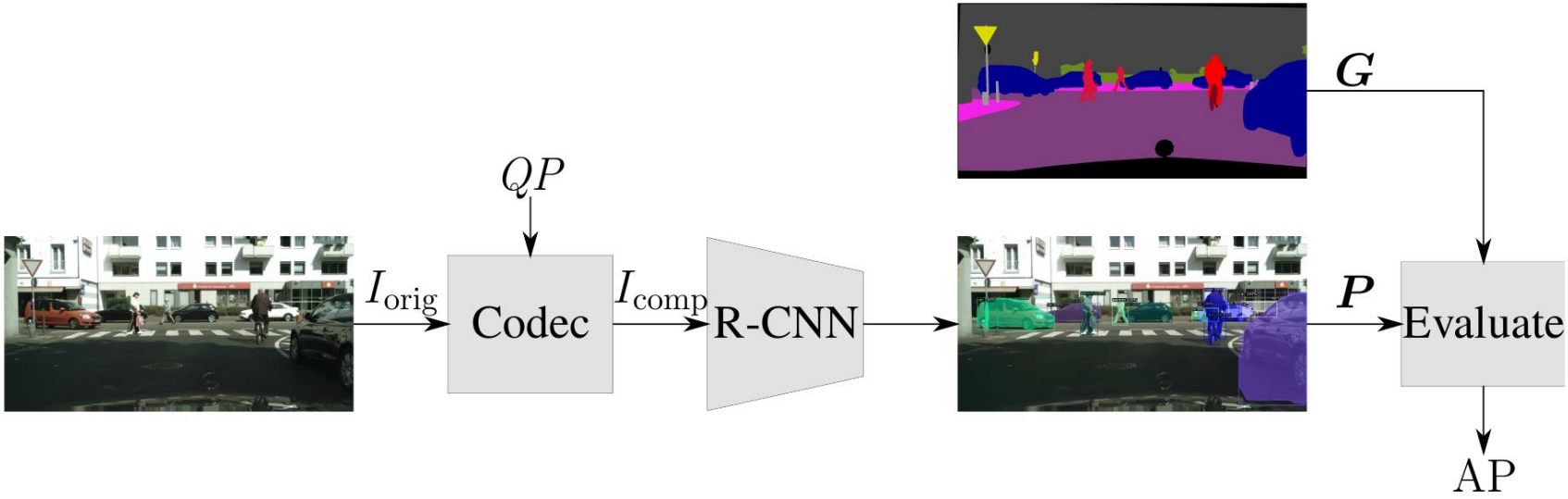


Mask R-CNN



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.

Signal Flow Video Coding for Machines

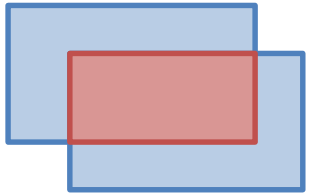


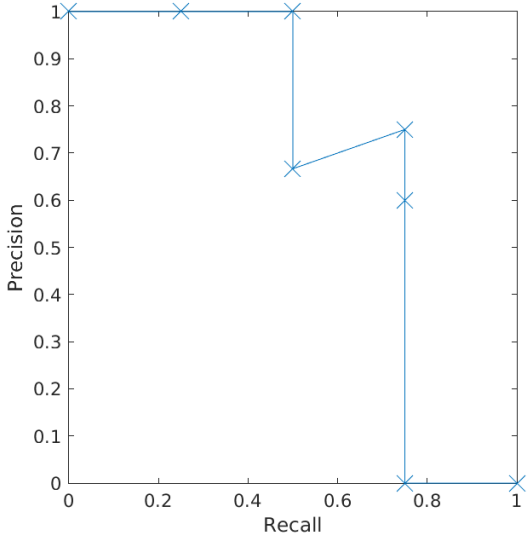
I_{orig}: Input image
I_{comp}: Compressed image
QP: Quantization parameter

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

Mean Average Precision (mAP)

- Metric to evaluate accuracy of object detection
- mAP used as proposed for Cityscapes
- Derive True Positives for certain Intersection over Union (IoU) thresholds
- Calculate precision and recall
- 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 ➔ weighted AP (wAP)



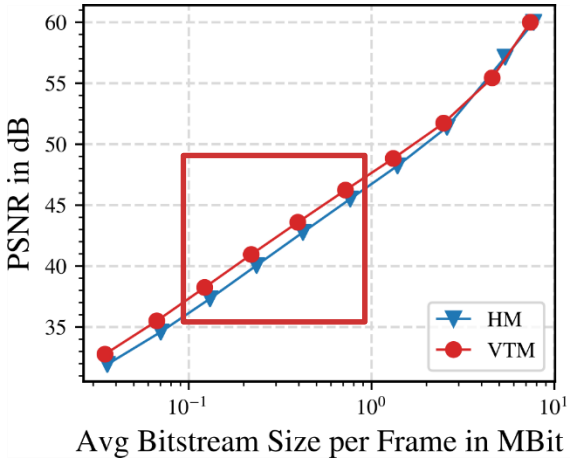


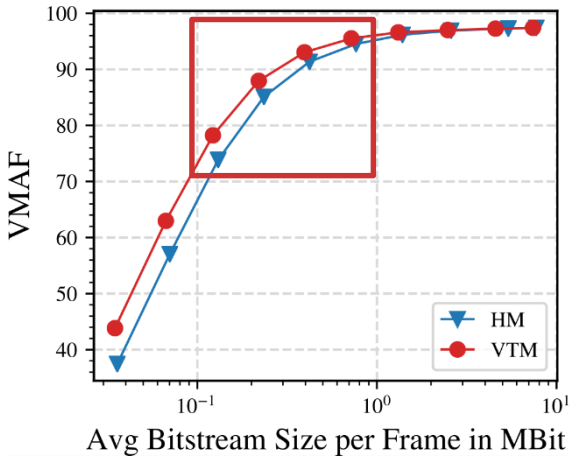
Cityscapes Scripts: Cordts et al., "The cityscapes dataset," https://github.com/mcordts/cityscapesScripts, 2017.

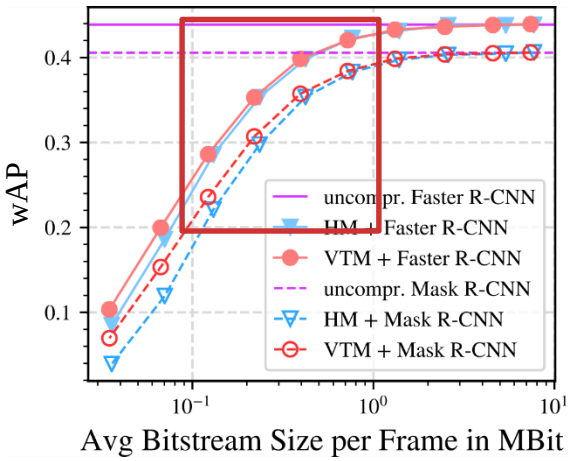
Agenda

1. Motivation Video Coding for Machines
2. Evaluation Framework
3. Classic Coding Performance vs. Video Coding for Machines
4. Feature-based Rate-Distortion Optimization
5. Conclusions

Comparison HEVC vs. VVC







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

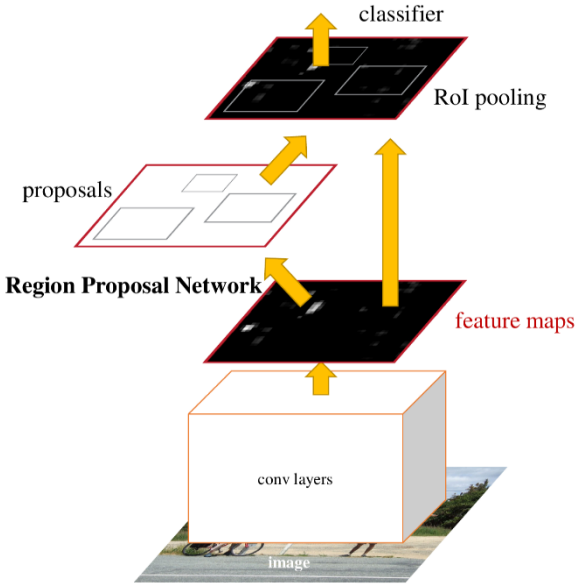
QP = {22, 27, 32, 37}
Anchor: HM

Fischer et al., "On intra video coding and in-loop filtering for neural object detection networks," *ICIP*, 2020.

Agenda

1. Motivation Video Coding for Machines
2. Evaluation Framework
3. Classic Coding Performance vs. Video Coding for Machines
4. Feature-based Rate-Distortion Optimization
5. Conclusions

Feature-Based Rate Distortion Optimization (FRDO)



- Feature space is created by convolution layers from input image
- For VCM, high feature fidelity is required
- High feature fidelity does not necessarily depend on a high pixel fidelity

→ This gap is exploited by standard-compliant FRDO for block partitioning in VVC

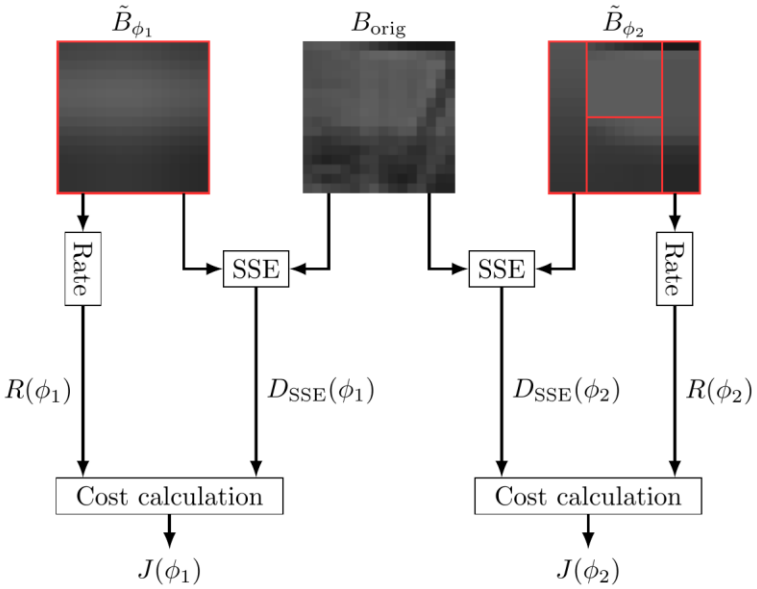
FRDO: Fischer et al., "Video Coding for Machines with Feature-Based Rate-Distortion Optimization," *MMSP*, 2020.
Image credit: Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *TPAMI*, 2017

Classic Rate-Distortion Optimization (RDO)

- Calculating costs J
  - For parameter configuration ϕ
  - Rate R
  - Distortion D
  - Lagrange multiplier λ

$$J(\phi) = D(\phi) + \lambda \cdot R(\phi)$$

- Commonly, sum of squared errors (SSE) used as distortion
- ϕ resulting in reconstructed block $\tilde{B}$ with lowest costs is selected



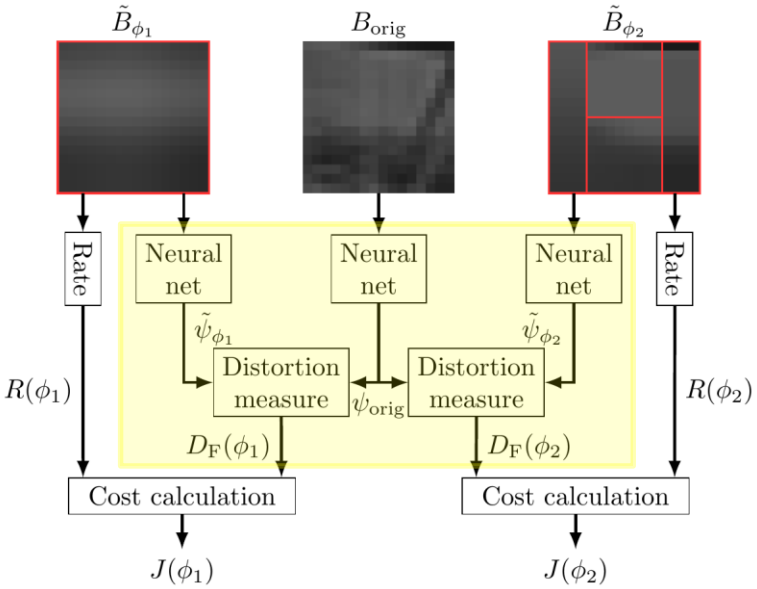
Feature-Based RDO

- Replacing SSE measure in pixel space with distortion measured in feature-space of each block ψ_ϕ

$$\psi_\phi = f(B_\phi)$$

$$D_{\mathrm{FSSE}}(\phi) = \sum_{(x,y,c)\in\psi} \left(\psi_{\mathrm{orig}}(x,y,c) - \tilde{\psi}_\phi(x,y,c)\right)^2$$

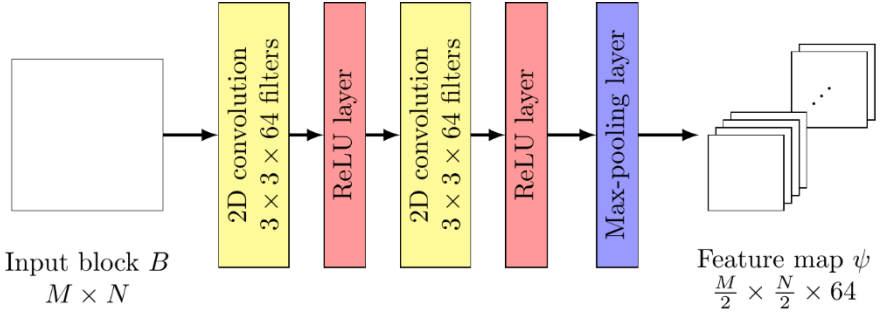
$$D_{\mathrm{FSAD}}(\phi) = \sum_{(x,y,c)\in\psi} \left|\psi_{\mathrm{orig}}(x,y,c) - \tilde{\psi}_\phi(x,y,c)\right|$$



FRDO: Fischer et al., "Video Coding for Machines with Feature-Based Rate-Distortion Optimization," *MMSP*, 2020.

Simulation Setup

- Added FRDO to VVC Test Model software VTM 8.0
- Hybrid FRDO (HFRDO) combining SSE and feature-based distortion
- QP from 12 to 27 in steps of 5
- First five layers of VGG-16 model used to obtain feature space



VTM: Bross et al., "JVET-Q2001: Versatile video coding (draft 8)," *JVET*, 2020.
VGG: Simonyan et al., "Very deep convolutional networks for large-scale image recognition," *Tech. Rep.*, 2015.

Simulation Results

Bjøntegaard delta rate with standard VTM-8.0 as anchor

	Quality Metric	
Optimization Method	**PSNR**	**wAP Mask R-CNN**
FRDO-SSE	4.30 %	-0.50 %
FRDO-SAD	3.02 %	-3.99 %
HFRDO-SSE	1.63 %	-4.55 %
HFRDO-SAD	1.26 %	-5.49 %

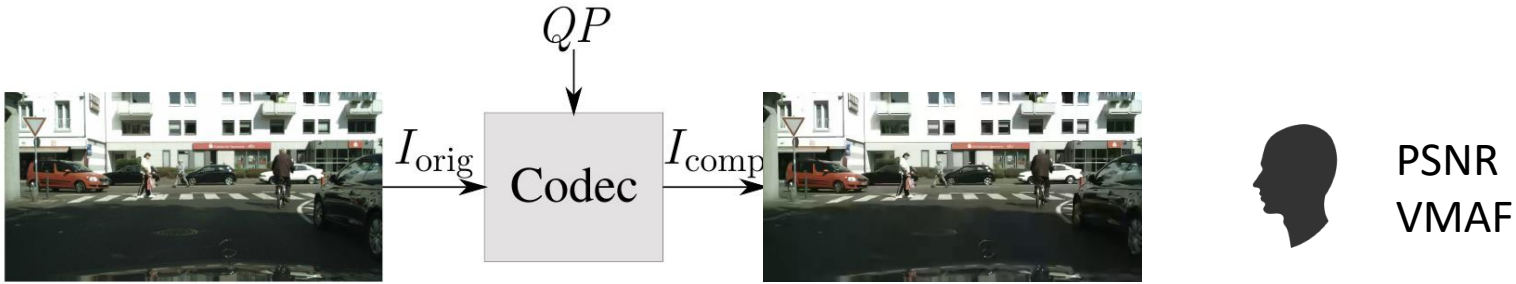
FRDO: Fischer et al., "Video Coding for Machines with Feature-Based Rate-Distortion Optimization," *MMSP*, 2020.

Conclusions

- Rising interest in video coding for machine scenarios
- Demands for optimized frameworks and codecs
- Coding gains of VTM over HM for VCM scenarios significantly lower
  - 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
- FRDO to obtain coding gains over standard VTM-8.0
  - Standard-compliant optimization for block partitioning
  - Saves up to 5.5% bitrate at the same Mask R-CNN detection accuracy
- Future work: confirm results for video data

BACKUP

Signal Flow Video Coding for Humans

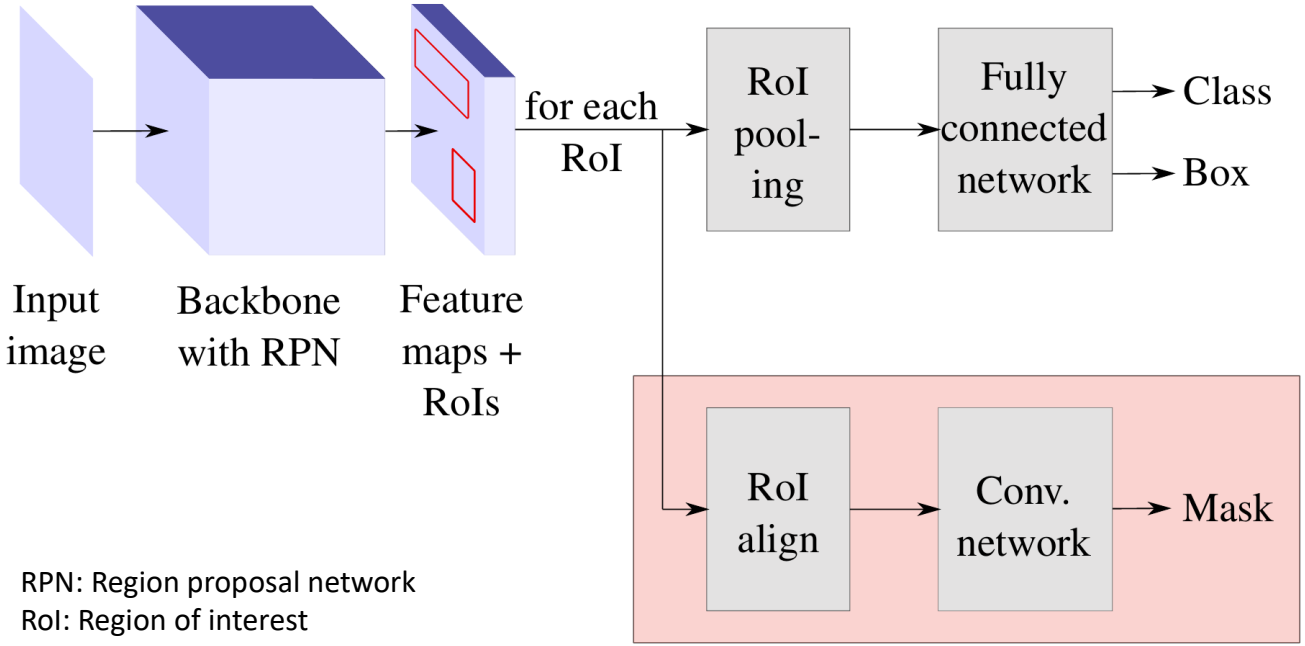


I_{orig}: Input image
I_{comp}: Compressed image
QP: Quantization parameter

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.

Investigated Object Detection R-CNNs



RPN: Region proposal network
RoI: Region of interest

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.

Normalizing Feature-Based Distortion with SSE

Classic RDO

$$J(\phi) = D_{\mathrm{SSE}}(\phi) + \lambda \cdot R(\phi)$$

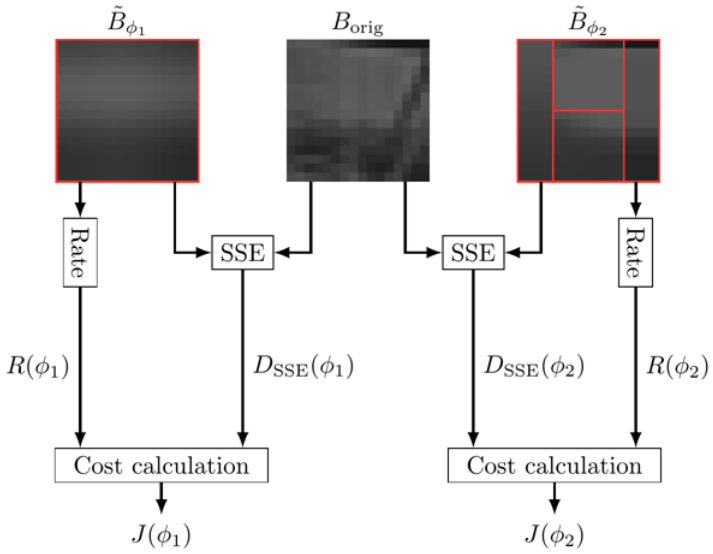
Feature-Based RDO

$$J(\phi) = D_F(\phi) + \lambda \cdot R(\phi)$$

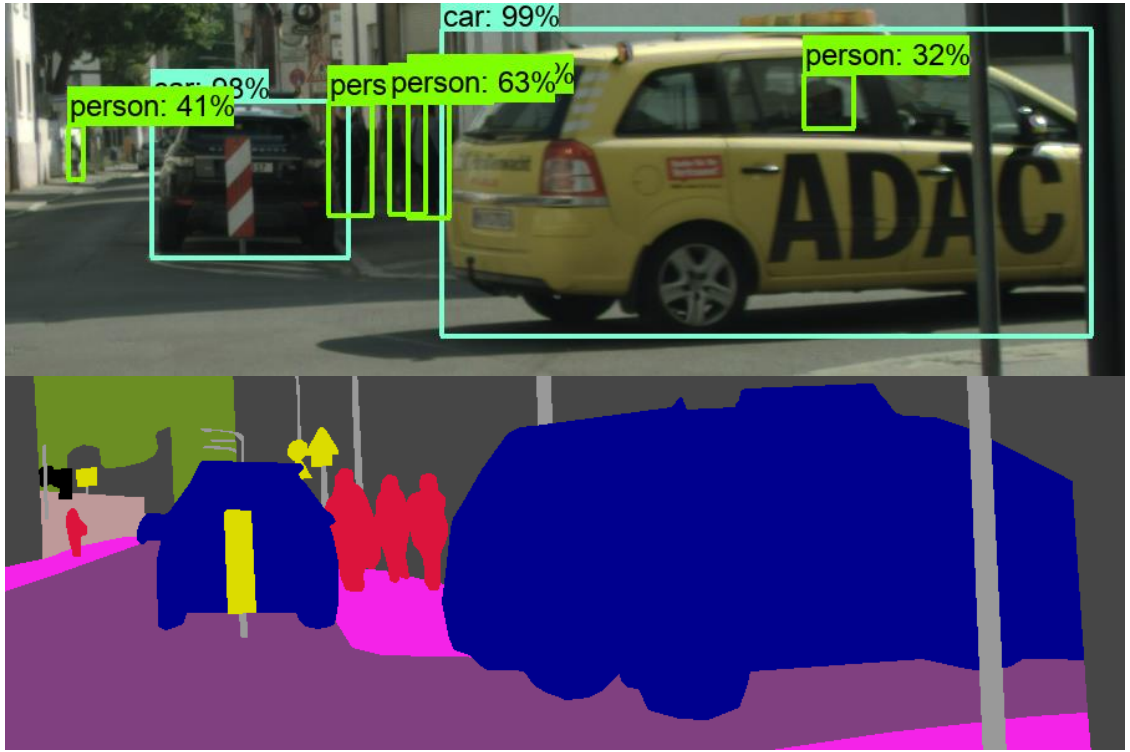
➔ $D_F(\phi)$ should be in same range as $D_{\mathrm{SSE}}(\phi)$

$$D_F^*(\phi_1) = D_F(\phi_1) \cdot \frac{D_{SSE}(\phi_1)}{D_F(\phi_1)} = D_{SSE}(\phi_1)$$

$$D_F^*(\phi_2) = D_F(\phi_2) \cdot \frac{D_{SSE}(\phi_1)}{D_F(\phi_1)}$$

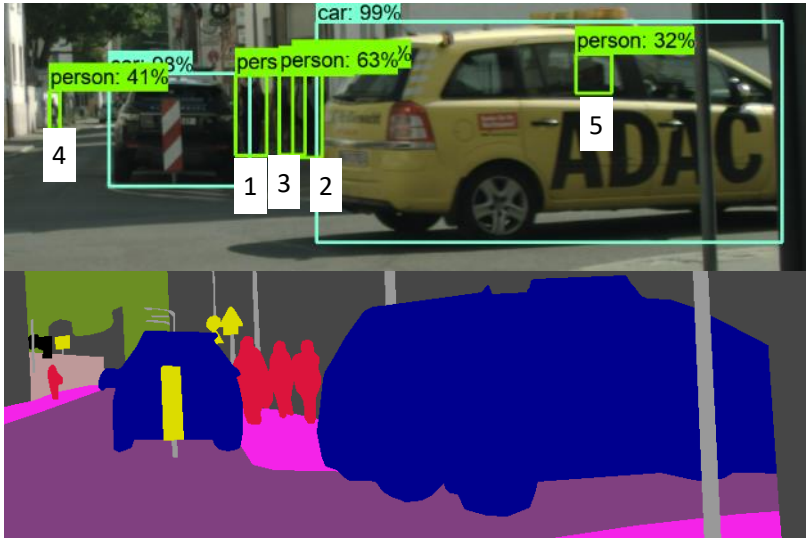


Example AP for Class Person and IoU=0.5



- 4 ground truth objects of class “Person”
- Convert ground truth data from pixelwise to boxes

Example AP for Class Person and IoU=0.5



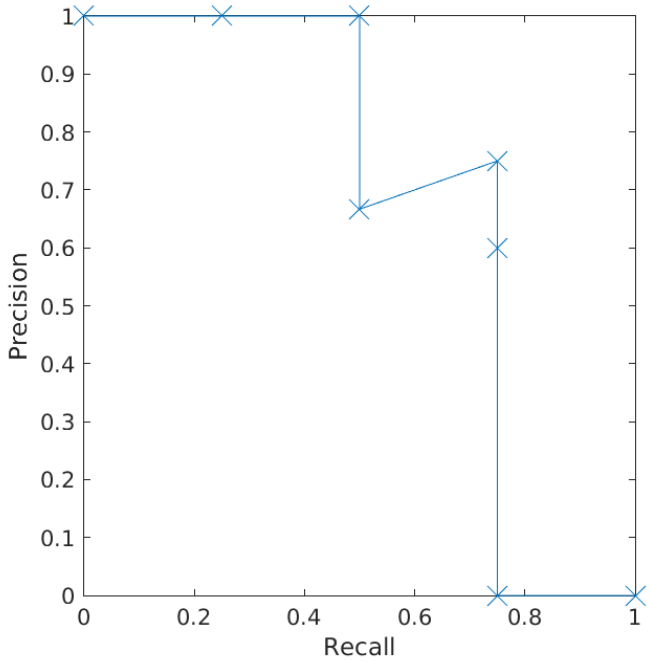
Rank	Score	Correct
1	0.93	True
2	0.91	True
3	0.63	False
4	0.41	True
5	0.32	False

$$Precision = \frac{true_pos}{true_pos + false_pos}$$

$$Recall = \frac{true_pos}{true_pos + false_neg}$$

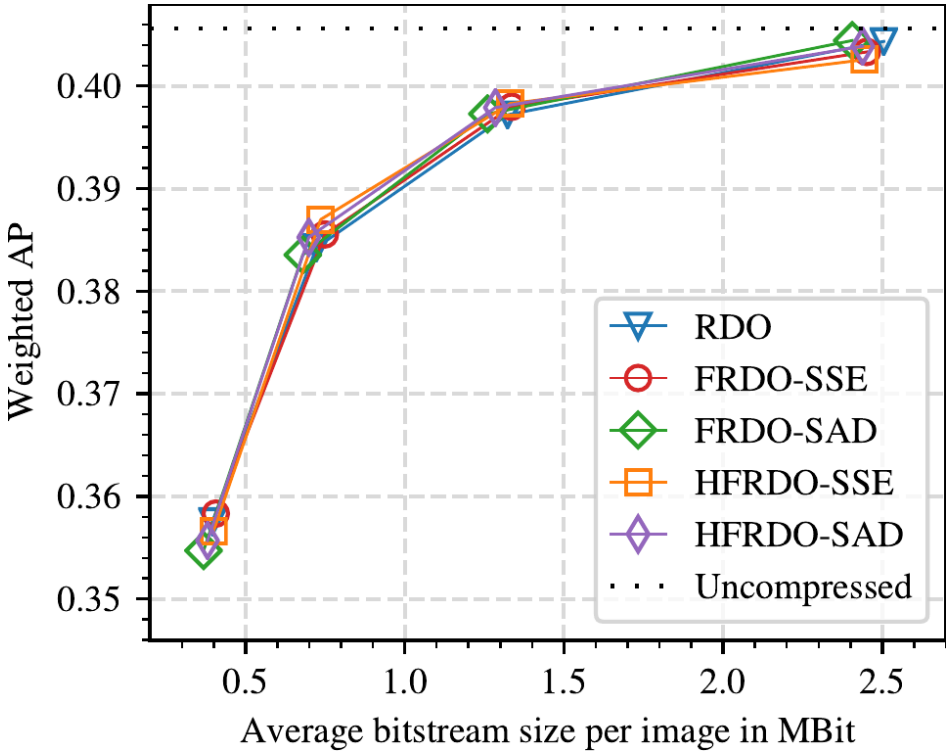
Example AP for Class Person and IoU=0.5

Rank	Score	Correct	Precision	Recall
1	0.93	True	$\frac{1}{1+0} = 1$	$\frac{1}{1+3} = 0.25$
2	0.91	True	$\frac{2}{2+0} = 1$	$\frac{2}{2+2} = 0.5$
3	0.63	False	$\frac{2}{2+1} = 0.67$	$\frac{2}{2+2} = 0.5$
4	0.41	True	$\frac{3}{3+1} = 0.75$	$\frac{3}{3+1} = 0.75$
5	0.32	False	$\frac{3}{3+2} = 0.6$	$\frac{3}{3+1} = 0.75$



➔ AP is area under Precision/Recall curve; here: AP=0.6771

Simulation Results



Simulation Results Allowing Delta QP

Bjøntegaard delta rate with standard VTM as anchor w/o delta QP

	Quality Metric	
Optimization Method	**PSNR**	**Weighted AP**
RDO with delta QP	-0.36 %	-1.15 %
FRDO-SAD with delta QP	4.55 %	-5.35 %
HFRDO-SAD with delta QP	1.01 %	-9.95 %

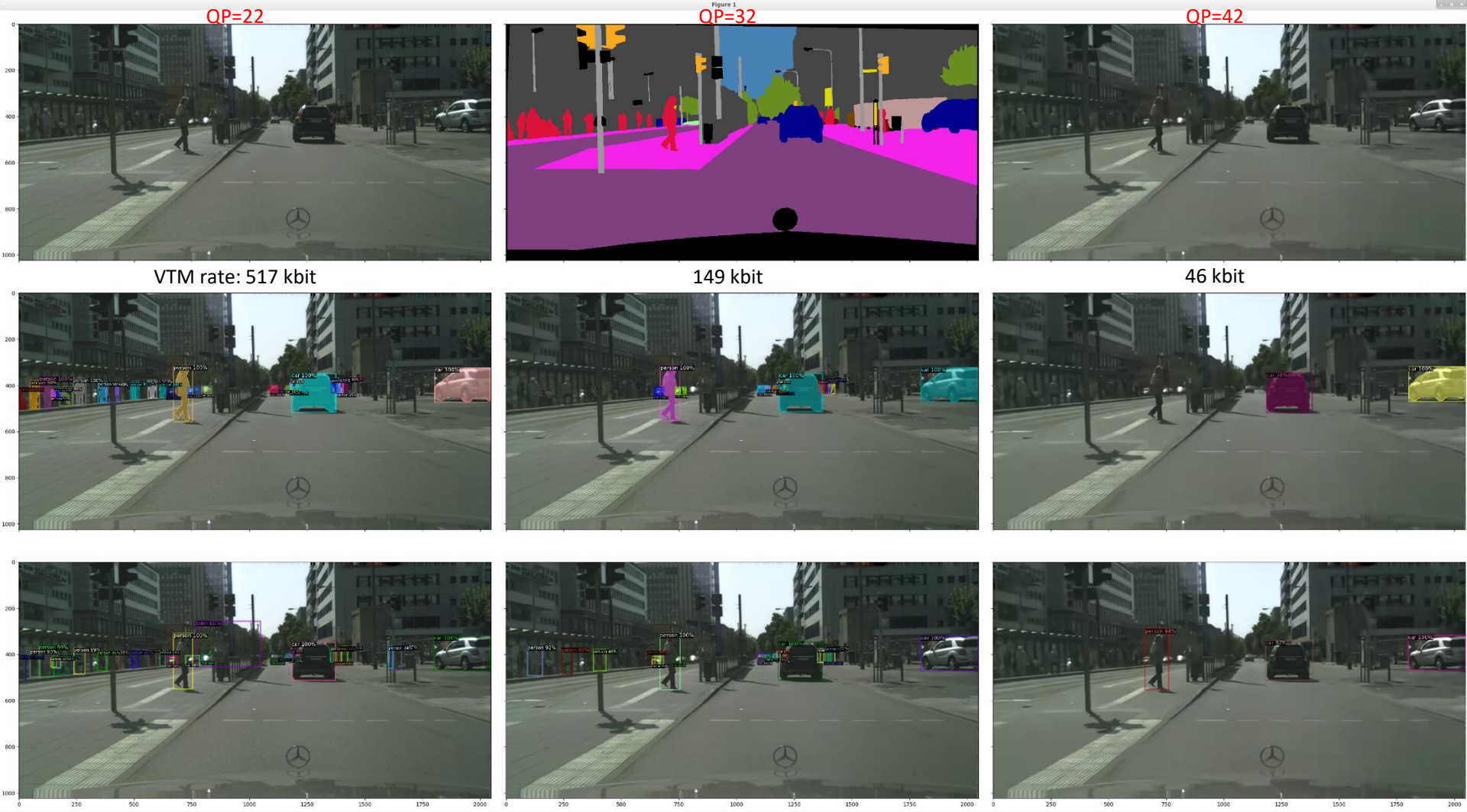
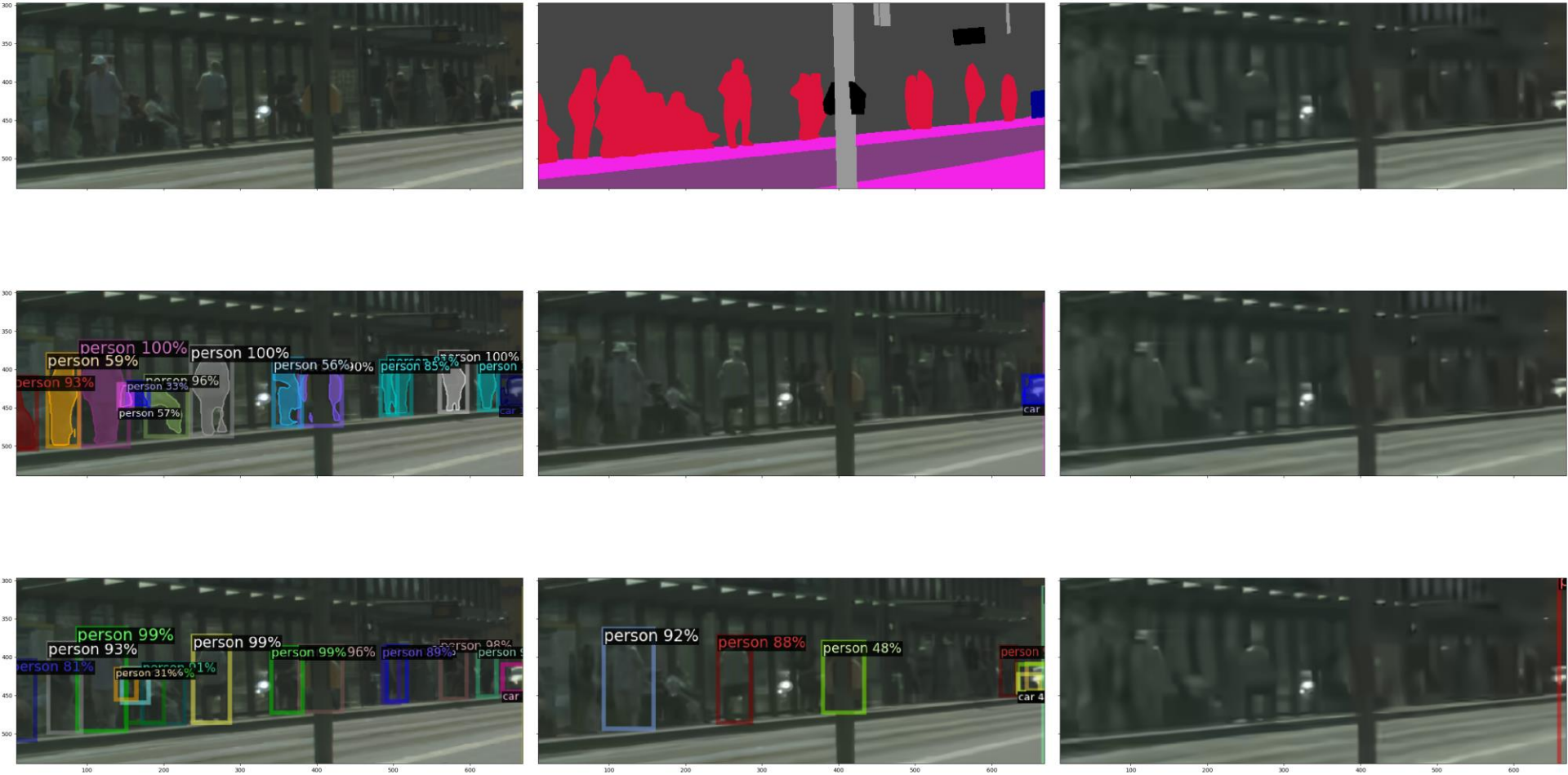
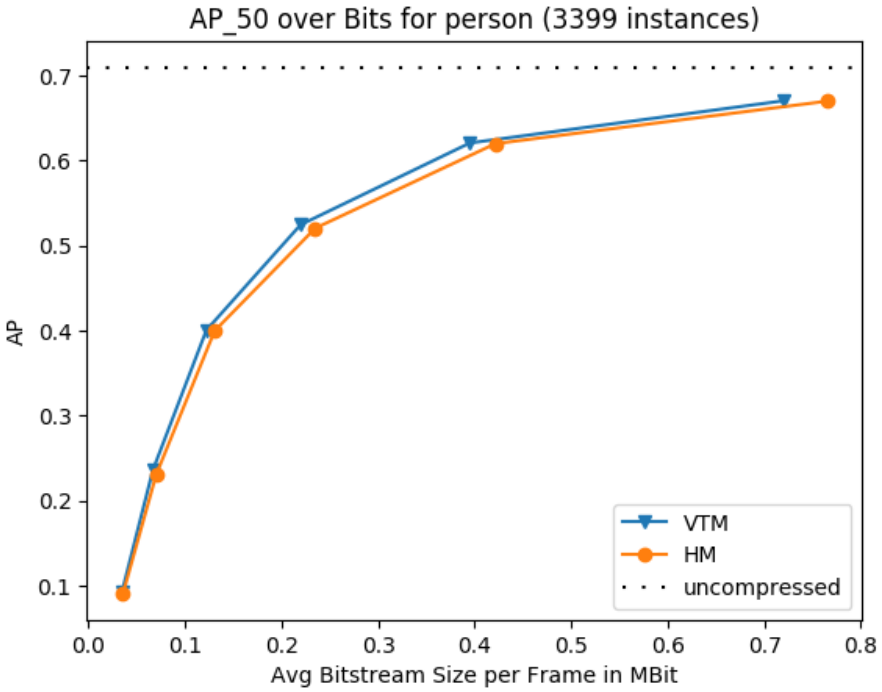
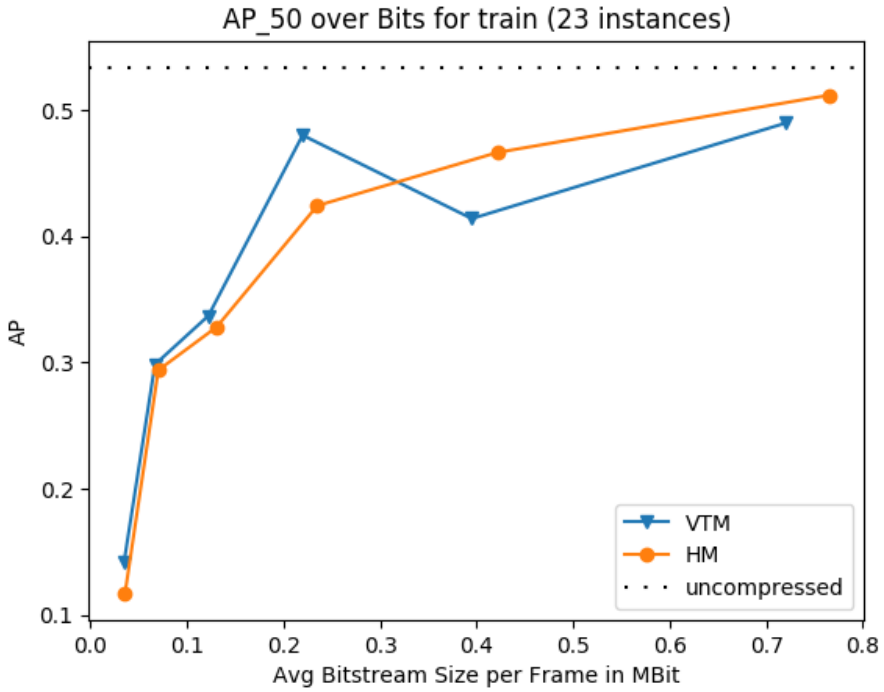


Figure 1
QP=22
QP=32
QP=42
VTM rate: 517 kbit
149 kbit
46 kbit

QP=22 QP=32 QP=42



Results AP_50 HM vs. VTM Faster R-CNN



Results AP_50 HM vs. VTM Faster R-CNN