Standard Codecs for Deep Vision Models

Christoph Reich
TU Munich, Computer Vision Group
TU Darmstadt, Visual Inference Lab
1st Workshop on AI for Streaming at CVPR
Seattle, USA, June 2024, 17th

A Perspective on Deep Vision Performance with Standard Image and Video Codecs





Christoph Reich[1,2,3,5] Oliver Hahn[1] Daniel Cremers[2] Stefan Roth[1,4] Biplob Debnath[3]

Deep Video Codec Control for Vision Models






Christoph Reich[1,2,3,5] Biplob Debnath[3] Deep Patel[3] Tim Prangemeier[1] Daniel Cremers[2] Srimat Chakradhar[3]

TECHNISCHE UNIVERSITÄT DARMSTADT
NEC Laboratories **America**
hessian.AI
Munich Center for Machine Learning

[1]TU Darmstadt [2]TU Munich [3]NEC Laboratories America, Inc. [4]hesssian.AI [5]Munich Center for Machine Learning

Computer Vision Group

TECHNISCHE UNIVERSITÄT DARMSTADT

A Perspective on Deep Vision Performance with Standard Image and Video Codecs





Christoph Reich[1,2,3,5] Oliver Hahn[1] Daniel Cremers[2] Stefan Roth[1,4] Biplob Debnath[3]

Deep Video Codec Control for Vision Models






Christoph Reich[1,2,3,5] Biplob Debnath[3] Deep Patel[3] Tim Prangemeier[1] Daniel Cremers[2] Srimat Chakradhar[3]

TECHNISCHE UNIVERSITÄT DARMSTADT

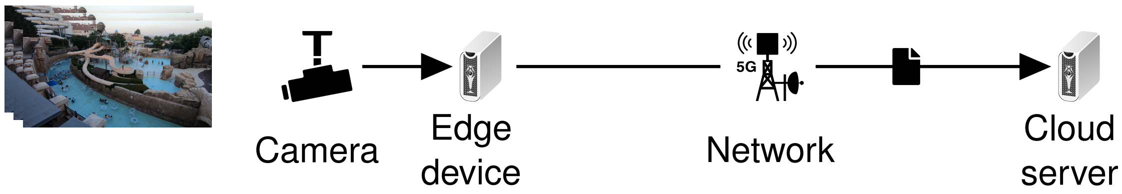
NEC Laboratories **America**

hessian.AI

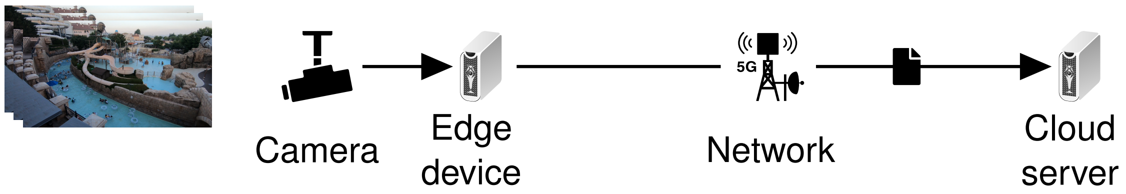
Munich Center for Machine Learning

[1]TU Darmstadt [2]TU Munich [3]NEC Laboratories America, Inc. [4]hesssian.AI [5]Munich Center for Machine Learning

Motivation

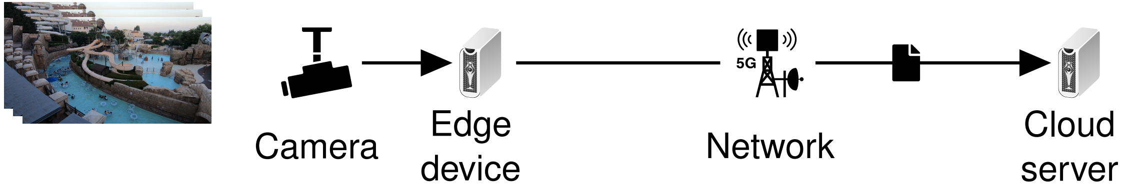


Motivation



- Standard image/video codecs (& rate control) used to compensate for **bandwidth** and **storage constrains**

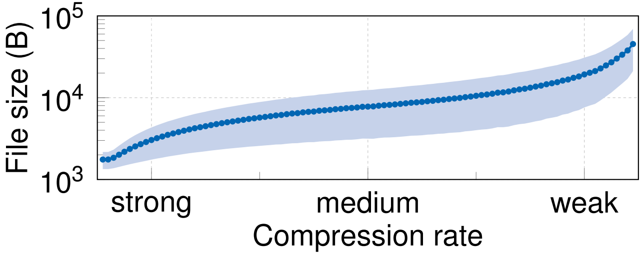
Motivation

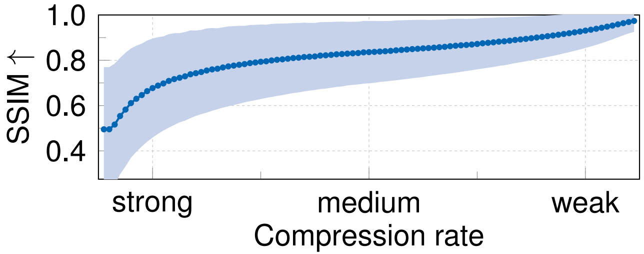


- Standard image/video codecs (& rate control) used to compensate for **bandwidth** and **storage constrains**
- Standardization required to ensure **interoperability** and **low costs**

Introduction

- Standard codecs been studied using Shannon's **rate-distortion theory** [1] and via **perceptual quality** [2]

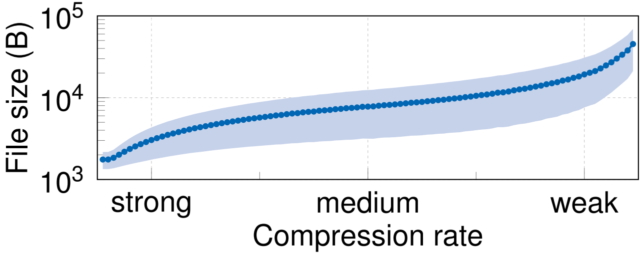


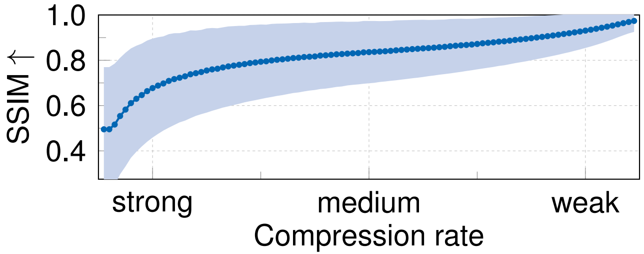


[1] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, 1949.
[2] Y. Blau *et al.*, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *ICML*, 2019.

Introduction

- Standard codecs been studied using Shannon's **rate-distortion theory** [1] and via **perceptual quality** [2]



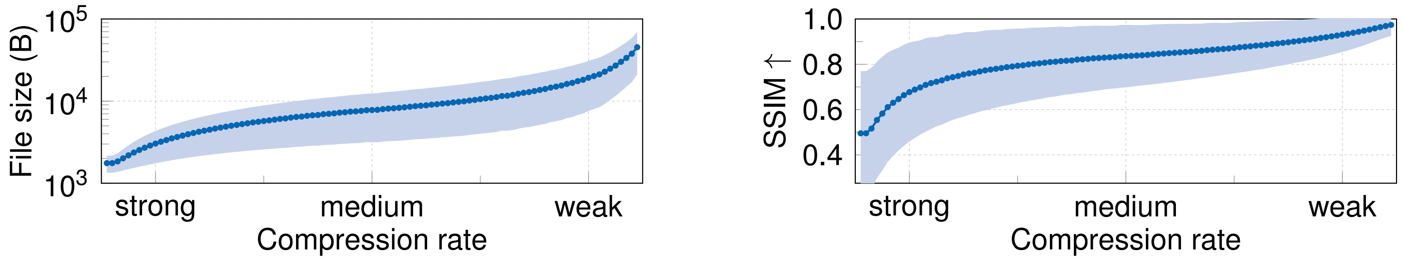


↯ **A significant and increasing amount of images and videos are analyzed by deep vision models**

[1] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, 1949.
[2] Y. Blau *et al.*, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *ICML*, 2019.

Introduction

- Standard codecs been studied using Shannon's **rate-distortion theory** [1] and via **perceptual quality** [2]



ϟ **A significant and increasing amount of images and videos are analyzed by deep vision models**

We examine the implications of using standard codecs within deep vision pipelines.

[1] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, 1949.
[2] Y. Blau *et al.*, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *ICML*, 2019.

Experiments

- **23 deep vision models** evaluated on coded images/videos

Experiments

- **23 deep vision models** evaluated on coded images/videos
- **JPEG** and **H.264** coding utilized

Experiments

- **23 deep vision models** evaluated on coded images/videos
- **JPEG** and **H.264** coding utilized
- Report results on a **wide range of different computer vision tasks**
  - Image classification
  - Object detection
  - Semantic segmentation
  - Optical flow estimation

Experiments

- **23 deep vision models** evaluated on coded images/videos
- **JPEG** and **H.264** coding utilized
- Report results on a **wide range of different computer vision tasks**
  - Image classification
  - Object detection
  - Semantic segmentation
  - Optical flow estimation

Evaluation approach

We measure the **relative vision performance** between the prediction obtained on the coded image/video and the prediction based on the original image/video (pseudo-label).

$$\mathrm{mIoU}\left(\mathrm{DeepLabV3}(\mathbf{I}_{\mathrm{coded}}), \mathrm{DeepLabV3}\left(\mathbf{I}_{\mathrm{original}}\right)\right)$$

Experiments

- **23 deep vision models** evaluated on coded images/videos
- **JPEG** and **H.264** coding utilized
- Report results on a **wide range of different computer vision tasks**
  - Image classification
  - Object detection
  - Semantic segmentation
  - Optical flow estimation

Evaluation approach

We measure the **relative vision performance** between the prediction obtained on the coded image/video and the prediction based on the original image/video (pseudo-label).

$$\mathrm{mIoU}\left(\mathrm{DeepLabV3}(\mathbf{I}_{\mathrm{coded}}), \mathrm{DeepLabV3}\left(\mathbf{I}_{\mathrm{original}}\right)\right)$$

- **Measures the effect of coding isolated from other factors**

Experiments

- **23 deep vision models** evaluated on coded images/videos
- **JPEG** and **H.264** coding utilized
- Report results on a **wide range of different computer vision tasks**
  - Image classification
  - Object detection
  - Semantic segmentation
  - Optical flow estimation

Evaluation approach

We measure the **relative vision performance** between the prediction obtained on the coded image/video and the prediction based on the original image/video (pseudo-label).

$$\mathrm{mIoU}\left(\mathrm{DeepLabV3}(\mathbf{I}_{\mathrm{coded}}), \mathrm{DeepLabV3}\left(\mathbf{I}_{\mathrm{original}}\right)\right)$$

- **Measures the effect of coding isolated from other factors**
- Interpretable and comparable results between models

Experiments

- **23 deep vision models** evaluated on coded images/videos
- **JPEG** and **H.264** coding utilized
- Report results on a **wide range of different computer vision tasks**
  - Image classification
  - Object detection
  - Semantic segmentation
  - Optical flow estimation

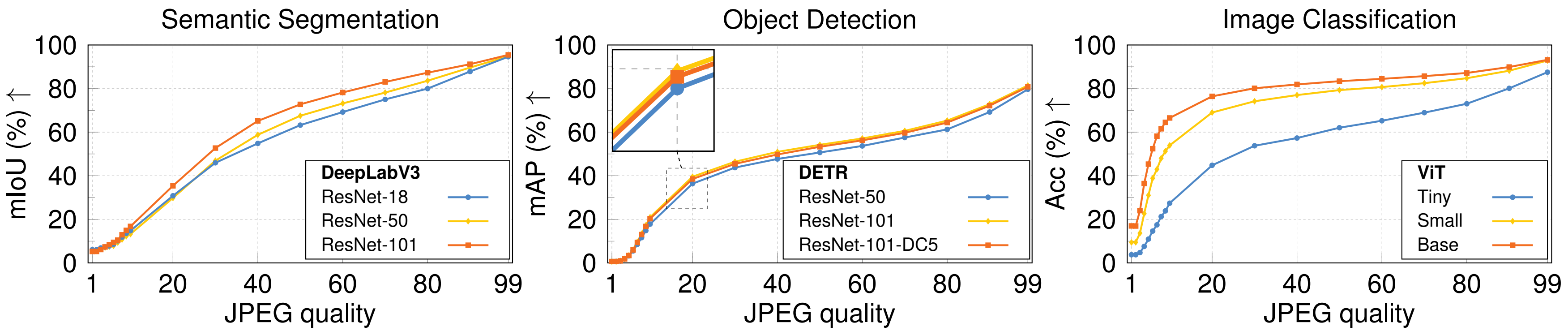
Evaluation approach

We measure the **relative vision performance** between the prediction obtained on the coded image/video and the prediction based on the original image/video (pseudo-label).

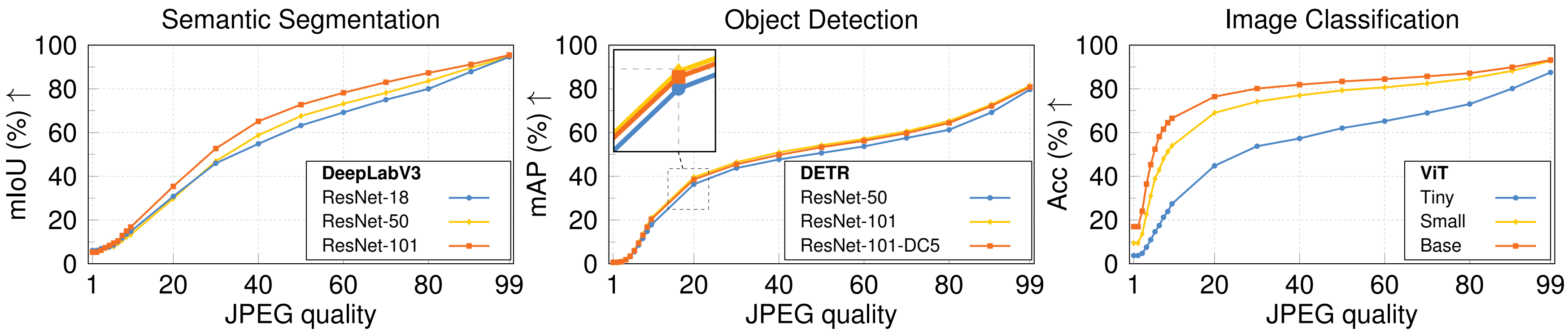
$$\mathrm{mIoU}\left(\mathrm{DeepLabV3}(\mathbf{I}_{\text{coded}}), \mathrm{DeepLabV3}\left(\mathbf{I}_{\text{original}}\right)\right)$$

- **Measures the effect of coding isolated from other factors**
- Interpretable and comparable results between models
- Paper presents also results w.r.t. ground truth labels

Results on JPEG-Coded Images I

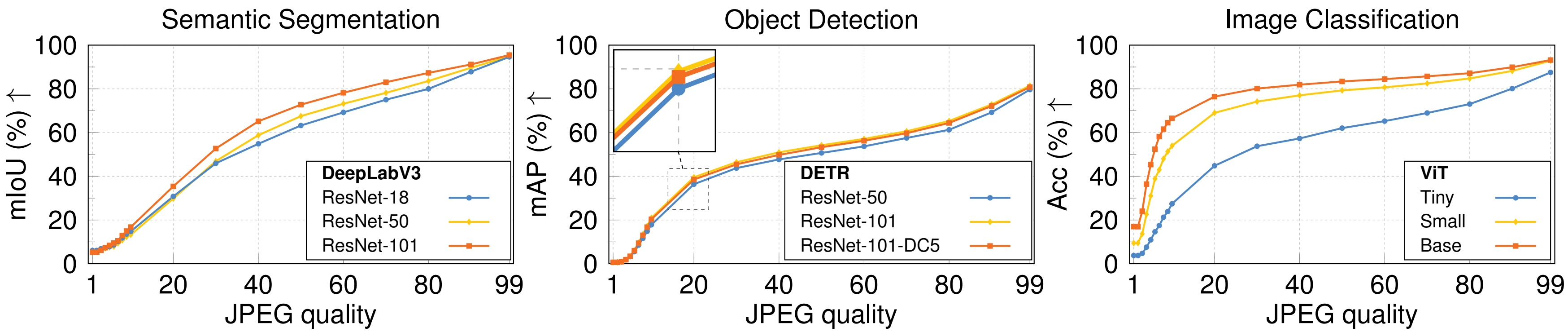


Results on JPEG-Coded Images I



Accuracy of deep vision models vastly deteriorates for small JPEG qualities.

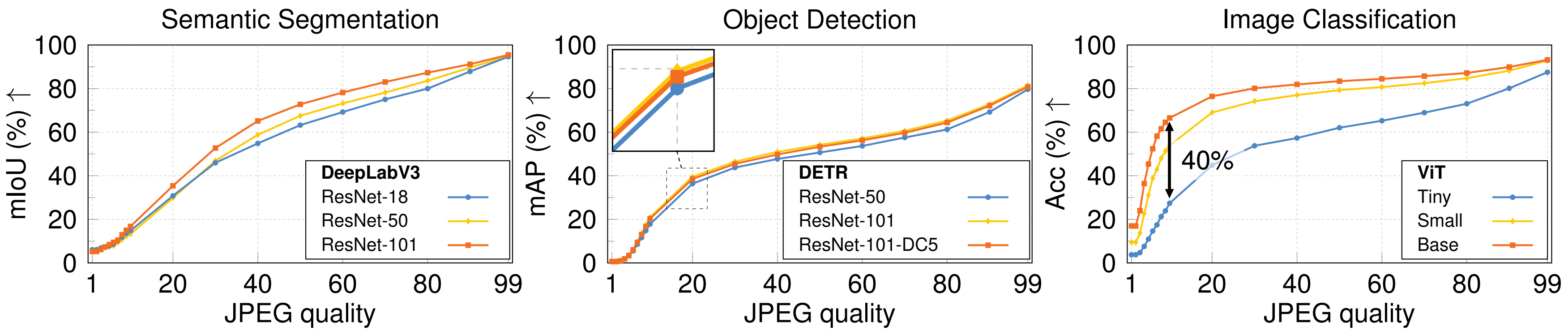
Results on JPEG-Coded Images I



Accuracy of deep vision models vastly deteriorates for small JPEG qualities.

- Dense prediction tasks are more sensitive to JPEG coding than image classification

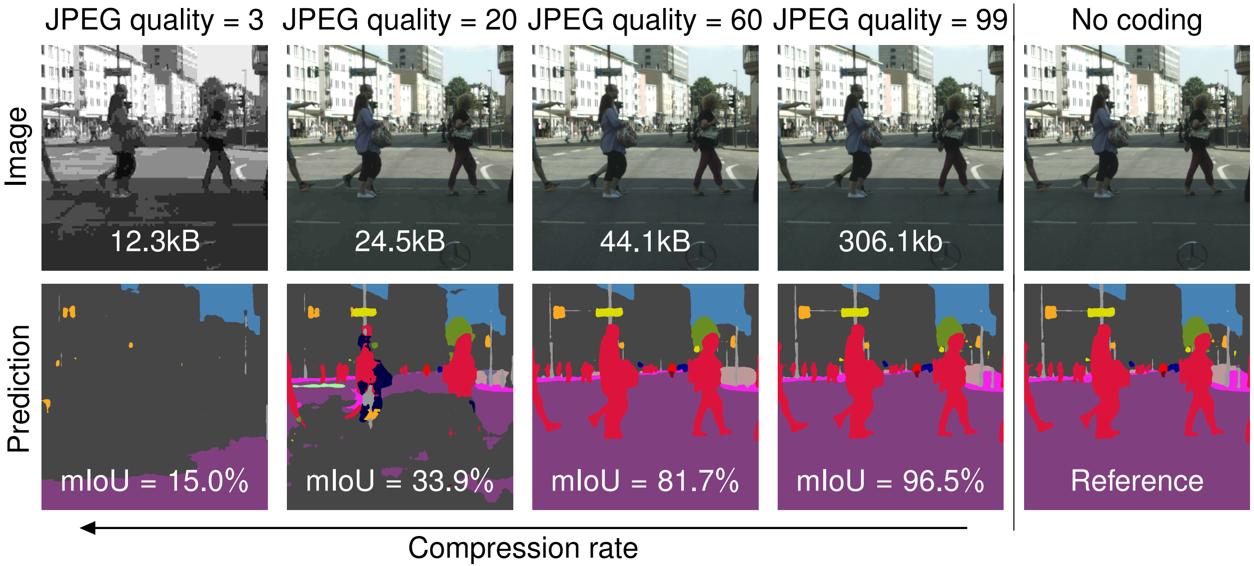
Results on JPEG-Coded Images I



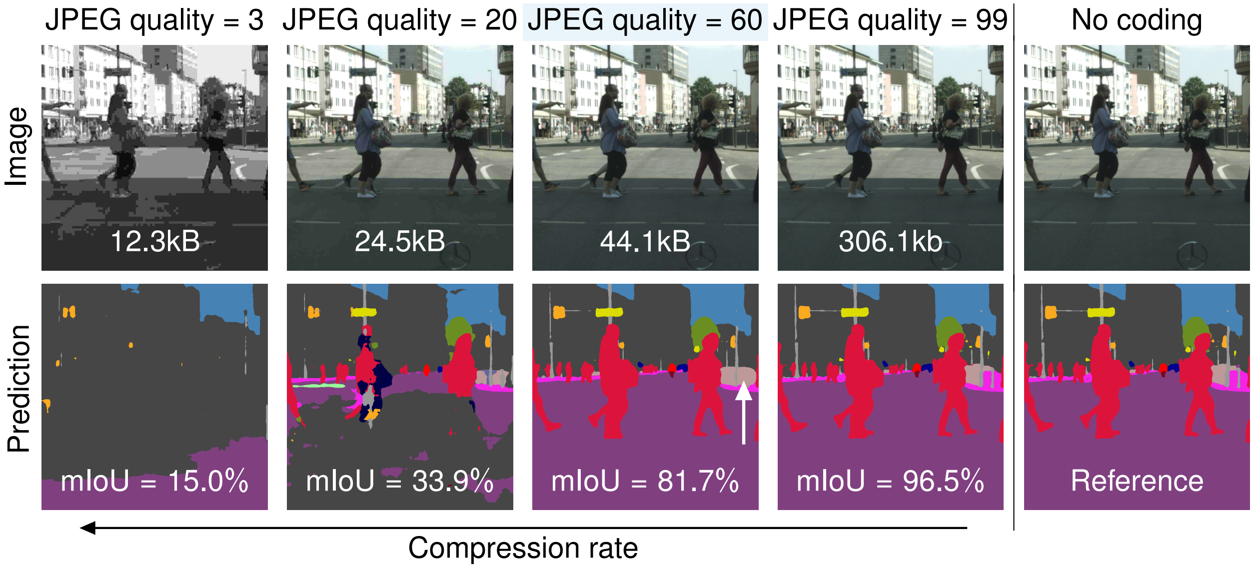
Accuracy of deep vision models vastly deteriorates for small JPEG qualities.

- Dense prediction tasks are more sensitive to JPEG coding than image classification
- Larger capacity models offer better robustness against JPEG coding

Results on JPEG-Coded Images II

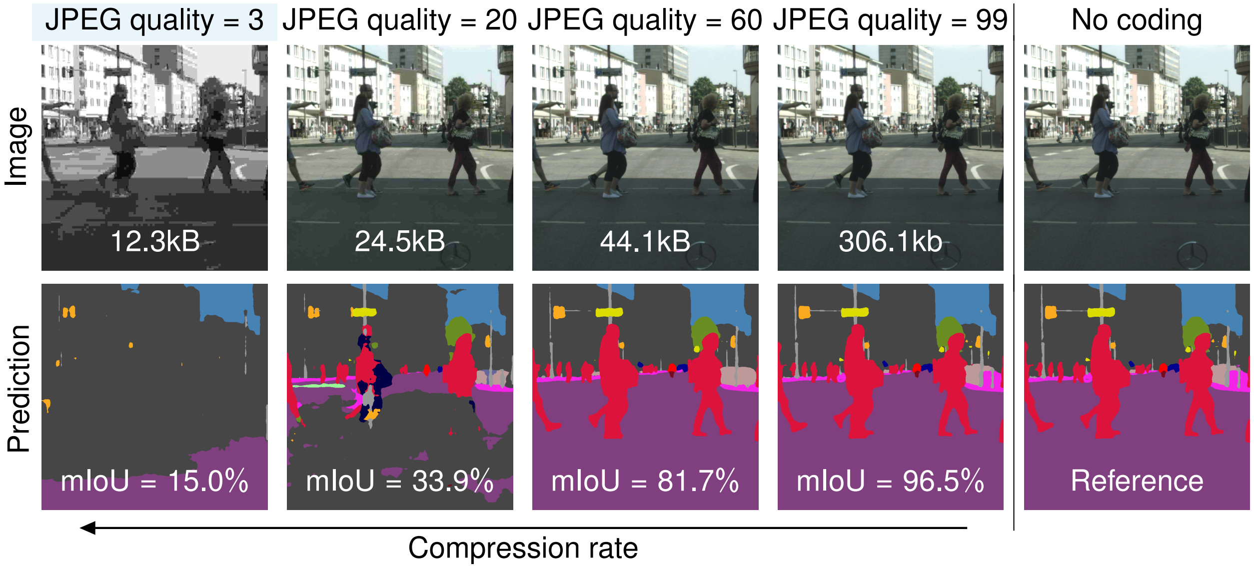


Results on JPEG-Coded Images II



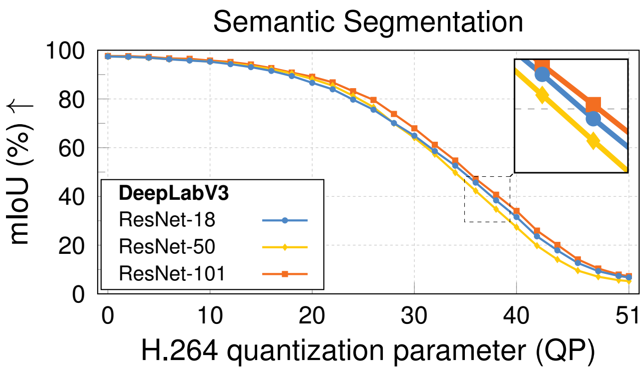
Weak compression rates can lead to wrong predictions

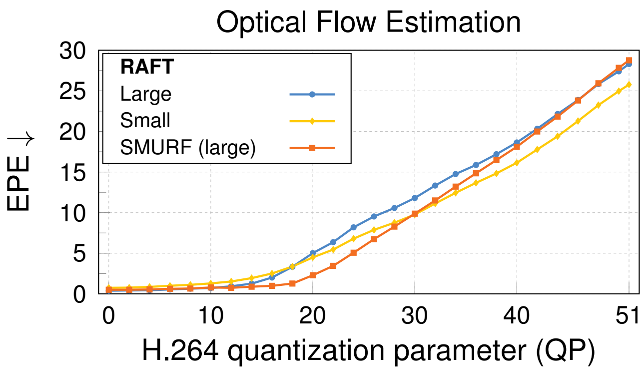
Results on JPEG-Coded Images II



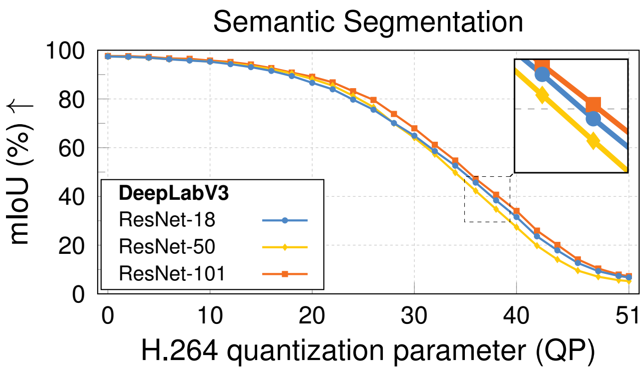
Weak compression rates can lead to wrong predictions – strong coding leads to a collapse in segmentation accuracy.

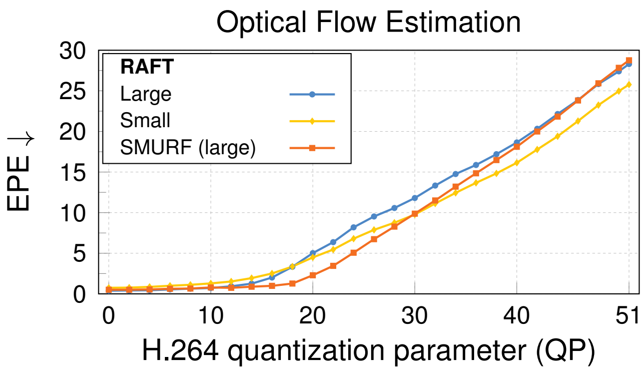
Results on H.264-Coded Videos





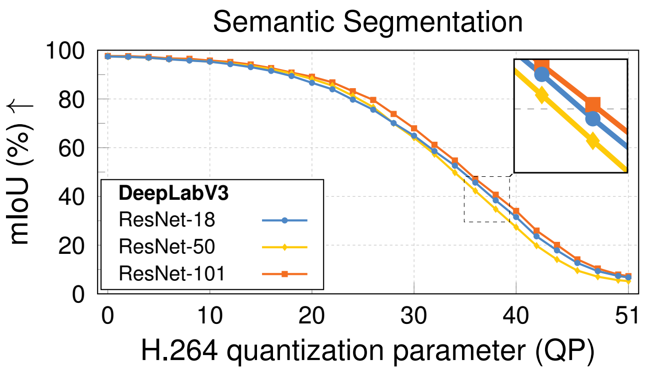
Results on H.264-Coded Videos

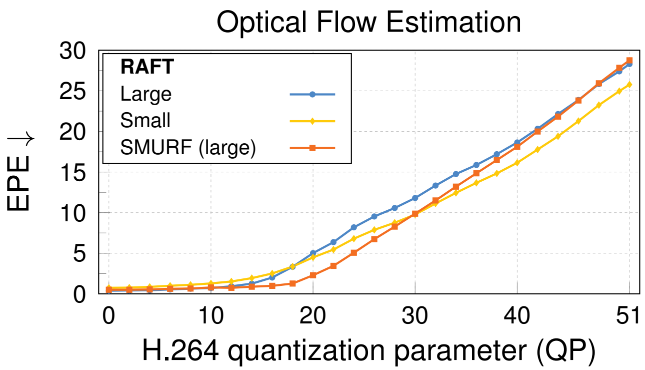




Accuracy of deep vision models vastly deteriorates for strong H.264 quantization.

Results on H.264-Coded Videos





Accuracy of deep vision models vastly deteriorates for strong H.264 quantization.

- Surprisingly, larger capacity models do not necessarily lead to more robustness against H.264 coding

Conclusion

- **Standard image and video coding significantly effects the accuracy of current deep vision models**

Conclusion

- **Standard image and video coding significantly effects the accuracy of current deep vision models**
- The accuracy of all 23 tested vision models deteriorated with standard coding

Conclusion

- **Standard image and video coding significantly effects the accuracy of current deep vision models**
- The accuracy of all 23 tested vision models deteriorated with standard coding
- Strong compression rates can lead to a complete collapse in accuracy

Conclusion

- **Standard image and video coding significantly effects the accuracy of current deep vision models**
- The accuracy of all 23 tested vision models deteriorated with standard coding
- Strong compression rates can lead to a complete collapse in accuracy

Acknowledgements

This work was supported by the European Research Council and by the State of Hesse through the cluster project “The Third Wave of Artificial Intelligence (3AI)“

Funded by the European Union

European Research Council
Established by the European Commission

3AI THE THIRD WAVE OF AI

Computer Vision Group

TECHNISCHE UNIVERSITÄT DARMSTADT

A Perspective on Deep Vision Performance with Standard Image and Video Codecs

Christoph Reich[1,2,3,5] Oliver Hahn[1] Daniel Cremers[2] Stefan Roth[1,4] Biplob Debnath[3]

Deep Video Codec Control for Vision Models

Christoph Reich[1,2,3,5] Biplob Debnath[3] Deep Patel[3] Tim Prangemeier[1] Daniel Cremers[2] Srimat Chakradhar[3]

TECHNISCHE UNIVERSITÄT DARMSTADT

NEC Laboratories **America**

hessian.AI

Munich Center for Machine Learning

[1]TU Darmstadt [2]TU Munich [3]NEC Laboratories America, Inc. [4]hesssian.AI [5]Munich Center for Machine Learning

Introduction

How can we optimize standard video codecs for deep vision models?

Introduction

How can we optimize standard video codecs for deep vision models?

More specifically, we want to consider the following conditions:

- ✓ Optimize downstream deep vision performance on coded videos
- ✓ Adapt to different bandwidth or storage constrains (rate control)
- ✓ Adhere to existing standards

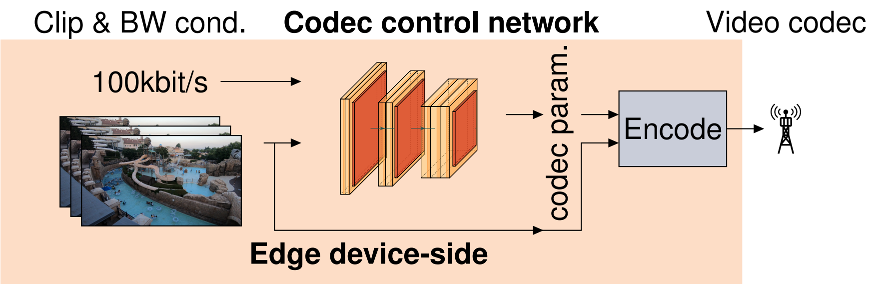
Related Work

	Optimize vision performance	Rate control	ISO
Deep video codecs [3]	✓	∼	✗
Standard video codecs (e.g., H.264 [4])	✗	✓	✓
Deep Video Codec Control	✓	✓	✓

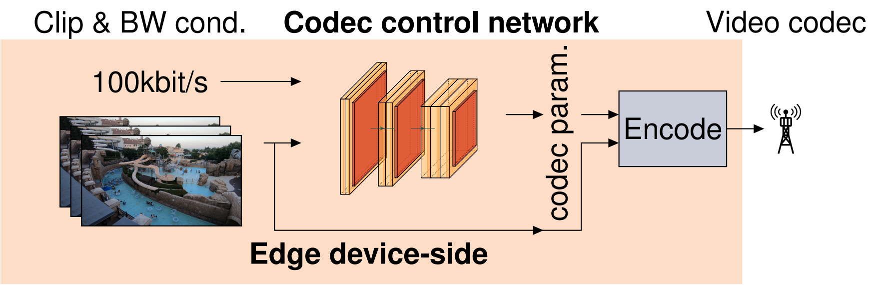
[3] Y. Zhang *et al.*, “A survey on perceptually optimized video coding,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–37, 2023.

[4] T. Wiegand *et al.*, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circ. Syst. Video Tech.*, vol. 13, no. 7, pp. 560–576, 2003.

Method

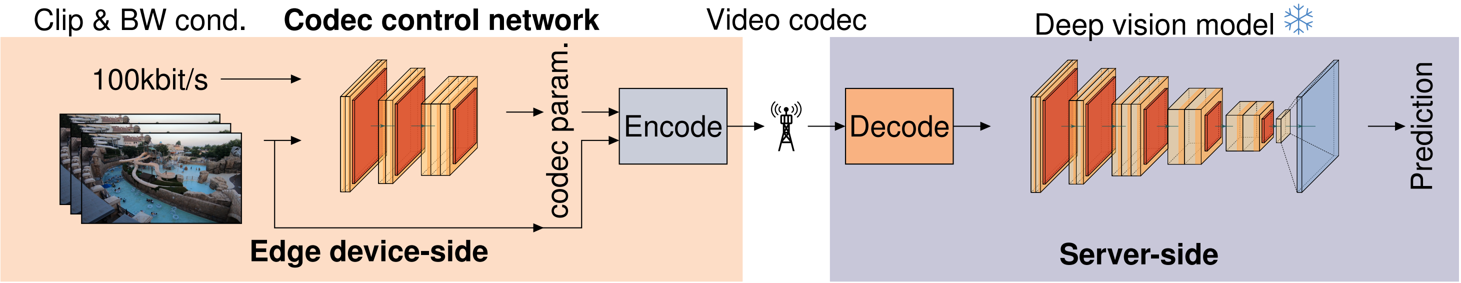

Clip & BW cond.
Codec control network
Video codec
100kbit/s
codec param.
Encode
Edge device-side

Method



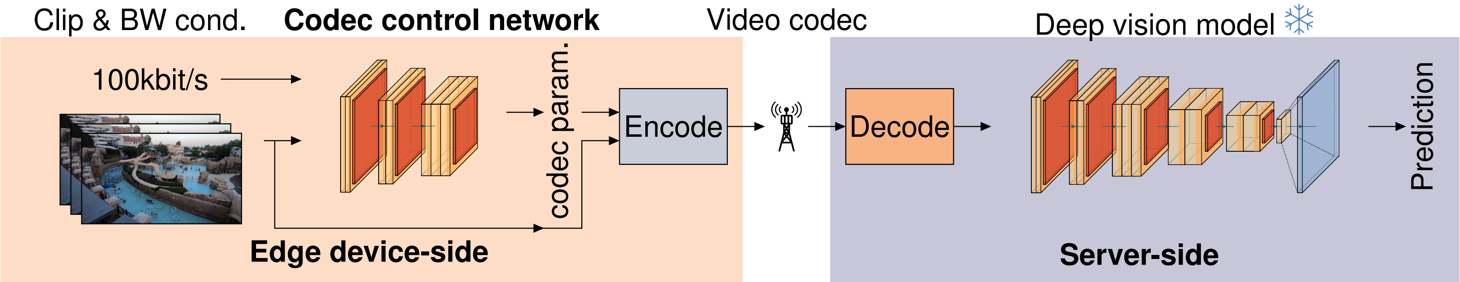
- Predict **high-dimensional codec parameters** s.t. vision performance is maximized

Method



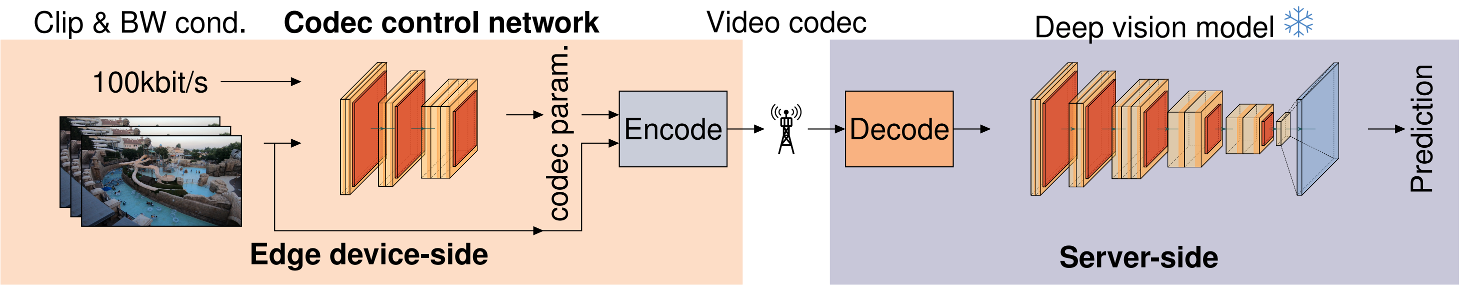
- Predict **high-dimensional codec parameters** s.t. vision performance is maximized

Method



- Predict **high-dimensional codec parameters** s.t. vision performance is maximized
- Encoded video bit-rate should not exceed bandwidth condition

Method



- Predict **high-dimensional codec parameters** s.t. vision performance is maximized
- Encoded video bit-rate should not exceed bandwidth condition
- Learn the control network in a **fully end-to-end setting**

Problem Formulation

$$\max_{\mathrm{QP}} \mathrm{M}(\qquad)$$

M Downstream metric (*e.g.*, mIoU)

Problem Formulation

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\qquad\qquad\qquad))$$

M	Downstream metric (*e.g.*, mIoU)
DNN	Downstream deep vision model (*e.g.*, DETR)

Problem Formulation

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \qquad)))$$

M	Downstream metric (*e.g.*, mIoU)
DNN	Downstream deep vision model (*e.g.*, DETR)
H.264	H.264 encoding-decoding mapping
V	Video clip to be coded of the shape $\mathbb{R}^{\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$

Problem Formulation

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_\theta(\mathbf{V}, b))))$$

M	Downstream metric (*e.g.*, mIoU)
DNN	Downstream deep vision model (*e.g.*, DETR)
H.264	H.264 encoding-decoding mapping
V	Video clip to be coded of the shape $\mathbb{R}^{\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$
C_θ	**Control network** (predicts macroblock-wise quantization parameters $\mathrm{QP} \in [0, 1, \ldots, 51]^{\mathrm{T}\times\mathrm{H}/16\times\mathrm{W}/16}$)

Problem Formulation

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_{\theta}(\mathbf{V}, b))))$$
$$\text{s.t.}\, \tilde{b} \leq b.$$

M	Downstream metric (*e.g.*, mIoU)
DNN	Downstream deep vision model (*e.g.*, DETR)
H.264	H.264 encoding-decoding mapping
$\mathbf{V}$	Video clip to be coded of the shape $\mathbb{R}^{\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$
C_{θ}	**Control network** (predicts macroblock-wise quantization parameters $\mathrm{QP} \in [0, 1, \dots, 51]^{\mathrm{T}\times\mathrm{H}/16\times\mathrm{W}/16}$)
b	Target bandwidth
$\tilde{b}$	Actual induced bandwidth

Problem with End-To-End Learning

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_{\theta}(\mathbf{V}, b))))$$
$$\text{s.t. } \tilde{b} \leq b.$$

Problem with End-To-End Learning

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_{\theta}(\mathbf{V}, b))))$$
$$\text{s.t.}\; \tilde{b} \leq b.$$

- ⚡ H.264 encoding-decoding is non-differentiable

Problem with End-To-End Learning

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_{\theta}(\mathbf{V}, b))))$$
$$\text{s.t. } \tilde{b} \leq b.$$

- ↯ H.264 encoding-decoding is non-differentiable
- ↯ Actual induced bandwidth is also non-differentiable

Problem with End-To-End Learning

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_\theta(\mathbf{V}, b))))$$
$$\text{s.t. } \tilde{b} \leq b.$$

- H.264 encoding-decoding is non-differentiable
- Actual induced bandwidth is also non-differentiable
- **Straight forward application of end-to-end learning not possible**

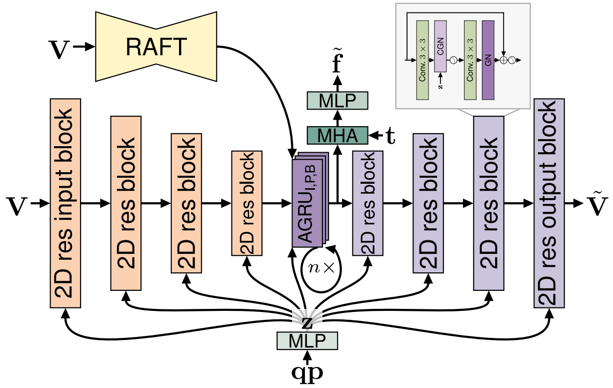
Differentiable Codec Surrogate Model

- Learn a differentiable surrogate model to approximate non-differentiable mappings

[5] W. Grathwohl *et al.*, "Backpropagation through the void: Optimizing control variates for black-box gradient estimation," in *ICLR*, 2018.

Differentiable Codec Surrogate Model

- Learn a differentiable surrogate model to approximate non-differentiable mappings

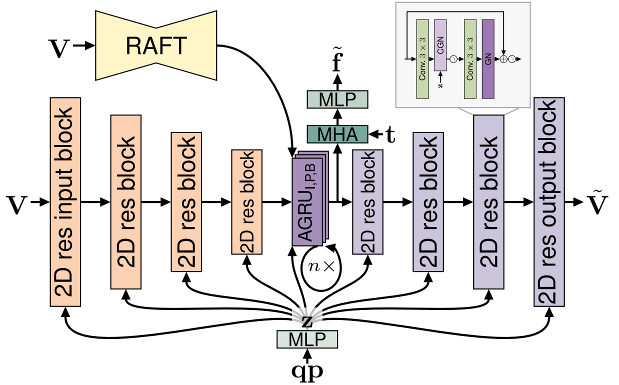


- We present a differentiable surrogate model predicting both the **coded video** and the **file size** (bandwidth)

[5] W. Grathwohl *et al.*, "Backpropagation through the void: Optimizing control variates for black-box gradient estimation," in *ICLR*, 2018.

Differentiable Codec Surrogate Model

- Learn a differentiable surrogate model to approximate non-differentiable mappings



- We present a differentiable surrogate model predicting both the **coded video** and the **file size** (bandwidth)
- Control variates theory used for learning the surrogate [5]

[5] W. Grathwohl *et al.*, "Backpropagation through the void: Optimizing control variates for black-box gradient estimation," in *ICLR*, 2018.

Surrogate Results

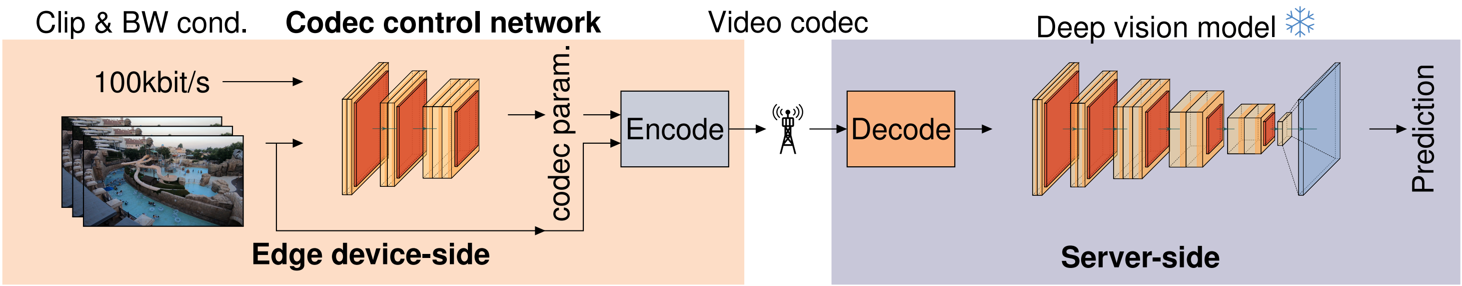
Surrogate Results

- **Our proposed surrogate approximates H.264 video distortion well**

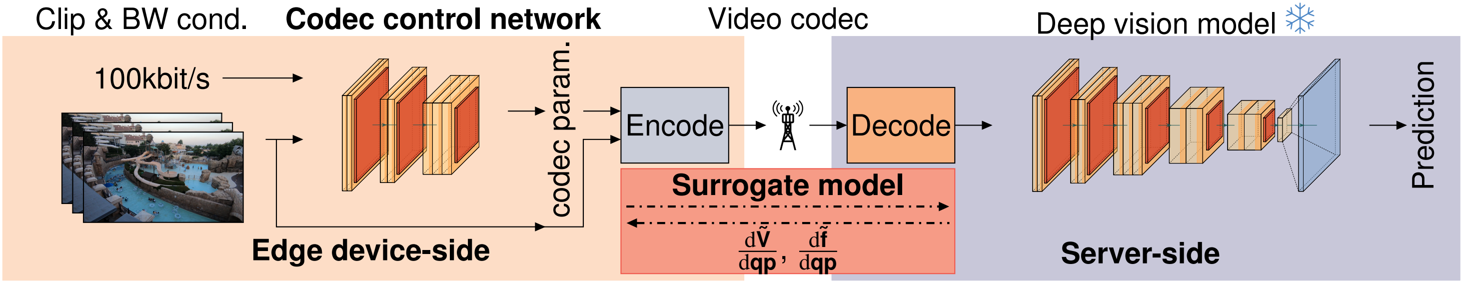
Surrogate Results

- **Our proposed surrogate approximates H.264 video distortion well**
- Relative file size (bandwidth) error typically **below 5%**

Deep Video Codec Control Pipeline

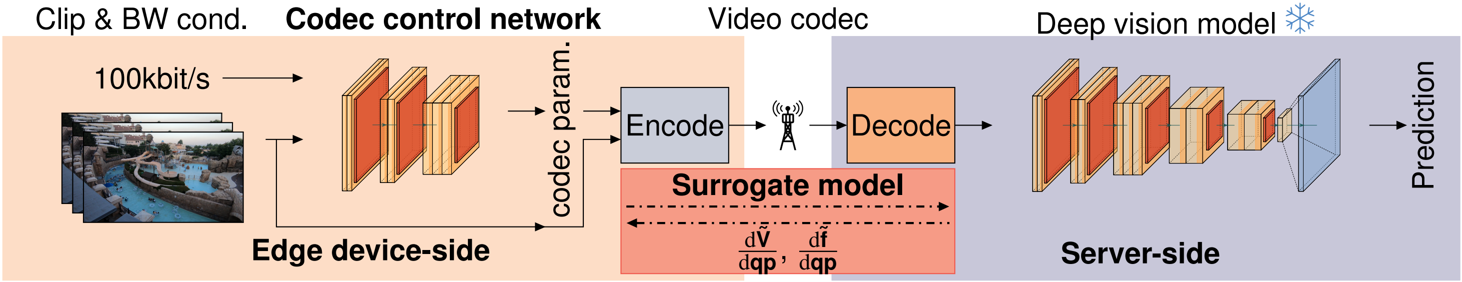


Deep Video Codec Control Pipeline



- Learn control network **end-to-end using the Lagrangian function** of the constrained optimization problem

Deep Video Codec Control Pipeline



- Learn control network **end-to-end using the Lagrangian function** of the constrained optimization problem
- We **regularize** the control network to generate a bandwidth close to the target bandwidth

Codec Control Results

Table: Semantic segmentation validation results on Cityscapes using a DeepLabV3 model.

Method	Bandwidth accuracy (%) ↑	Segmentation accuracy (%) ↑
	Cityscapes	
2-pass ABR (H.264)	68.13	64.29
Deep Video Codec Control	**96.22**	**84.79**
	CamVid	
2-pass ABR (H.264)	63.91	54.06
Deep Video Codec Control	**94.64**	**65.70**

Codec Control Results

Table: Semantic segmentation validation results on Cityscapes using a DeepLabV3 model.

Method	Bandwidth accuracy (%) ↑	Segmentation accuracy (%) ↑
	Cityscapes	
2-pass ABR (H.264)	68.13	64.29
Deep Video Codec Control	**96.22**	**84.79**
	CamVid	
2-pass ABR (H.264)	63.91	54.06
Deep Video Codec Control	**94.64**	**65.70**

- **Our Deep Codec Control consistently outperformed 2-pass ABR**

Codec Control Results

Table: Semantic segmentation validation results on Cityscapes using a DeepLabV3 model.

Method	Bandwidth accuracy (%) ↑	Segmentation accuracy (%) ↑
	Cityscapes	
2-pass ABR (H.264)	68.13	64.29
Deep Video Codec Control	**96.22**	**84.79**
	CamVid	
2-pass ABR (H.264)	63.91	54.06
Deep Video Codec Control	**94.64**	**65.70**

- **Our Deep Codec Control consistently outperformed 2-pass ABR**
- We preserve up to **20% more semantic accuracy** than 2-pass ABR

Downstream Task Transfer Result

Table: Transfer results of our Deep Video Codec Control from **optical flow estimation** $\rightarrow$ **semantic segmentation** on Cityscapes.

Training task	Bandwidth accuracy (%) $\uparrow$	Segmentation accuracy (%) $\uparrow$
Optical flow estimation	97.79	75.03
Trained on target task	96.22	84.79

Downstream Task Transfer Result

Table: Transfer results of our Deep Video Codec Control from **optical flow estimation** → **semantic segmentation** on Cityscapes.

Training task	Bandwidth accuracy (%) ↑	Segmentation accuracy (%) ↑
Optical flow estimation	97.79	75.03
Trained on target task	96.22	84.79

↯ Transferring between downstream task during inference leads to a drop in vision performance

Downstream Task Transfer Result

Table: Transfer results of our Deep Video Codec Control from **optical flow estimation** $\rightarrow$ **semantic segmentation** on Cityscapes.

Training task	Bandwidth accuracy (%) $\uparrow$	Segmentation accuracy (%) $\uparrow$
Optical flow estimation	97.79	75.03
Trained on target task	96.22	84.79

- ⚡ Transferring between downstream task during inference leads to a drop in vision performance
- Our end-to-end learned codec control learns a task-specific behavior

Conclusion

- **We present the first end-to-end learnable codec control for a standard codec**

Conclusion

- **We present the first end-to-end learnable codec control for a standard codec**
- Our Deep Video Codec Control adheres to existing **standardizations**, **optimizes vision performance**, and **performs rate control**

Conclusion

- **We present the first end-to-end learnable codec control for a standard codec**
- Our Deep Video Codec Control adheres to existing **standardizations**, **optimizes vision performance**, and **performs rate control**

Future research questions:

- How to support multiple downstream tasks with a single codec control?
- How to generalize our Deep Video Codec Control to other standard codecs (*e.g.*, H.265)?

A Perspective on Deep Vision Performance with Standard Image and Video Codecs

Deep Video Codec Control for Vision Models

Questions?

References

[1] C. E. Shannon, “Communication in the Presence of Noise,” *Proceedings of the IRE*, 1949.

[2] Y. Blau *et al.*, “Rethinking lossy compression: The rate-distortion-perception tradeoff,” in *ICML*, 2019.

[3] Y. Zhang *et al.*, “A survey on perceptually optimized video coding,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–37, 2023.

[4] T. Wiegand *et al.*, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circ. Syst. Video Tech.*, vol. 13, no. 7, pp. 560–576, 2003.

[5] W. Grathwohl *et al.*, “Backpropagation through the void: Optimizing control variates for black-box gradient estimation,” in *ICLR*, 2018.