CVPR 2020 Workshop and Challenge on Learned Image Compression

Learning-Based Image/Video Coding

Lu Yu

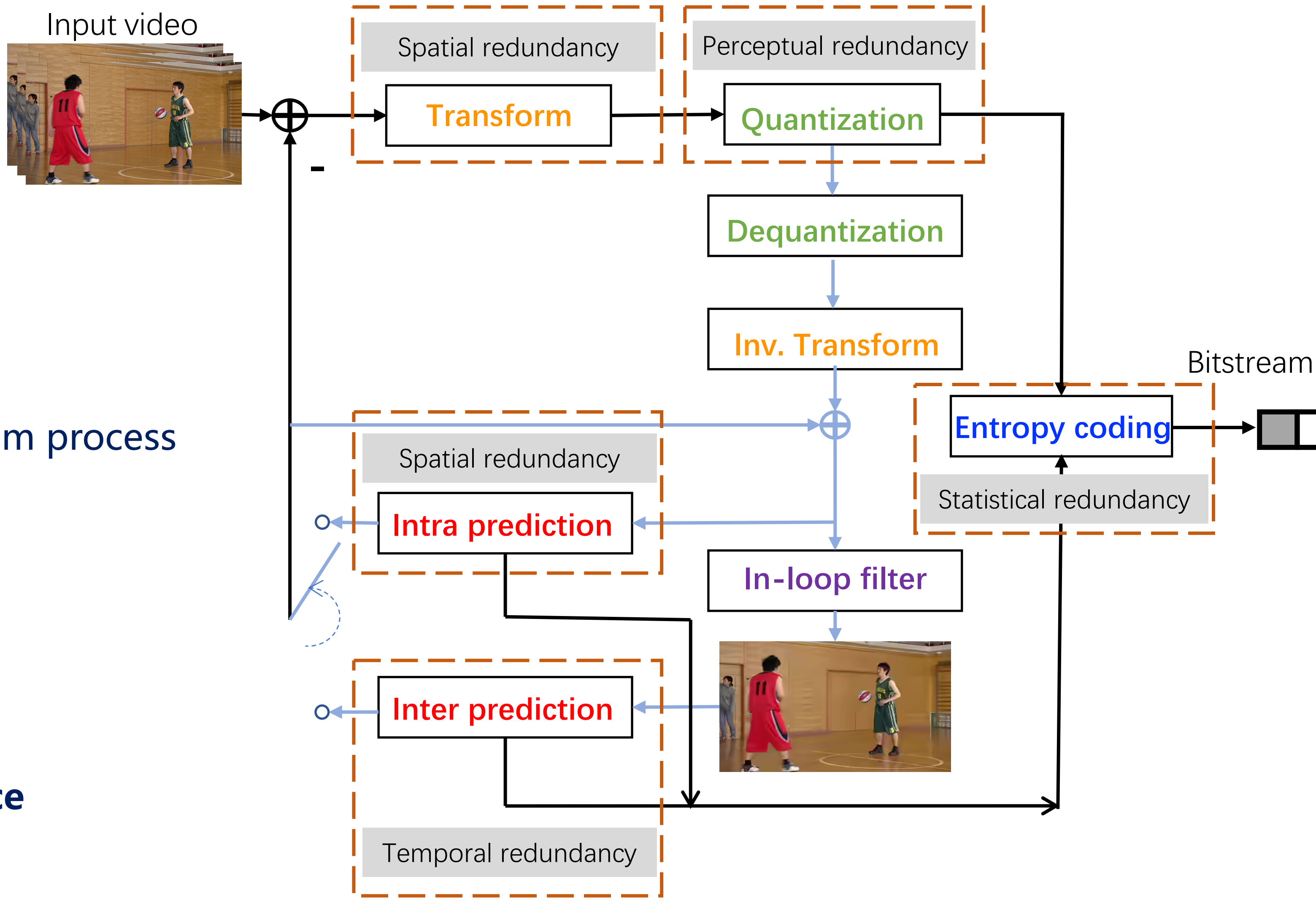
Zhejiang University

Outlines

- **System architecture of learning based image/video coding**
  - Learning based modules embedded into traditional hybrid coding frameworks
    - In-loop filter, Intra prediction, Inter prediction, Entropy coding, etc.
    - Transform, quantization
    - Encoder optimization
  - End-to-end image and video coding
- **Coding for human vision vs. coding for machine intelligence**

Theory of Source Coding and Hybrid Coding Framework

- **Two threads of image/video coding**
  - **Characteristics of source signal**
    - **Spatial-temporal correlation**
      - Intra and inter prediction
      - transform
    - **Statistical correlation**
      - Symbols: stationary random process
      - Entropy coding
  - **Characteristics of human vision**
    - **Limited sensitivity**
      - Quantization
- **Balance between cost and performance**
  - **Rate-distortion theory**



In-Loop Filter

Filtering

- **Network input**
  - Current compressed frame
- **Network output**
  - Filtered frame
- **Network structure**
  - 22-layer CNN with inception structure

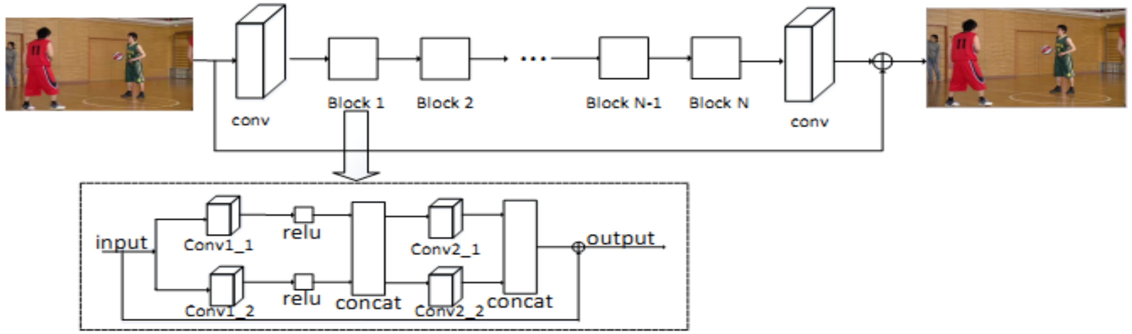


TABLE I
CONFIGURATION OF VR BLOCK

Layer	Layer 1		Layer 2	
Conv. module	conv1_1	conv1_2	conv2_1	conv2_2
Filter size	3×3	1×1	3×3	1×1
# filters	32	32	32	32

- **Integration into coding system**
  - Same model for Luma and chroma component
  - Different model for different QP
  - For I-frame: replace Deblocking filter (DB) and Sample Adaptive Offset (SAO)
  - For B/P-frame: added between DB and SAO, switchable at CTU-level
- **Performance** (anchor: HM16.0)

	All-Intra			Low-Delay B			Random-Access		
	Y (%)	U (%)	V (%)	Y (%)	U (%)	V (%)	Y (%)	U (%)	V (%)
Class A	−5.4	−6.2	−5.5	–	–	–	−5.3	−11.2	−9.6
Class B	−7.3	−7.5	−9.0	−7.3	−11.4	−12.1	−8.0	−11.1	−11.1
Class C	−9.9	−10.4	−13.4	−8.8	−11.2	−13.5	−8.7	−11.9	−14.9
Class D	−10.0	−10.4	−13.4	−8.1	−9.0	−12.0	−7.8	−9.0	−12.4
Class E	−13.4	−10.0	−9.5	−14.2	−14.9	−13.2	–	–	–
Overall	**−9.2**	**−8.9**	**−10.2**	**−9.6**	**−11.6**	**−12.7**	**−7.4**	**−10.8**	**−12.0**
Enc. Time	4710%			1818%			1835%		
Dec. Time	267686%			756074%			727860%		

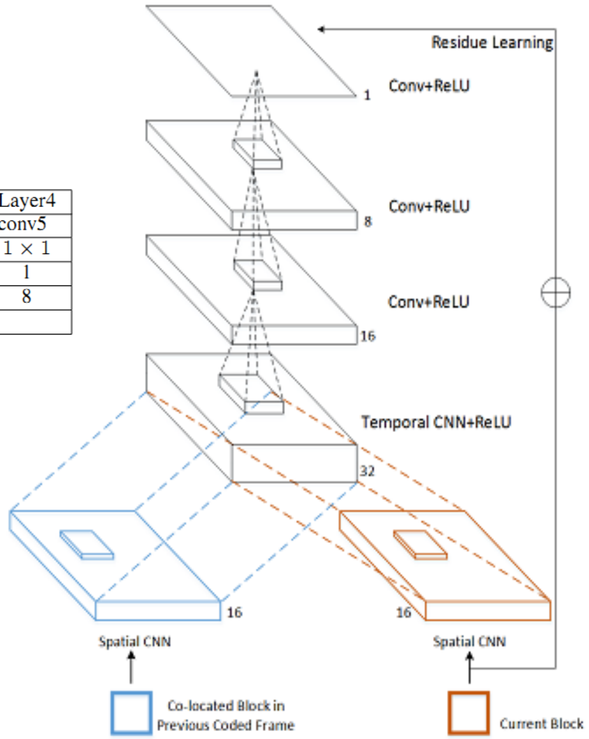
[1] Dai Y, Liu D, Zha Z J, et al. A CNN-Based In-Loop Filter with CU Classification for HEVC[C]//2018 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2018: 1-4.

In-Loop Filter

Filtering with spatial and temporal information

- **Network input**
  - Current compressed frame
  - Previous reconstructed frame
- **Network output**
  - Filtered frame
- **Network structure**
  - 4-layer CNN

	Layer1		Layer2	Layer3	Layer4
	conv1	conv2	conv3	conv4	conv5
Filter Size	5 × 5	3 × 3	3 × 3	3 × 3	1 × 1
Feature Map Number	32	32	16	8	1
Param Number	800	288	9216	1152	8
Total Param Number	11464				



- **Integration into coding system**
  - Same model for Luma and chroma component
  - Different model for different QP
  - Used in I/P/B frames
  - After DB and SAO
  - Switchable at CTU-level
- **Performance** (anchor: RA, HM16.15)

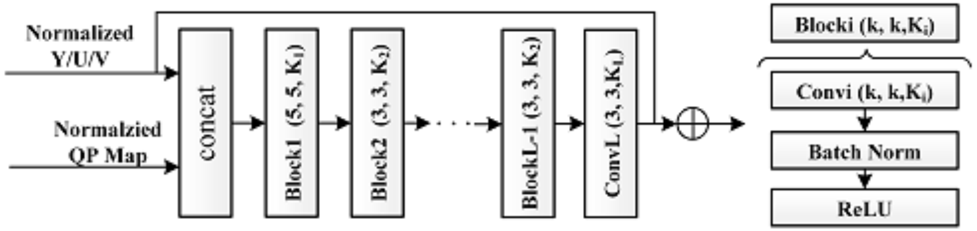
Sequences		Random Access
		Y
Class B	Kimono	-0.7%
	ParkScene	-0.8%
	Cactus	-0.3%
	BasketballDrive	-1.0%
	BQTerrace	-0.1%
Class C	BasketballDrill	-1.0%
	BQMall	-1.1%
	PartyScene	-1.2%
	RaceHorsesC	-1.5%
Class D	BasketballPass	-2.0%
	BQSquare	-1.8%
	BlowingBubbles	-2.1%
	RaceHorses	-2.2%
Class E	FourPeople	-5.1%
	Johnny	-0.8%
	KristenAndSara	-1.5%
Overall		-1.3%

[2] Jia C, Wang S, Zhang X, et al. Spatial-temporal residue network based in-loop filter for video coding[C]//2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017: 1-4.

In-Loop Filter

Filtering with quantization information

- **Network input**
  - Current compressed frame
  - Normalized QP map
- **Network output**
  - Filtered frame
- **Network structure**
  - 8-layer CNN



(K_L=64)

- **Integration into coding system**
  - Same model for Luma and chroma component
  - Same model for all QPs
  - Replace bi-lateral filter, DB and SAO, and before ALF
  - Only used on I frames
  - No RDO

- **Network compression**
  - Pruning:
    - ✓ Operate during training
    - ✓ Filters pruned based on absolute value of the scale parameter in its corresponding BN layer
    - ✓ Loss function: additional regularizers for efficient compression
  - Low rank approximation:
    - ✓ Operate after pruning
  - Dynamic fixed point adoption

- **Performance** (anchor: RA, JEM7.0)

Table 6. Test results of AI configuration with ALF on

				CPU+GPU		CPU	
	Y	U	V	EncT	DecT	EncT	DecT
ClassA1	-2.26%	-6.21%	-5.05%	93%	157%	109%	15360%
ClassA2	-3.58%	-6.33%	-7.02%	92%	158%	112%	16312%
ClassB	-3.08%	-5.06%	-6.27%	94%	148%	108%	15360%
ClassC	-3.88%	-6.98%	-9.11%	94%	158%	103%	11139%
ClassD	-4.13%	-5.63%	-8.20%	94%	214%	102%	7256%
ClassE	-4.93%	-7.41%	-6.88%	94%	169%	111%	15441%
Overall	**-3.57%**	**-6.17%**	**-7.06%**	**93%**	**157%**	**109%**	**12887%**

Table 7. Test results of RA configuration with ALF on

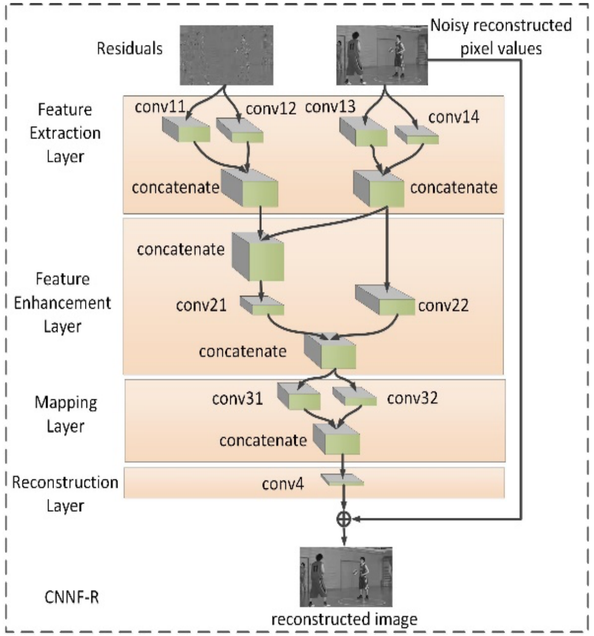
				CPU	
	Y	U	V	EncT	DecT
ClassA1	-0.39%	-1.96%	-1.93%	99%	275%
ClassA2	-1.76%	-3.70%	-4.29%	99%	303%
ClassB	-1.46%	-4.65%	-4.14%	99%	339%
ClassC	-1.28%	-4.40%	-4.75%	99%	289%
ClassD	-1.22%	-3.28%	-4.20%	99%	219%
Overall	**-1.23%**	**-3.65%**	**-3.88%**	**99%**	**284%**

[3] Song X, Yao J, Zhou L, et al. A practical convolutional neural network as loop filter for intra frame[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 1133-1137.

In-Loop Filter

Filtering with high-frequency information

- **Network input**
  - Current compressed frame
  - Reconstructed residual values
- **Network output**
  - Filtered frame
- **Network structure**
  - 4-layer CNN



Layer	Layer1				Layer2		Layer3		Layer4
Conv.module	conv11	conv12	conv13	conv14	conv21	conv22	conv31	conv32	conv4
Filter size	3x3	5x5	3x3	5x5	3x3	3x3	1x1	3x3	3x3
Filter number	32	16	32	16	16	32	32	16	1
Parameters	320	416	320	416	13,840	13,856	1,568	6,928	433
Total parameters	**38,097**								

- **Integration into coding system**
  - Same model for Luma and chroma component
  - Different model for different QP
  - Replace DB and SAO
  - Only used on I frames
  - No RDO
- **Performance** (anchor: HM16.15)

TABLE III
AI CONFIGURATION, BD-RATE RESULTS OF CNNF-R AND VRCNN COMPARED WITH HM16.15

Class	Sequence	BD-rate(VRCNN)			BD-rate(CNNF-R)		
		Y	U	V	Y	U	V
ClassA	Traffic	-4.7%	-2.9%	-3.4%	-5.6%	-3.7%	-4.1%
	PeopleOnStreet	-4.6%	-5.0%	-4.5%	-5.6%	-5.6%	-5.1%
	Nebuta	-0.3%	-3.9%	-3.0%	-0.9%	-5.3%	-4.0%
	SteamLocomotive	-0.8%	-0.7%	-0.5%	-2.0%	-2.1%	-1.8%
ClassB	Kimono	-2.5%	-2.1%	-1.8%	-3.5%	-2.8%	-2.3%
	ParkScene	-3.6%	-3.4%	-2.6%	-4.7%	-3.7%	-2.9%
	Cactus	-3.1%	-3.0%	-5.1%	-4.7%	-4.1%	-6.4%
	BasketballDrive	-1.1%	-2.2%	-4.0%	-3.6%	-3.9%	-6.1%
	BQTerrace	-0.7%	-2.9%	-1.8%	-2.5%	-4.4%	-2.7%
ClassC	BasketballDrill	-5.3%	-4.5%	-4.8%	-6.4%	-5.7%	-6.7%
	BQMall	-4.1%	-4.1%	-3.9%	-5.2%	-5.0%	-4.9%
	PartyScene	-2.7%	-3.4%	-3.3%	-3.5%	-4.0%	-3.9%
	RaceHorses	-3.7%	-5.8%	-8.1%	-4.4%	-6.5%	-9.0%
ClassD	BasketballPass	-3.8%	-4.1%	-7.1%	-5.3%	-5.1%	-8.6%
	BQSquare	-2.5%	-2.6%	-4.0%	-3.5%	-3.1%	-5.1%
	BlowingBubbles	-3.4%	-6.0%	-6.0%	-4.2%	-6.7%	-6.6%
	RaceHorses	-6.0%	-7.7%	-9.3%	-6.6%	-8.7%	-10.4%
ClassE	FourPeople	-5.8%	-4.4%	-4.5%	**-6.8%**	-5.5%	-5.4%
	Johnny	-4.4%	-5.1%	-5.6%	-5.6%	-6.6%	-6.7%
	KristenAndSara	-5.5%	-5.4%	-5.7%	-6.4%	-6.7%	-6.7%
Summary	**Class A**	-2.6%	-3.1%	-2.9%	-3.5%	-4.2%	-3.8%
	Class B	-2.2%	-2.7%	-3.1%	-3.8%	-3.8%	-4.1%
	Class C	-3.9%	-4.4%	-5.0%	-4.9%	-5.3%	-6.1%
	Class D	-3.9%	-5.1%	-6.6%	-4.9%	-5.9%	-7.7%
	Class E	-5.2%	-5.0%	-5.3%	-6.2%	-6.2%	-6.3%
Average		**-3.6%**	**-4.2%**	**-4.9%**	**-4.8%**	**-5.2%**	**-5.9%**

TABLE IV
RA CONFIGURATION, BD-RATE RESULTS OF CNNF-R AND VRCNN COMPARED WITH HM16.15

Class	BD-rate(VRCNN)			BD-rate(CNNF-R)		
	Y	U	V	Y	U	V
ClassA	-1.4%	-0.9%	-0.6%	-1.7%	-2.3%	-1.9%
ClassB	-1.9%	-0.1%	0.5%	-2.5%	-1.1%	-0.4%
ClassC	-0.5%	-0.7%	-0.8%	-1.0%	-1.5%	-1.6%
ClassD	-0.5%	0.0%	-0.8%	-1.0%	-0.8%	-1.5%
ClassE	-4.5%	-4.0%	-4.0%	-5.4%	-5.7%	-5.2%
Average	**-1.7%**	**-1.0%**	**-1.0%**	**-2.3%**	**-2.0%**	**-1.9%**

TABLE V
AI CONFIGURATION, COMPUTATION COMPLEXITY OF CNNF-R AND VRCNN COMPARED WITH HM16.15 ON CPUs

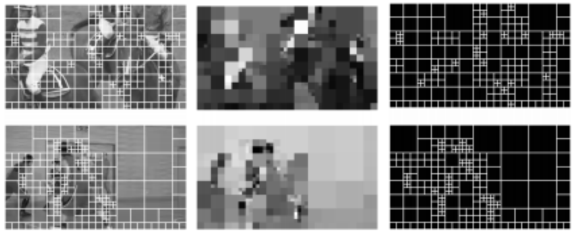
Class	VRCNN vs. HM16.15		CNNF-R vs. HM16.15	
	EncT/times	DecT/times	EncT/times	DecT/times
ClassA	44	3680	13	1082
ClassB	45	3990	13	1137
ClassC	35	2768	10	802
ClassD	37	2094	11	588
ClassE	48	3313	12	935
Average	**42**	**3169**	**12**	**909**

[4] Li D, Yu L. An In-Loop Filter Based on Low-Complexity CNN using Residuals in Intra Video Coding[C]//2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019: 1-5.

In-Loop Filter

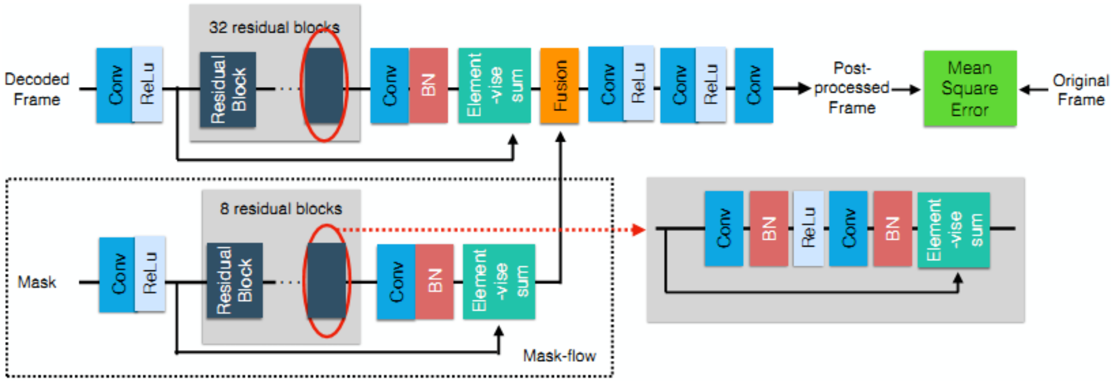
Filtering with block partition information

- **Network input**
  - Current compressed frame
  - Block partition information: CU size



(a) Original frame with partition information (b) Local Mean-based mask (c) Boundary-based mask

- **Network output**
  - Filtered frame
- **Network structure**
  - deep CNN



- **Integration into coding system**
  - Different model for different video content in an Exhaustive search way
  - Different model for different QP
  - Used on I/P/B frames
  - After DB and SAO
  - CTU-level switchable
- **Performance** (anchor: HM16.0)

Conf.	Seq.	VRCNN [18]			Our 2-in+MM+AF (D^*)			Our ASN@4D^*		
		Y	U	V	Y	U	V	Y	U	V
LP	Class A	-7.10	-2.41	-1.97	-10.04	-6.04	-5.72	-12.02	-6.63	-6.33
	Class B	-4.57	-4.13	-5.44	-10.10	-9.48	-11.94	-11.14	-10.64	-13.56
	Class C	-0.21	-2.73	-4.12	-7.39	-8.56	-11.07	-8.26	-9.66	-12.51
	Class D	0.49	-1.88	-2.66	-7.53	-6.76	-8.36	-8.31	-8.23	-8.21
	Class E	-7.11	-11.25	-12.24	-14.97	-17.70	-17.75	-15.14	-18.52	-18.41
	Average	-3.57	-4.13	-4.95	-9.76	-9.30	-10.68	**-10.97**	**-10.34**	**-11.56**
LB	Class A	-4.83	-1.13	-0.76	-7.74	-5.17	-4.88	-9.28	-5.57	-5.31
	Class B	-2.26	-2.68	-3.90	-7.58	-8.33	-10.68	-8.33	-9.21	-11.95
	Class C	0.46	-2.07	-3.39	-6.86	-7.86	-10.20	-7.46	-8.87	-11.50
	Class D	1.18	-1.55	-2.25	-7.31	-6.37	-7.98	-7.67	-6.85	-9.14
	Class E	-5.57	-10.52	-11.38	-13.04	-16.80	-16.73	-13.52	-17.56	-17.29
	Average	-2.04	-3.20	-3.96	-8.23	-8.48	-9.79	**-8.99**	**-9.20**	**-10.77**
RA	Class A	-4.64	-0.42	-0.03	-7.36	-5.13	-4.81	-8.89	-5.39	-5.05
	Class B	-2.11	-1.39	-2.33	-7.53	-7.81	-9.88	-8.29	-8.23	-10.57
	Class C	0.63	-1.48	-2.65	-6.48	-7.84	-10.14	-7.11	-8.72	-11.33
	Class D	1.73	-0.74	-1.43	-6.95	-6.04	-7.88	-7.26	-7.32	-7.25
	Class E	-4.81	-9.48	-10.19	-12.54	-16.01	-15.75	-12.29	-16.05	-15.79
	Average	-1.71	-2.30	-2.93	-7.92	-8.16	-9.40	**-8.57**	**-8.75**	**-9.73**
AI	Class A	-3.61	-1.69	-1.74	-6.41	-3.74	-3.55	-7.27	-3.73	-3.79
	Class B	-1.35	-2.33	-3.80	-5.87	-5.74	-8.32	-5.91	-5.21	-8.21
	Class C	1.00	-2.48	-3.80	-6.11	-6.85	-9.39	-5.92	-7.41	-10.02
	Class D	1.34	-2.46	-3.55	-6.35	-6.11	-8.32	-6.05	-6.06	-6.08
	Class E	-5.13	-8.43	-9.08	-11.68	-12.88	-12.73	-12.29	-13.39	-12.74
	Average	-1.36	-3.17	-4.13	-6.99	-6.71	**-8.24**	**-7.17**	**-6.75**	-7.94

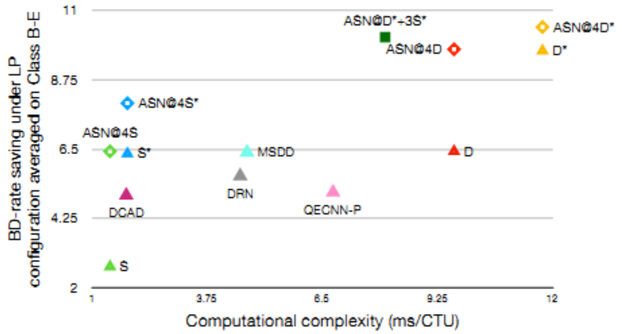


Figure 14. Comparison of different methods on computational time per CTU in decoder side versus BD-rate saving over HEVC baseline.

(1) VRCNN (S) [18] which is a baseline CNN-based compressed-video post-processing method; (2) QECNN-P [20] which is a compressed-video post-processing method for P frames in HEVC; (3) DRN [21], which is another state-of-the-art compressed-video post-processing method. (4) VR-CNN+MM+AF (S^*), which integrates our partition-aware-based approach into the existing baseline VRCNN method; (5) DRN+MM+AF, which integrates our partition-aware-based approach into the existing DRN method; (6) Our 2-in+MM+AF (D^*), which is the full version of our partition-aware-based approach with local mean-based mask and add-based fusion; (7) Our ASN@4D^*, which is the adaptive-switching scheme with the deep CNN model. From the table,

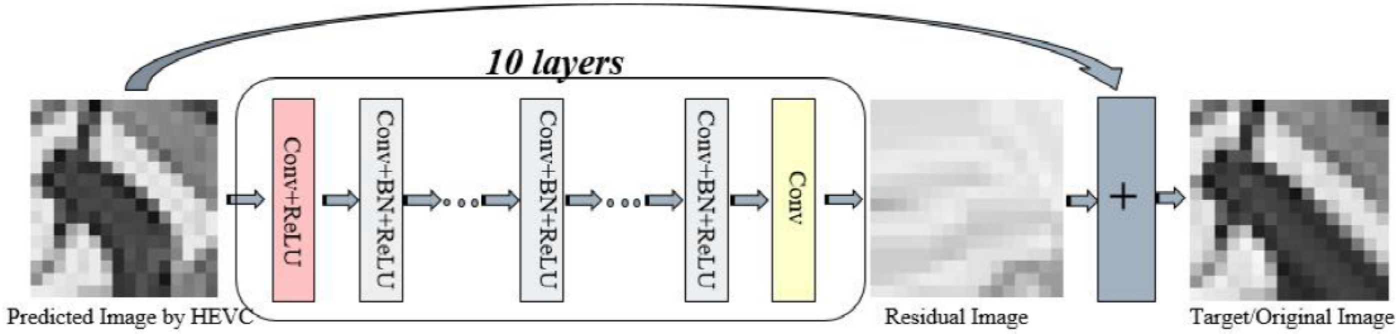
[5] Lin W, He X, Han X, et al. Partition-Aware Adaptive Switching Neural Networks for Post-Processing in HEVC[J]. IEEE Transactions on Multimedia, 2019.

In-Loop Filter

- **Content adaptive filtering**
  - Filtering for reconstructed pixels
    - Inserted into diff. position of in-loop filtering chain: deblocking → SAO → ALF
    - Replace some filters in the chain
  - Information utilized
    - Reconstructed pixels in current frame
    - Temporal neighboring pixels
    - QP map, blocksize, prediction residuals, …
  - Network
    - From 4-layer to deep

Spatial-Temporal Prediction: Intra

Prediction block refinement using CNN



- **Integration into coding system**
  - Replace all existing intra modes
  - Fixed block size
- **Performance** (anchor: AI, HM14.0)

Table 1 BD-rate saving for The Proposed Scheme with Ranges of Sequences

Sequences	BD-rate	Sequences	BD-rate
Traffic	-0.9%	PartyScene	-0.5%
PeopleOnStreet	-1.2%	RaceHorses	-0.7%
Kimono	-0.2%	BasketballPass	-0.4%
ParkScene	-0.8%	BQSquare	-0.1%
Cactus	-0.8%	BlowingBubbles	-0.7%
BasketballDrive	-0.6%	RaceHorses	-0.7%
BQTerrace	-0.8%	FourPeople	-0.3%
BasketballDrill	-0.5%	Johnny	-1.0%
BQMall	-0.6%	KristenAndSara	-0.8%
All average	-0.70%		

- **Network input**
  - 8x8 PU and its three nearest 8x8 reconstruction blocks
- **Network output**
  - Refined PU
- **Network Structure:** composed of 10 weight layers
  - Conv+ ReLU: for the first layer, 64 filters of size 3×3×c
  - Conv + BN + ReLU: for layers 2 ~ 9, 64 filters of size 3×3×64
  - Conv: for the last layer, c filters of size 3x3x64

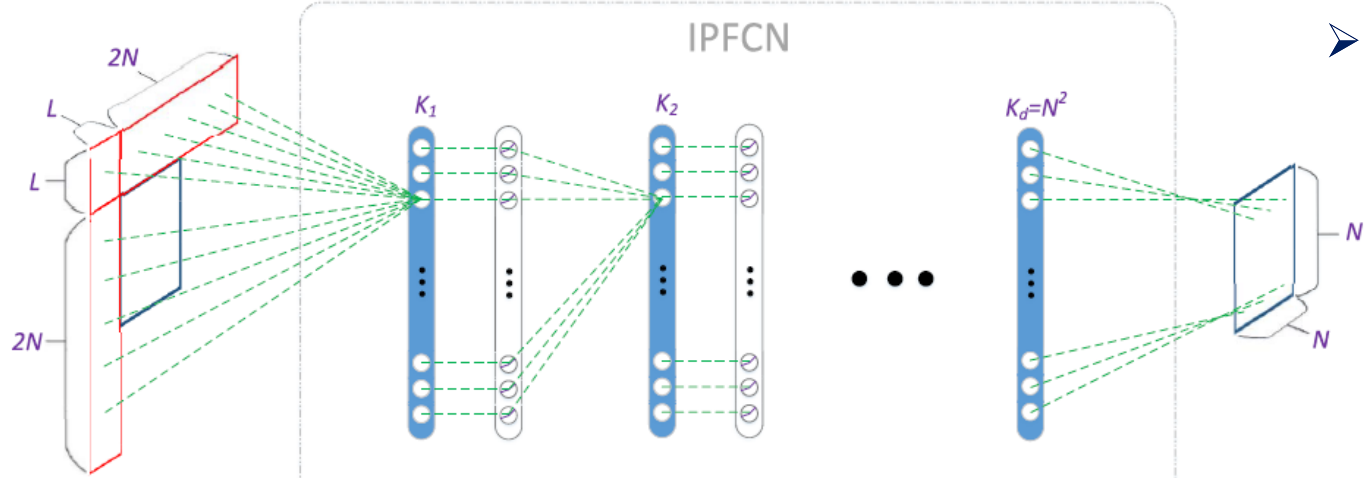
  *c: c represents the number of image channels

[1] Cui W, Zhang T, Zhang S, et al. Convolutional neural networks based intra prediction for HEVC[J]. arXiv preprint arXiv:1808.05734, 2018.

Spatial-Temporal Prediction: Intra

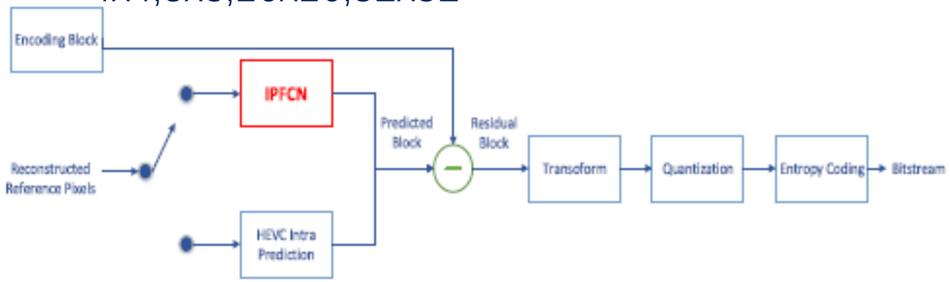
Prediction Block Generation Using CNN

- **Network input**
  - 8 rows and 8 columns reference pixels
- **Network output**
  - prediction block
- **Network Structure:**
  - 4 fully connected networks with PReLU



Fully Connected Layer   Non-Linear Activation Layer

- **Integration into coding system**
  - As an additional intra mode
  - CU-level selective
  - Different models for all TU size in HEVC：4x4,8x8,16x16,32x32



- **Performance** (anchor: AI, HM16.9)

Sequence	IPFCN-D			IPFCN-S		
	Small QPs	Normal QPs	Large QPs	Small QPs	Normal QPs	Large QPs
Class A (4K)	−2.2%	−4.5%	−5.0%	−1.8%	−3.8%	−4.5%
Class B (1080P)	−1.9%	−3.1%	−3.9%	−1.5%	−2.7%	−3.0%
Class C (WVGA)	−1.1%	−2.1%	−3.3%	−0.9%	−1.8%	−2.6%
Class D (WQVGA)	−0.9%	−1.8%	−3.0%	−0.8%	−1.5%	−2.9%
Class E (720P)	−2.3%	−4.5%	−4.2%	−1.8%	−4.2%	−3.8%
Average of All Classes	−1.8%	−3.4%	−4.0%	−1.4%	−2.9%	−3.5%

*Small QPs: {11, 16, 21, 26}, Normal QPs: {22, 27, 32, 37}, Large QPs: {33, 38, 43, 48}.

IPFCN-D: different model for angular intra modes and non-angular intra modes, respectively
IPFCN-S: same model for angular intra modes and non-angular intra modes

[2] Li J, Li B, Xu J, et al. Fully connected network-based intra prediction for image coding[J]. IEEE Transactions on Image Processing, 2018, 27(7): 3236-3247.

Spatial-Temporal Prediction: Intra

Prediction Block Generation Using RNN

- Network input
  - neighboring reconstructed pixels and current PU

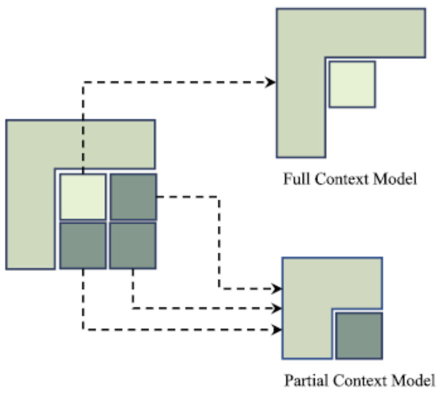
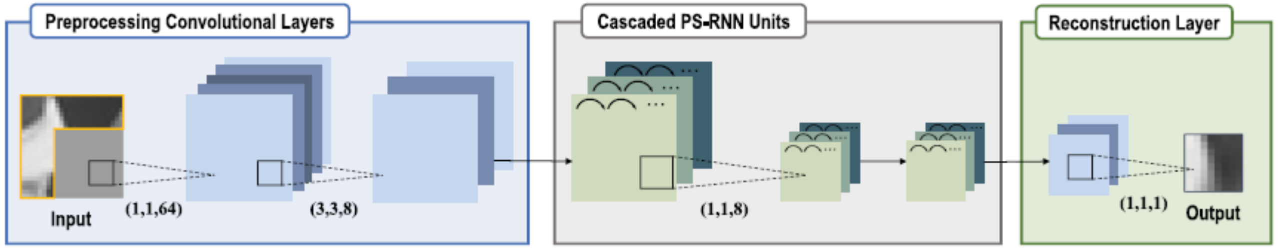


Fig. 4. Different availability of reference samples in a coding unit. Blocks with two different colors are processed using two different models.

- Network output
  - prediction block
- Training strategy:
  - Loss Function : MSE/SATD

- Network Structure:
  - Overall structure: CNN + RNN
    - ✓ using CNN to extract local features of the input context block and transform the image to feature space.
    - ✓ using PS-RNN units to generate the prediction of the feature vectors.



  - PS-RNN:

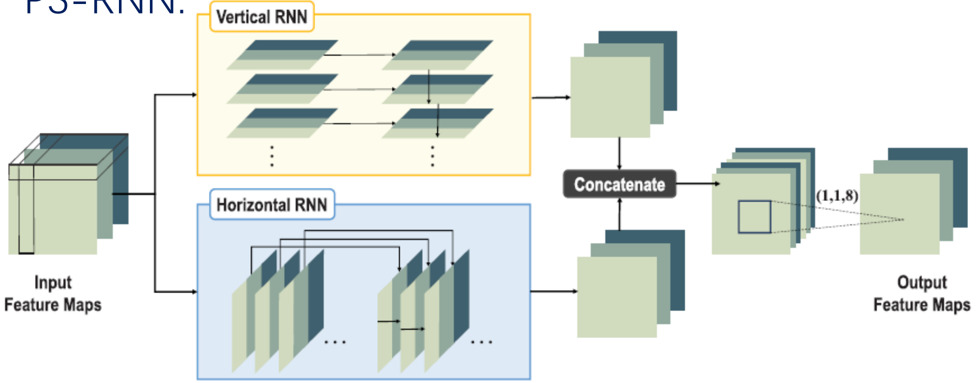


Fig. 2. Structure of a PS-RNN unit. It splits a stack of feature maps into vertical and horizontal planes. Each plane represents a feature map of a vertical line or a horizontal line in the original grey-scale image. After the progressive prediction, these planes are concatenated to reconstruct the feature maps. A convolutional layer is used to fuse the predictions from the vertical and horizontal feature maps.

[3] Hu Y, Yang W, Li M, et al. Progressive spatial recurrent neural network for intra prediction[J]. IEEE Transactions on Multimedia, 2019, 21(12): 3024-3037.

Spatial-Temporal Prediction: Intra

Prediction block generation using RNN

- **Performance** (anchor: AI, HM16.15)

TABLE I
QUANTITATIVE ANALYSIS OF SELECTED METHODS. THE RESULTS ARE SHOWN IN BD-RATE USING HEVC (HM 16.15) AS THE ANCHOR. PU SIZE IS SET TO 8 × 8 IN BOTH THE PROPOSED MODEL AND THE ANCHOR

Class	Sequence	PS-RNN-SATD	PS-RNN-MSE	FC-SATD	Li [16]
Class A	Traffic	**-3.8%**	-2.3%	-3.1%	-1.0%
	PeopleOnStreet	**-3.8%**	-2.2%	-3.1%	-1.3%
	Nebuta(10bit)	**-1.9%**	**-1.9%**	**-1.9%**	-1.6%
	SteamLocomotive(10bit)	**-3.2%**	-2.8%	**-3.2%**	-1.7%
	Class A Average	**-3.2%**	-2.3%	-2.8%	-1.4%
Class B	Kimono	**-6.6%**	-3.6%	-6.4%	-3.2%
	ParkScene	**-3.4%**	-1.9%	-2.9%	-1.1%
	Cactus	**-3.3%**	-1.8%	-2.2%	-0.9%
	BasketballDrive	**-7.8%**	-3.2%	-3.7%	-0.9%
	BQTerrace	**-2.6%**	-1.8%	-1.6%	-0.5%
	Class B Average	**-4.7%**	-2.5%	-3.4%	-1.3%
Class C	BasketballDrill	**-2.9%**	-1.5%	-1.9%	-0.3%
	BQMall	**-2.9%**	-1.9%	-1.4%	-0.3%
	PartyScene	**-2.3%**	-1.8%	-1.1%	-0.4%
	RaceHorses	**-2.8%**	-2.1%	-2.3%	-0.8%
	Class C Average	**-2.7%**	-1.8%	-1.7%	-0.5%
Class D	BasketballPass	**-2.5%**	-1.7%	-1.4%	-0.4%
	BQSquare	**-1.8%**	-1.2%	-0.8%	-0.2%
	BlowingBubbles	**-2.3%**	-1.6%	-1.7%	-0.6%
	RaceHorses	**-2.6%**	-2.5%	-2.2%	-0.6%
	Class D Average	**-2.3%**	-1.8%	-1.5%	-0.5%
Class E	Johnney	**-6.8%**	-3.8%	-4.7%	-1.0%
	FourPeople	**-5.6%**	-2.8%	-4.1%	-0.8%
	KristenAndSara	**-6.6%**	-2.9%	-4.0%	-0.8%
	Class E Average	**-6.3%**	-3.2%	-4.3%	-0.9%
Average		**-3.8%**	-2.3%	-2.7%	-0.9%

[3] Hu Y, Yang W, Li M, et al. Progressive spatial recurrent neural network for intra prediction[J]. IEEE Transactions on Multimedia, 2019, 21(12): 3024-3037.

Spatial-Temporal Prediction: Intra

Prediction Block Generation Using Single Layer Network

- **Network input**
  - R rows and R columns reference pixels
    - ✓ Height/width of current block smaller than 32: R = 2
    - ✓ Otherwise: R =1
  - Mode:
    - ✓ Height/width of current block smaller than 32: 35 modes
    - ✓ Otherwise: 11 modes

- **Network output**
  - prediction block

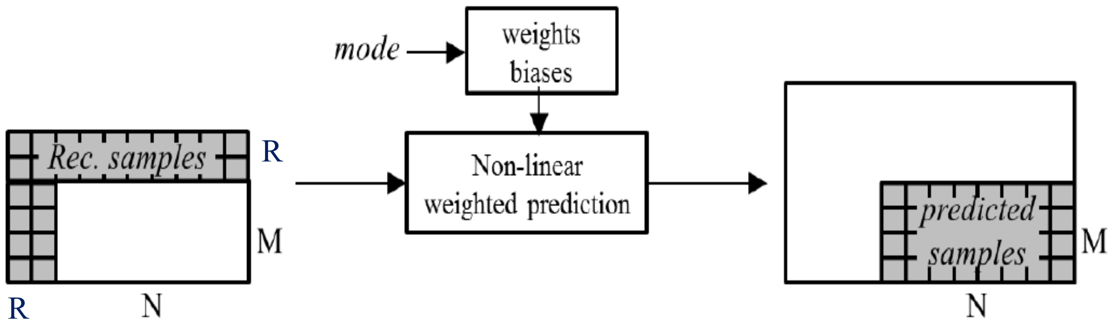


Figure 1. Prediction of MxN intra block from reconstructed samples using a neural network.

- **Network Structure:**
  - 2-layer neural network during training
    - ✓ Layer1: feature extraction, same for all modes
    - ✓ Layer2: prediction, different for different modes

$$f(\mathbf{x})_i = \max(-1, \mathbf{x}_i),$$
$$\mathbf{t}_1 = f(\mathbf{A}_1\mathbf{r} + \mathbf{b}_1)$$
$$\mathbf{p}_k(\mathbf{r}) = \mathbf{A}_{2,k}\mathbf{t}_1 + \mathbf{b}_{2,k}.$$

R: reference samples
$\{A_{i,k}, b_i\}$ = network parameters
i = network layer index , k = mode index
$P_k(r)$ = output prediction results

  - Network Simplification:
    - ✓ Pruning: compare the predictor network and the zero predictor in terms of loss function in frequency domain. If loss decrease is smaller than threshold, use zero predictor instead.
    - ✓ Affine linear predictors: removing the activation function, using a single matrix multiplication and bias instead.

[4] Helle P, Pfaff J, Schäfer M, et al. Intra picture prediction for video coding with neural networks[C]//2019 Data Compression Conference (DCC). IEEE, 2019: 448-457.

Spatial-Temporal Prediction: Intra

Prediction Block Generation Using Single Layer Network

- **Signaling mode index**
  - Use a two-layer network to predict the conditional probability of each mode
  - The outputs from step#1 are sorted to obtain an MPM-list and an index is signaled in the same way as a conventional intra prediction mode index.

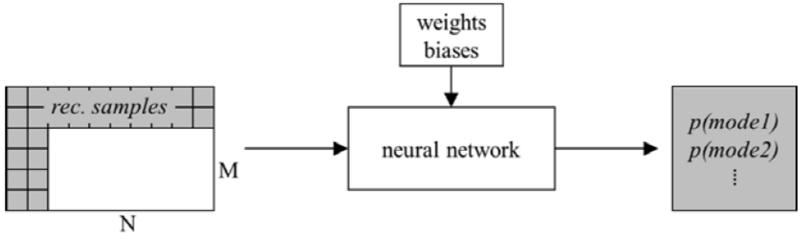


Figure 2. Prediction of mode probabilities from reconstructed samples using a neural network.

- **Integration into coding system**
  - Network generated prediction as an additional intra mode
  - RDO to choose intra mode

- **Performance** (anchor: AI, VTM1.0)

Sequence class	Sequence name	BD-Rate Y'	BD-Rate Cb	BD-Rate Cr
A1	Tango2	-5.20	-4.31	-3.42
	FoodMarket4	-5.67	-2.90	-3.06
	Campfire	-1.44	-0.88	-1.29
A2	CatRobot1	-3.66	-2.69	-1.96
	DaylightRoad2	-4.01	-1.60	-2.47
	ParkRunning3	-1.93	-1.81	-2.24
B	MarketPlace	-3.11	-1.33	-0.48
	RitualDance	-5.49	-2.89	-2.68
	Cactus	-3.88	-2.17	-1.99
	BasketballDrive	-2.92	-1.95	-2.00
	BQTerrace	-2.60	-0.60	0.10
C	RaceHorses	-2.78	-1.63	-1.88
	BQMall	-4.40	-2.47	-2.29
	PartyScene	-2.89	-1.56	-1.62
	BasketballDrill	-2.51	-2.32	-2.61
D	RaceHorses	-3.74	-2.43	-2.00
	BQSquare	-2.84	-0.79	-1.05
	BlowingBubbles	-2.71	-1.63	-2.00
	BasketballPass	-3.43	-2.22	-2.52
E	FourPeople	-6.23	-2.93	-3.44
	Johnny	-5.80	-2.80	-2.85
	KristenAndSara	-6.12	-3.23	-3.88
average		**-3.79**	**-2.14**	**-2.17**

[4] Helle P, Pfaff J, Schäfer M, et al. Intra picture prediction for video coding with neural networks[C]//2019 Data Compression Conference (DCC). IEEE, 2019: 448-457.

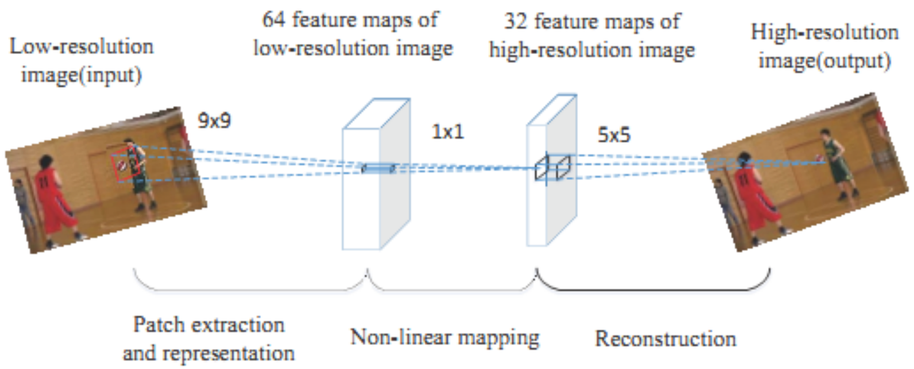
Spatial-Temporal Prediction: Intra

- **Prediction for block of pixel values**
  - Refinement of traditional prediction: content adaptive filtering
  - Prediction by extrapolation
    - Prediction domain: spatial domain, frequency domain
    - Supplement or replace to traditional modes
    - Network architecture: CNN, RNN, FCN and their combinations
    - Reference pixels: one or multiple raw(s)/column(s)
    - Loss function: Energy of residuals in spatial domain (MSE), Hardmad transform domain (SATD), DCT domain
- **Prediction of intra mode**
  - Probability estimation for all modes: Most Probability Modes list

Spatial-Temporal Prediction: Inter

Subpixel Interpolation

- **Network input**
  - Integer-pixel frame
- **Network output**
  - Half-pixel Interpolated frame
- **Network Structure:**
  - SRCNN : 4-layer CNN



- **Integration into coding system**
  - Different model for different QP
  - Directly replace ½ DCTIF
- **Performance** (anchor: LDP, HM16.7)

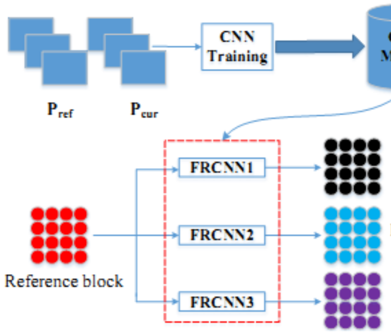
Class	Sequence	BD-rate Y (%)	U (%)	V (%)
Class B	Kimono	-1.1	0.1	0.2
	ParkScene	-0.4	-0.3	-0.3
	Cactus	-0.8	0.0	0.3
	BasketballDrive	-1.3	-0.2	-0.1
	BQTerrace	-3.2	-1.6	-1.6
Class C	BasketballDrill	-1.2	-0.6	0.2
	BQMall	-0.9	0.2	0.7
	PartyScene	0.2	0.5	0.3
	RaceHorses	-1.5	-0.5	-0.1
Class D	BasketballPass	-1.3	-0.4	0.3
	BQSquare	1.2	2.9	3.1
	BlowingBubbles	-0.3	0.4	0.8
	RaceHorses	-0.8	-0.9	0.0
Class E	FourPeople	-1.3	-0.4	0.1
	Johnny	-1.2	-0.4	-0.7
	KristenAndSara	-1.0	0.3	0.2
Class F	BasketballDrillText	-1.4	-0.2	0.1
	ChinaSpeed	-0.6	-0.5	-0.3
	SlideEditing	0.0	0.3	0.4
	SlideShow	-0.7	-0.1	-0.2
Class Summary	Class B	-1.4	-0.4	-0.3
	Class C	-0.9	-0.1	0.3
	Class D	-0.3	0.5	1.0
	Class E	-1.2	-0.2	-0.1
	Class F	-0.7	-0.1	0.0
Overall	**All**	**-0.9**	**-0.1**	**0.2**

[1] Yan N, Liu D, Li H, et al. A convolutional neural network approach for half-pel interpolation in video coding[C]//2017 IEEE international symposium on circuits and systems (ISCAS). IEEE, 2017: 1-4..

Spatial-Temporal Prediction: Inter

Subpixel Interpolation

- **Network input**
  - Integer-pixel position samples
- **Network output**
  - Half-pixel position samples of each sub-pixel position
- **Network Structure:**
  - Different FRCNN for different half-pixel position
  - FRCNN: 4-layer CNN with Inception structure



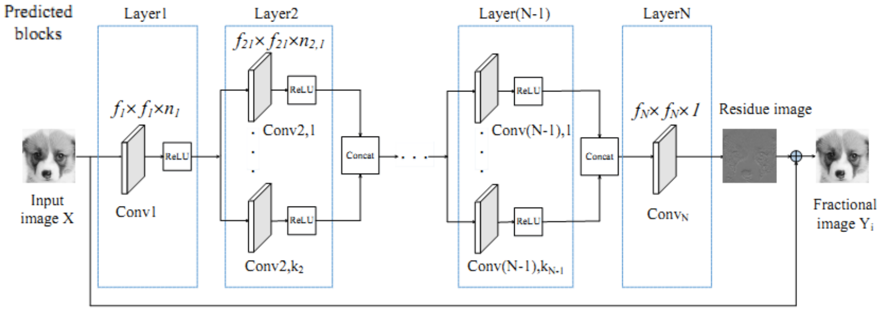


Fig. 3. Schematic illustration of FRCNN, which consists of a series of convolutional modules and non-linear units (ReLU [20] in this paper). The integer-pixel values of a reference picture are taken as input to the network, and the output is the predicted pixel values corresponding to a specific fractional-pixel position (i.e. $a_{i,j} - r_{i,j}$ in Fig. 1).

- **Integration into coding system**
  - Different model for different QP, different half-pixel position and different inter-prediction direction
  - Use as an additional interpolation filter: CU-level selection between CNN,½ DCTIF and ¼ DCTIF
- **Performance** (anchor: HM16.7)

TABLE II

BD-Rate Results of Our Scheme Compared to HEVC (Entire Sequence, Training Data Generated by BlowingBubbles at Different QPs)

Class	Sequence	BD-rate of LDP Y (%)	BD-rate of LDP U (%)	BD-rate of LDP V (%)	BD-rate of LDB Y (%)	BD-rate of LDB U (%)	BD-rate of LDB V (%)	BD-rate of RA Y (%)	BD-rate of RA U (%)	BD-rate of RA V (%)
Class A	Traffic	–	–	–	–	–	–	−0.7	−0.0	−0.0
	PeopleOnStreet	–	–	–	–	–	–	−0.9	−1.2	−0.8
	Nebuta	–	–	–	–	–	–	−0.8	−0.9	−1.0
	SteamLocomotive	–	–	–	–	–	–	−1.4	−1.2	−1.4
Class B	Kimono	−4.3	0.2	0.1	−1.2	0.9	0.5	−0.5	−0.2	−0.3
	ParkScene	−1.9	−1.0	−1.3	−0.9	−0.6	−0.6	−0.3	−0.1	−0.1
	Cactus	−3.8	−0.9	−1.1	−2.2	−0.3	−1.3	−1.2	−0.5	−0.4
	BasketballDrive	−5.0	−1.8	−1.7	−2.5	−0.5	−0.7	−1.5	−0.8	−0.7
	BQTerrace	−6.5	−4.1	−4.7	−2.9	−0.8	−0.8	−1.5	−0.2	−0.2
Class C	BasketballDrill	−4.0	−1.1	1.1	−3.8	−0.8	−0.2	−1.8	−0.3	−0.4
	BQMall	−4.8	2.3	−1.7	−3.8	−0.8	−0.2	−1.9	−0.5	−0.6
	PartyScene	−3.2	−1.6	−2.0	−3.3	−0.9	−0.8	−2.3	−0.7	−0.8
	RaceHorses	−3.0	−1.8	−1.9	−1.5	−0.9	−1.3	−1.1	−0.8	−1.2
Class D	BasketballPass	−3.3	−1.8	−1.4	−2.1	−0.5	−1.2	−1.1	−1.0	−0.9
	BQSquare	−4.2	−0.7	−1.1	−4.6	−2.1	−2.7	−2.9	−1.1	−1.6
	BlowingBubbles	−4.7	−1.0	−0.9	−5.4	−0.9	0.6	−2.4	−0.9	−0.6
	RaceHorses	−1.9	−1.7	−1.6	−1.4	−0.8	−0.6	−0.8	−0.6	−0.5
Class E	FourPeople	−5.7	−1.9	−1.9	−3.6	−0.6	−0.4	–	–	–
	Johnny	−6.2	−1.8	−2.7	−3.8	0.0	−1.0	–	–	–
	KristenAndSara	−6.3	−1.3	−1.4	−4.9	−0.9	0.2	–	–	–
Class F	BasketballDrillText	−4.1	−2.1	−0.8	−3.5	−0.6	−0.5	−1.7	−0.5	−0.6
	ChinaSpeed	−2.0	−1.7	−1.2	−1.4	−1.2	0.0	−1.3	−1.3	−1.1
	SlideEditing	−0.7	−0.2	−0.3	−0.5	−0.2	−0.3	−0.1	−0.1	−0.1
	SlideShow	−2.3	−1.8	−2.2	−1.9	−2.7	−1.8	−0.7	−0.4	−0.5
Class Summary	Class A	–	–	–	–	–	–	−0.9	−0.8	−0.8
	Class B	−4.3	−1.5	−1.8	−1.9	−0.3	−0.4	−1.0	−0.4	−0.3
	Class C	−3.8	−1.7	−1.1	−3.1	−0.8	−0.7	−1.8	−0.6	−0.8
	Class D	−3.5	−1.3	−1.3	−3.4	−1.1	−0.9	−1.8	−0.9	−0.9
	Class E	−6.1	−1.7	−2.0	−4.1	−0.5	−0.4	–	–	–
	Class F	−2.3	−1.5	−1.4	−1.8	−1.2	−0.7	−0.7	−0.4	−0.5
Overall	**All**	**−3.9**	**−1.5**	**−1.4**	**−2.7**	**−0.7**	**−0.6**	**−1.3**	**−0.6**	**−0.7**

[2] Yan N, Liu D, Li H, et al. Convolutional neural network-based fractional-pixel motion compensation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(3): 840-853..

Spatial-Temporal Prediction: Inter

Subpixel Interpolation

- **Network input**
  - Integer-pixel position samples
- **Network output**
  - Quarter/half-pixel position samples of each sub-pixel position
- **Network Structure:**
  - Grouped variation neural network:
    - ✓ one model can generate all sub-pixel positions at one sub-pixel level and deal with frames coded with different QPs.
    - ✓ Shared feature map is generated and then used to infer sub-pixel samples at different locations.

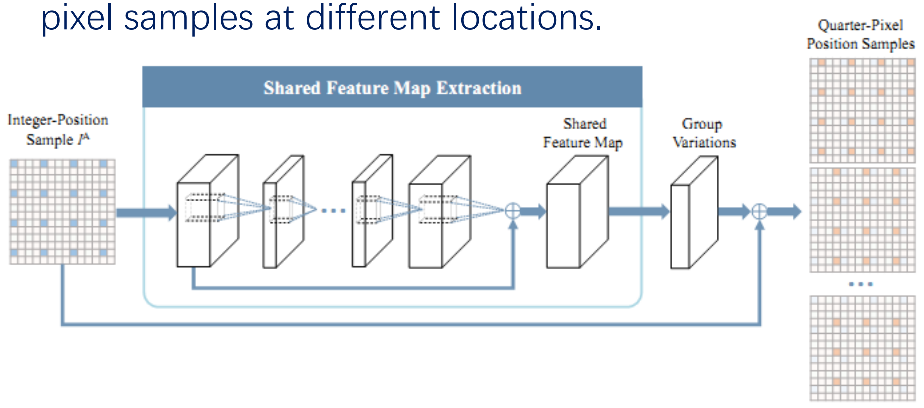


Fig. 2. Framework of the proposed GVCNN. The network first extracts feature maps from the integer-position sample. Then the group variations that identify the differences between different sub-pixel position samples and the integer-position sample are inferred using the same feature maps. Final results of sub-pixel position samples are naturally obtained by adding the variations back to the integer-position sample.

- **Integration into coding system**
  - Different model for different sub-pixel level
  - Use as an additional interpolation filter: CU-level selection between CNN,½ DCTIF and ¼ DCTIF
- **Performance** (anchor: HM16.4)

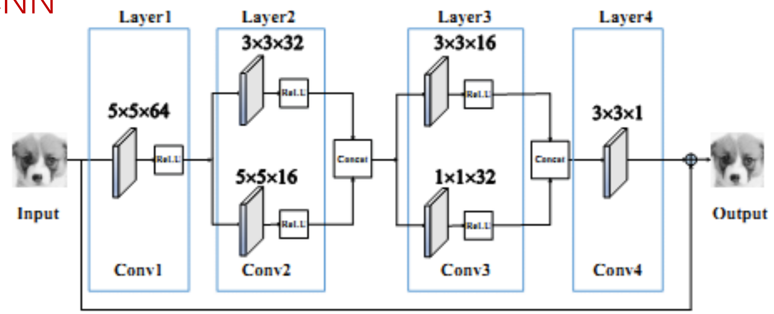
Class	Sequence	BD-rate of LDP			BD-rate of LDB			BD-rate of RA		
		Y	U	V	Y	U	V	Y	U	V
Class A	Traffic	-	-	-	-	-	-	-1.1%	0.1%	0.3%
	PeopleOnStreet	-	-	-	-	-	-	-0.9%	-0.1%	-0.8%
	Nebuta	-	-	-	-	-	-	-0.1%	-0.3%	-0.4%
	SteamLocomotive	-	-	-	-	-	-	-0.2%	-0.5%	-0.5%
	Average	-	-	-	-	-	-	-0.6%	-0.2%	-0.3%
Class B	Kimono	-4.1%	2.1%	1.6%	-1.7%	0.9%	0.4%	-1.3%	0.2%	-0.2%
	BQTerrace	-5.2%	-3.4%	-3.9%	-1.3%	0.3%	0.2%	-2.5%	0.0%	-1.1%
	BasketballDrive	-3.3%	0.2%	-0.5%	-0.5%	0.5%	0.1%	-1.4%	-0.8%	-0.8%
	ParkScene	-1.3%	0.0%	-0.5%	-0.8%	0.7%	0.2%	-0.7%	0.3%	-0.5%
	Cactus	-2.5%	-0.5%	-0.8%	-1.0%	0.0%	-0.1%	-1.1%	-0.2%	-1.0%
	Average	-3.3%	-0.3%	-0.8%	-1.1%	0.8%	0.4%	-1.4%	-0.1%	-0.7%
Class C	BasketballDrill	-2.2%	-1.3%	-0.6%	-1.0%	0.0%	-0.1%	-0.7%	0.0%	0.1%
	BQMall	-2.9%	-2.1%	-1.7%	-1.3%	-0.9%	-1.0%	-1.1%	-0.2%	-1.0%
	PartyScene	-1.6%	-0.5%	-1.4%	-0.9%	-0.2%	-0.4%	-0.7%	-0.4%	-1.1%
	RaceHorsesC	-2.0%	-1.4%	-1.6%	-1.5%	-0.4%	-0.8%	-1.6%	-1.5%	-1.3%
	Average	-2.2%	-1.3%	-1.3%	-1.1%	-0.4%	-0.6%	-1.0%	-0.5%	-0.8%
Class D	BasketballPass	-3.3%	-1.7%	-1.4%	-1.8%	-0.9%	-1.5%	-0.9%	-1.0%	-1.5%
	BlowingBubbles	-2.1%	-0.9%	-0.3%	-1.0%	0.2%	-0.2%	-0.8%	-0.7%	0.2%
	BQSquare	-0.6%	1.2%	3.1%	-0.7%	1.6%	0.9%	-0.7%	-0.3%	-0.5%
	RaceHorses	-2.7%	-1.7%	-0.8%	-1.9%	0.0%	-0.9%	-1.4%	-0.9%	-1.1%
	Average	-2.2%	-0.7%	0.2%	-1.4%	0.2%	-0.4%	-1.0%	-0.7%	-0.7%
Class E	FourPeople	-1.6%	-0.5%	-0.2%	-2.8%	5.9%	0.0%	-	-	-
	Johnny	-2.9%	0.2%	0.4%	-1.2%	1.1%	1.4%	-	-	-
	KristenAndSara	-2.2%	1.1%	0.2%	-1.3%	1.6%	1.2%	-	-	-
	Average	-2.2%	0.3%	0.2%	-1.6%	2.5%	0.8%	-	-	-
Class F	BasketballDrillText	-1.8%	-0.9%	0.1%	-1.0%	-0.6%	-0.8%	-0.9%	0.2%	-0.5%
	ChinaSpeed	-1.4%	-1.9%	-1.4%	-1.0%	-1.3%	-1.5%	-1.4%	-2.0%	-1.8%
	SlideEditing	0.0%	-0.1%	-0.2%	0.0%	-0.1%	-0.1%	0.6%	0.9%	0.9%
	SlideShow	-0.5%	0.2%	-0.6%	-0.9%	0.5%	1.0%	-0.8%	0.3%	-0.9%
	Average	-0.9%	-0.7%	-0.5%	-0.7%	-0.9%	-0.6%	-0.6%	-0.2%	-0.6%
All Sequences	Overall	-2.2%	-0.6%	-0.5%	-1.2%	0.4%	-0.1%	-0.9%	-0.3%	-0.6%

[3] Liu J, Xia S, Yang W, et al. One-for-all: Grouped variation network-based fractional interpolation in video coding[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2140-2151.

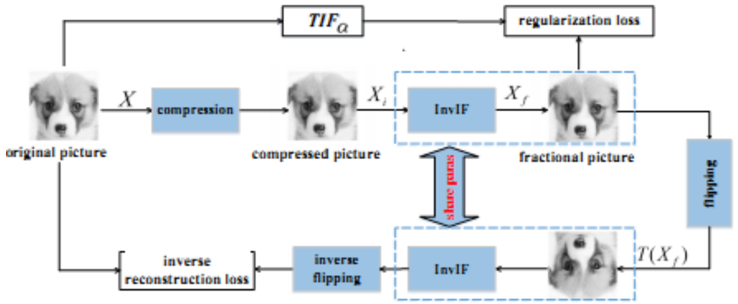
Spatial-Temporal Prediction: Inter

Subpixel Interpolation

- **Network input**
  - Integer-pixel position samples
- **Network output**
  - Half-pixel position samples of each sub-pixel position
- **Network Structure:**
  - 4-layer CNN



- **Training Scheme:**
  - Interpolate sub-pixel samples from integer-pixel samples
  - Recover integer-pixels samples from sub-pixel samples



- **Integration into coding system**
  - Different model for different QP, different sub-pixel position
  - Additional mode and replacement mode are studied

- **Performance** (anchor: HM16.7)

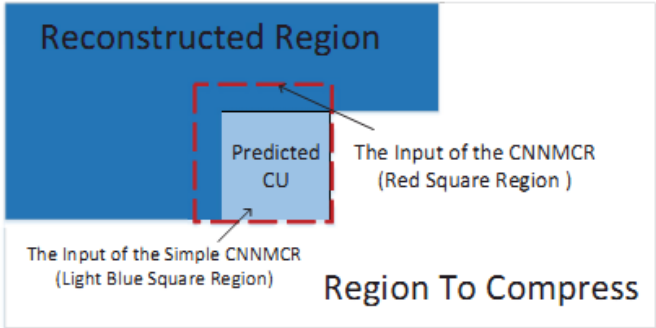
Class	Sequence	Choose between DCTIF/InvIF						InvIF Only					
		LDB			RA			LDB			RA		
		Y (%)	U (%)	V (%)	Y (%)	U (%)	V (%)	Y (%)	U (%)	V (%)	Y (%)	U (%)	V (%)
Class A	Traffic	–	–	–	−4.1	−0.2	−0.2	–	–	–	−3.3	0.2	−0.3
	PeopleOnStreet	–	–	–	−3.2	−0.7	−0.7	–	–	–	−3.0	−0.5	−0.3
	Nebuta	–	–	–	−0.5	−0.7	−0.5	–	–	–	0.6	0.5	0.2
	SteamLocomotive	–	–	–	−2.4	−0.6	−1.2	–	–	–	−0.6	0.5	−0.9
Class B	Kimono	−2.8	0.3	0.3	−2.4	−0.2	−0.3	−1.4	1.0	0.7	−1.8	0.3	0.2
	ParkScene	−2.2	−0.5	−0.7	−2.6	−0.2	−0.2	0.7	0.3	0.1	−2.0	0.2	0.1
	Cactus	−4.6	−0.3	−1.3	−4.1	−0.7	−0.5	−2.0	0.7	1.6	−3.6	−0.3	−0.1
	BasketballDrive	−4.0	−0.3	−0.5	−2.9	−0.6	−0.6	−2.6	1.1	0.7	−1.8	0.1	0.3
	BQTerrace	−5.4	−1.6	−0.7	−6.7	−1.1	−1.5	1.9	1.5	2.3	−4.0	−0.3	−0.7
Class C	BasketballDrill	−5.1	0.4	0.9	−3.7	−0.1	−0.1	−4.1	1.2	−1.9	−3.4	0.4	0.2
	BQMall	−5.9	−1.0	−1.0	−4.7	−0.6	−1.1	−3.9	−0.4	−0.7	−3.9	−0.3	−0.5
	PartyScene	−4.7	−2.2	−2.4	−5.7	−1.8	−1.8	−2.7	−2.0	−2.0	−5.2	−1.5	−1.5
	RaceHorses	−3.1	−0.4	−0.9	−2.4	−0.5	−0.6	−2.1	−0.0	0.6	−1.9	0.2	0.0
Class D	BasketballPass	−3.8	−0.7	−0.7	−2.6	−0.6	−0.7	−3.3	−0.1	−0.1	−2.3	−0.3	−0.2
	BQSquare	−10.3	−4.7	−4.9	−13.1	−5.4	−5.2	−8.0	−3.3	−3.2	−12.5	−5.1	−5.1
	BlowingBubbles	−5.3	−2.0	−0.9	−5.3	−1.1	−1.2	−3.6	−1.0	−0.1	−5.0	−0.9	−1.0
	RaceHorses	−4.2	−0.9	−0.3	−2.9	−0.2	−0.1	−3.5	−0.2	0.2	−2.5	0.0	0.2
Class E	FourPeople	−7.9	−0.3	−0.4	–	–	–	−5.4	0.6	0.6	–	–	–
	Johnny	−6.8	1.8	0.3	–	–	–	−1.3	5.8	6.4	–	–	–
	KristenAndSara	−8.8	0.8	1.4	–	–	–	−6.4	2.6	3.8	–	–	–
Class F	BasketballDrillText	−4.9	−0.3	−0.4	−3.8	−0.5	−0.6	−4.0	0.3	1.0	−3.4	0.2	−0.2
	ChinaSpeed	−1.4	−1.0	−0.9	−1.0	−1.0	−0.9	−0.1	0.5	0.4	−0.3	−0.2	0.0
	SlideEditing	−0.4	−0.1	−0.2	−0.3	−0.2	−0.2	0.0	0.3	0.5	−0.0	−0.1	−0.1
	SlideShow	−3.2	−3.1	−1.5	−0.9	−0.3	−0.4	−2.1	−2.3	−2.3	−0.4	0.1	0.0
Summary	Class A	–	–	–	−2.5	−0.5	−0.6	–	–	–	−1.6	0.2	−0.2
	Class B	−3.8	−0.6	−0.3	−3.7	−0.6	−0.6	−0.7	0.9	1.1	−2.6	0.0	0.0
	Class C	−4.7	−0.8	−0.8	−4.2	−0.7	−0.9	−3.2	−0.3	−0.1	−3.6	−0.3	−0.4
	Class D	−5.9	−2.1	−1.7	−6.0	−1.8	−1.8	−4.6	−1.1	−0.8	−5.6	−1.6	−1.5
	Class E	−7.9	0.5	0.4	–	–	–	−4.4	3.0	3.6	–	–	–
	Class F	−2.5	−1.1	−0.7	−1.5	−0.5	−0.5	−1.6	−0.3	−0.1	−1.0	−0.1	−0.1
Overall	**All**	**−4.7**	**−0.8**	**−0.7**	**−3.6**	**−0.8**	**−0.9**	**−2.7**	**0.3**	**0.6**	**−2.9**	**−0.3**	**−0.4**

[4] Yan N, Liu D, Li H, et al. Invertibility-Driven Interpolation Filter for Video Coding[J]. IEEE Transactions on Image Processing, 2019, 28(10): 4912-4925.

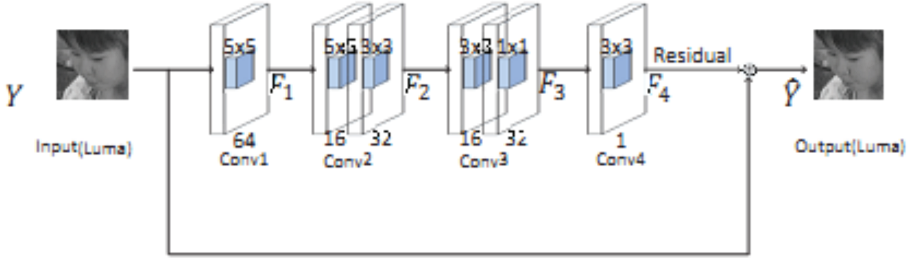
Spatial-Temporal Prediction: Inter

Block Refinement of Uni-Prediction

- **Network input**
  - Predicted CU by conventional methods
  - L-shape neighboring pixels of current CU



- **Network output**
  - Refined predicted block
- **Network Structure:**
  - VRCNN：4-layer CNN



- **Integration into coding system**
  - Different model for different QP
  - Switchable at CU-level

- **Performance** (anchor: LDP, HM12.0)

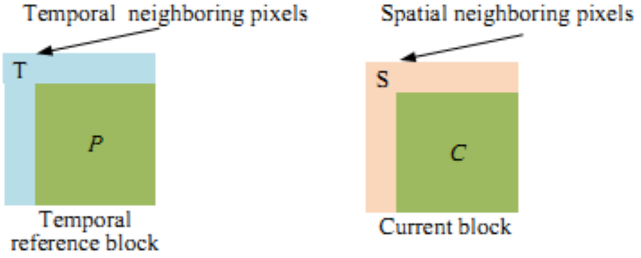
Class	Sequence	BD-Rate (Simple CNNMCR)			BD-Rate (CNNMCR)			BD-Rate (OBMC)			BD-Rate (OBMC + CNNMCR)		
		Y	U	V	Y	U	V	Y	U	V	Y	U	V
Class B	Kimono	−1.9%	0.7%	−0.3%	−2.7%	0.6%	−0.3%	−1.8%	−2.4%	−2.2%	−4.3%	−2.4%	−2.6%
	ParkScene	0.5%	0.0%	−0.2%	0.4%	0.3%	0.0%	−2.7%	−3.0%	−2.9%	−2.5%	−2.7%	−2.9%
	Cactus	−1.8%	−0.6%	−0.3%	−2.5%	−0.9%	−0.6%	−4.4%	−4.3%	−4.2%	−6.5%	−5.1%	−4.2%
	BasketballDrive	−1.9%	0.2%	0.1%	−2.6%	−0.2%	0.1%	−2.3%	−3.4%	−3.0%	−4.6%	−3.1%	−2.9%
	BQTerrace	−5.0%	−2.4%	−2.3%	−6.0%	−3.2%	−3.0%	−6.8%	−5.5%	−4.8%	−11.2%	−7.0%	−6.9%
Class C	BasketballDrill	−2.7%	1.0%	1.4%	−3.3%	0.8%	1.0%	−4.3%	−3.7%	−4.6%	−7.5%	−2.7%	−3.9%
	BQMall	−1.9%	0.3%	0.0%	−2.9%	−0.6%	0.2%	−4.5%	−5.2%	−5.5%	−6.9%	−5.0%	−5.4%
	PartyScene	−2.2%	0.1%	−0.2%	−2.7%	−0.5%	−0.3%	−3.8%	−2.8%	−3.4%	−6.3%	−3.8%	−3.8%
	RaceHorsesC	−0.5%	0.8%	0.0%	−0.9%	0.3%	−0.1%	−4.0%	−4.8%	−5.2%	−4.7%	−4.3%	−4.1%
Class D	BasketballPass	−1.4%	1.1%	1.5%	−2.3%	0.6%	0.7%	−4.4%	−3.3%	−4.1%	−5.9%	−3.2%	−3.9%
	BQSquare	−6.0%	−2.2%	−1.9%	−6.7%	−2.3%	−1.7%	−7.0%	−4.6%	−6.0%	−12.4%	−6.3%	−7.2%
	BlowingBubbles	−1.9%	0.0%	−0.1%	−2.6%	0.1%	0.8%	−3.4%	−2.7%	−2.8%	−5.7%	−1.9%	−3.2%
	RaceHorses	−0.8%	0.3%	0.3%	−1.5%	0.3%	0.2%	−3.9%	−4.0%	−3.4%	−5.1%	−3.6%	−3.6%
Class E	FourPeople	−1.1%	0.7%	0.7%	−2.1%	0.8%	1.0%	−2.2%	−2.5%	−2.1%	−4.7%	−1.7%	−1.8%
	Johnny	−1.3%	2.4%	1.6%	−2.5%	1.0%	1.8%	−3.7%	−2.4%	−2.4%	−5.5%	−1.1%	−0.5%
	KristenAndSara	−1.7%	1.9%	1.8%	−2.7%	1.7%	1.3%	−2.0%	−1.6%	−1.5%	−5.0%	−0.7%	−0.4%
Class F	BasketballDrillText	−2.4%	0.3%	0.6%	−2.6%	0.3%	1.0%	−4.0%	−3.7%	−4.5%	−6.4%	−3.3%	−3.5%
	ChinaSpeed	0.0%	0.6%	0.8%	0.0%	0.2%	0.2%	0.9%	−0.2%	−0.1%	0.8%	−0.5%	−0.1%
	SlideEditing	0.5%	0.2%	0.3%	0.3%	0.1%	0.0%	0.9%	1.0%	0.9%	2.0%	1.7%	1.7%
	SlideShow	−1.5%	0.5%	−0.9%	−1.1%	0.6%	−0.9%	0.0%	−0.2%	−1.6%	−1.0%	0.1%	−0.7%
Class Summary	Class B	−2.0%	−0.4%	−0.6%	−2.7%	−0.7%	−0.8%	−3.6%	−3.7%	−3.4%	−5.8%	−4.1%	−3.9%
	Class C	−1.8%	0.5%	0.3%	−2.5%	0.0%	0.2%	−4.2%	−4.1%	−4.7%	−6.4%	−3.9%	−4.3%
	Class D	−2.5%	−0.2%	−0.1%	−3.3%	−0.3%	0.0%	−4.7%	−3.7%	−4.1%	−7.2%	−3.7%	−4.5%
	Class E	−1.4%	1.7%	1.4%	−2.5%	1.2%	1.4%	−2.6%	−2.2%	−2.0%	−5.1%	−1.2%	−0.9%
	Class F	−0.8%	0.4%	0.2%	−0.8%	0.3%	0.1%	−0.6%	−0.8%	−1.3%	−1.2%	−0.5%	−0.7%
Average of Classes B–F		**−1.8%**	**0.3%**	**0.1%**	**−2.3%**	**0.0%**	**0.1%**	**−3.2%**	**−3.0%**	**−3.2%**	**−5.2%**	**−2.8%**	**−3.0%**

[5] Huo S, Liu D, Wu F, et al. Convolutional neural network-based motion compensation refinement for video coding[C]//2018 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2018: 1-4.

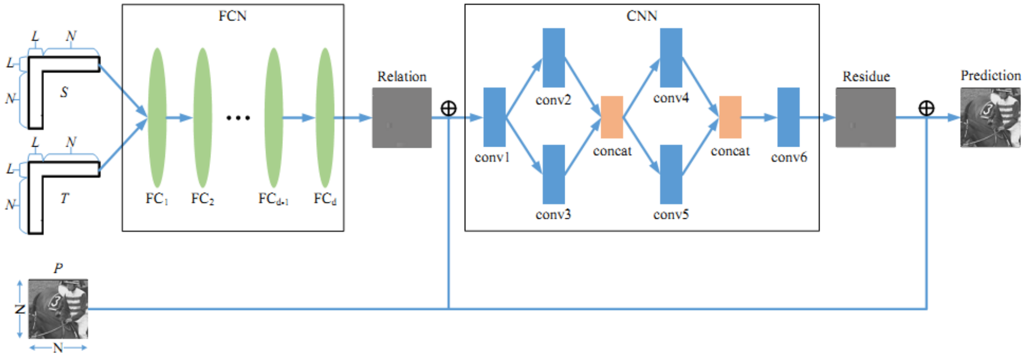
Spatial-Temporal Prediction: Inter

Block Refinement of Uni-Prediction

- **Network input**
  - Prediction CU of conventional methods
  - L-shape neighboring reconstructed pixels of both current predicted block and temporal reference block



- **Network output**
  - Refined predicted block
- **Network Structure:**
  - Fully connected network + CNN



- **Integration into coding system**
  - Different model for different QP and different blocksize
  - Switchable at CU-level
- **Performance** (anchor: LDP, HM16.9)

Table 2. The BD-rate of NNIP for luma component compare to HM 16.9

Class	Resolution	Sequence	BD-Y
Class A	2560x1600	*Traffic*	-1.5%
		PeopleOnstreet	-0.6%
Class B	1920x1080	*Kimono*	-1.9%
		ParkScene	-0.3%
		Cactus	-2.3%
		*BasketballDrive**	-3.8%
		BQTerrace	-8.6%
Class C	832x480	*BasketballDrill*	-1.3%
		*BQMall**	-2.2%
		PartyScene	-0.7%
		RaceHorses	-0.6%
Class D	416x240	*BasketballPass*	-0.9%
		BQSquare	-1.3%
		*BlowingBubbles**	-0.7%
		RaceHorses	-0.6%
Class E	1280x720	*FourPeople*	-1.5%
		Johny	-2.0%
		KristenAndSara	-2.1%
Average			-1.7%

Table 3. The computational complexity of NNIP

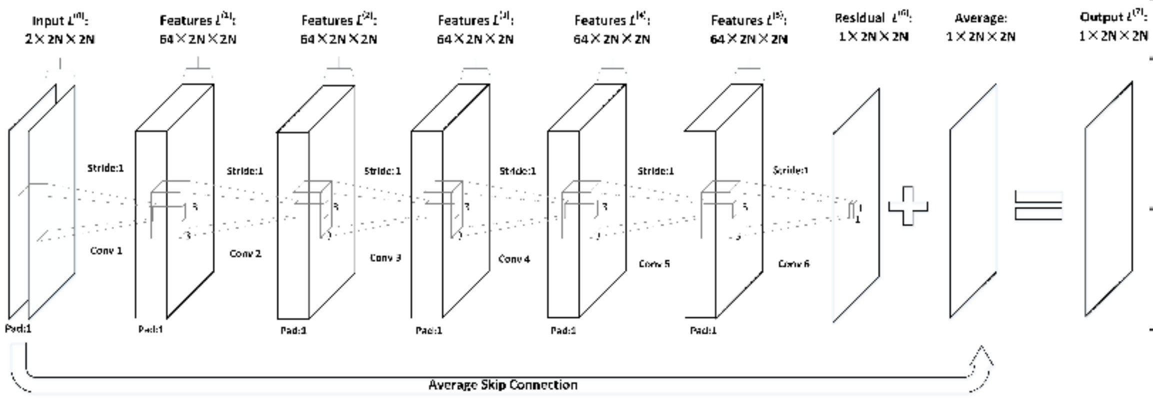
	ΔT_{enc}	ΔT_{dec}
Class A	3273%	1700%
Class B	3314%	3301%
Class C	2479%	2416%
Class D	2842%	1578%
Class E	5310%	1113%
Average	3444%	2022%

[6] Y. Wang, X. Fan, C. Jia, D. Zhao and W. Gao, "Neural Network Based Inter Prediction for HEVC," 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, 2018, pp. 1-6, doi: 10.1109/ICME.2018.8486600.

Spatial-Temporal Prediction: Inter

Bi-prediction Block Generation

- **Network input**
  - 2 reference blocks
- **Network output**
  - Bi-directional prediction block
- **Network Structure:**
  - CNN



- **Integration into coding system**
  - Different model for different QP and different block size
  - Directly replace the traditional simple average of bi-prediction reference blocks
- **Performance** (anchor: RA, HM16.15)

TABLE II
BD-RATE REDUCTIONS IN THE DIFFERENT CONFIGURATIONS.

		Random Access			Low Delay B		
	Sequences	BD-rate Y	BD-rate U	BD-rate V	BD-rate Y	BD-rate U	BD-rate V
Class A	Traffic	-2.6 %	0.4 %	0.4 %	-2.1 %	1.6 %	1.7 %
	PeopleOnStreet	-1.7 %	-1.0 %	-1.1 %	-0.6 %	0.3 %	-0.1 %
Class B	Kimono	-2.5 %	-0.1 %	-0.1 %	-1.7 %	0.4 %	1.0 %
	ParkScene	-2.7 %	-0.2 %	-0.4 %	-1.5 %	0.8 %	0.2 %
	Cactus	-3.5 %	-0.1 %	-0.6 %	-1.6 %	0.1 %	0.1 %
	BasketballDrive	-2.6 %	-0.6 %	-0.5 %	-1.3 %	1.0 %	0.8 %
	BQTerrace	-6.2 %	-0.8 %	-0.9 %	-3.3 %	1.0 %	-0.8 %
Class C	BasketballDrill	-2.1 %	0.4 %	0.4 %	-2.0 %	1.9 %	1.7 %
	BQMall	-2.7 %	-0.3 %	-0.5 %	-1.5 %	0.6 %	1.0 %
	PartyScene	-2.7 %	-0.5 %	-0.7 %	-0.5 %	0.2 %	0.6 %
	RaceHorses	-1.0 %	-0.5 %	-0.5 %	-0.1 %	0.3 %	0.5 %
Class D	BasketballPass	-1.6 %	-0.7 %	-0.3 %	-0.5 %	0.5 %	0.5 %
	BQSquare	-8.8 %	-4.1 %	-4.0 %	-1.8 %	-0.2 %	1.1 %
	BlowingBubbles	-2.6 %	-0.6 %	-0.4 %	-1.3 %	0.5 %	1.2 %
	RaceHorses	-1.4 %	-0.4 %	-0.7 %	-0.4 %	0.5 %	0.0 %
Class E	FourPeople	-	-	-	-3.7 %	0.4 %	0.4 %
	Johnny	-	-	-	-2.8 %	2.6 %	3.4 %
	KristenAndSara	-	-	-	-2.6 %	1.6 %	2.5 %
Average		-3.0 %	-0.6 %	-0.6 %	-1.6 %	0.8 %	0.9 %

TABLE I
COMPUTATIONAL COMPLEXITY ON RA CONFIGURATION OF DIFFEREN
SEQUENCES

	Sequences	ΔT_{Enc}	ΔT_{Dec}
Class A	Traffic	184.0 %	3632.7 %
	PeopleOnStreet	152.1 %	3093.1 %
Class B	Kimono	159.4 %	4866.2 %
	ParkScene	173.8 %	3678.2 %
	Cactus	162.8 %	4627.5 %
	BasketballDrive	158.4 %	4081.1 %
	BQTerrace	190.8 %	3516.7 %
Class C	BasketballDrill	153.7 %	3623.0 %
	BQMall	161.3 %	4555.7 %
	PartyScene	163.1 %	4704.9 %
	RaceHorses	140.0 %	3675.2 %
Class D	BasketballPass	180.4 %	6078.2 %
	BQSquare	206.6 %	6759.1 %
	BlowingBubbles	162.4 %	5590.5 %
	RaceHorses	122.8 %	4567.3 %
Average		164.9 %	4470.0 %

[7] Zhao Z, Wang S, Wang S, et al. Enhanced Bi-Prediction With Convolutional Neural Network for High-Efficiency Video Coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(11): 3291-3301.

Spatial-Temporal Prediction: Inter

Refinement of Bi-prediction Block

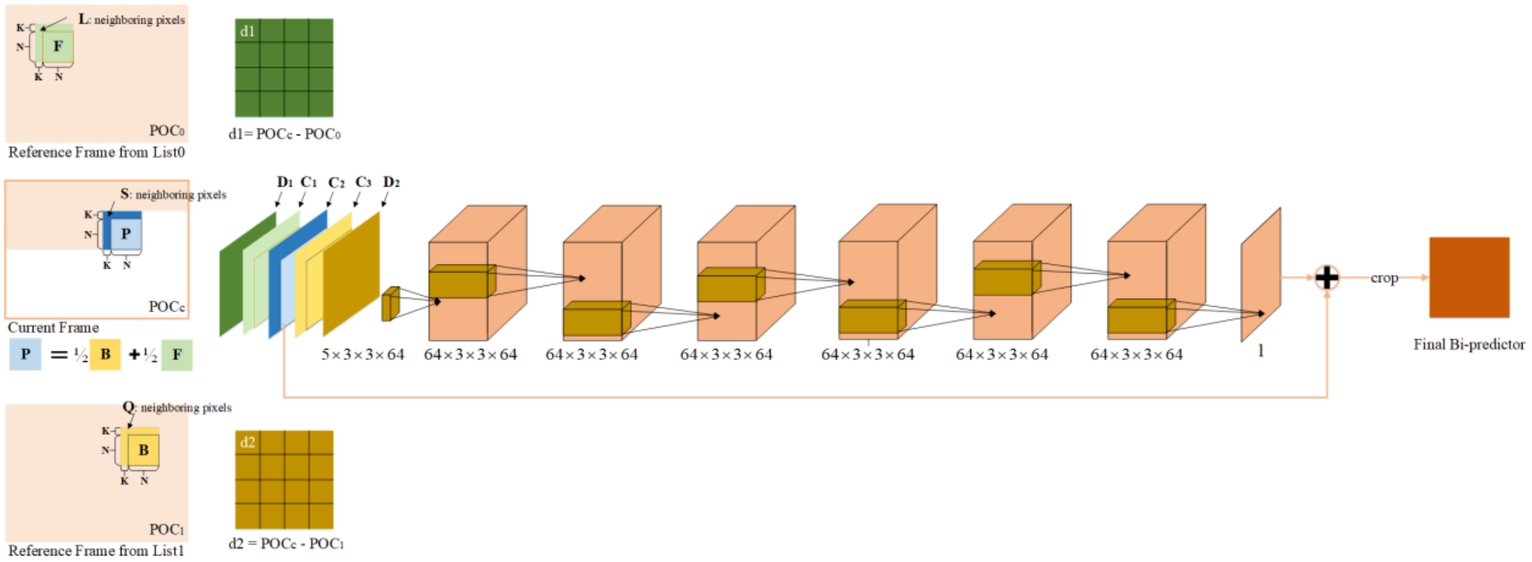
➢ Network input

- 2 reference blocks, together with L-shape neighboring pixels of the 2 reference blocks
- Predicted block by averaging of 2 reference blocks, together with L-shape neighboring pixels of current block
- Temporal distances between each reference block and current block

➢ Network output

- Current bi-predicted block

➢ Network Structure



➢ Integration into coding system

- Different model for different QP and different block size
- Replace traditional averaging bi-prediction in AMVP mode
- Switchable in Merge mode

➢ Performance

(anchor: RA, HM16.15)

PERFORMANCE COMPARISON WITH THE FIRST 2 SECONDS COMPRESSED UNDER RA CONFIGURATION(ANCHOR: HM-16.15)

Sequence	LWP [10]	Zhao [14]	LWP + Zhao	Our STCNN
Kimono	-0.19%	-2.06%	-2.18%	-3.00%
BQTerrace	-0.45%	-6.79%	-6.79%	-7.86%
Cactus	-0.91%	-4.58%	-5.15%	-5.33%
BasketballDrive	-1.59%	-3.04%	-4.48%	-3.51%
ParkScene	-0.02%	-3.01%	-2.95%	-3.68%
Class B	**-0.63%**	**-3.90%**	**-4.31%**	**-4.68%**
BasketballDrill	-0.54%	-2.54%	-2.78%	-2.72%
BQMall	-0.06%	-2.81%	-2.86%	-3.42%
PartyScene	-0.11%	-2.60%	-2.67%	-3.85%
RaceHorsesC	-0.23%	-0.96%	-1.07%	-1.51%
Class C	**-0.24%**	**-2.23%**	**-2.35%**	**-2.88%**
BasketballPass	0.14%	-2.49%	-2.55%	-3.32%
BQSquare	0.09%	-8.10%	-7.91%	-10.77%
BlowingBubbles	-0.53%	-2.10%	-2.29%	-2.72%
RaceHorses	-0.26%	-1.65%	-1.37%	-2.21%
Class D	**-0.14%**	**-3.59%**	**-3.53%**	**-4.76%**
FourPeople	-0.40%	-7.11%	-7.43%	-8.73%
Johnny	-0.41%	-5.89%	-5.89%	-7.05%
KristenAndSara	-0.08%	-7.03%	-6.99%	-8.02%
Class E	**-0.30%**	**-6.68%**	**-6.77%**	**-7.94%**
CanotSTA	-1.77%	-2.70%	-3.82%	-3.48%
MilkyWay	-3.50%	-5.14%	-7.35%	-7.43%
TPMSTA	-0.72%	-2.86%	-3.51%	-3.29%
WAmoving	-0.66%	-3.03%	-3.61%	-4.06%
NewSeq	**-1.66%**	**-3.43%**	**-4.57%**	**-4.56%**
AverageAllSeq	**-0.61%**	**-3.83%**	**-4.18%**	**-4.80%**

COMPUTATION COMPLEXITY ON RA CONFIGURATION

Class	LWP [10]		Zhao [14]		LWP [10]+Zhao [14]		proposed STCNN	
	Enc(%)	Dec(%)	Enc(%)	Dec(%)	Enc(%)	Dec(%)	Enc(%)	Dec(%)
Class B	123	102	165	750	191	750	179	1425
Class C	121	98	157	843	180	821	168	1396
Class D	116	104	160	1429	181	1462	169	2209
Class E	124	106	181	947	209	953	195	1462
Overall	**121**	**102**	**165**	**980**	**189**	**984**	**177**	**1620**

[8] Mao J, Yu L. Convolutional Neural Network Based Bi-prediction Utilizing Spatial and Temporal Information in Video Coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(7), 1856-1870.

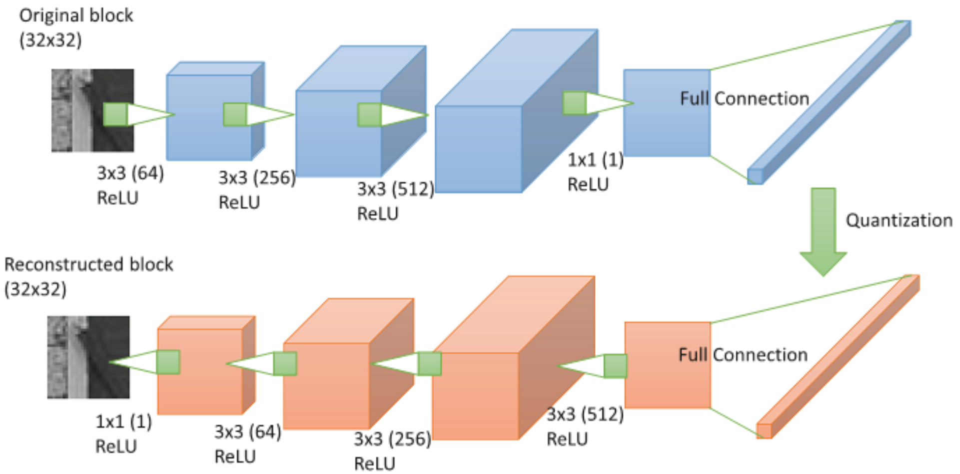
Spatial-Temporal Prediction: Inter

- **Prediction of block of pixel values**
  - Fractional pixel interpolation
    - Super-resolution: position-aware model
  - Refinement of traditional prediction or directly generation of prediction
    - Content adaptive temporal filtering to replace simple average
    - Generalize of bi-hypothesis uni-directional and bi-directional by introduce temporal distances: temporal interpolation and extrapolation
    - With/without motion vector
  - As supplementing inter modes or replacing to traditional ones
- **Prediction of motion/optical flow**

Transform

➢ Network structure:

- CNN Layers: feature analysis
- Fully Connection Layer: fulfill the transform



➢ Training method:

- Initialization: FC Layer is initialized by transform matrix of DCT/IDCT
- Joint training of FC and CNN
- Loss: joint rate-distortion cost
  - Rate estimated by the l1-norm of the quantized coefficients
  - Distortion estimated by MSE

▪ **How good will it be for prediction residuals?**

➢ Performance

Table 1. BD-rate results of our symmetric network compared with different anchors

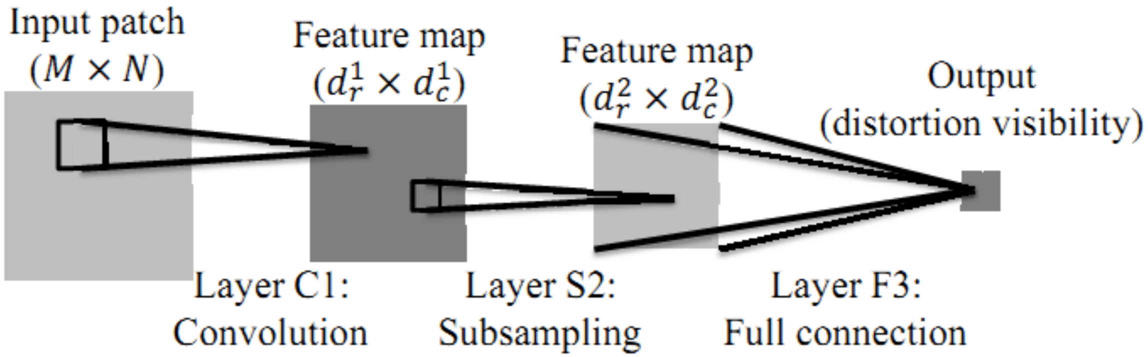
	Ours vs. DCT (32 × 32)	Ours vs. JPEG	Ours vs. Toderici *et al.*
kodim01	−17.74%	−28.45%	−36.71%
kodim02	−3.15%	−56.38%	−79.21%
kodim03	−10.28%	−47.22%	−67.99%
kodim04	−2.86%	−50.96%	−65.11%
kodim05	−18.31%	−24.65%	−28.60%
kodim06	−13.29%	−35.05%	−59.77%
kodim07	−11.10%	−39.13%	−54.94%
kodim08	−11.63%	−24.42%	−36.11%
kodim09	−8.09%	−41.15%	−61.45%
kodim10	−5.46%	−42.18%	−61.39%
kodim11	−10.91%	−33.09%	−55.52%
kodim12	−8.84%	−43.60%	−69.24%
kodim13	−13.35%	−23.27%	−41.25%
kodim14	−15.91%	−30.20%	−46.09%
kodim15	*7.17%*	−37.82%	−60.90%
kodim16	−7.95%	−46.79%	−67.92%
kodim17	*11.88%*	−35.22%	−44.26%
kodim18	−15.79%	−34.46%	−48.31%
kodim19	−9.84%	−47.52%	−65.65%
kodim20	−3.26%	−35.89%	−62.65%
kodim21	−17.32%	−34.89%	−60.27%
kodim22	−12.52%	−39.39%	−57.85%
kodim23	−3.48%	−54.09%	−77.02%
kodim24	−10.00%	−26.80%	−51.71%
Average	−8.83%	−38.03%	−56.66%

[1] Liu D, Ma H, Xiong Z, et al. CNN-based DCT-like transform for image compression[C]//International Conference on Multimedia Modeling. Springer, Cham, 2018: 61-72.

Quantisation

Content-adaptive QP selection

- **Local visibility threshold prediction-** VNet-2
  - Convolution layer: 362 trainable parameters (19*19 kernel + 1bias)
  - Subsampling layer: scale=2, 2 trainable parameters(1 weight + 1 bias)
  - Full connection layer: 530 trainable parameters(23*23 weight + 1 bias)

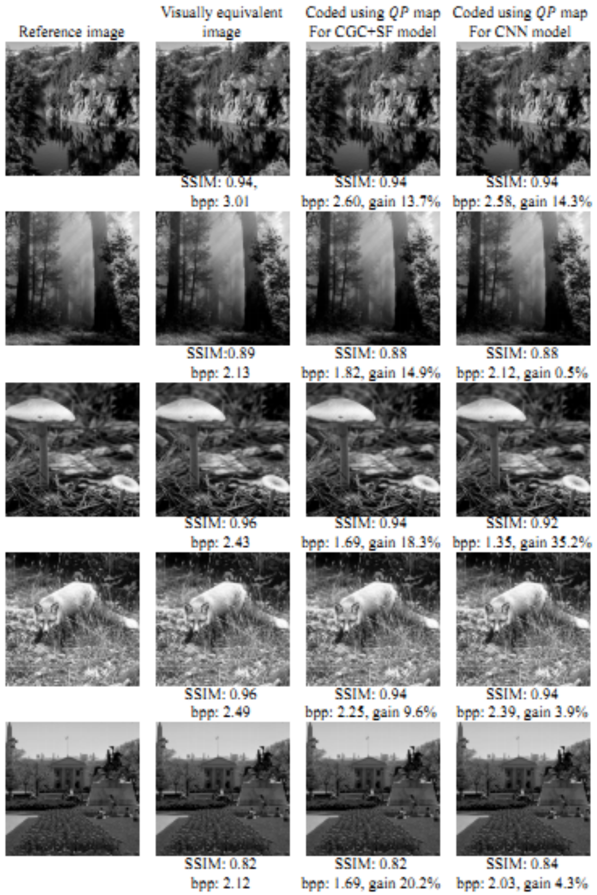


- **Quantization steps derivation for CTU**

$$\log(Q_{step}) = \alpha C^2 + \beta C + \gamma$$

  - C : predicted local visibility threshold
  - $\{\alpha,\beta,\gamma\}$: model coefficients depend on patch features, predicted from 3 separate NNs.

- **Performance**
  - 11% bitrate saving for luma channel against HEVC at same SSIM.

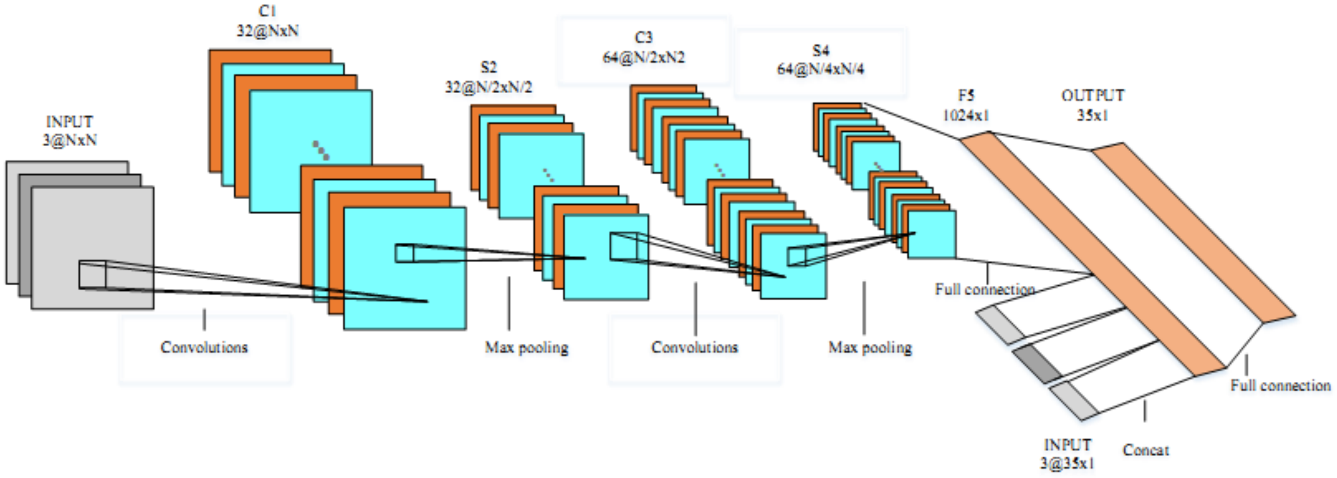


[1] Alam M M, Nguyen T D, Hagan M T, et al. A perceptual quantization strategy for HEVC based on a convolutional neural network trained on natural images[C]//Applications of Digital Image Processing XXXVIII. International Society for Optics and Photonics, 2015, 9599: 959918.

Entropy coding

Probability Estimation of Intra Prediction Mode

- **Network inputs**
  - **Reconstructed pixels**: above-left, above and left blocks with the same size of current coding block
  - **Prediction modes of 3 neighboring blocks**: one 35-D one-hot binary vector for each neighboring block
- **Network output**
  - 35-D probability vector of 35 intra prediction modes
- **Network structure**
  - Based on LeNet-5



- **Integration into coding system**

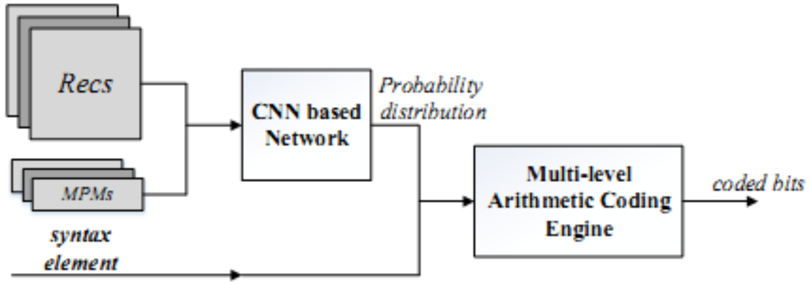


Fig. 3. The scheme of CNN-based arithmetic coding.

- **Performance** (anchor: AI, HM12.0)

TABLE I
BITS SAVINGS FOR INTRA PREDICTION MODES IN HM-INTRA-8

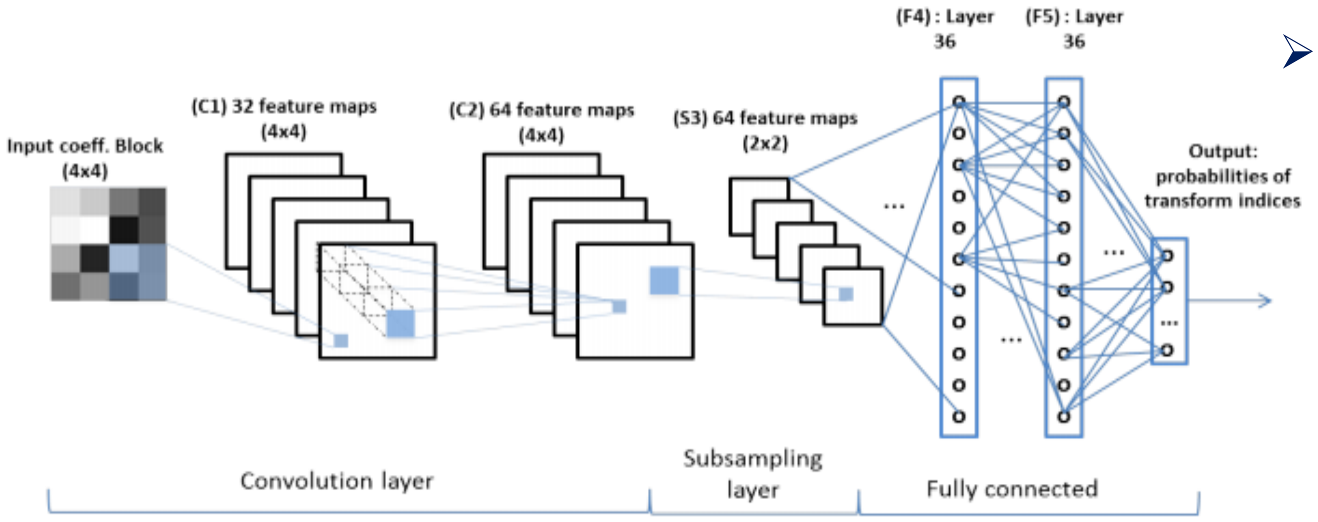
QP	22	27	32	37
ClassA	−9.9%	−9.8%	−9.6%	−8.0%
ClassB	−8.9%	−9.1%	−8.7%	−6.3%
ClassC	−10.0%	−10.2%	−9.7%	−7.1%
ClassD	−7.0%	−8.0%	−8.7%	−6.6%
ClassE	−9.7%	−11.5%	−13.0%	−12.0%
ClassF	−8.8%	−9.9%	−9.7%	−9.3%
Average	−9.0%	−9.8%	−9.9%	−8.2%

[1] Song R, Liu D, Li H, et al. Neural network-based arithmetic coding of intra prediction modes in HEVC[C]//2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017: 1-4.

Entropy coding

Probability Estimation of Transform Kernel Index

- **Network input**
  - Transform coefficients block
- **Network output**
  - Probability vector of transform kernel indexes
- **Network structure**
  - Convolution layer
  - Subsampling layer: scale=2
  - Fully connected layer



- **Integration into coding system**
  - Utilize the probability to reorder transform kernel indexes
  - Binarize the index with truncated unary code

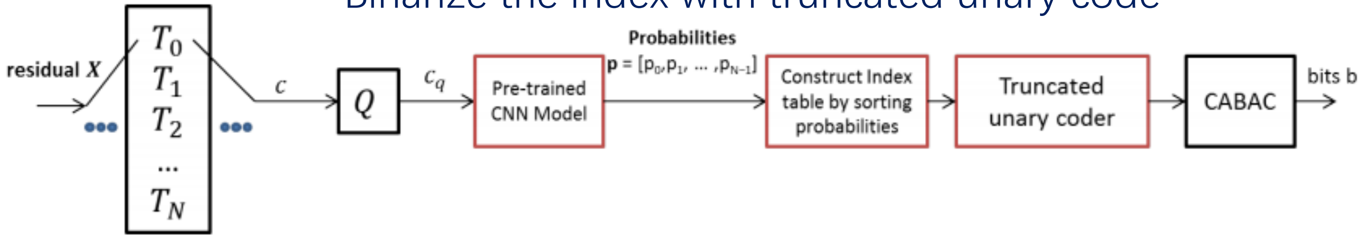


Fig. 1: Block Diagram of proposed CNN-based transform index coding

- **Performance** (anchor: AI, HM15.0)

Class	Sequence	EP	CTXT	CNN	NoIndex
A	Nebuta	-0.37	-0.38	-0.40	-0.37
	PeopleOnStreet	-0.75	-0.69	-0.90	-1.75
	SteamLocomotive	0.03	-0.03	-0.03	-0.19
	Traffic	-0.85	-0.83	-1.10	-1.82
	Overall	-0.49	-0.48	-0.61	-1.03
B	BasketballDrive	-0.22	-0.42	-0.48	-0.47
	BQTerrace	-1.44	-1.56	-1.70	-3.11
	Cactus	-1.02	-1.07	-1.25	-2.48
	Kimono	0.18	-0.27	-0.05	-0.08
	ParkScene	-0.45	-0.59	-0.74	-2.81
	Overall	-0.59	-0.67	-0.84	-1.79
C	BasketballDrill	-3.14	-3.36	-3.34	-3.29
	BQMall	-1.74	-1.81	-1.91	-3.33
	PartyScene	-2.03	-2.14	-2.15	-3.89
	RaceHorses	-1.59	-1.57	-1.82	-3.38
	Overall	-2.12	-2.22	-2.31	-3.47

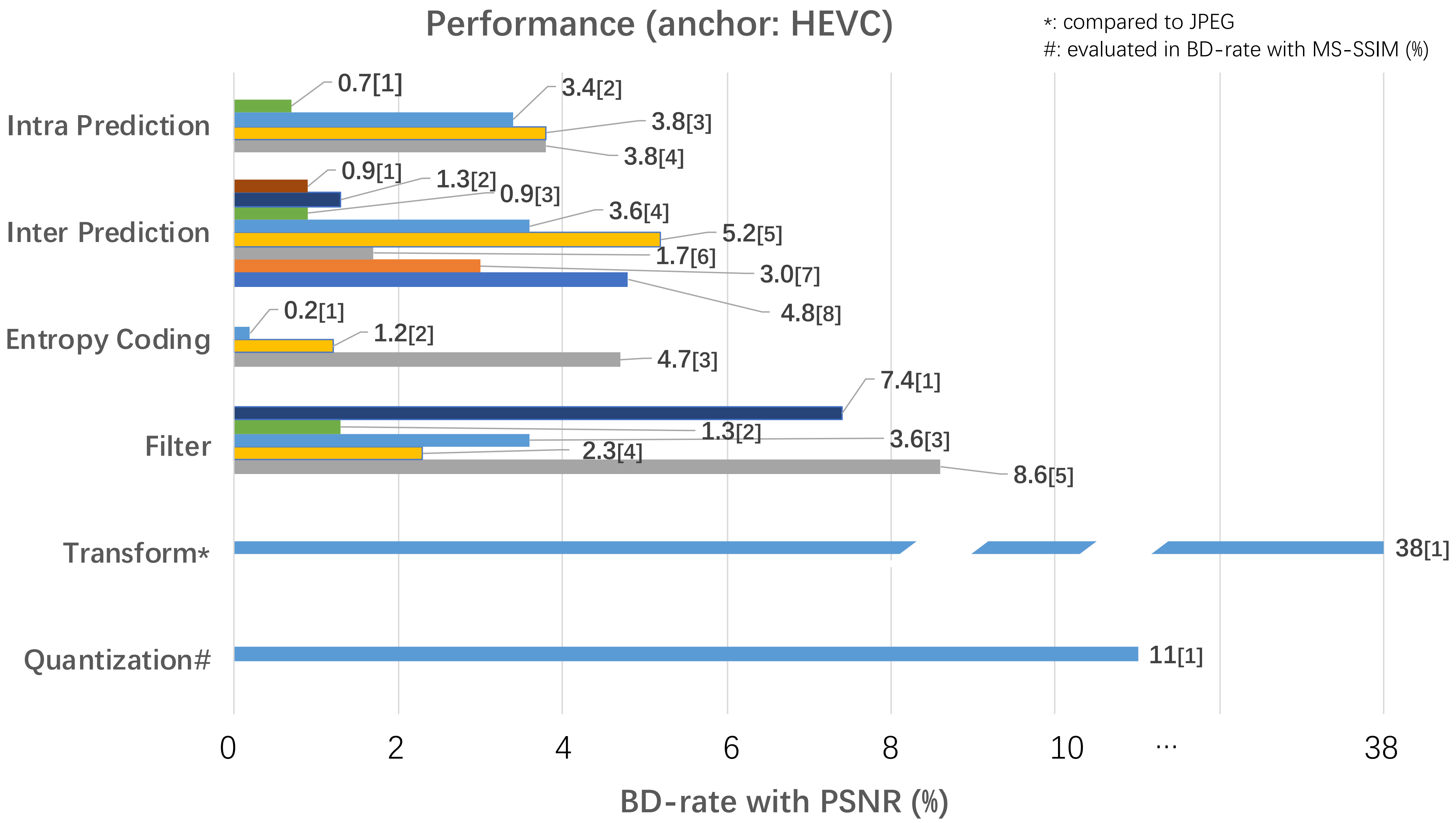
[2] Puri S, Lasserre S, Le Callet P. CNN-based transform index prediction in multiple transforms framework to assist entropy coding[C]//2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017: 798-802.

Entropy coding

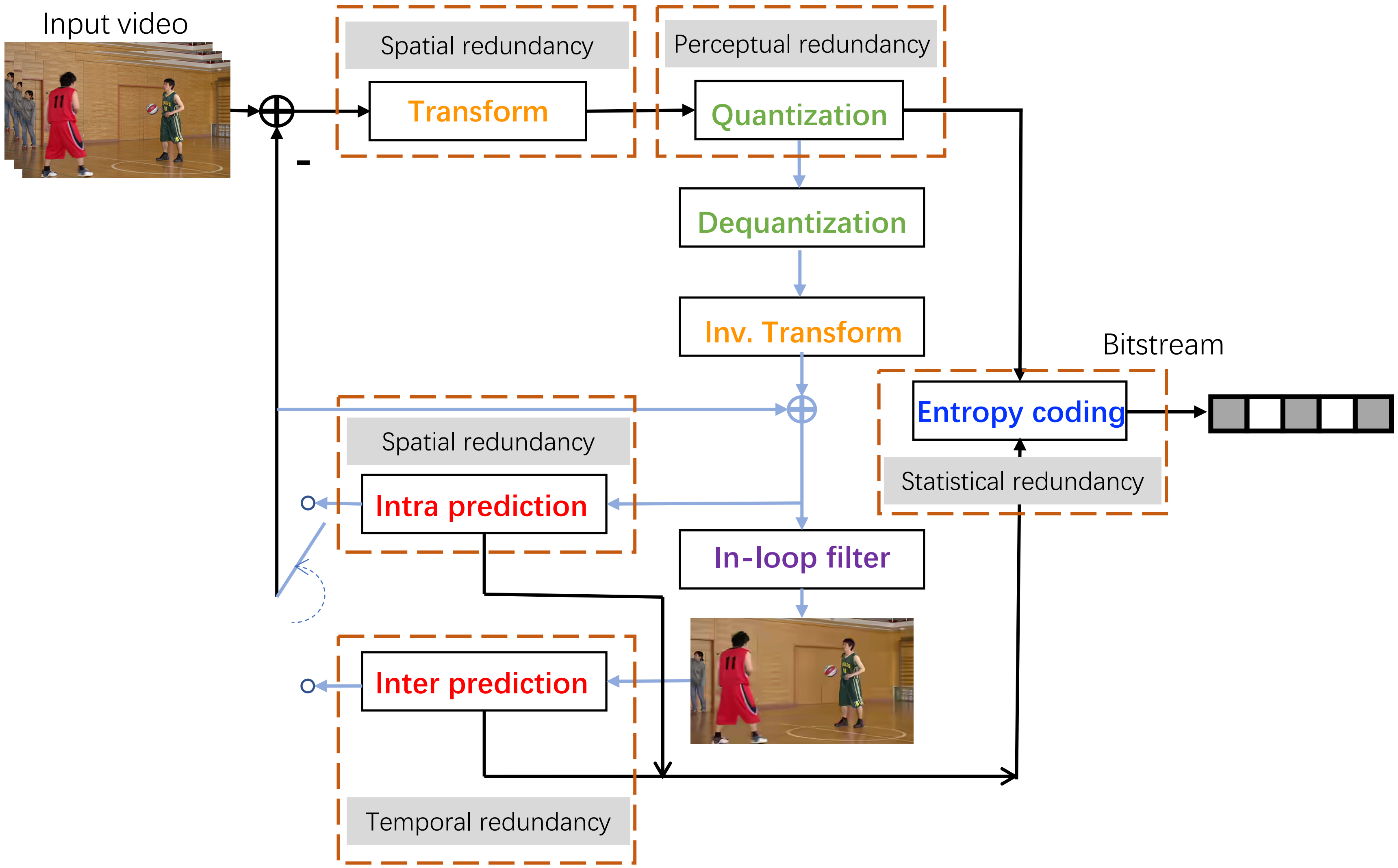
- **Probability estimation** – **Possibility estimation**
  - For different syntaxes
    - Mode indexes, coefficients values, …
  - Using correlated information
    - Reconstructed pixels, intermediate reconstructed pixels
    - Decoded neighboring modes
  - Labels
    - Happened or not – POSSIBILITY instead of probability
      - *Possibility* describes the likelihood of a value happening in one symbol while *probability* describe the frequency of a value happening in an infinite string of symbols
      - *Possibility* is a more suitable descriptor for non-stationary process

- Z. He, L. Yu, Possibility distribution based lossless coding and its optimization, *Signal Processing,* Vol. 150, pp 122-134, Sep. 2018

Performance



Hybrid or End-to-End ?

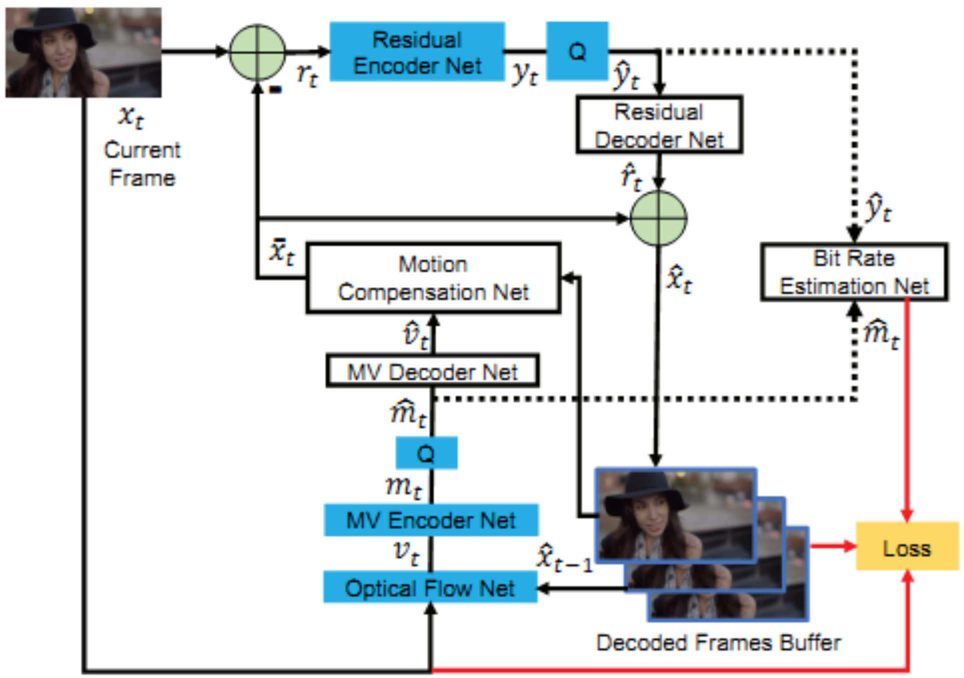


End-to-End Video Coding

➢ **Overall Network Structure**

limited temporal information utilization

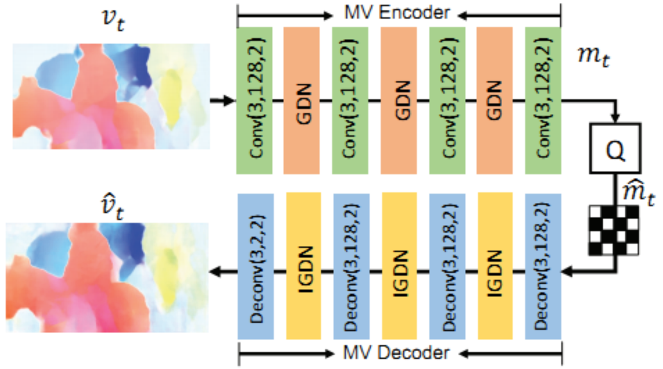
End-to-end Video Compression Framework



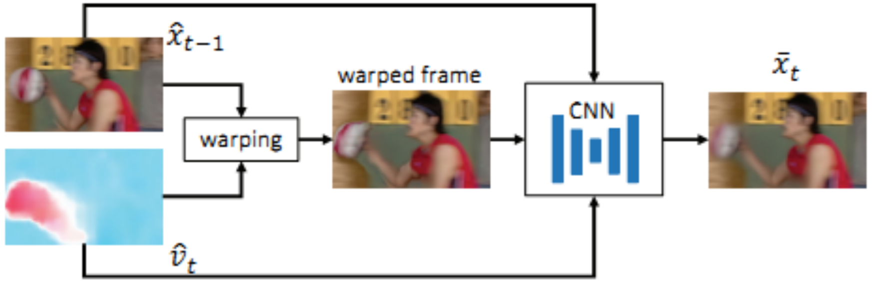
➢ **Loss function**

$$L = \lambda D + R = \lambda d(x_t, \bar{x}_t) + R(\hat{m}_t) + R(\hat{y}_t)$$

- **Optical Flow Net**
  - ✓ An optical flow estimation network
- **MV Encoder & MV Decoder**



- **Motion Compensation Network**



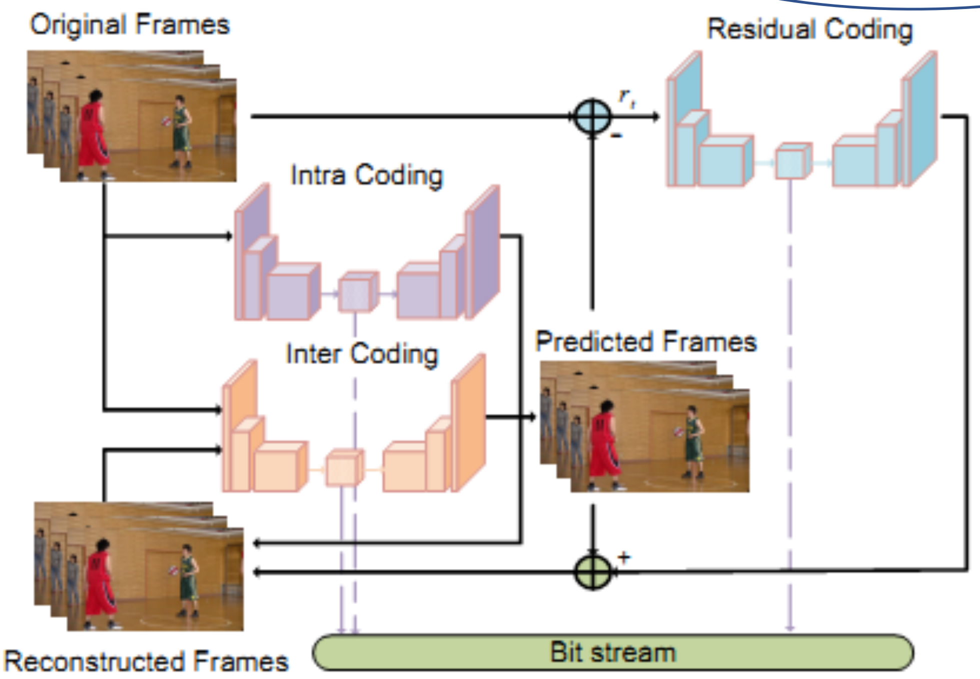
- **Residual Encoder Net & Decoder Net**
  - ✓ An end-to-end image compression network
- **Bit Rate Estimation Net**
  - ✓ Bit rate estimation part of an end-to-end image compression network

[1] Lu G, Ouyang W, Xu D, et al. Dvc: An end-to-end deep video compression framework[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 11006-11015.

End-to-End Video Coding

➢ **Overall Network Structure**

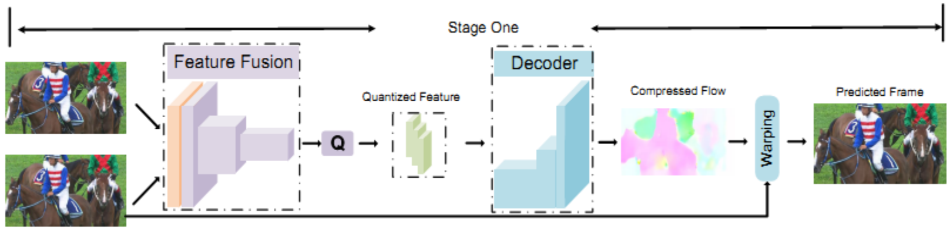
temporal information utilization by using LSTM



➢ **Loss function**

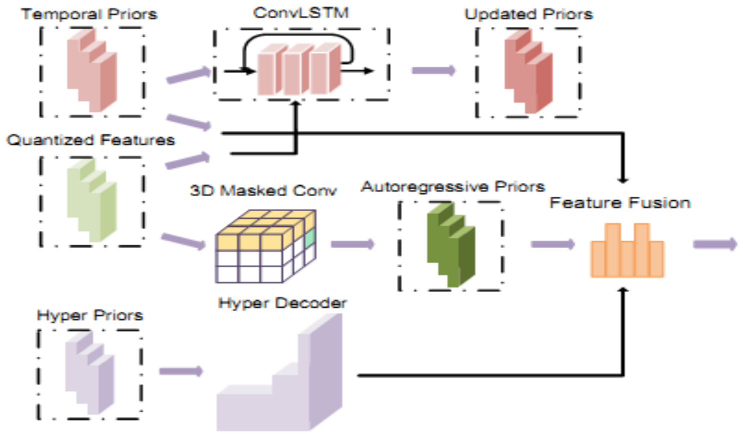
$$L = \lambda D + R = \lambda d(x_t, \bar{x}_t) + R(\hat{m}_t) + R(\hat{y}_t)$$

- **Intra Coding & Residual Coding**
  - ✓ An end-to-end image compression network
- **Inter Coding**
  - ✓ **One-stage Unsupervised Flow Learning:**



Optical flow estimation and compression realized in one stage

  - ✓ **Context Adaptive Flow Compression**

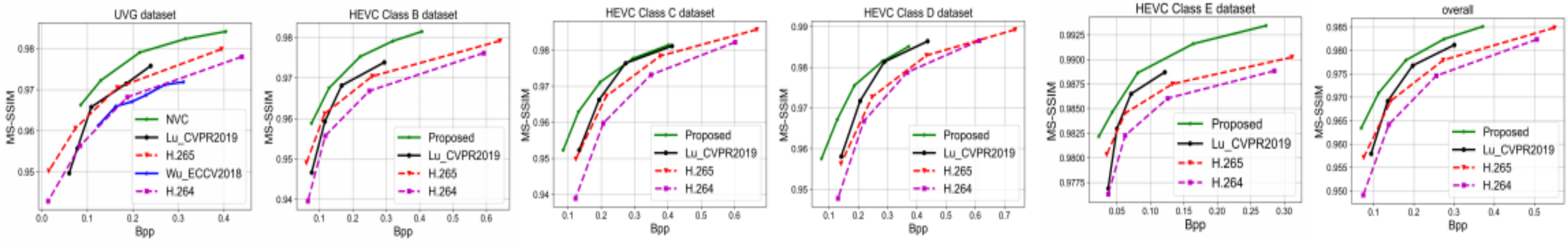


For entropy model, besides using spatial features and hyperpriors, temporal priors generated by ConvLSTM are used.

[2] Liu H, Huang L, Lu M, et al. Learned Video Compression via Joint Spatial-Temporal Correlation Exploration[C]//AAAI. 2020.

End-to-End Video Coding

➢ Performance



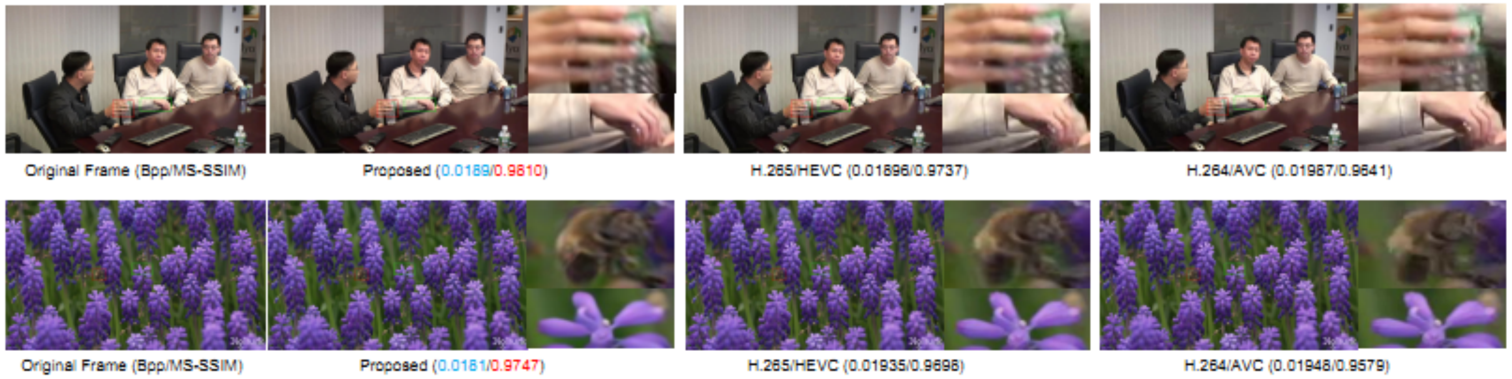


Figure 8: **Visual Comparison.** Reconstructed frames of our method, H.265/HEVC and H.264/AVC. We avoid blocky artifacts and provide better quality of reconstructed frame at low bit rate.

[1] Liu H, Huang L, Lu M, et al. Learned Video Compression via Joint Spatial-Temporal Correlation Exploration[C]//AAAI. 2020.

Conclusion

- **All roads lead to Rome**
  - NN modules embedded into hybrid video coding frameworks can bring significant coding gains
  - End-to-end image and video coding – still follow the source coding theory
  - **Training:** separately or jointly

- Performance of learning based coding comes from
  - Re-organization of information: non-linear transform to independent symbol
  - Quantization: scalar vs. vector quantization
  - Entropy coding: hyperprior to estimate of possibility + arithmetic coding

Latest Publications on Learning-based Coding

- **SPECIAL SECTION ON LEARNING-BASED IMAGE AND VIDEO CODING, IEEE TCSVT 2020. Jul**

  12 papers:

  - End-to-end image compression (1)
  - Intra prediction (3)
  - Inter prediction (2)
  - Filtering (2)
  - Arithmetic coding (1)
  - Encoder optimization (3)

IEEE TRANSACTIONS ON
CIRCUITS AND SYSTEMS FOR
VIDEO TECHNOLOGY

A PUBLICATION OF THE IEEE CIRCUITS AND SYSTEMS SOCIETY

CAS

WWW.IEEE-CAS.ORG

JULY 2020 VOLUME 30 NUMBER 7 ITCTEM (ISSN 1051-8215)

SPECIAL SECTION ON LEARING-BASED IMAGE AND VIDEO COMPRESSION

GUEST EDITORIAL

(Contents Continued on Back Cover)

IEEE

Deep Neural Network Based Video Coding

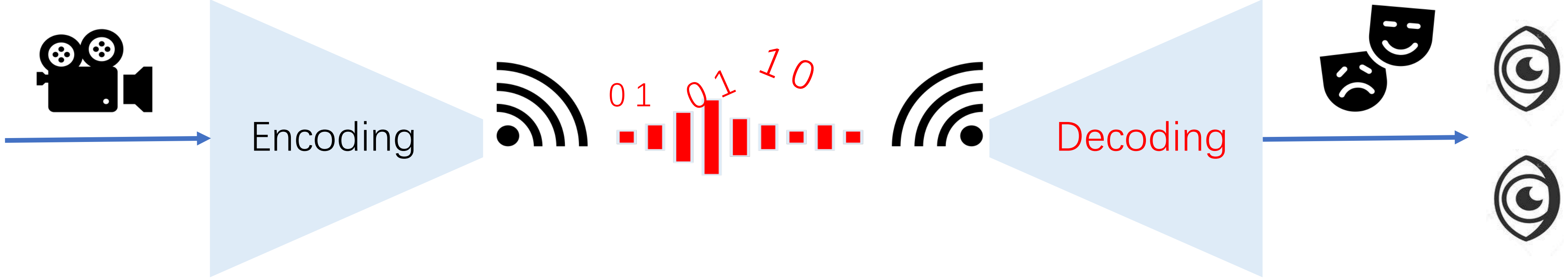
- **AhG on DNNVC established in 130th MPEG meeting in Apr. 2020**
- **Mandates**
  - Evaluate and quantify **performance improvement potential of DNN based video coding technologies** (including hybrid video coding system with DNN modules and end-to-end DNN coding systems) compared to existing MPEG standards such as HEVC and VVC, considering various quality metrics;
  - Study **quality metrics for DNN based video coding**;
  - Solicit input contributions on DNN based video coding technologies;
  - Analyze the **encoding and decoding complexity** of NN based video coding technologies by considering software and hardware implementations, including impact on power consumption;
  - Investigate technical aspects specific to NN-based video coding, such as design network representation, operation, tensor, on-the-fly network adaption (e.g. updating during encoding) etc

Subscribe mailing list:

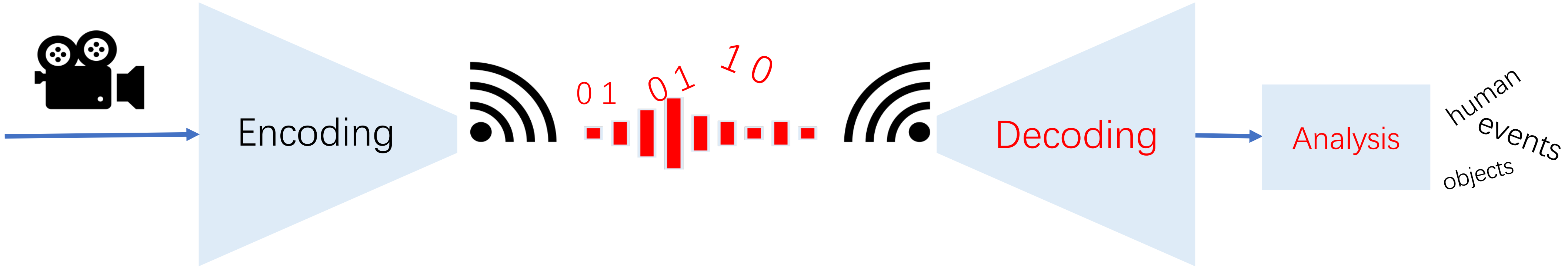
https://lists.aau.at/mailman/listinfo/mpeg-dnnvc

Image/Video Coding for …

- Reconstruction image/video for **human vision** -- yes, but not the only target



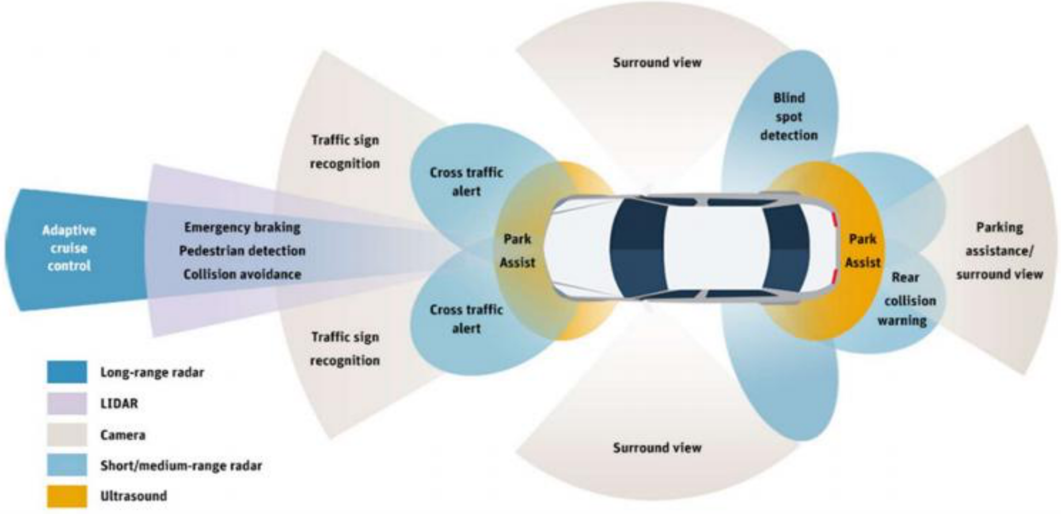
- Coding image/video for **machine understanding**

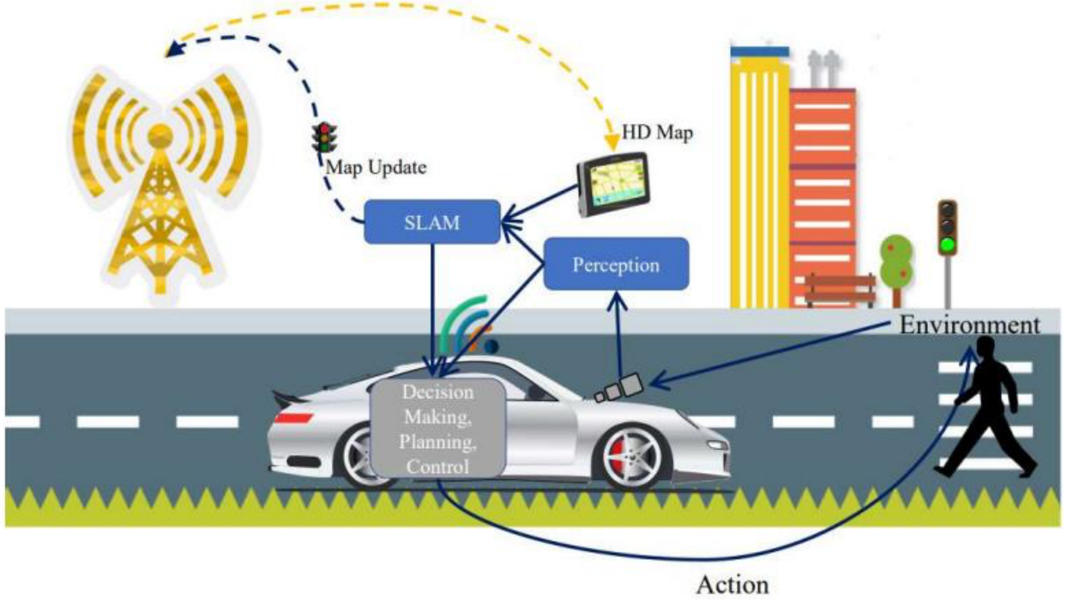


Video Coding for Machine: Use Cases

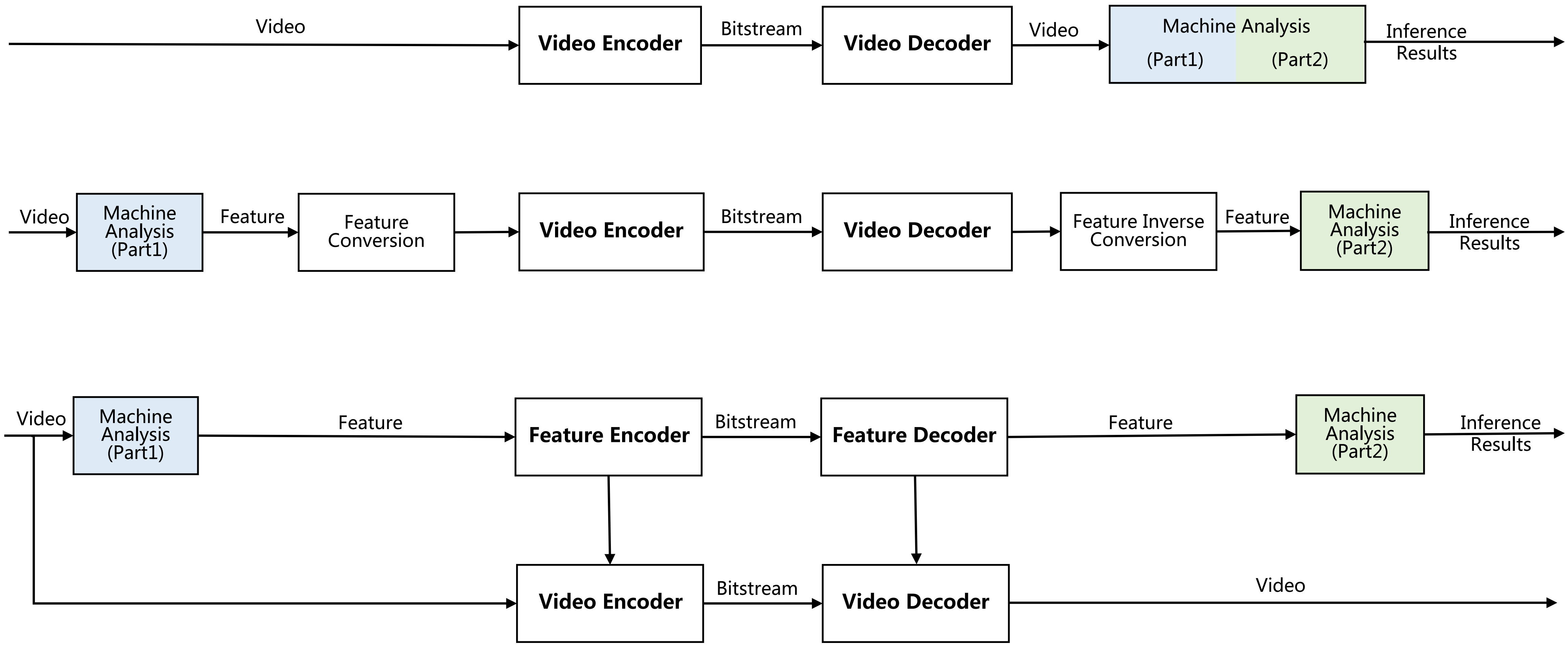
- **6 major application areas**
  - Smart Industry
  - Intelligent Transportation
  - Smart Retailer
  - Smart City
  - Smart Sensors Networks
  - Immersive Video / HD Entertainment
  - Smart Media Editing and Creation

- **Use Cases:**
  - **machine-oriented analysis**
  - **hybrid machine/human representation**

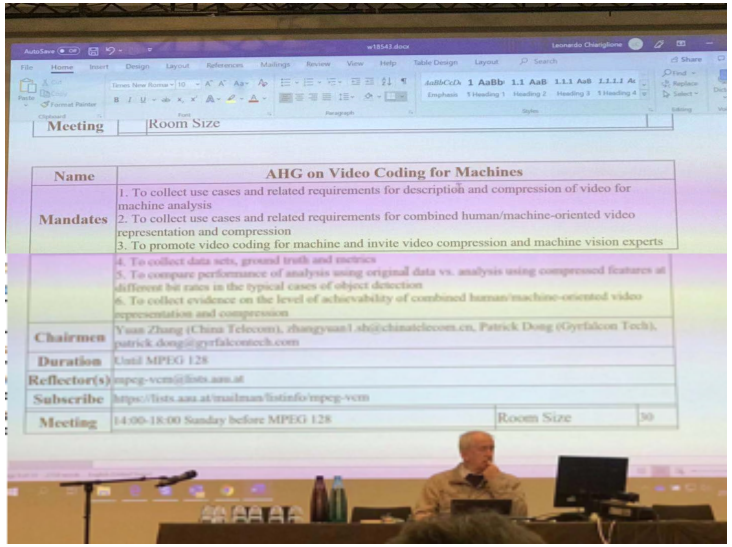




Video Coding for Machine: Potential Pipelines



Video Coding for Machine



VCM mailing list

- **AhG on VCM established in 127th MPEG meeting in July, 2019**
- **Mandates**
  - To create and evaluate anchors for object detection, object segmentation and object tracking
  - To collect data sets, ground truth
  - To define metrics for object detection, object segmentation and object tracking
  - To compare **performance of analysis using original data vs. analysis using compressed features at different bit rates** in the typical cases of object detection
  - To collect evidence on the **level of achievability of combined human/machine-oriented video representation and compression**
  - To encourage experts to provide **feature stream codecs**
  - To encourage experts to provide uncompressed bitstream from feature extractor
- **Preliminary Timeline**
  - 2019.07 Establish VCM, set up mailing list, release use cases
  - 2020.01 Release requirements, provide evidences on Mandate 5 and 6
  - *2020.07 Call for evidence*

Thanks!

Contact me: yul@zju.edu.cn