Boosting Neural Image Compression for Machines Using Latent Space Masking

Kristian Fischer, Fabian Brand, and André Kaup

Multimedia Communications and Signal Processing
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Cauerstr. 7, 91058 Erlangen, Germany

Friedrich-Alexander-Universität
Technische Fakultät

1. Introduction

- Today, rising interest in image/video coding for machines where accuracy of analysis network defines coding quality
- Also, tremendous progress in field of learned image compression
- Learning weights $\boldsymbol{\theta}$ for human visual system (HVS):

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} D_{\mathrm{HVS}}(\boldsymbol{x}, f_{\mathrm{NCN}}(\boldsymbol{x}|\boldsymbol{\theta})) + \lambda \cdot R(f_{\mathrm{NCN}}(\boldsymbol{x}|\boldsymbol{\theta}))$$

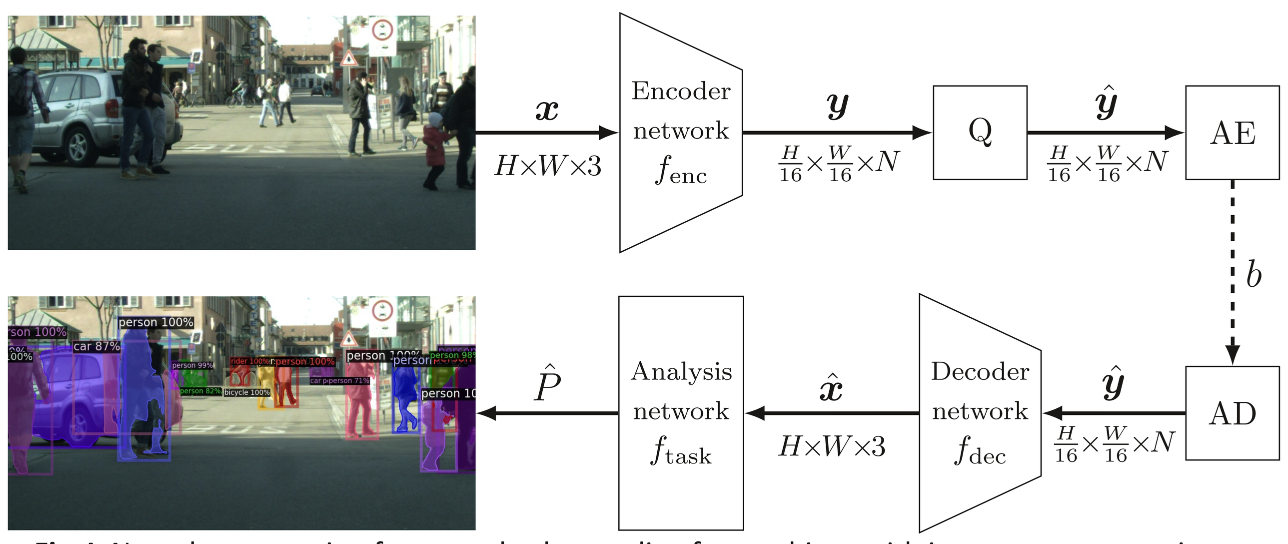


Fig. I: Neural compression framework when coding for machines with instance segmentation as analysis task. Upper and lower branch symbolize encoder and decoder side, respectively.

- Possibility to train the coding chain in end-to-end manner with task loss L_{task}

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} L_{\mathrm{task}}(f_{\mathrm{task}}(f_{\mathrm{NCN}}(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{\phi})) + \lambda \cdot R(f_{\mathrm{NCN}}(\boldsymbol{x}|\boldsymbol{\theta}))$$

- Problem: Saliency has to be learned implicitly by the neural image compression network (NCN)
- Proposal: latent space masking network (*LSMnet*) to mask out less salient elements of the latent representation $\boldsymbol{y}$

2. Latent Space Masking by LSMnet

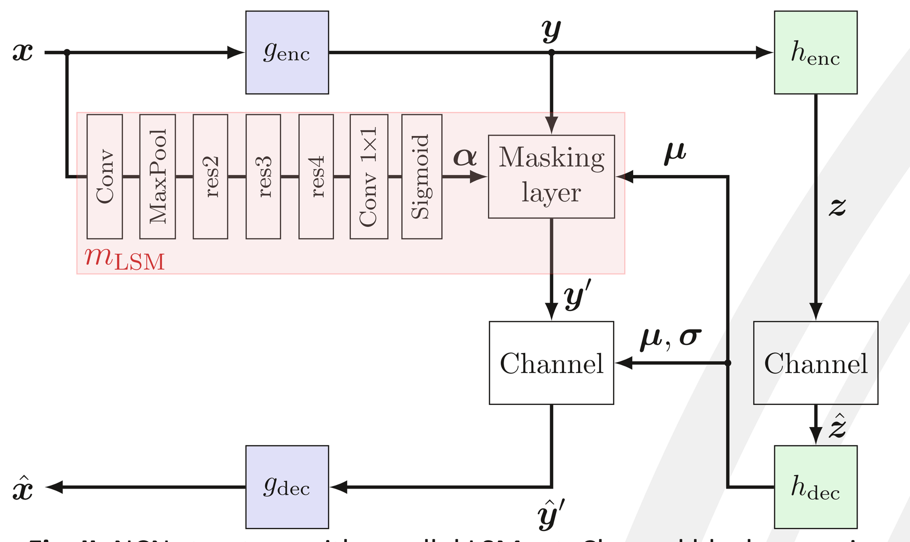


Fig. II: NCN structure with parallel LSMnet. Channel block comprises quantization and arithmetic coding.

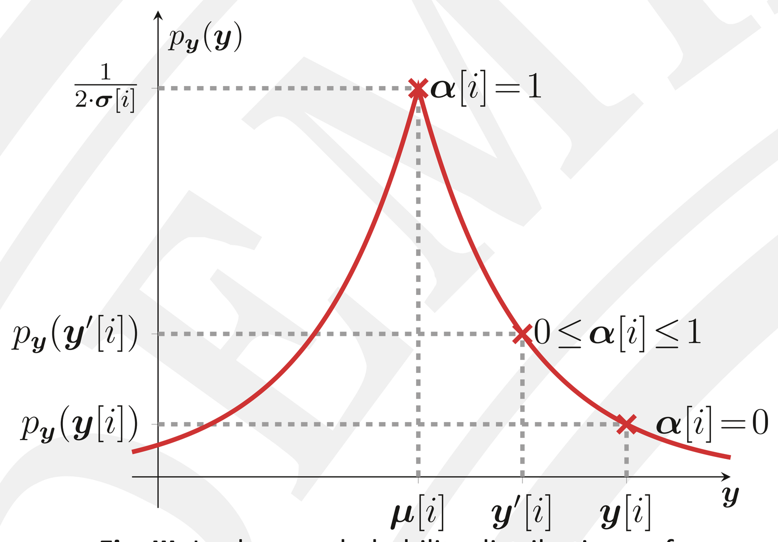


Fig. III: Laplace probability distribution p_y for a latent representation $\boldsymbol{y}$ at position i.

Concept

- LSMnet m_{LSM} generates features $\boldsymbol{\alpha}$ to soft mask the latent representation
- Elements that do not hold information for task of analysis network are transmitted with less accuracy to reduce bitrate
- Proposed soft masking scheme shifts the non-salient latents towards the estimated mean value $\boldsymbol{\mu}$ of Laplace distribution

$$\boldsymbol{y}'[i] = \boldsymbol{y}[i] - \boldsymbol{\alpha}[i] \cdot (\boldsymbol{y}[i] - \boldsymbol{\mu}[i])$$

Implementation

- Backbone features of analysis network already contain saliency information
- Thus, LSMnet consists of fixed backbone structure plus trainable 1x1 convolution and sigmoid layer
- Runs in parallel to NCN encoder g_{enc}
- Conjunct fine-tuning of NCN weights with LSMnet possible but not necessary

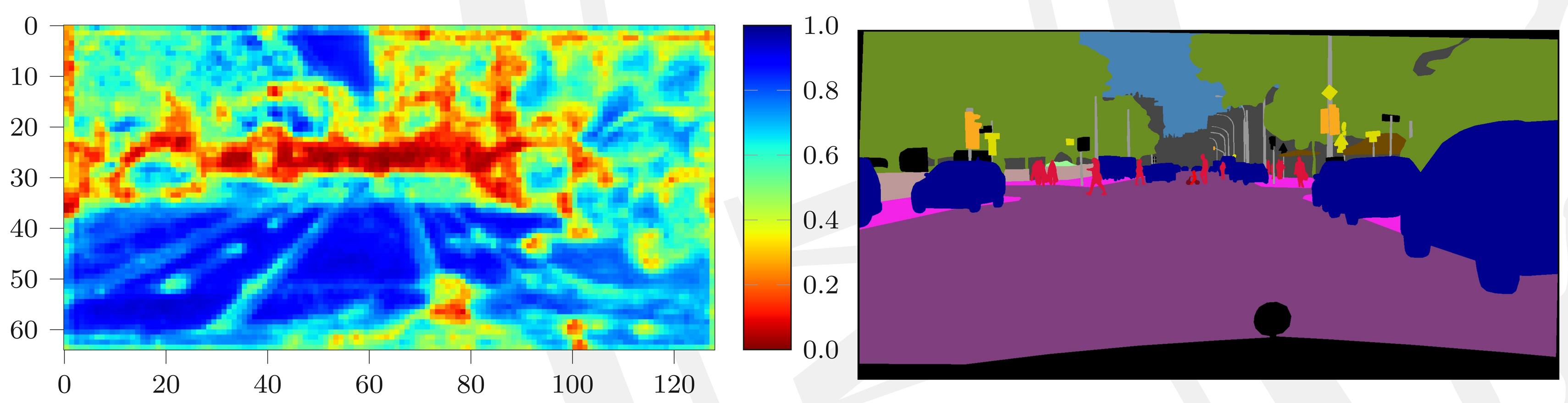


Fig. IV: Masking features $\boldsymbol{\alpha}$ generated by LSMnet (left) averaged over all channels for the Cityscapes input image frankfurt_000000_001236_leftImg8bit. Higher values with blue colors correspond to areas that are considered to be less important by LSMnet. Corresponding ground truth annotations are depicted on the right.

3. Analytical Methods

Training Procedure

- Basic NCN without LSMnet similar to [1] trained for 1000 epochs on Cityscapes (CS) training dataset [2] end-to-end with analysis network
- Training of LSMnet 1x1 convolution for 100 additional epochs
- Tested different backbone structures trained on different tasks and datasets

Experimental Setup

- Compression of 500 Cityscapes validation images
- Instance segmentation network Mask R-CNN [3] with ResNet50 FPN backbone as analysis network
- Detection accuracy is measured with weighted average precision (wAP) [4]
- VVC [5] test model (VTM-10.0) as reference codec

[1] D. Minnen, J. Ballé, and G. D. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression," NIPS, Dec. 2018.
[2] M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. CVPR, Jun. 2016,
[3] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in Proc. ICCV, Oct. 2017
[4] K. Fischer, C. Herglotz, and A. Kaup, "On Intra Video Coding and In-loop Filtering for Neural Object Detection Networks," in Proc ICIP, Oct. 2020
[5] B. Bross et al., "Overview of the Versatile Video Coding (VVC) Standard and its Applications," TCSVT, Oct. 2021

4. Experimental Results

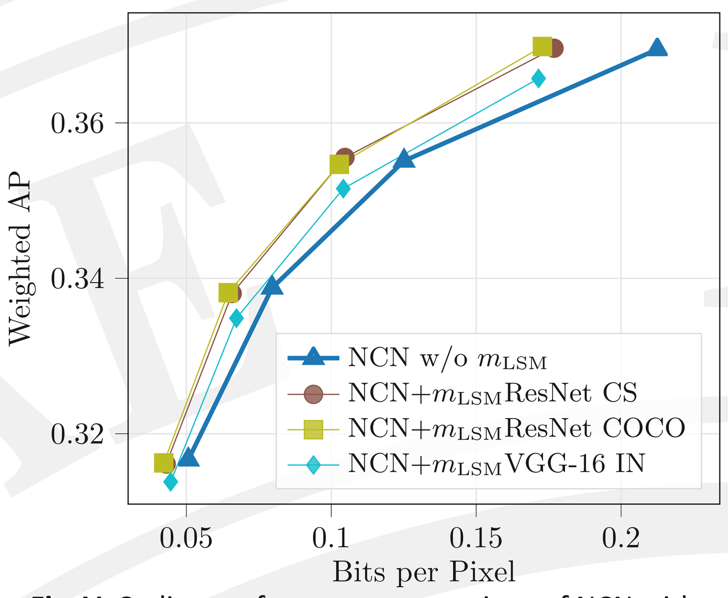


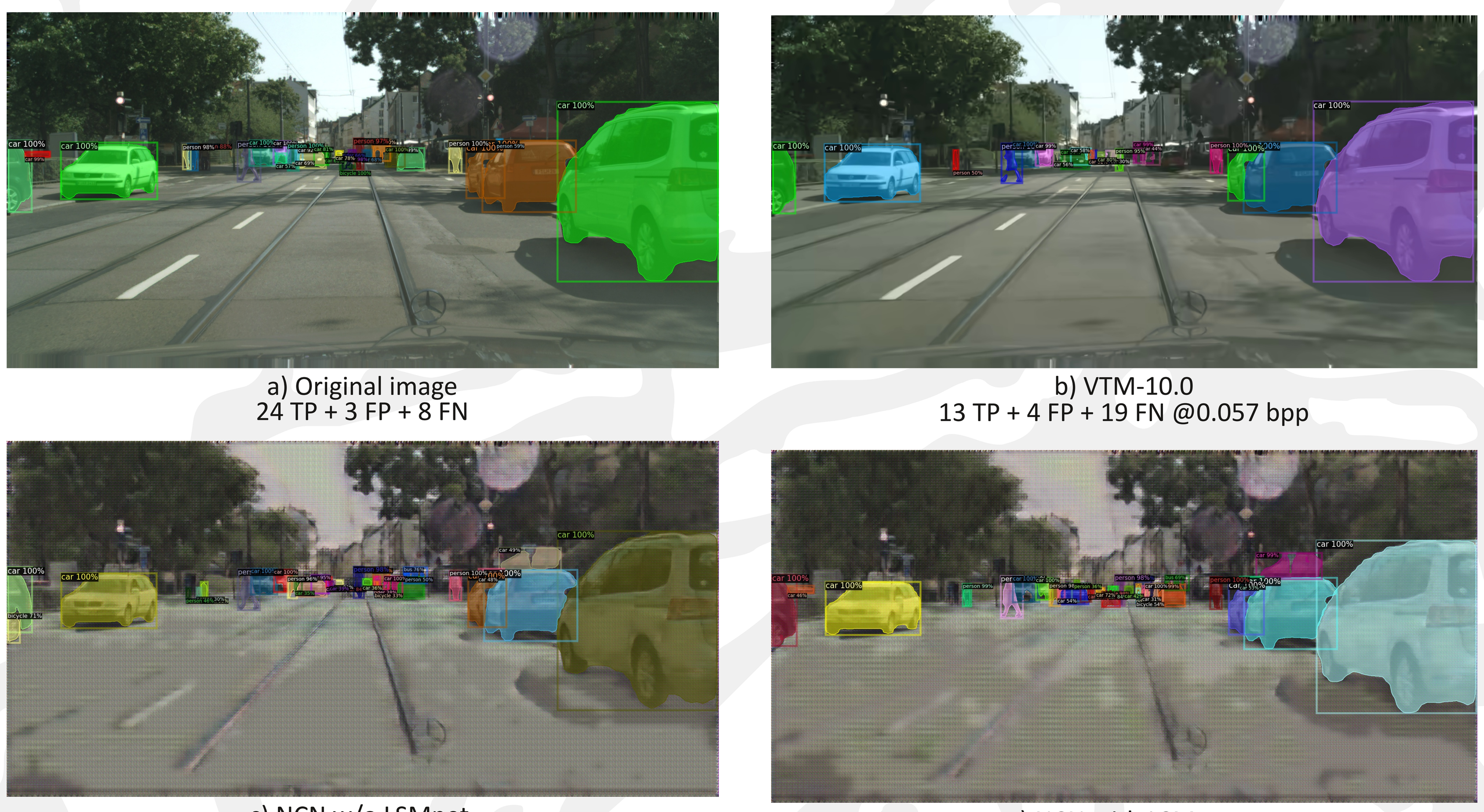
Fig. V: Coding performance comparison of NCN with or without LSMnet. Here: only the 1x1 convolution layer of LSMnet was trained.

Backbone m_{LSM}	Freeze NCN	BDR wAP (NCN w/o m_{LSM})	BDR wAP (VTM-10.0)
ResNet CS	yes	-16.3 %	-51.0 %
ResNet CS	no	**-27.3 %**	**-54.3 %**
ResNet COCO	yes	-17.4 %	-51.6 %
ResNet COCO	no	-25.5 %	**-54.3 %**
VGG-16 IN	yes	-7.1 %	-47.8 %
VGG-16 IN	no	3.7 %	-40.6 %

Tab. I: Bjøntegaard delta rate values (BDR) for comparing coding performance of NCNs with additional LSMnet. Anchor method is given in parentheses. Best values are set in bold.

- All NCNs with LSMnet outperform the reference model without LSMnet
- Masking latents reduces bitrate while maintaining detection accuracy
- Improved performance if LSMnet backbone has been trained on same task and dataset as analysis network
- Fine-tune the NCN weights with LSMnet results in even higher coding gains of 27.3 % over the NCN without LSMnet and 54.3 % over VTM-10.0

5. Conclusion



a) Original image
24 TP + 3 FP + 8 FN

b) VTM-10.0
13 TP + 4 FP + 19 FN @0.057 bpp

c) NCN w/o LSMnet
18 TP + 8 FP + 14 FN @0.045 bpp

c) NCN with LSMnet
18 TP + 9 FP + 14 FN @0.039 bpp

Fig. VI: Visual Example for coding frankfurt_000000_001236_leftImg8bit.

- Adding LSMnet to existing NCN architecture results in superior coding performance when coding for an analysis network
- This does not necessarily require a complete re-training of the NCN
- Decoder structure remains untouched
- Visual quality is strongly degraded in non-salient areas
- Possible application of LSMnet also when coding for human visual system

SVCP, July 2022