Deep Video Codec Control for Vision Models

Christoph Reich[1,2,3,4], **Biplob Debnath**[1], **Deep Patel**[1], **Tim Prangemeier**[3], **Daniel Cremers**[2,4], **and Srimat Chakradhar**[1]

[1] NEC Laboratories America, Inc., [2] TU Munich, [3] TU Darmstadt, [4] Munich Center for Machine Learning (MCML)

Introduction

- Standard lossy video codecs are part of almost all real-world video processing pipelines
- Standardization is key to ensuring interoperability & low costs in real-world applications
- Existing standard codecs are *not optimized* for current deep vision models
- **We aim to optimize standard video codecs (e.g., H.264) for deep vision models**

Tab. 1. High-level comparison of our Deep Video Codec Control with existing approaches.

	Optimize vision performance	Rate control	ISO
Deep video codecs	✓	~	✗
Standard video codecs (e.g., H.264 [1])	✗	✓	✓
Deep video codec control	✓	✓	✓

Problem Formulation

Goal: Optimize deep vision performance for a given rate *w/o breaking standardization*

- Learn control network C_θ to predict macroblock-wise quantization parameters QP
- Stay within the available bandwidth budget (rate control)
- Maximize downstream performance of a deep vision model DNN (e.g., DeepLabV3)

$$\max_{\mathrm{QP}} \mathrm{M}(\mathrm{DNN}(\mathrm{H.264}(\mathbf{V}, \mathrm{C}_\theta(\mathbf{V}, b)))) \quad \text{s.t. } \hat{b} \leq b. \tag{1}$$

- Control problem can be formulated as a constrained optimization problem (*cf.* Eq. (1))

Motivation

How can we learn our Deep Video Codec Control such that performance (Eq. (1)) is maximized?

- ⚡ Video encoding-decoding is not differentiable
- ⚡ Reinforcement learning does not scale to large action spaces (high-dimensional QP)
- ~ Learning on a proxy task is suboptimal [3]
- ✓ **Learn a differentiable surrogate model of the video codec**
- 🚀 **Learn control by using end-to-end learning, optimizing Lagrange function of Eq. (1)**

tl;dr:

Standard codecs are not optimized for current deep vision models.

We present the first end-to-end learnable **Deep Video Codec Control** to optimize standard codecs for vision models w/o breaking standardization.

Deep Video Codec Control

We train our Deep Video Codec Control end-to-end using a downstream model and our differentiable codec surrogate.

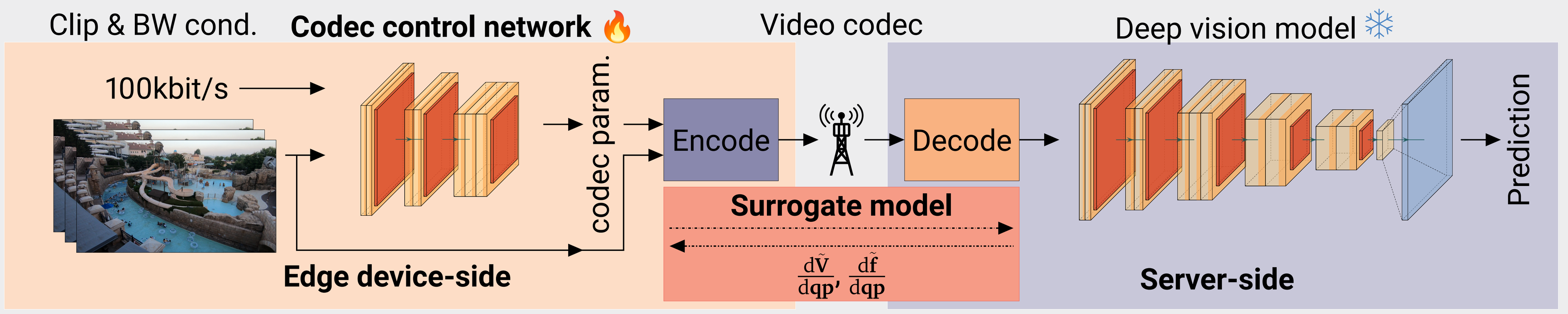

Fig. 1 The control network predicts high-dimensional codec parameters for an input clip and a given dynamic bandwidth condition.

- We use a loss between the downstream prediction on the *coded* and the *original video* to **maximize vision performance**
- For **rate control**, we penalize the control network using a bandwidth loss
- We regularize the control network to generate a bandwidth close to the target bandwidth

Differentiable Codec Surrogate

We learn a differentiable H.264 surrogate predicting both the coded clip and generated file size/bandwidth.

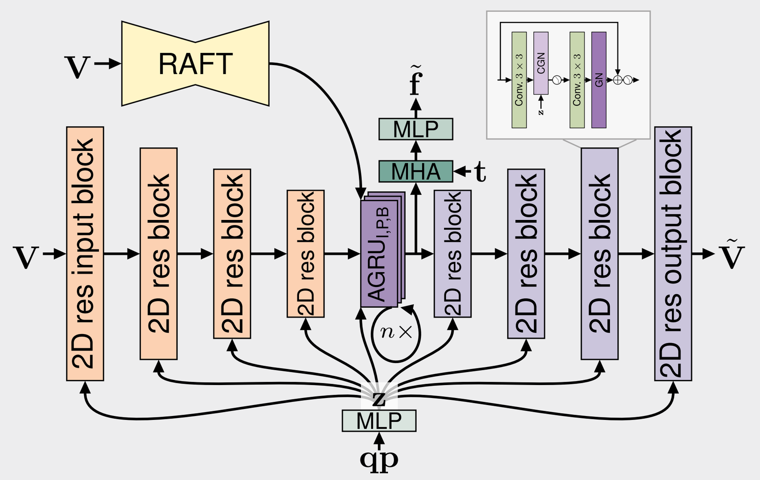

Fig. 2 Differentiable H.264 codec surrogate model.

- Control variates theory used for learning the surrogate [2]

Surrogate Results

Fig. 3 Qualitative surrogate model results.

- Our surrogate approximates H.264 coding well
- Relative file size error typically below 5%

Codec Control Results

We demonstrated the effectiveness of our Deep Video Codec Control on the tasks of semantic segmentation and optical flow estimation (see paper).

Tab. 2 Semantic segmentation validation results. BW ($\mathrm{acc}_{\mathrm{bw}}$) & segmentation accuracies ($\mathrm{acc}_{\mathrm{seg}}$) for difference BW tolerances reported. Metrics averaged over ten BW conditions.

	$\mathrm{acc}_{\mathrm{bw}}$ (%) ↑			$\mathrm{acc}_{\mathrm{seg}}$ (%) ↑		
Method	Δ0%	Δ2%	Δ5%	Δ0%	Δ2%	Δ5%
Cityscapes						
2-pass ABR (H.264)	68.13	74.98	82.31	64.29	70.57	77.07
Deep Video Codec Control	**96.22**	**97.05**	**97.91**	**84.79**	**85.50**	**86.28**
CamVid						
2-pass ABR (H.264)	63.91	74.43	85.36	54.06	62.49	**71.53**
Deep Video Codec Control	**94.64**	**95.61**	**96.46**	**65.70**	**62.52**	59.01

- 🚀 **Our Deep Codec Control consistently outperformed 2-pass ABR**
- 🚀 We are able to preserve up to 20% in semantic segmentation accuracy

We also analyzed the control performance when transferred between vision tasks

Tab. 3 Transfer results of our codec control from optical flow estimation to semantic segmentation on Cityscapes. We also report results when directly trained on semantic segmentation.

	$\mathrm{acc}_{\mathrm{bw}}$ ↑			$\mathrm{acc}_{\mathrm{seg}}$ ↑		
Training task	Δ0%	Δ2%	Δ5%	Δ0%	Δ2%	Δ5%
Optical flow estimation	97.79	98.31	98.90	75.03	75.37	75.76
Semantic segmentation	96.22	97.05	97.91	84.79	85.50	86.28

- Transferring between tasks leads to a drop in downstream performance
- 🚀 This demonstrates that out control learns a task-specific behavior

Conclusion

- **We demonstrate that learning an end-to-end deep codec control is feasible**
- Our Deep Video Codec Control outperforms traditional rate control modules

Future research questions:

- How to facilitate multiple downstream models and tasks [4]
- How to generalize our Deep Video Codec Control to other standard codecs (e.g., H.265 [5])

References

[1] 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.

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

[3] K. Du *et al.*, "AccMPEG: Optimizing video encoding for accurate video analytics," in *MLSys*, vol. 4, 2022, pp. 450–466.

[4] Y.-H. Chen *et al.*, "TransTIC: Transferring transformer-based image compression from human perception to machine perception," in *ICCV*, 2023.

[5] G. J. Sullivan *et al.*, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 22, no. 12, pp. 1649–1668, 2012.

Paper: https://arxiv.org/abs/2308.16215

Correspondence: c.reich@tum.de