Towards Image Understanding from Deep Compression Without Decoding

Robert Torfason, Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, Luc Van Gool

ICLR 2018

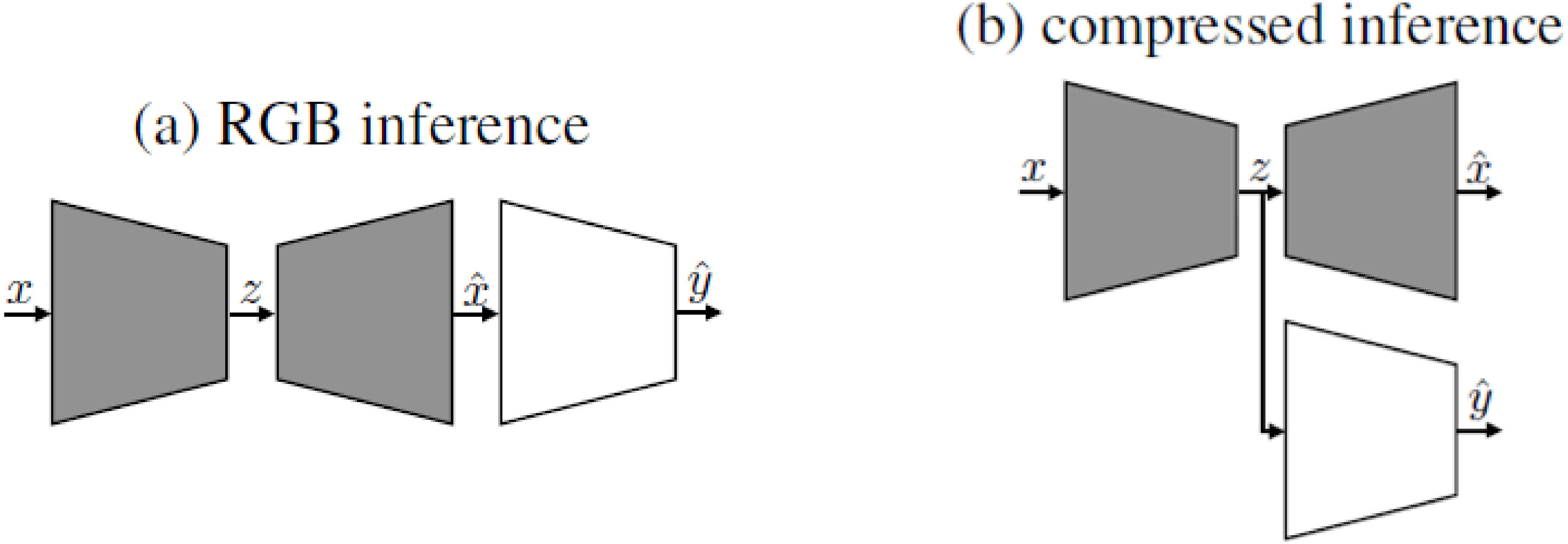
STAT 946 – Deep Learning – Fall 2018
November 06th, 2018

Presented By:
Aravind Ravi (20752644)
MASc, Systems Design Engineering
Engineering Bionics Lab

Motivation

Authors propose to perform **inference** from **compressed representations** without decoding the RGB image

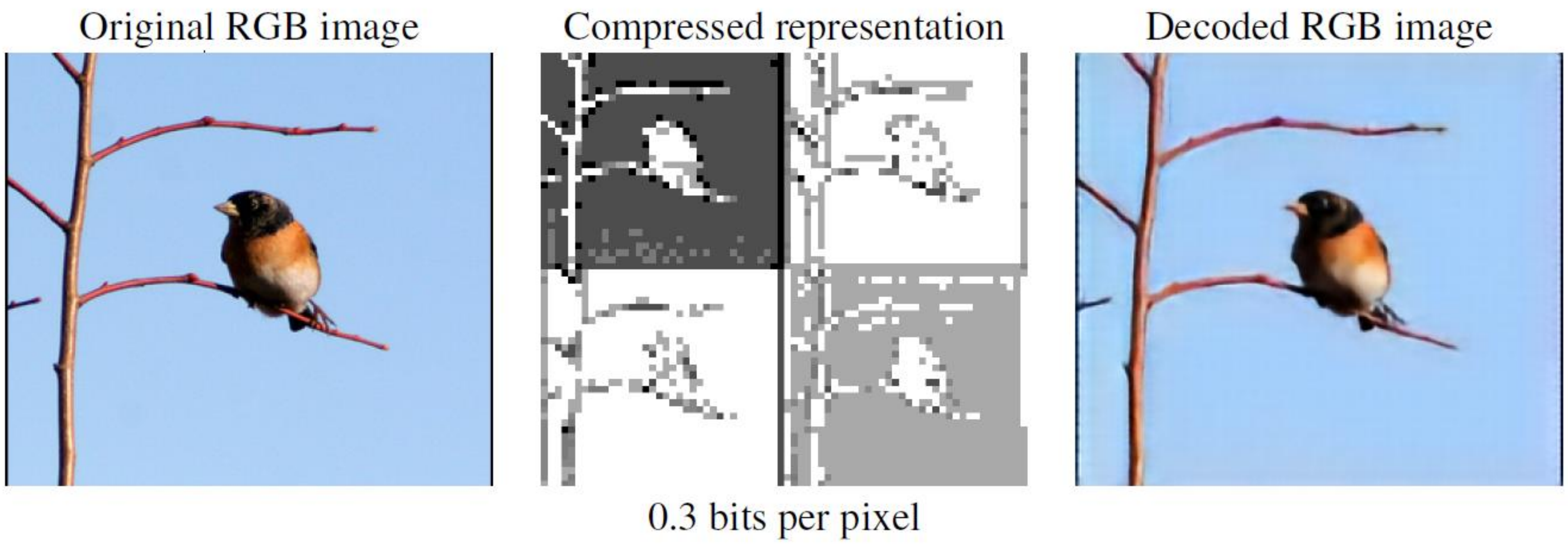
- Bypasses the process of decoding the image into the RGB space before classification
- Reduces the overall computational complexity up to 2 times



Contributions

- Image Classification and Semantic Segmentation from Compressed representations
  - Reducing the computational complexity by 2 times

- Joint training for image compression and classification
  - Improves quality of the image and increase in accuracy of classification and segmentation



Related Work

Prior work - **Uses engineered codecs for inference tasks**

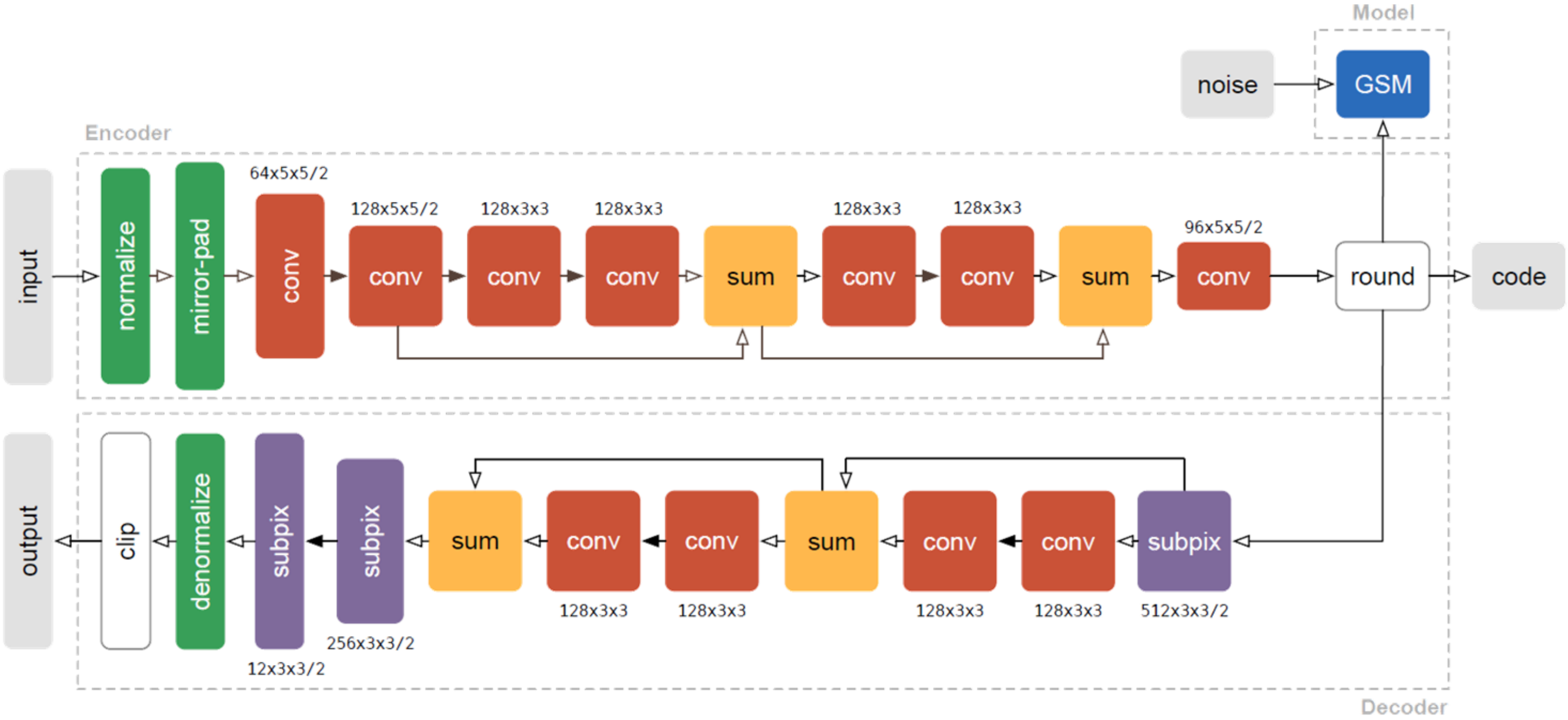
- Classification of compressed hyperspectral images
- Discrete Cosine Transform based compression performed on images before feeding into a neural network, which shows an improvement in training speed by up to 10 times
- Video analysis on compressed video (using engineered codecs)

Proposed Method - Perform inference from learned feature representation

Authors Claim

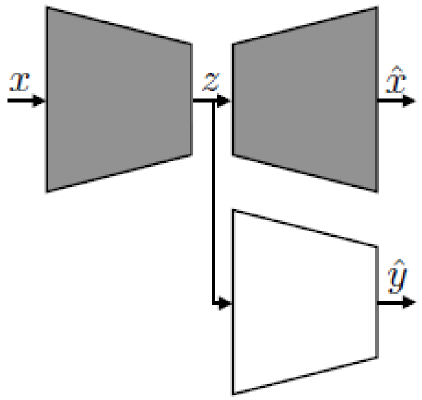
So far there hasn't been any work **using learned compressed representations** for **image classification and segmentation**

Learned Deeply Compressed Representations



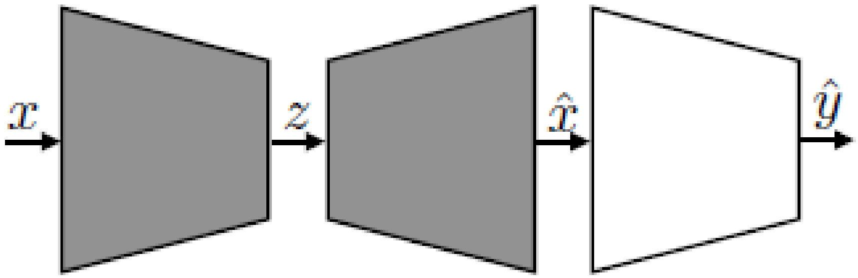
Compression Architecture (Theis et al. 2017)

(b) compressed inference



- Input – 224x224 Images
- Compressed Output – 28x28xC
  C - Number of Channels
- Z – Quantized compressed representation

Learned Deeply Compressed Representations



- Quantization introduces a distortion D on $\hat{x}$ with respect to x
- Length of the bitstream is measured by the rate R (Also measured in terms of Entropy)
- To Train, Rate-Distortion Trade-Off is minimized, given as:

$$D + \beta R$$

Learned Deeply Compressed Representations

The loss function is thus formulated as:

$$\mathcal{L}_c = \underbrace{\mathrm{MSE}(x, \hat{x})}_{Distortion} + \beta \underbrace{\max(H(q) - H_t, 0)}_{Rate}$$

- Metric for D is the mean squared error (MSE) between x and $\hat{x}$
- R is estimated using H(q) where H(q) is the entropy of the probability distribution over the symbols
- The trade-off is controlled by adjusting β
- For each β an operating point is derived for which the images have a certain bitrate (measured as bits per pixel - BPP)
- Three operating points at 0.0983 bpp (C=8), 0.330 bpp (C=16), and 0.635 bpp (C=32) are obtained empirically

Ref.: Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., & Gool, L. V. (2017). Soft-to-hard vector quantization for end-to-end learning compressible representations. In Advances in Neural Information Processing Systems (pp. 1141-1151).

Image Classification from Compressed Representations

Classification on RGB Images

The authors use Residual Networks (ResNet-50) architecture to perform image classification on RGB images. The authors modify the **ResNet-50** to obtain **ResNet-71**

Classification on Compressed Representations

Three other architectures are created (removing blocks larger than spatial dimensions of 28x28)

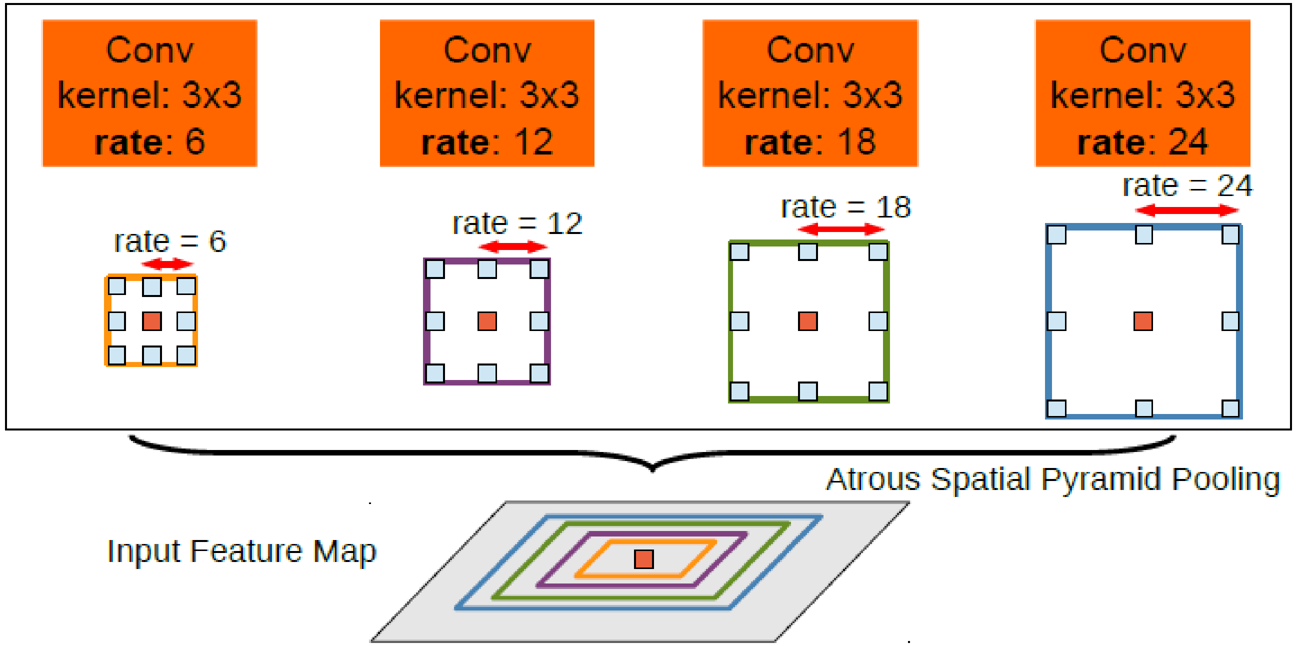
- **cResNet-39** for compressed representations as input
- To match the computational complexity of ResNet-50 and ResNet-71 - **cResNet-51** and **cResNet-72** are created

Network	**root**	**conv2_x** 56×56	**conv3_x** 28×28	**conv4_x** 14×14	**conv5_x** 7×7	**FLOPs** $[\times 10^9]$
ResNet-50	yes	3	4	*6*	3	3.86
ResNet-71	yes	3	4	*13*	3	5.38
cResNet-39	no	none	4	*6*	3	2.95
cResNet-51	no	none	4	*10*	3	3.83
cResNet-72	no	none	4	*17*	3	5.36

Semantic Segmentation from Compressed Representations

The ResNet based Deep Lab architectures are adapted in this paper as follows:

- Atrous Convolutions – Filter with holes
- Atrous Spatial Pyramid Pooling



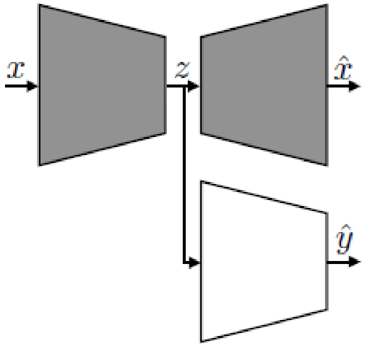
- The filters are upsampled instead of downsampling the feature maps.
- This is done to increase their receptive field and to prevent aggressive subsampling of the feature maps
- Rate corresponds to the number of zeros between the filter values
- Extract features in separate branches and fuse them to generate final result

Source: Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

Joint Training for Compression and Image Classification

Joint training strategy - Combine **compression and classification** tasks
Combines the compression network and the cResNet-51 architecture

(b) compressed inference



All parts, encoder, decoder, and inference network, are trained at the same time

Loss Function for joint training:

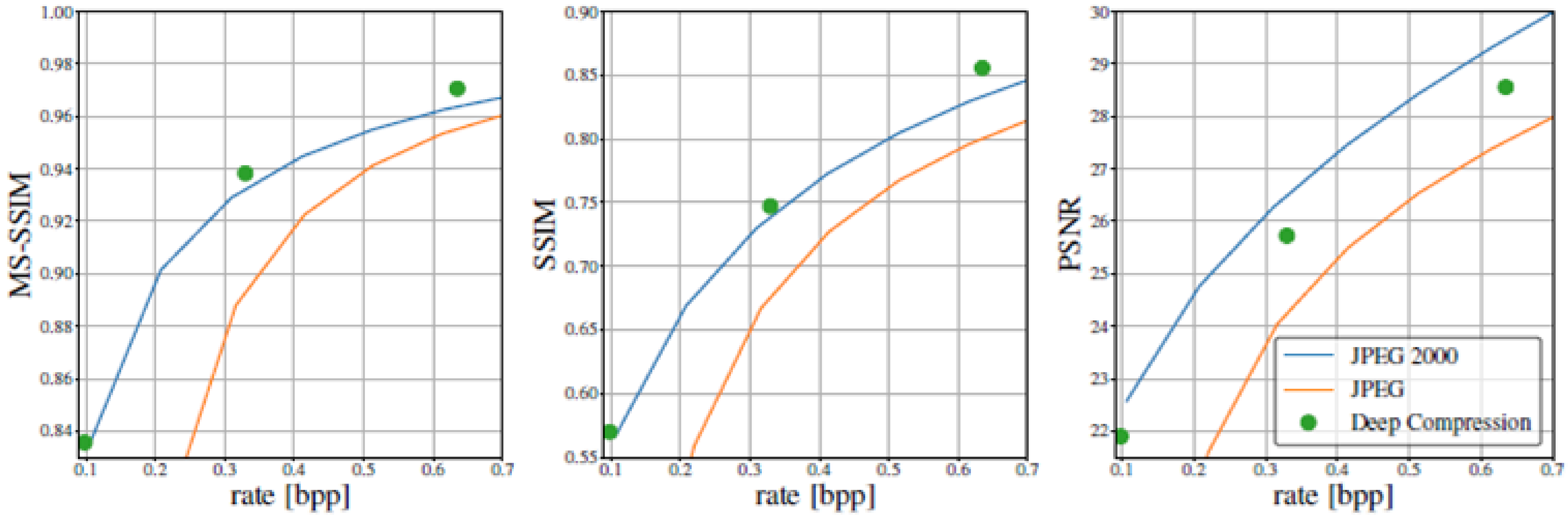
$$\mathcal{L}_c = \gamma\underbrace{(\mathrm{MSE}(x,\hat{x}) + \beta\max(H(q)-H_t,0))}_{Rate - Distortion\ TradeOff\ - Compression} + \overbrace{l_{ce}(y,\hat{y})}^{Categorical\ Cross - Entropy\ Loss\ - Classification}$$

γ – trade-off between compression loss and classification loss

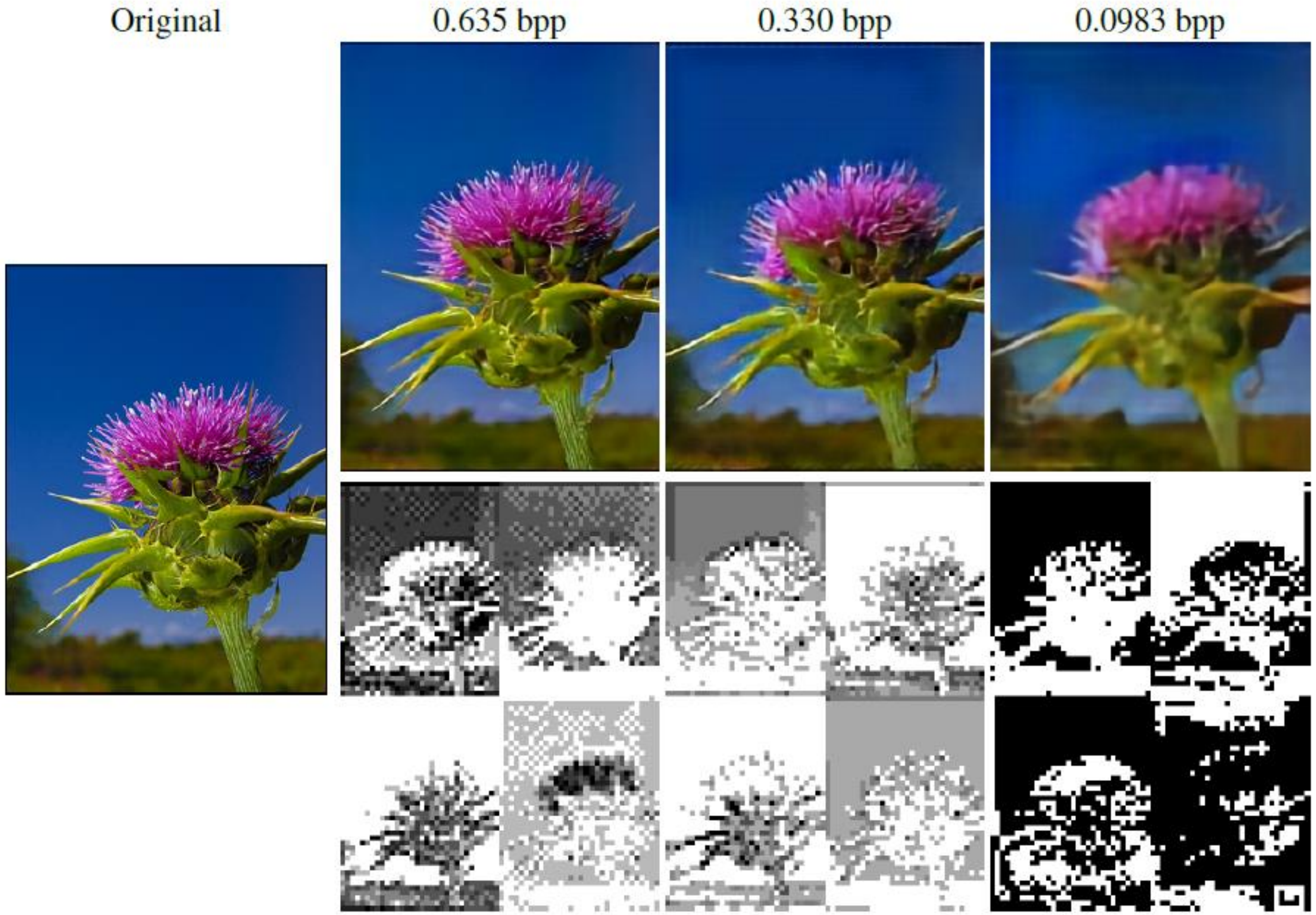
Learned Deeply Compressed Representations Results

PSNR – Peak Signal to Noise Ratio
SSIM – Structural Similarity Index
M-SSIM – Multi-Scale Structural Similarity Index

Dataset: ILSVRC2012 dataset



Learned Deeply Compressed Representations Results



- 4 channels with highest entropy
- As the rate gets lower the entropy cost forces the compressed representations to use fewer quantization levels
- Most aggressive rates, the channels map to only 2 levels of quantization

Classification on Compressed Representations Results

Dataset: ILSVRC2012

bpp	Network architecture	Top 5 acc. [%]	Top 1 acc. [%]
	Resnet-50	89.96	71.06
0.635	ResNet-50	88.34	68.26
	cResNet-51	87.85	67.68
	cResNet-39	87.47	67.17
0.330	ResNet-50	86.25	65.18
	cResNet-51	85.87	64.78
	cResNet-39	85.46	64.14
0.0983	ResNet-50	78.52	55.30
	cResNet-51	78.20	55.18
	cResNet-39	77.65	54.31
	ResNet-71	79.28	56.23
	cResNet-72	79.02	55.82

Classification:
Similar to that on RGB images

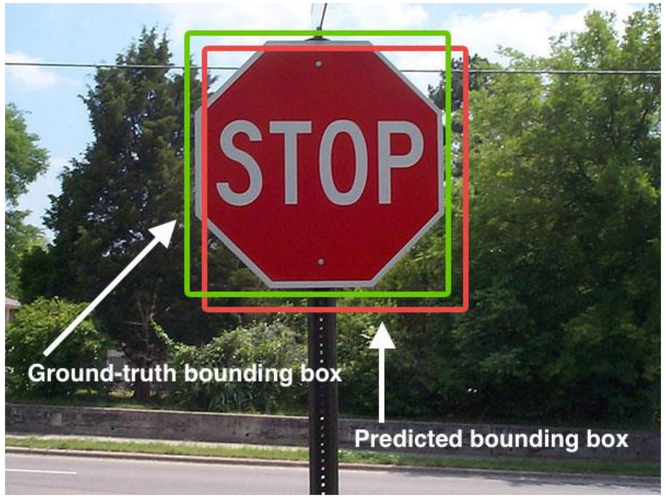
Computational Gains:
At 0.635 bpp the ImageNet dataset requires 24.8 GB of storage space instead of 144 GB for the original version, a reduction by a factor 5.8 times

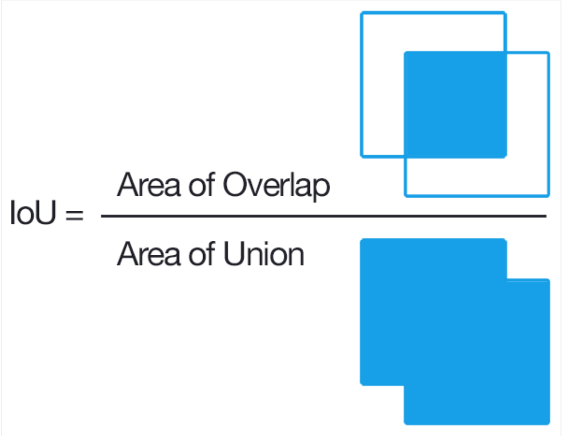
Segmentation Results

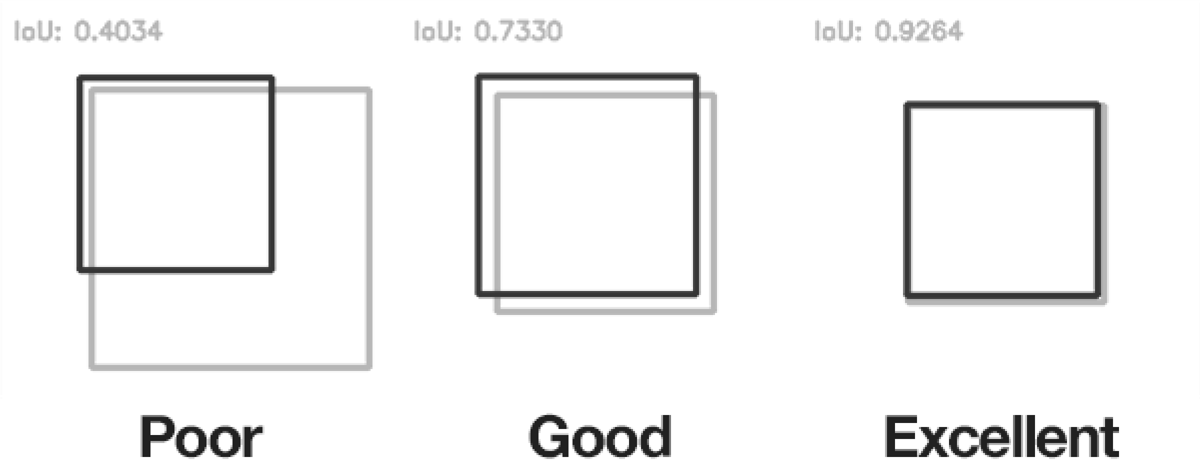
Dataset: PASCAL VOC-2012 dataset

bpp	Network architecture	mIoU [%]
	Resnet-50	65.75
0.635	ResNet-50	62.97
0.635	cResNet-51	62.86
0.635	cResNet-39	61.85
0.330	ResNet-50	60.75
0.330	cResNet-51	61.12
0.330	cResNet-39	60.78
0.0983	ResNet-50	52.97
0.0983	cResNet-51	54.62
0.0983	cResNet-39	53.51
0.0983	ResNet-71	54.55
0.0983	cResNet-72	55.78

mIoU – Mean Intersection over Union

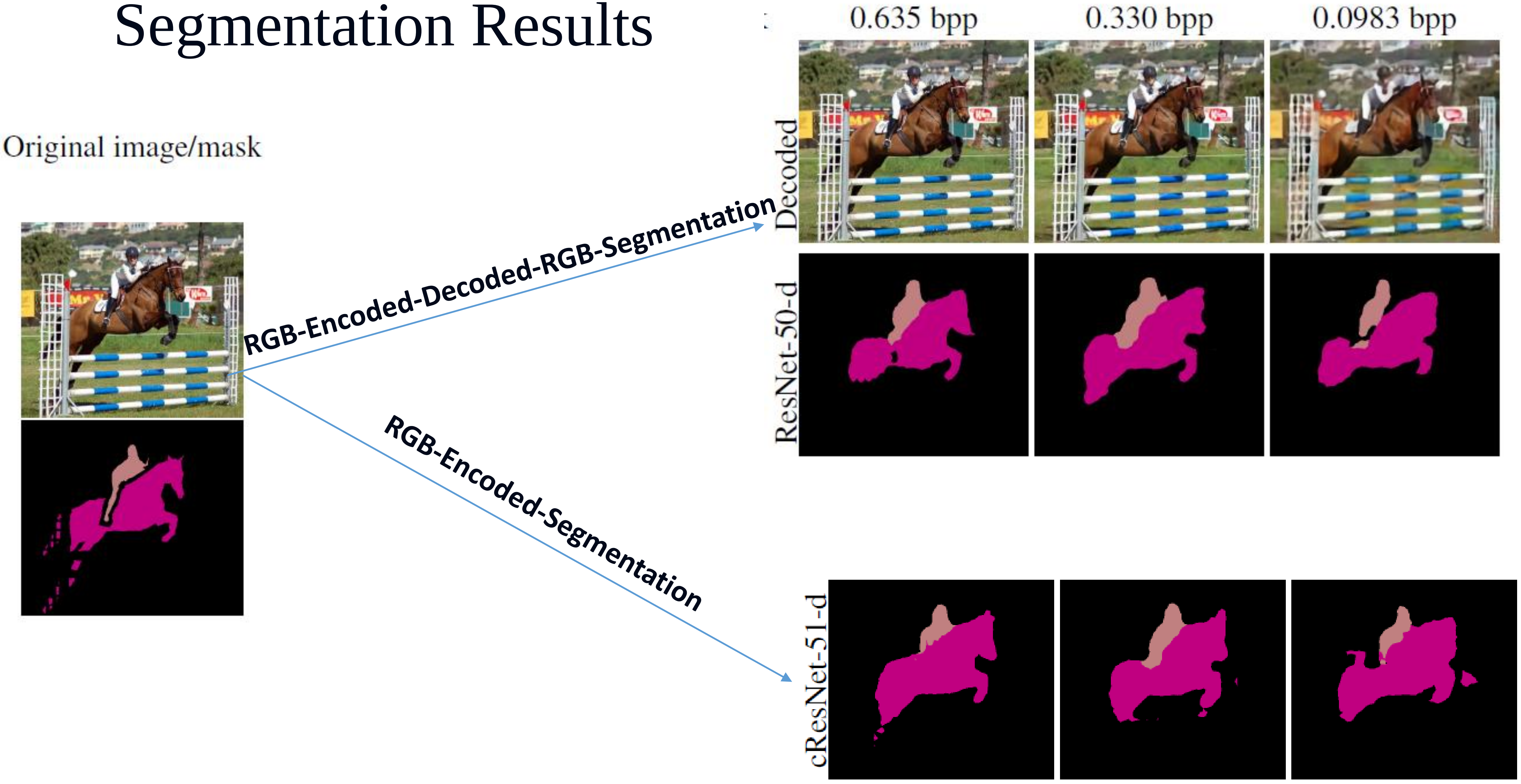




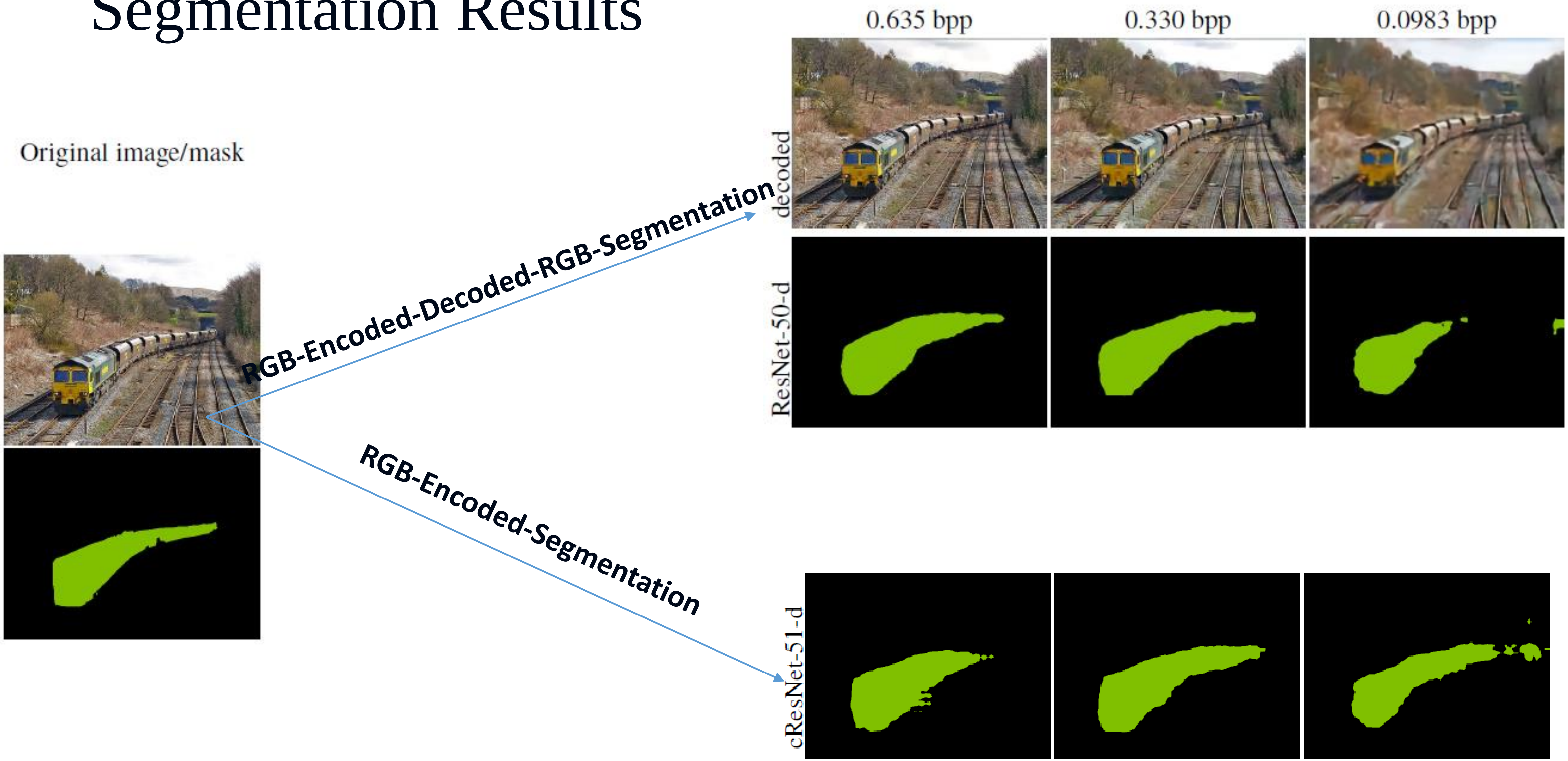


Source: https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

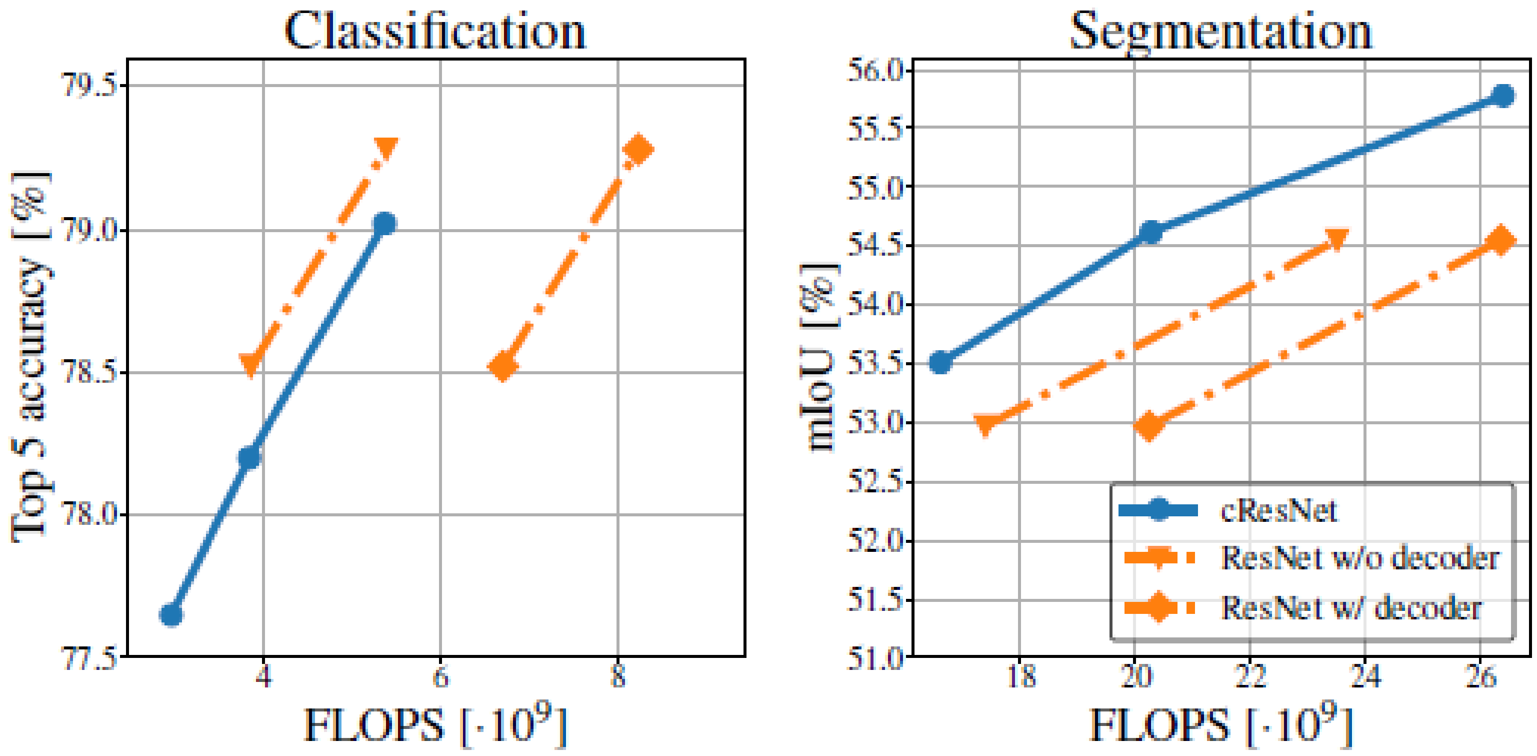
Segmentation Results



Segmentation Results

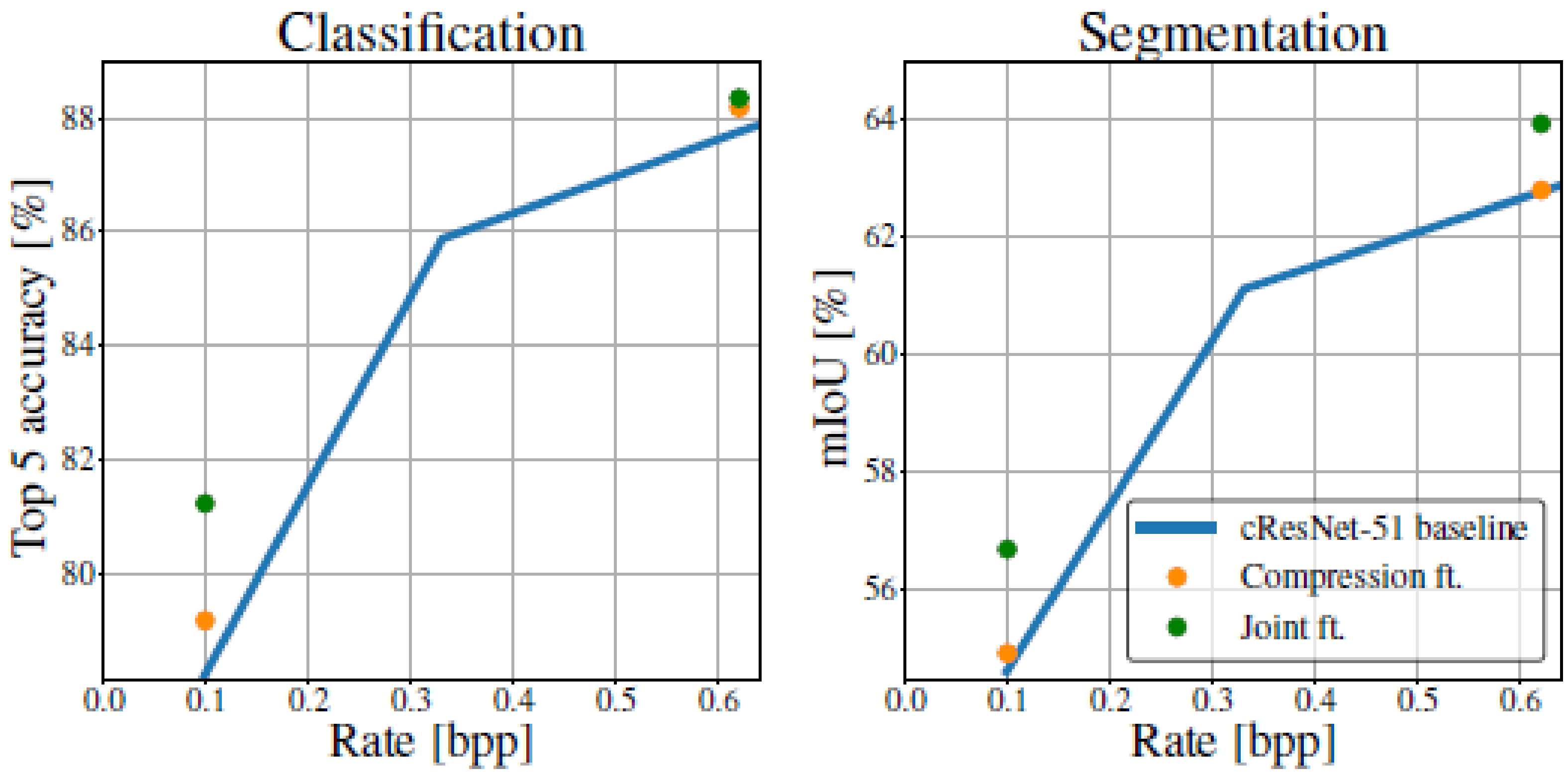


Computational Gains Results



Operating Point: 0.0983 bpp

Joint Training for Compression and Image Classification Results



Critique

Positive Points

The work has provided **extensive experimental evaluation** and evidence that suggests that learned compressed representations can be effective in classification and segmentation tasks

Applications of this can be in multimedia communication, wireless transmission of images, video surveillance on the mobile edge, conserve wireless bandwidth, savings on storage while retaining the perceptual quality of images

Drawbacks

The authors mention that the **complexity of the current approach** is still high in comparison with methods like JPEG or JPEG2000. Can be **overcome** when the networks are trained and **run on dedicated GPUs**.

Providing extensive experimental contributions, the authors have **written a quite lengthy paper**. There are parts of the paper where the ideas have been repeated frequently, and **could've compressed the paper** for a more well balanced length.

Thank You

Any Questions?

References

- Torfason, R., Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., & Van Gool, L. (2018). Towards image understanding from deep compression without decoding. arXiv preprint arXiv:1803.06131.
- Theis, L., Shi, W., Cunningham, A., & Huszár, F. (2017). Lossy image compression with compressive autoencoders. arXiv preprint arXiv:1703.00395.
- Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., & Gool, L. V. (2017). Soft-to-hard vector quantization for end-to-end learning compressible representations. In Advances in Neural Information Processing Systems (pp. 1141-1151).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.