# Deep Learning for Image and Video Compression

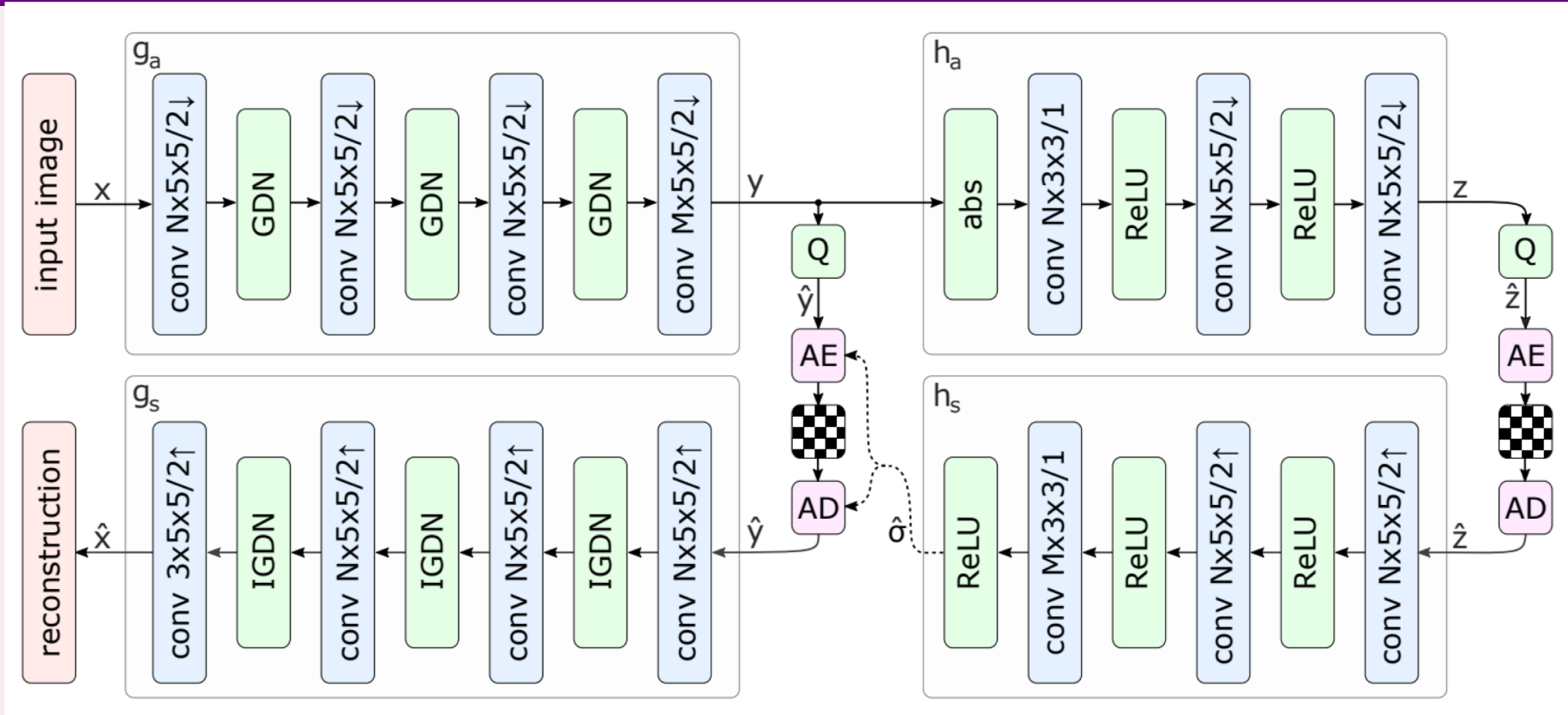
Yao Wang
Dept. of Electrical and Computer Engineering
NYU Wireless
Tandon School of Engineering
New York University
wp.nyu.edu/videolab


AOMedia Research Symposium, Oct, 2019, San Francisco

# Outline

- Learnt image compression using variational encoders
  - Framework of Balle et al.
  - Improvement using nonlocal attention maps and masked 3D convolution for conditional entropy coding (with Zhan Ma, Nanjing Univ.)
  - Scalable extension
- Learnt video compression (with Zhan Ma, Nanjing Univ.)
- Exploratory work:
  - Video prediction using dynamic deformable filters
  - Block-based image compression by denoising with side information

# Image Compression Using Variational Autoencoder (General Framework)

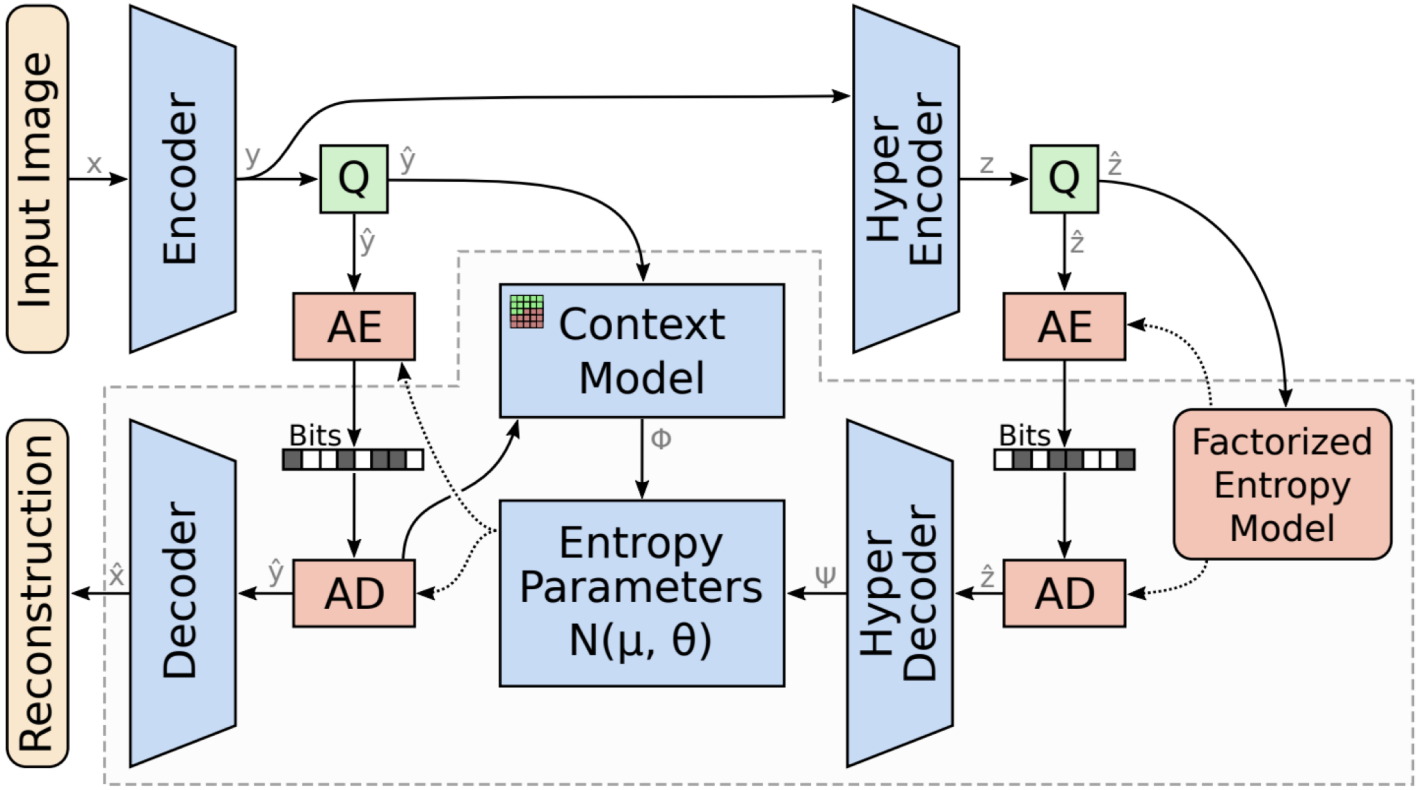


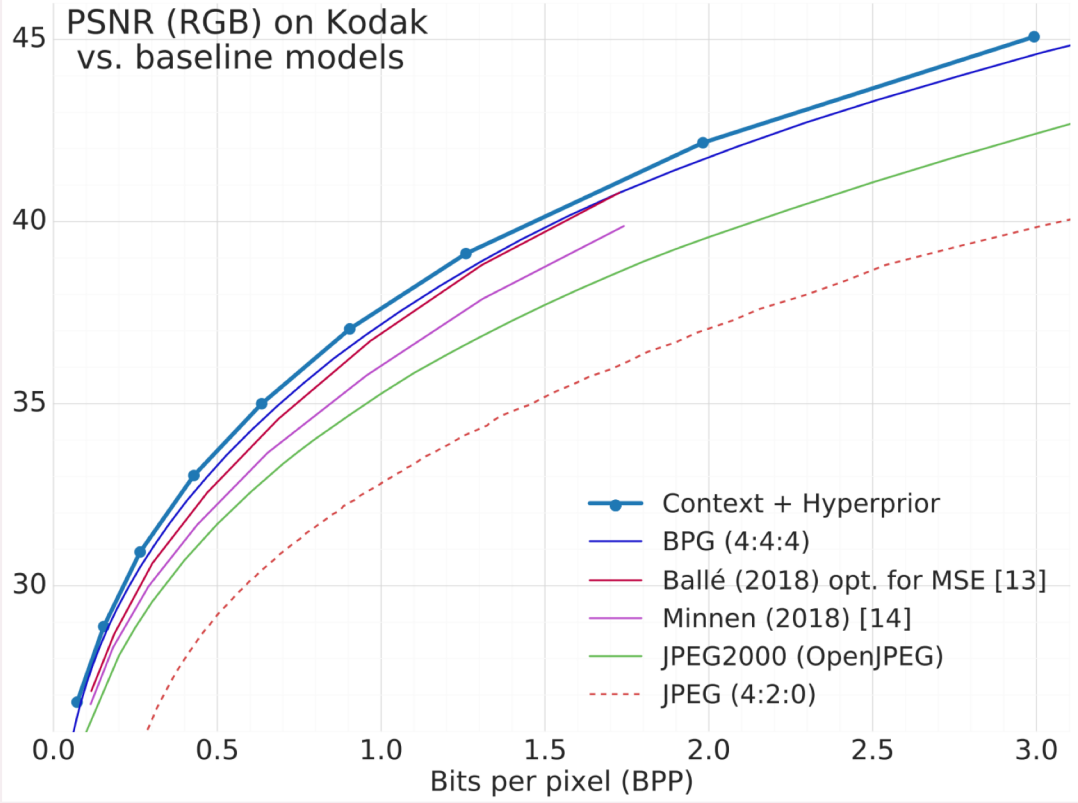

y: features describing image

z (hyper priors): features for estimating marginal probability model parameters for y (STD of Gaussian)

[Balle2018] J. Balle, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," ICLR 2018

# VAE Using Autoregressive Context Model



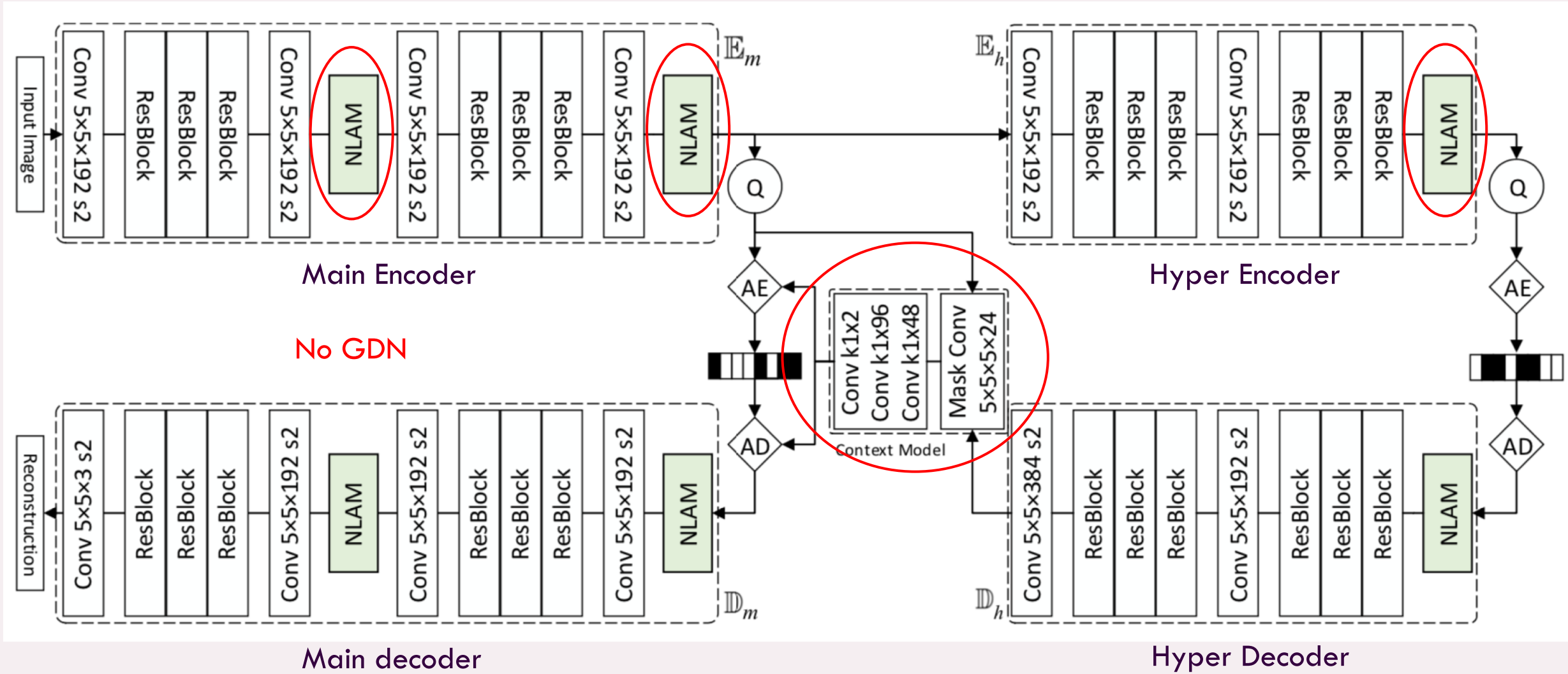




[Minnen2018] D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," NIPS 2018.

Context model: adjacent previously coded pixels in the current channel, and all previously coded channels

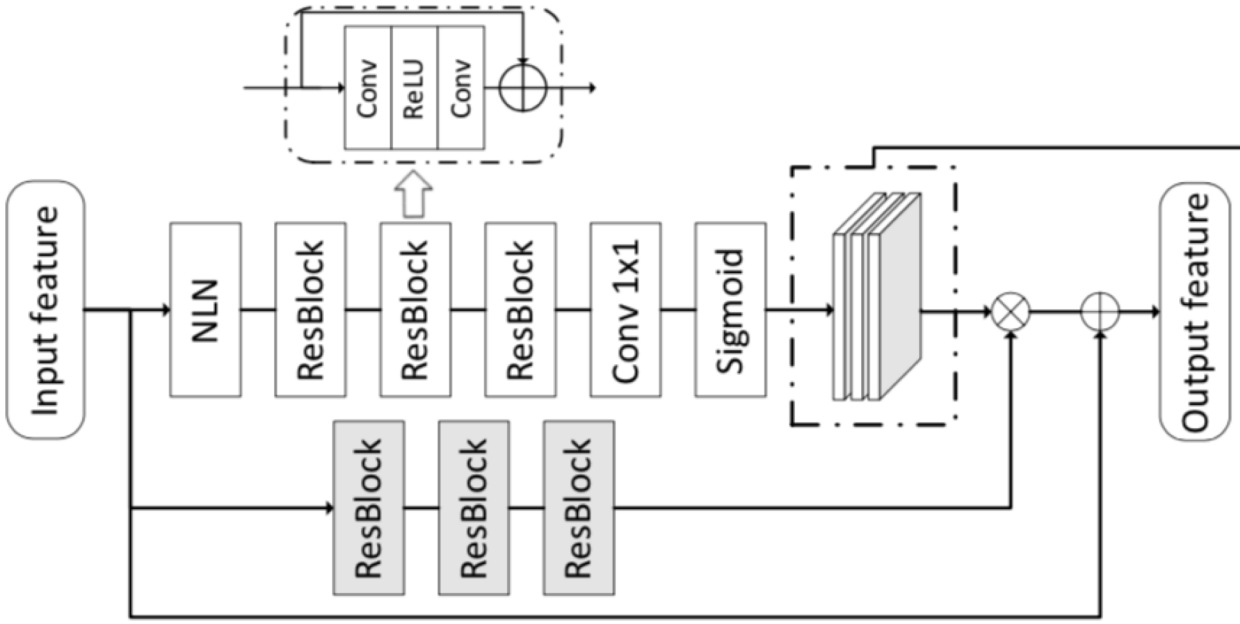
Using hyperprior and context to estimate probability model (mean and STD)

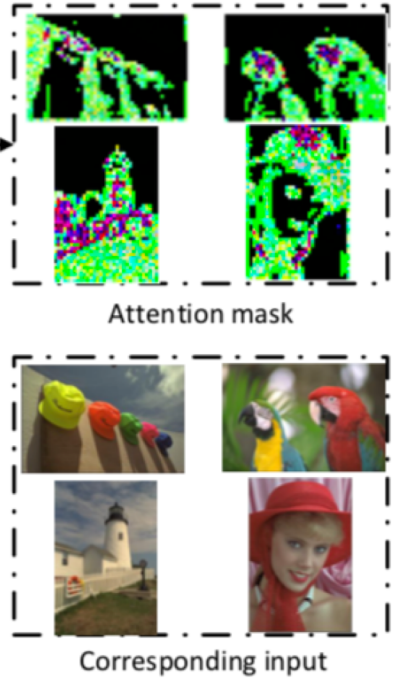
# NLAIC: Non-Local Attention Optimized Image Compression (Collaborator: Zhan Ma, Nanjing Univ.)

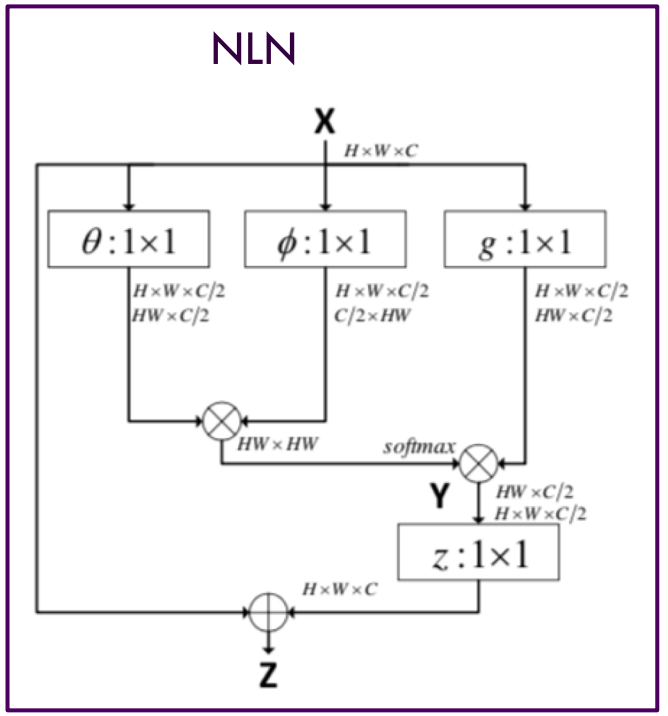



Liu, H.; Chen, T.; Guo, P.; Shen, Q.; Cao, X.; Wang, Y.; and Ma, Z. 2019. Non-local attention optimized deep image compression. arXiv:1904.09757.

# Non-Local Attention Module (NLAM)

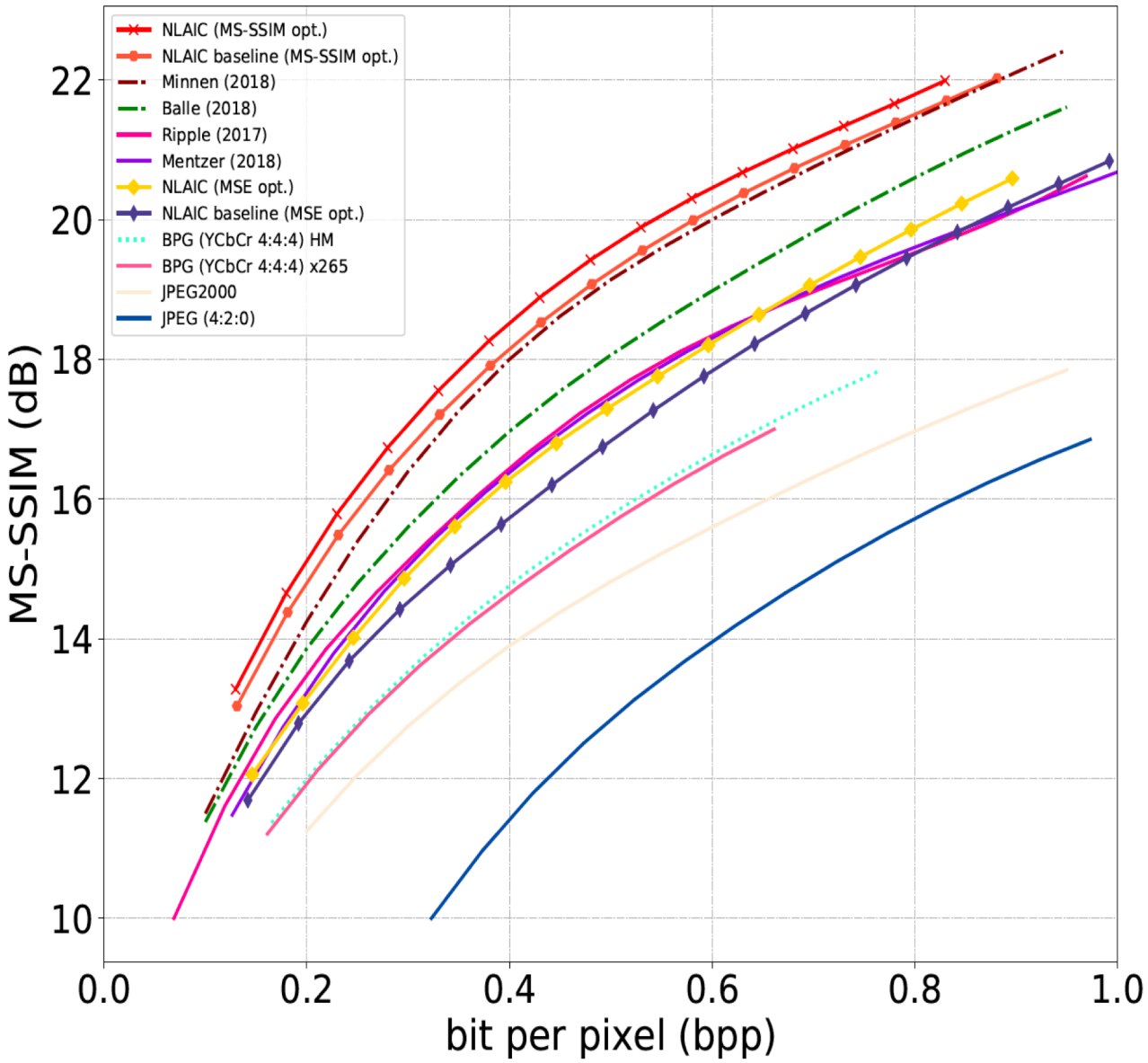


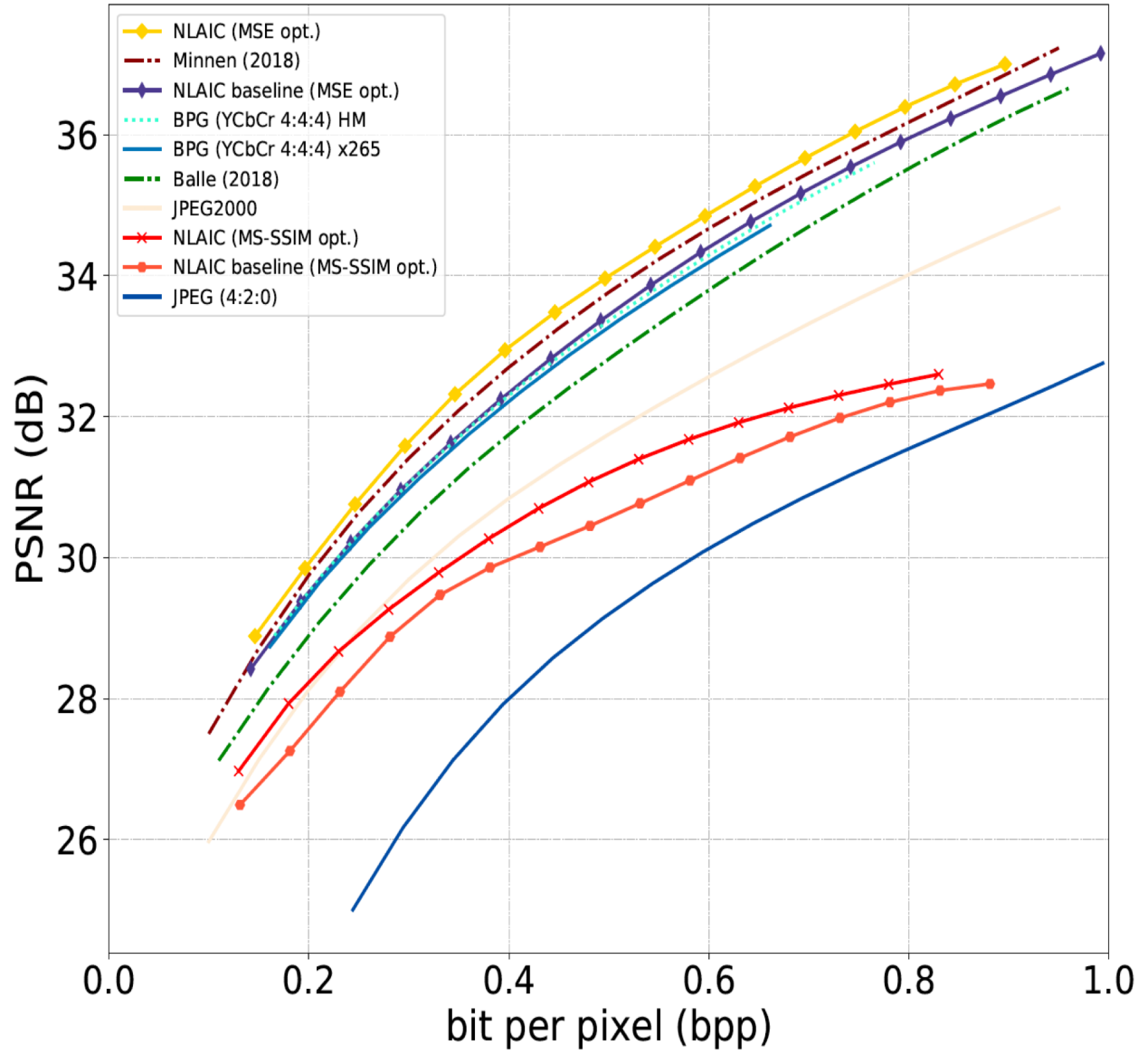




- NLAM generates attention weights, which allows non-salient regions be quantized more heavily
- NLAM uses both local and non-local neighbors (using NLN) to generate the attention maps
- X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," CVPR2018

# Performance on Kodak Dataset



(a) JPEG:0.3014bpp
PSNR:21.23 MS-SSIM:0.8504

(b) BPG:0.3464bpp
PSNR:24.84 MS-SSIM:0.9270

(c) NLAIC MSE opt.:0.2929bpp
PSNR:24.71 MS-SSIM:0.9277

(d) NLAIC MS-SSIM opt.:0.3087bpp
PSNR:23.57 MS-SSIM:0.9551

(e) Original

(a) JPEG:0.2127bpp
PSNR:25.17 MS-SSIM:0.8629

(b) BPG:0.1142bpp
PSNR:31.97 MS-SSIM:0.9581

(c) NLAIC MSE opt.:0.1276bpp
PSNR:34.63 MS-SSIM:0.9738

(d) NLAIC MS-SSIM opt.:0.1074bpp
PSNR:32.54 MS-SSIM:0.9759

(e) Original

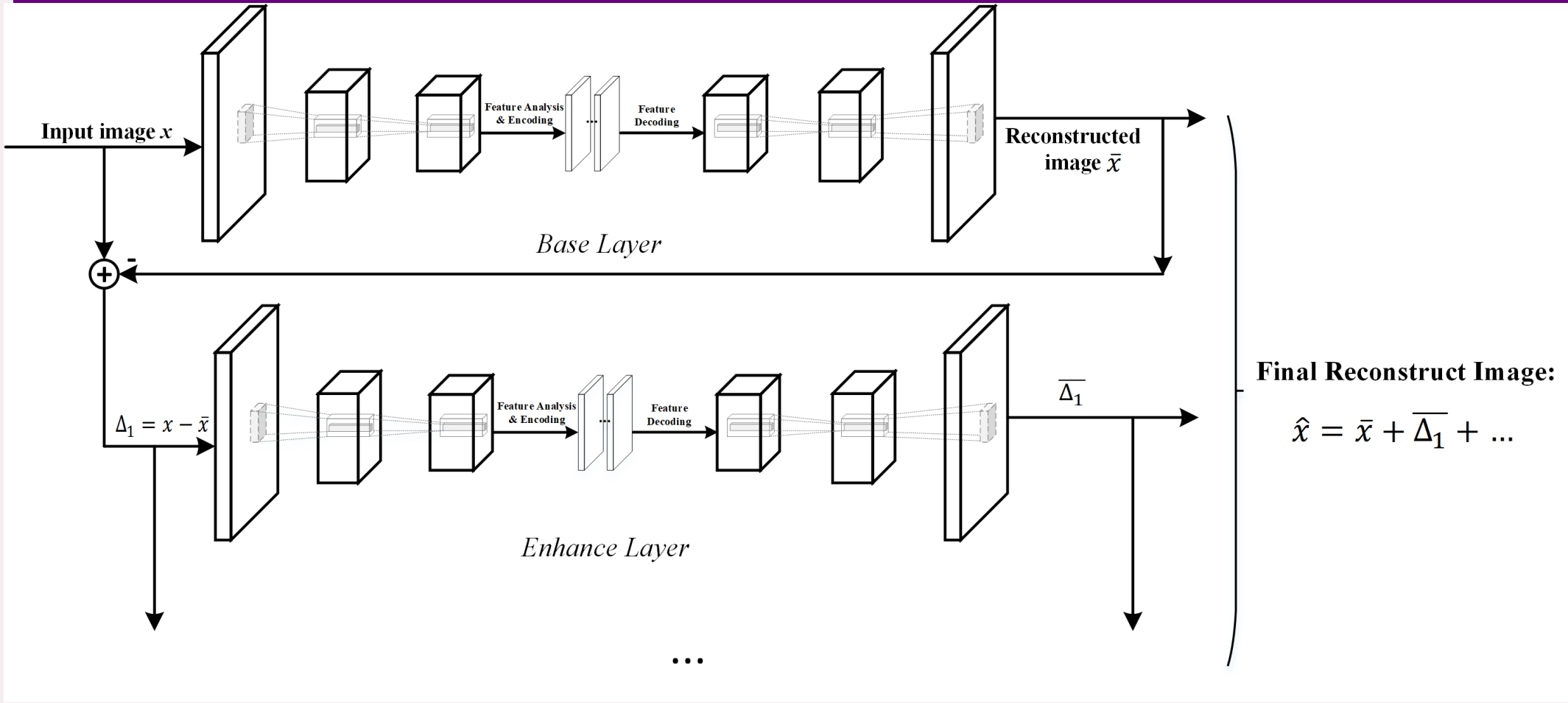
# Problems with Previous Framework

- Train a different model for each bit-rate point using a particular $\lambda$

$$Loss = ||x - \hat{x}||_2^2 + \lambda * R(\hat{y})$$

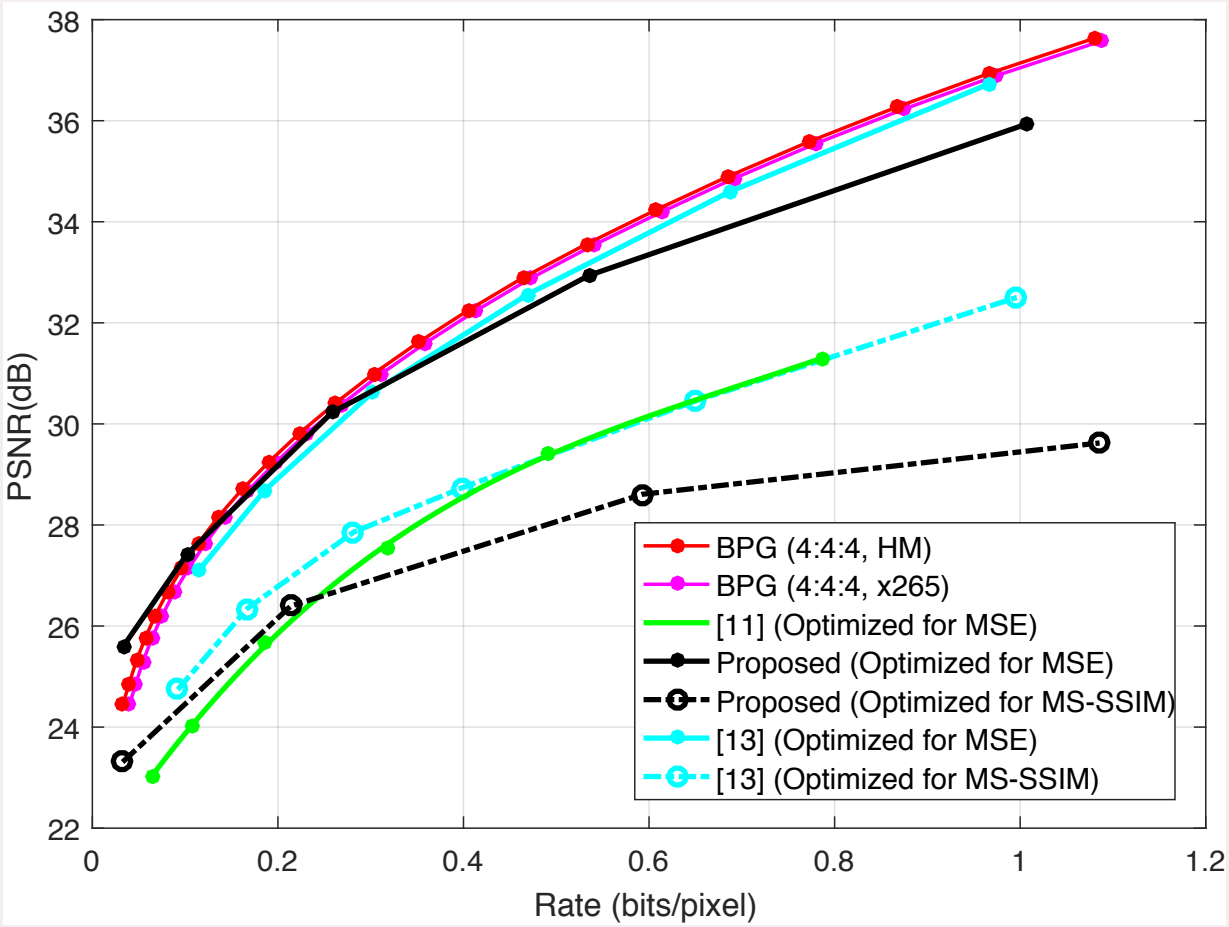
- Hard to deploy in networked applications
  - Need to have multiple encoder/decoder pairs to meet different bandwidths
  - Not scalable: low rate bit streams cannot be shared among users with different bandwidths

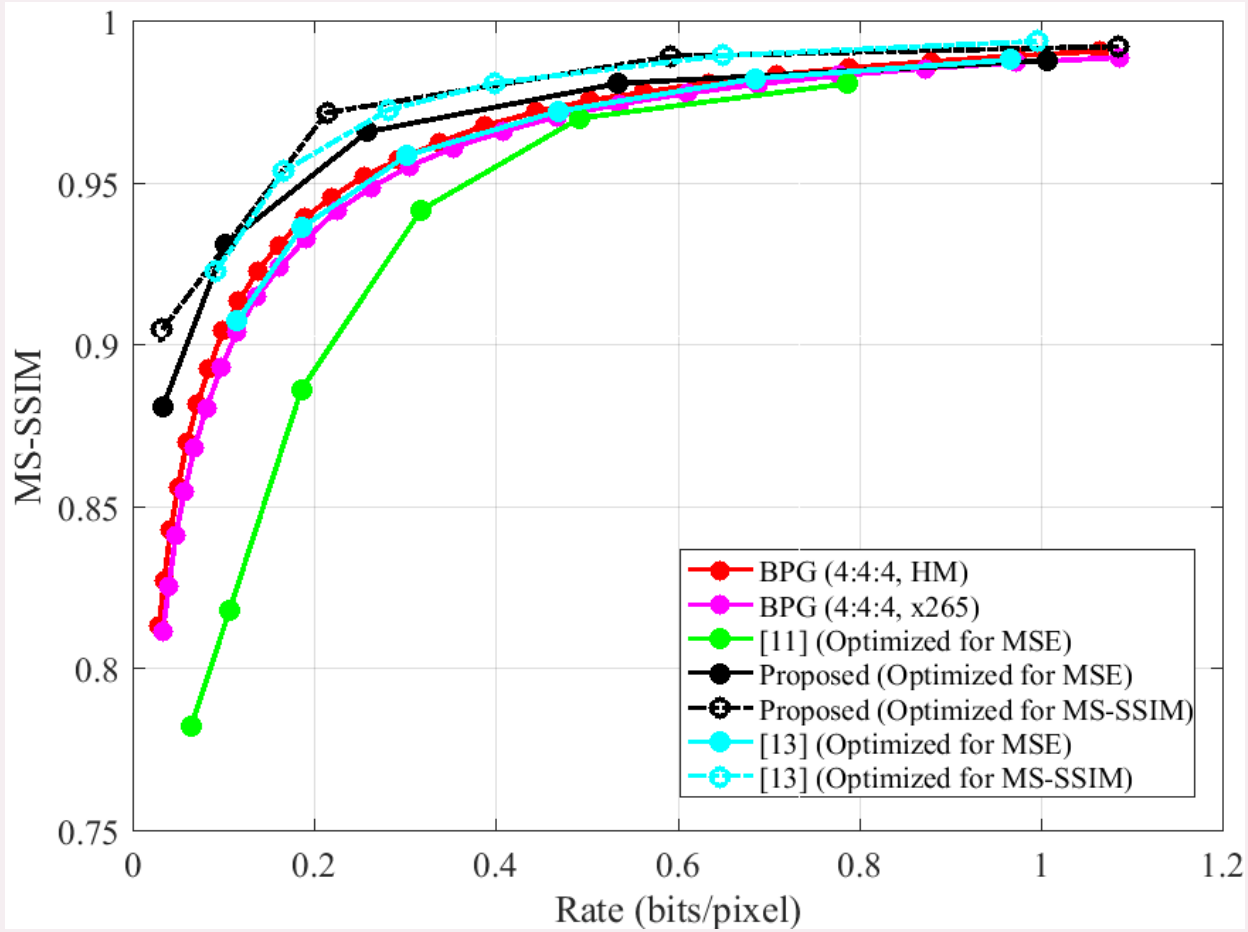
# Layered/Variable Rate Image Compression Using a Stack of Auto-Encoders




- Each layer uses the structure of [Balle2018], but with different number of latent feature maps.
- Chuanmin Jia, Zhaoyi Liu, Yao Wang, Siwei Ma, Wen Gao, Layered Image Compression Using Scalable Auto-Encoder, MIPR 2019. Best student paper award

# Experimental Results (PSNR and MS-SSIM)

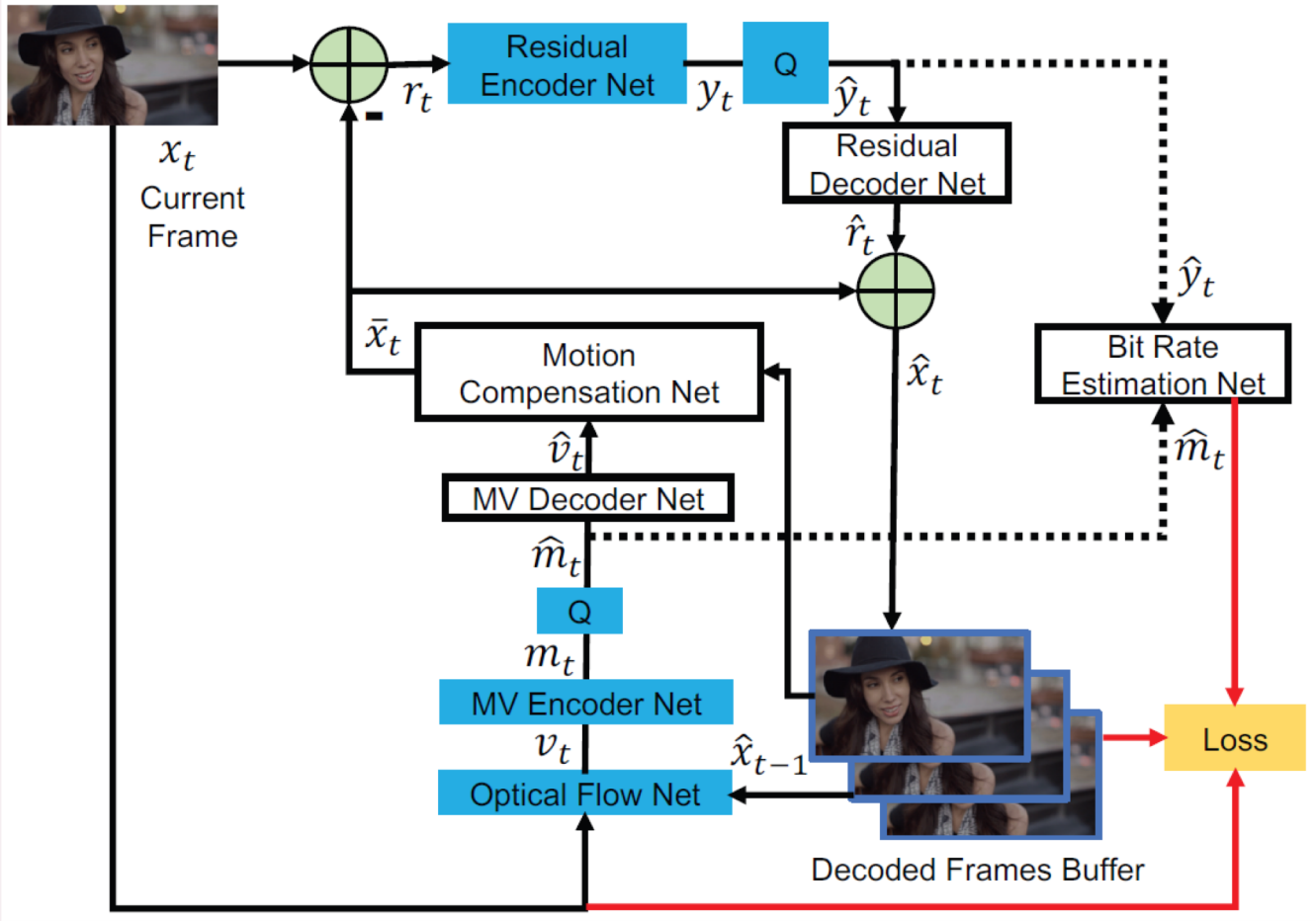



[11]: Balle et al, ICLR 2017
[13]: Balle et al, ICLR 2018

Scalable coding performance similar to non-scalable [11] over entire range for MS-SSIM, competitive or better at lower rate in terms of PSNR

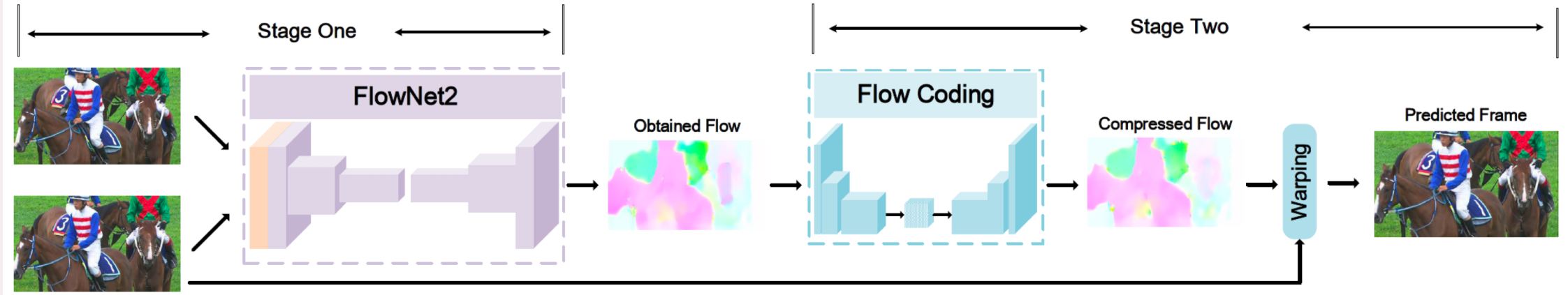
# End-to-End Learnt Video Coding [Lu2019]



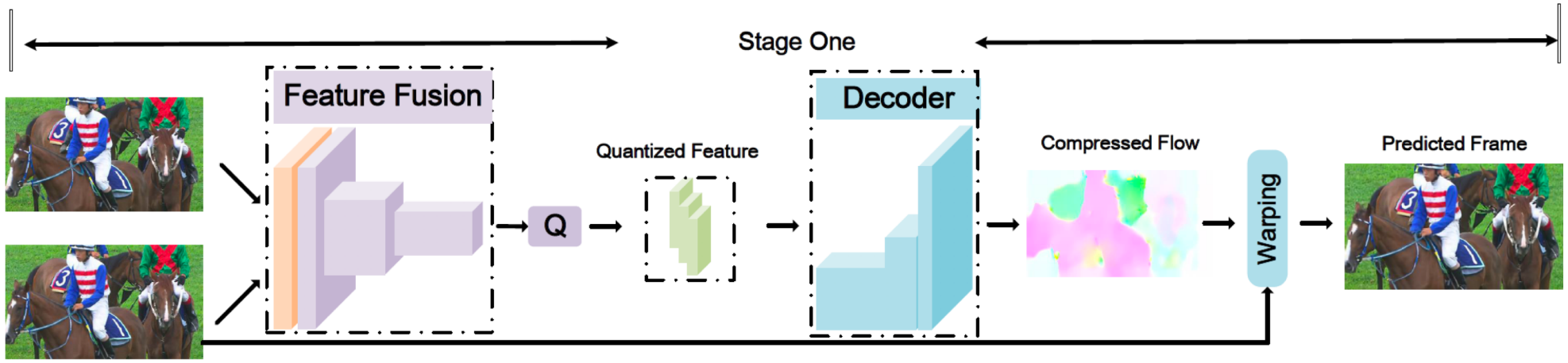

- Implement every part in traditional video coding framework with neural network
- Jointly optimize rate-distortion trade-off through a single loss function.
- The first end-to-end model that jointly learns motion estimation, motion compression, and residual compression.
- Outperforms H.264 in PSNR and MS-SSIM, and on par or better than H.265 in MS-SSIM at high rates.

Guo Lu, et al. "DVC: An End-to-End Deep Video Compression Framework", CVPR2019. https://github.com/GuoLusjtu/DVC

# Frame Prediction Using Implicit Flow Estimation (Collaborator: Zhan Ma)



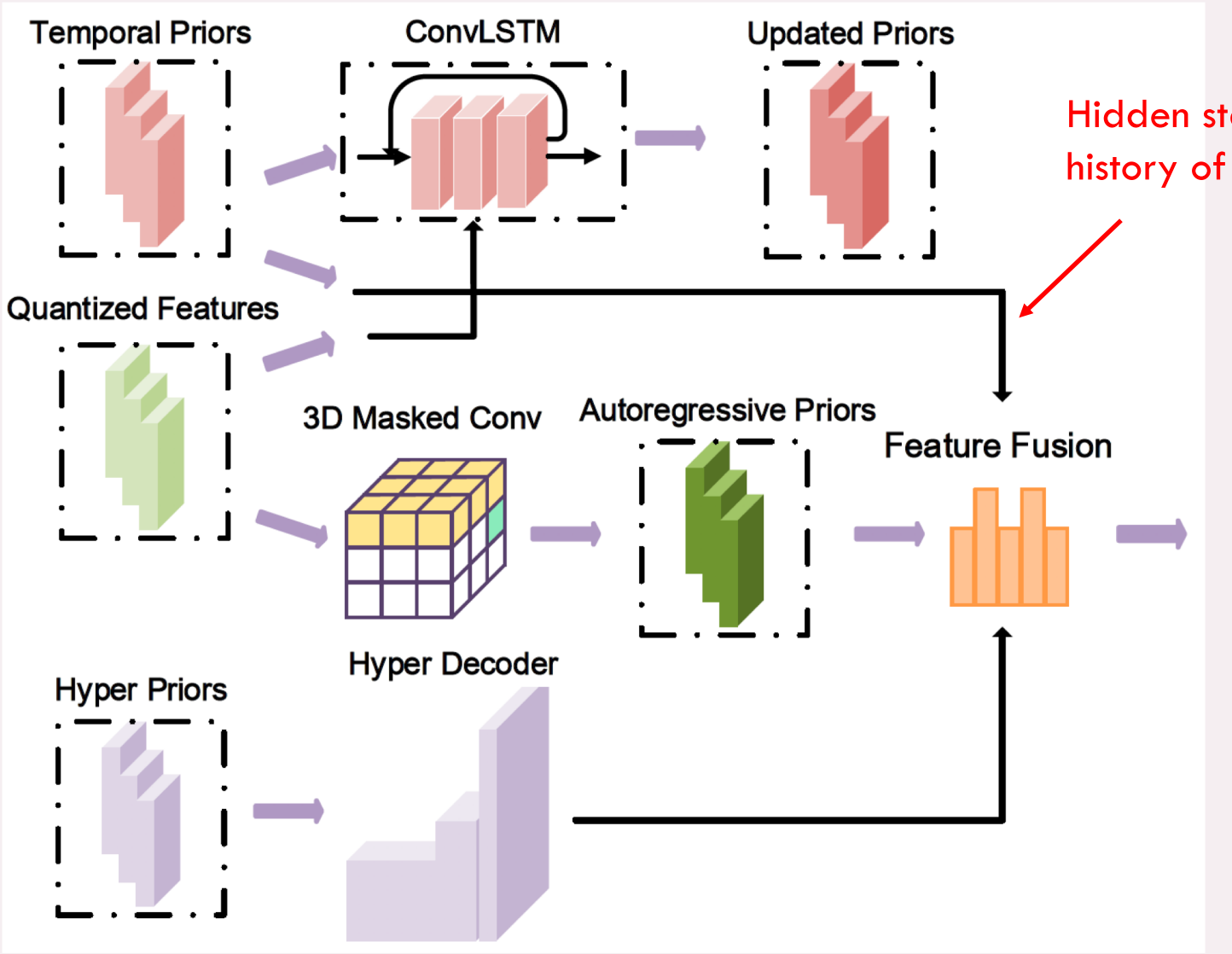

(a) [Lu2019]

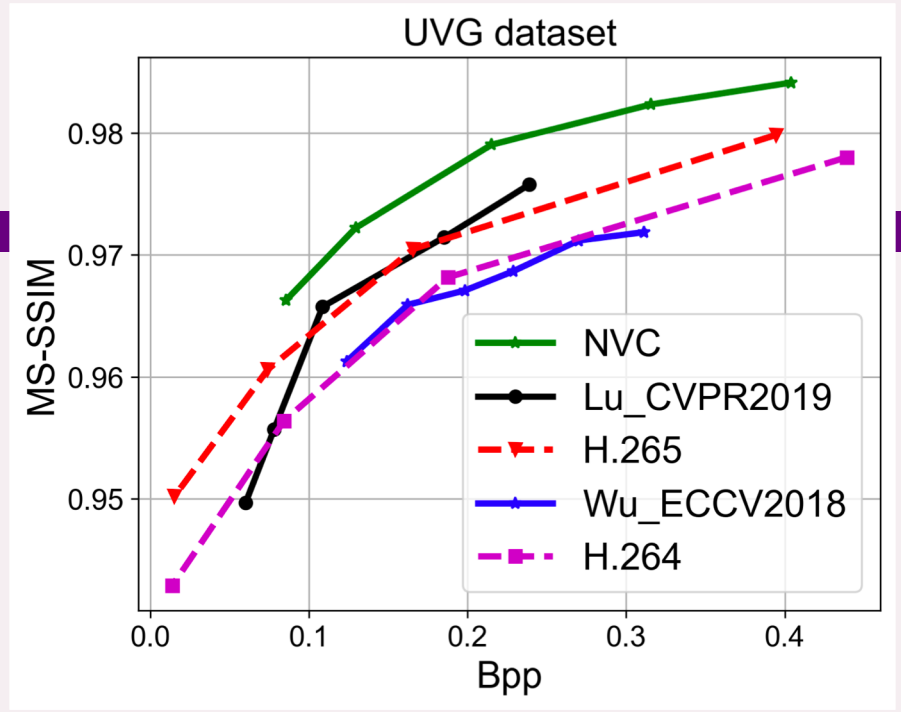


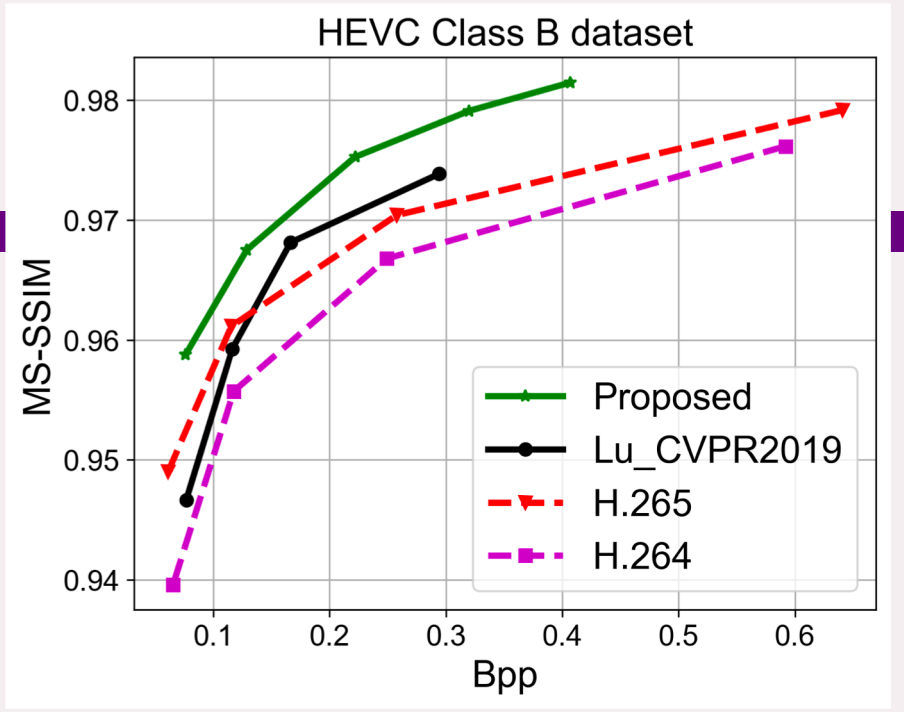

(b) Proposed approach

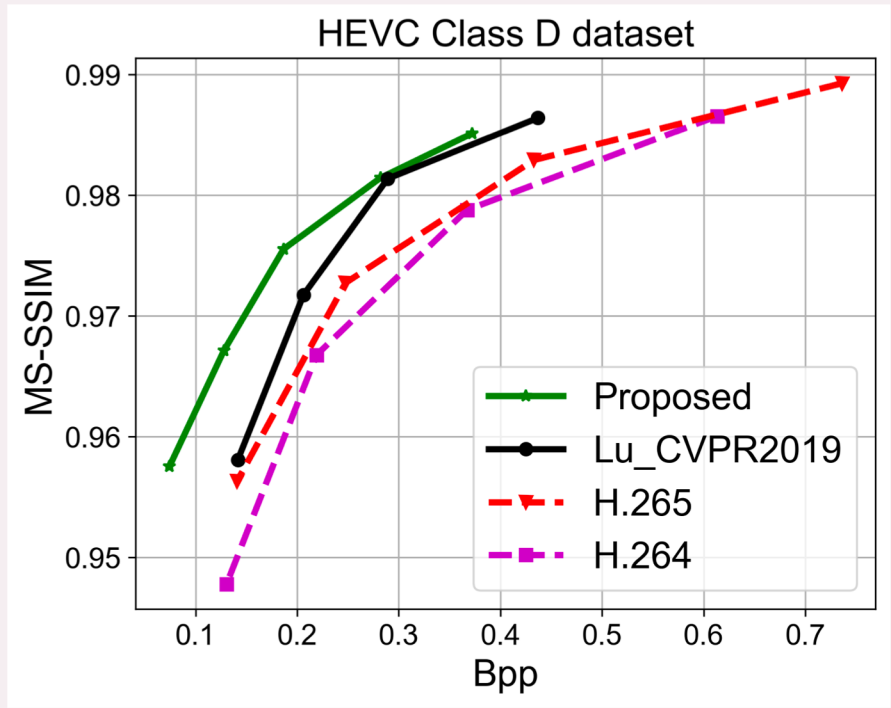
Liu, H., Chen, T., Lu, M., Shen, Q., & Ma, Z. (2019). Neural Video Compression using Spatio-Temporal Priors. *arXiv preprint arXiv:1902.07383*. (Preliminary version)

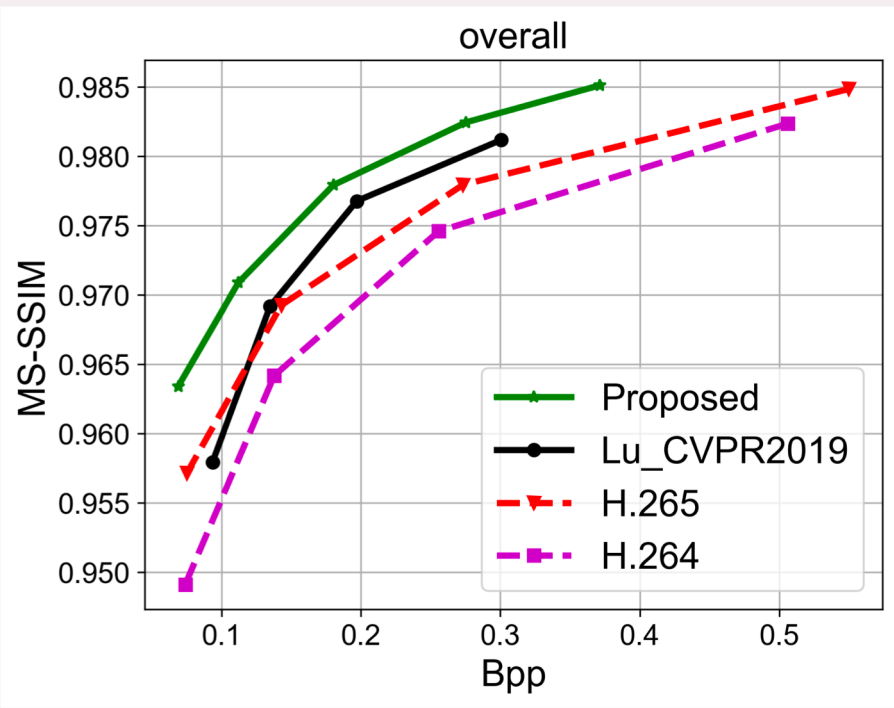
# Entropy coding for flow features




UVG dataset
MS-SSIM
Bpp
NVC
Lu_CVPR2019
H.265
Wu_ECCV2018
H.264


HEVC Class B dataset
MS-SSIM
Bpp
Proposed
Lu_CVPR2019
H.265
H.264

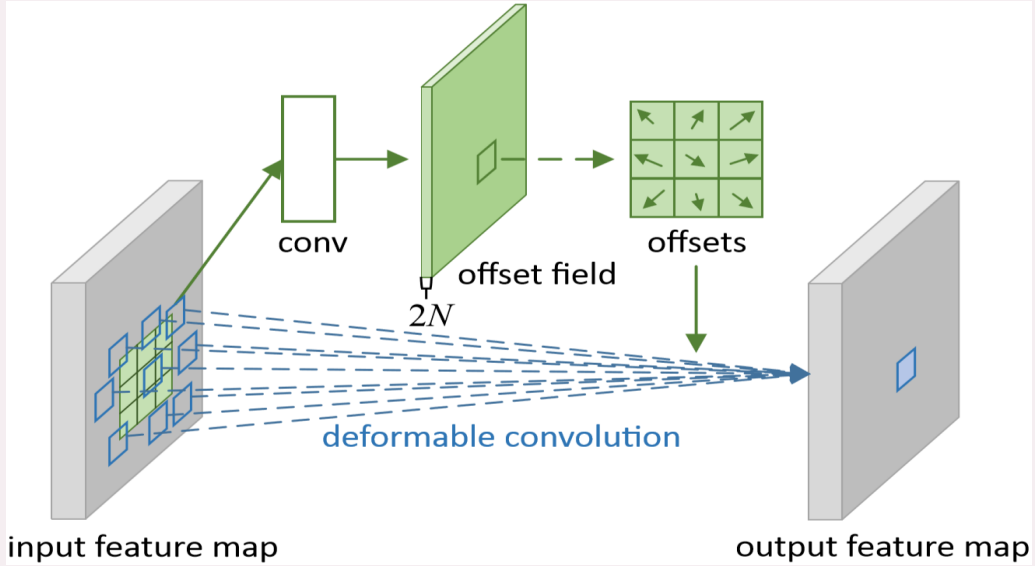

HEVC Class D dataset
MS-SSIM
Bpp
Proposed
Lu_CVPR2019
H.265
H.264

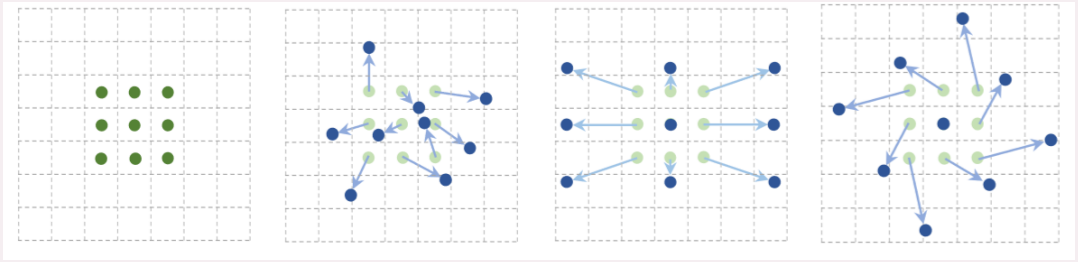

overall
MS-SSIM
Bpp
Proposed
Lu_CVPR2019
H.265
H.264

# Video Prediction Using Dynamic deformable filters

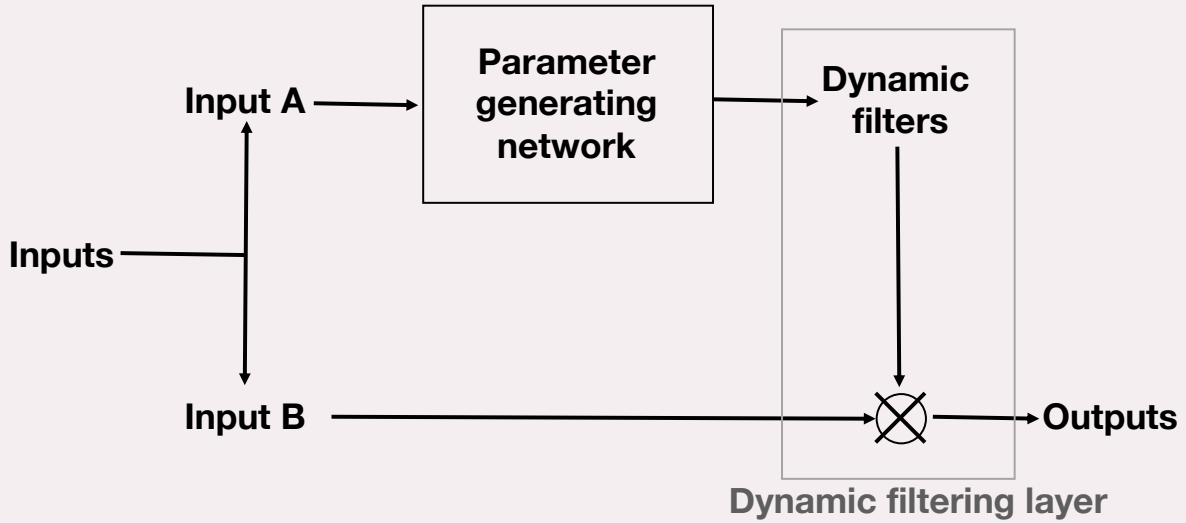
- Deformable filters
- Dynamic filters
- Dynamic deformable filters
- Zhiqi Chen, NYU

# Deformable vs. Dynamic Filter



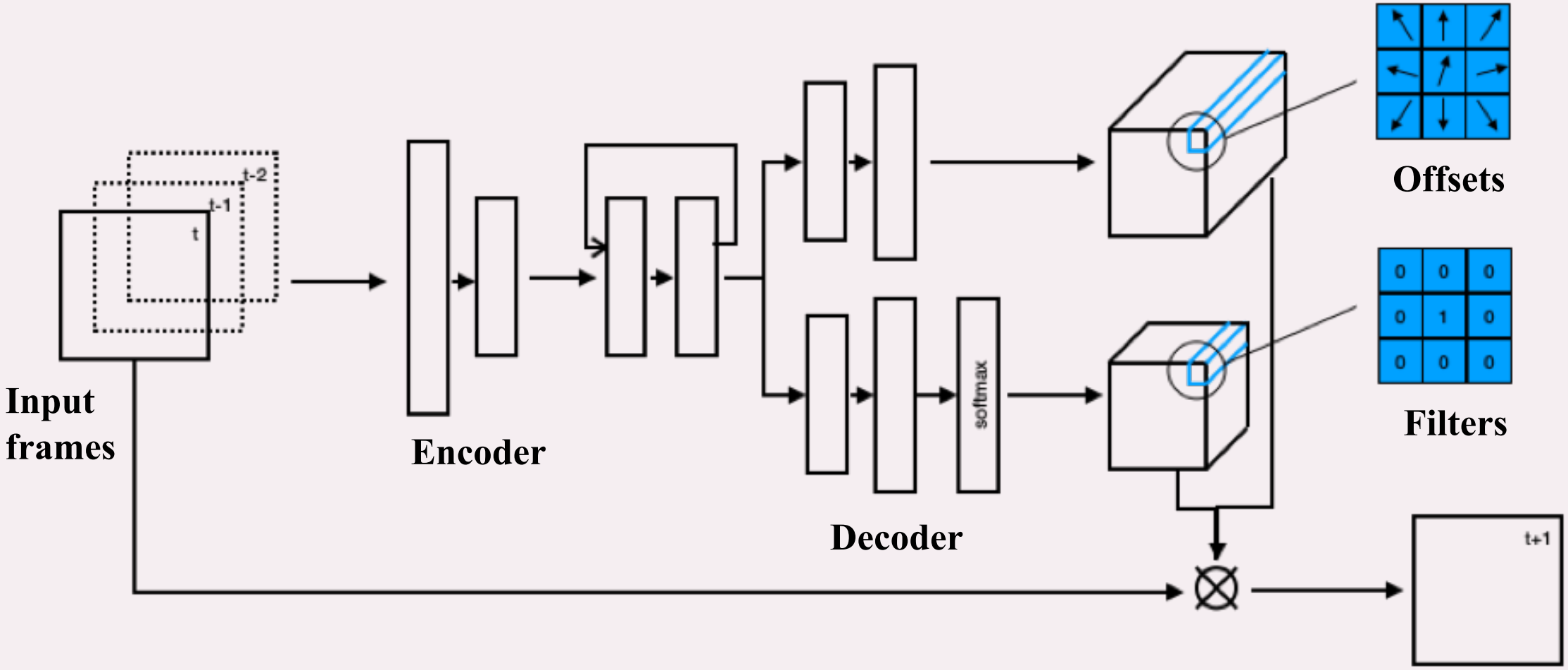
Dai, Jifeng, et al. "Deformable convolutional networks." CVPR 2017.




Jia, Xu, et al. "Dynamic filter networks." NIPS 2016. (DFN)

Using a very large filter size could have the same effect as deformable filter

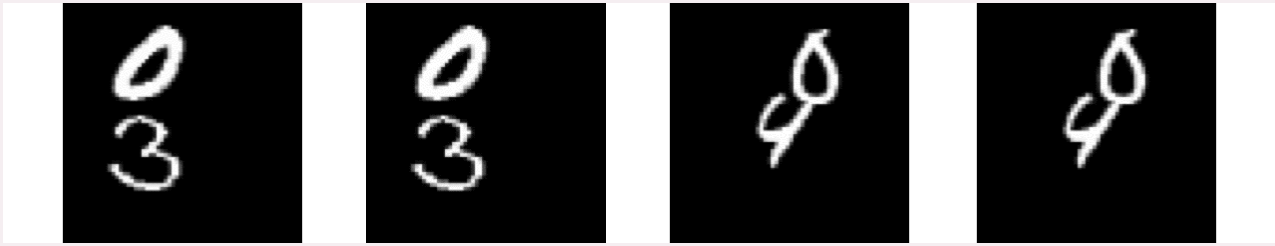
# Video Prediction Using Dynamic Deformable Filters




- Use past frames for generating deformable filters (no need to send side info)
- Each pixel is predicted from weighted average of multiple displaced pixels

# Prediction Results for Moving MNIST

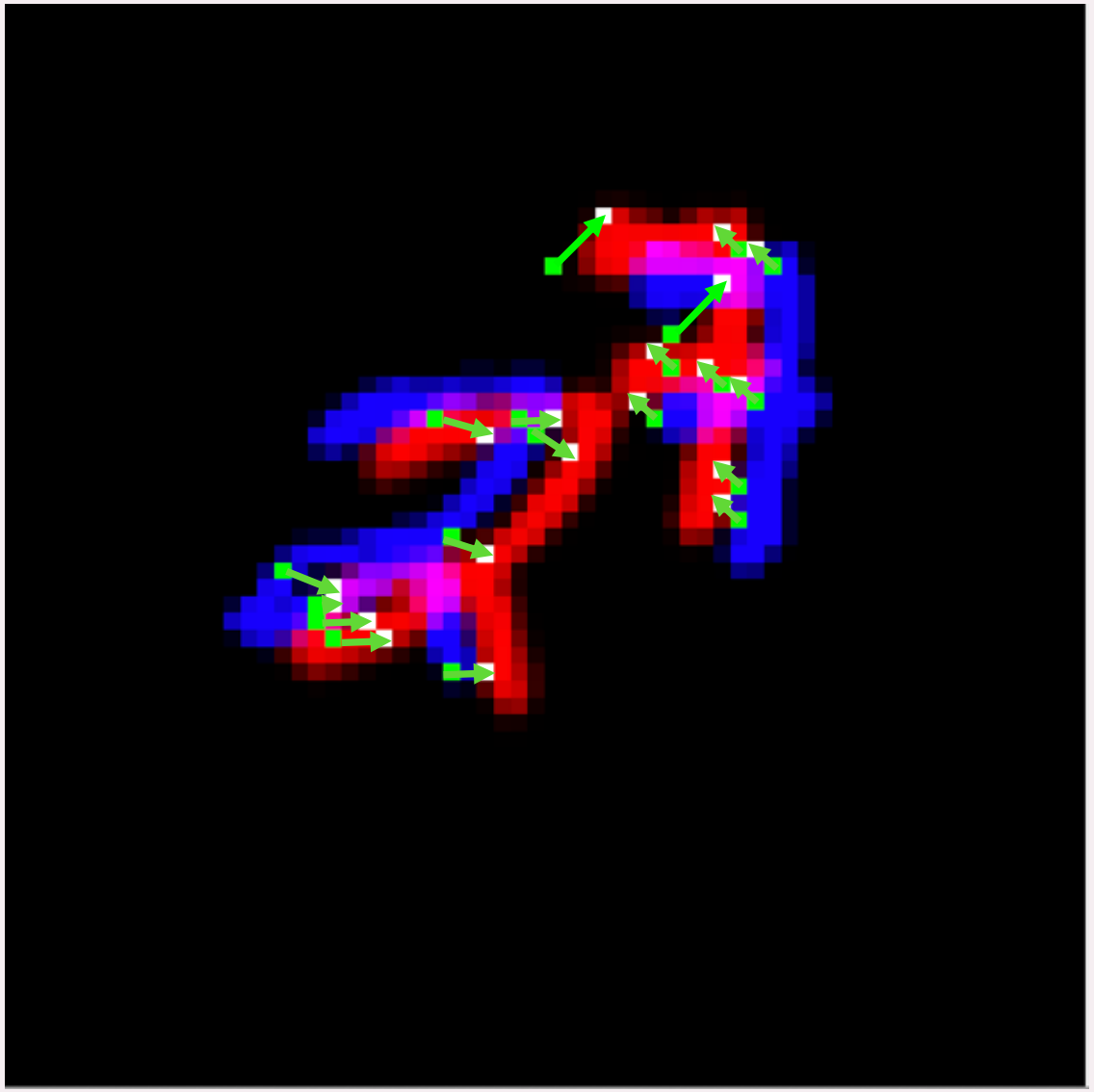
Ground Truth | Deform-DFN (kernel size 3) | Deform-DFN (kernel size 5) | DFN (kernel size 3) | DFN (kernel size 5) | DFN (kernel size 9)



Ground Truth | Deform-DFN (kernel size 3) | Ground Truth | Deform-DFN (kernel size 3)

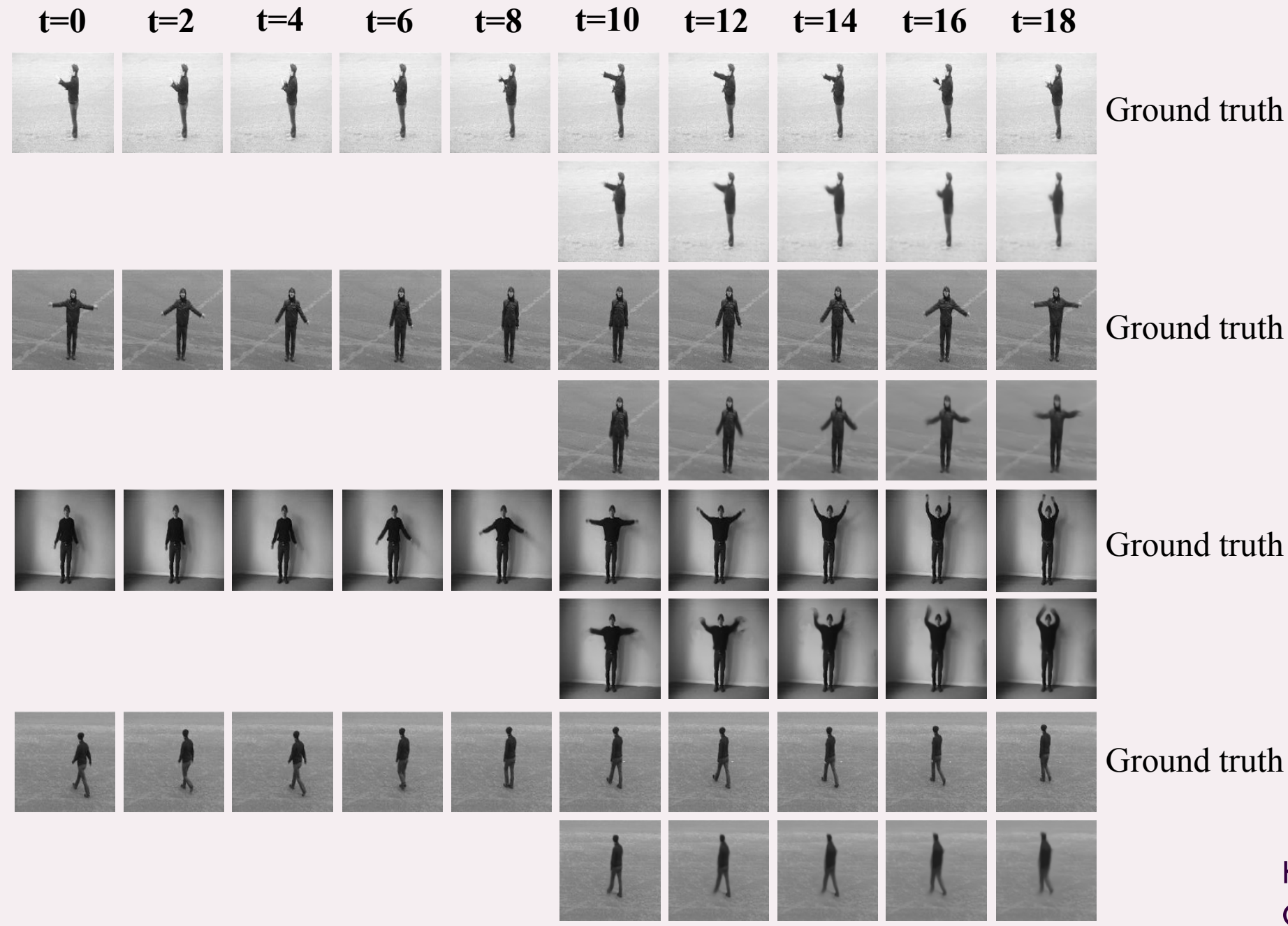
Use past 10 frames to predict future 10 frames recursively

# Visualization of the Offset



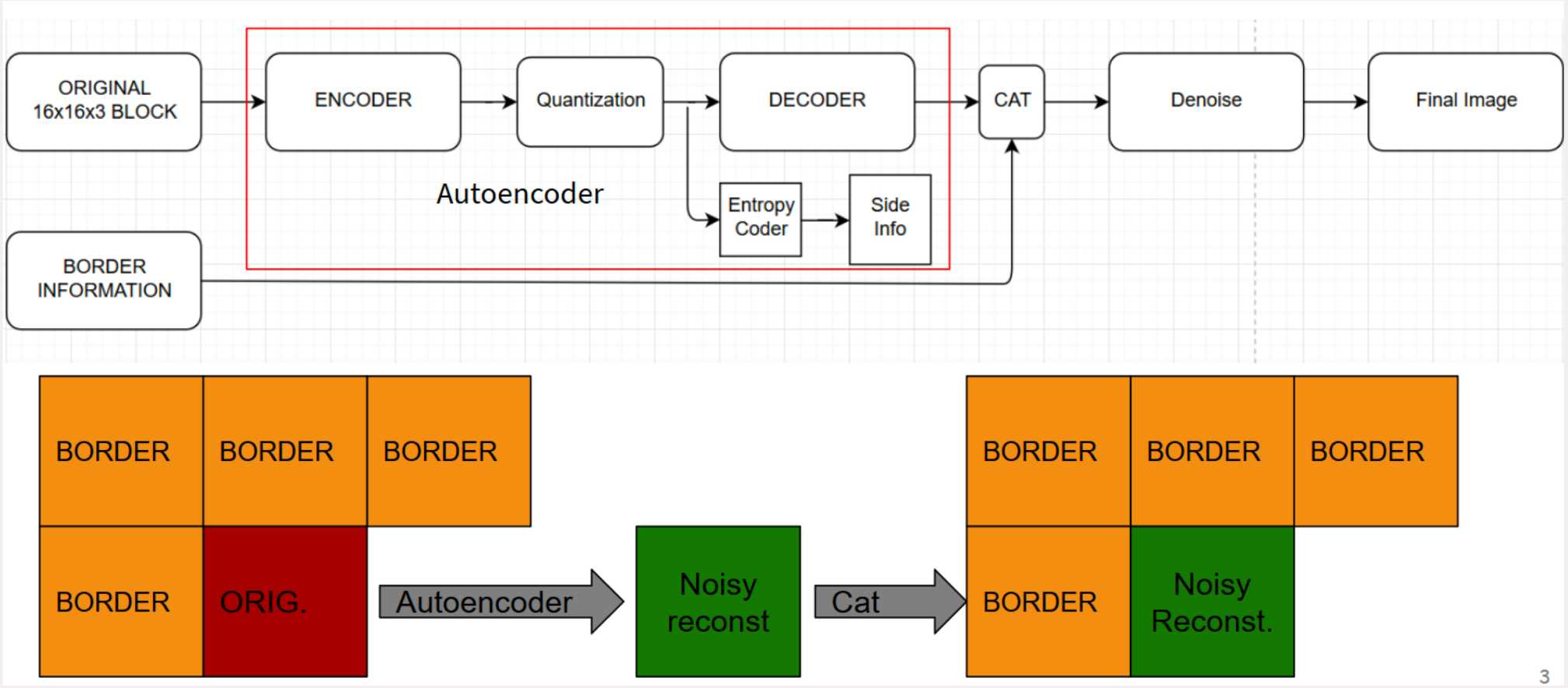
- Blue: last frame
- Red: prediction

- Arrow indicates offset with max filter weight (mapping from green spot in last frame to the white spot in the next frame)




KTH Action
Classification dataset

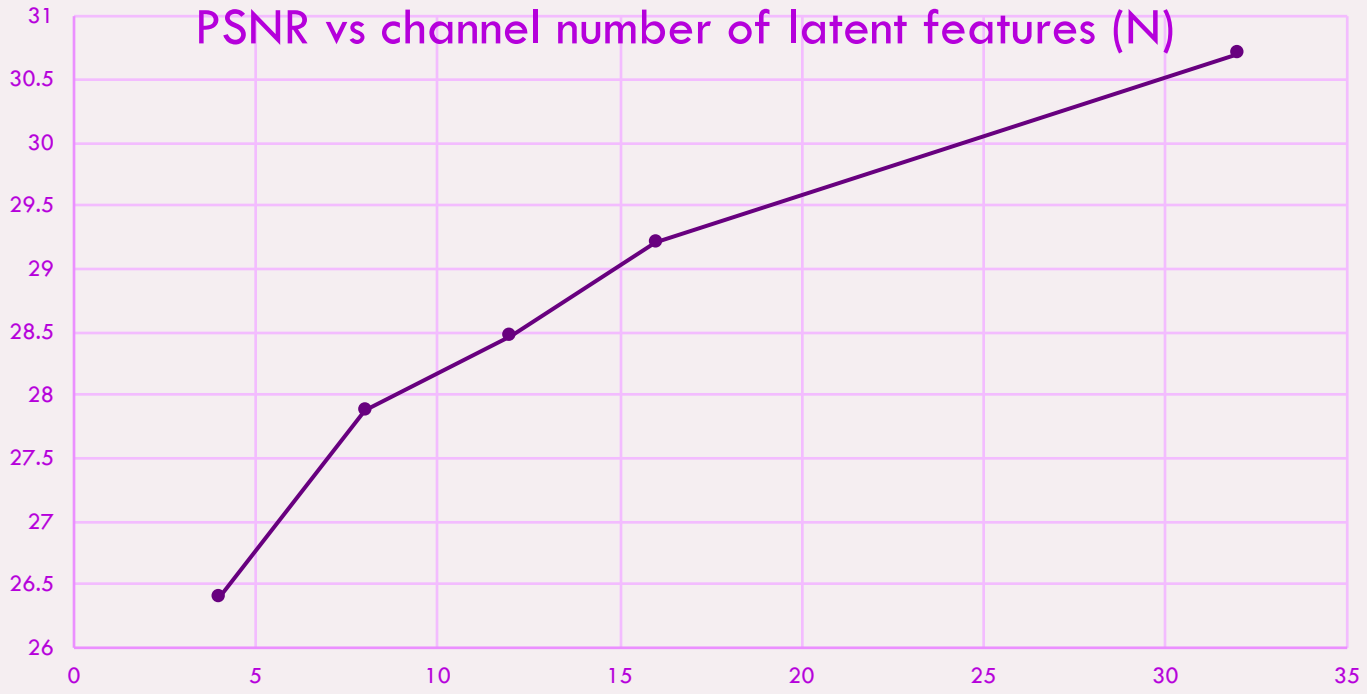
# Block-Based Compression by Denoising with Side Information




- Idea inspired from Debargha Mukherjee, Google
- Students: Jeffrey Mao and Jacky Yuan, NYU

# Performance (Very Preliminary)

| bpp | N values | PSNR |
|---|---|---|
| 0.06 | 4 | 26.4 |
| 0.12 | 8 | 27.87 |
| 0.18 | 12 | 28.47 |
| 0.25 | 16 | 29.2 |
| 0.5 | 32 | 30.7 |




- Quantize latent features to binary. Rate obtained by assuming 1 bit per feature.
- Context-based entropy coding will reduce the bit rate significantly.
- Future work: consider the rate of the side information in the loss function for training to enable end-to-end RD optimization

# Acknowledgement


- Students at Video lab at NYU
  - https://wp.nyu.edu/videolab/
- Vision lab at Nanjing University, led by Zhan Ma
  - http://vision.nju.edu.cn/index.php
- Work on scalable image compression
  - Chuanmin Jia, visiting student from Beijing Univ.
- Thanks for Google Faculty Research Award!