Overfitting the Data: Compact Neural Video Delivery via Content-aware Feature Modulation

Jiaming Liu[1*], Ming Lu[2*†], Kaixin Chen[1*], Xiaoqi Li[1*], Shizun Wang[1*], Zhaoqing Wang[1], Enhua Wu[3], Yurong Chen[2], Chuang Zhang[1‡], and Ming Wu[1]

[1]Beijing University of Posts and Telecommunications
[2]Intel Labs China
[3]State Key Lab of Computer Science, IOS, CAS & FST, University of Macau
{liujiaming,zhangchuang}@bupt.edu.cn,lu199192@gmail.com

In Sec. 1, we demonstrate the details of our VSD4K dataset. As reported in Sec. 2, we also evaluate content-aware learning and external learning on public datasets like Vid4 [3] and REDS [4]. In Sec. 3, we apply our method to lightweight architecture (ESPCN [5]) and compare with H.264/H.265 standard. We evaluate our method on public vimeo dataset [7] in Sec. 4. All the results use PSNR as the evaluation metric.

1. Details of VSD4K

As shown in Tab. 1, we download the original 4K videos from YouTube as our source videos. Due to computational limitation, we resize the source videos to 1080p as our ground-truth. According to FFmpeg [1], we resize the 4k video by bicubic interpolation and alternate bit-rate based on [2].

2. Content-aware learning on public datasets

We present the benefit of utilizing DNN's overfitting property for video delivery on public dataset. As shown in Tab. 2, we compare content-aware learning and external learning on public datasets like Vid4 [3] and REDS [4]. As can be seen, EDSR with content-aware learning significantly outperforms EDVR with external learning. These results prove that content-aware learning is more suitable for video delivery compared with external learning.

*Equal Contribution.

†This work was done when Jiaming Liu was an intern at Intel Labs China supervised by Ming Lu

‡Chuang Zhang is responsible for correspondence.

3. H.264/H.265 against Ours (ESPCN)

In this section, we adopt ESPCN [5] to compare our method with H.264/H.265 standard under same storage cost. The quantitative results are shown in Tab. 3. Our results still outperform H.264 and H.265 in most cases. We also show the qualitative comparison in Fig. 1.

4. Evaluation on Vimeo90k[7]

In this section, we conduct experiments on public Vimeo90k[7] to present the universality of our method. We randomly selection two videos from `http://data.csail.mit.edu/tofu/dataset/original_video_list.txt`. As shown in Tab. 4, our method outperforms S_{1-n} to some extent. We also compare our method with standard H.264 and H.265. For a particular LR video, we set the sum of (LR video and SR model) as constant value. Then, we decrease the bit-rate of H.264 and H.265 video to reach the same storage as the former. Under some storage cost, our method shows promising results.

Acknowledgements. This work was supported by the NSFC Grant No. 62072449.

References

[1] FFmpeg. Ffmpeg git repo. https://git.ffmpeg.org/ffmpeg.git.

[2] YouTube help. Choose live encoder settings bitrate. http://support.google.com/youtube/answer.

[3] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013.

[4] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee.

Category	Source	Highest Resolution	Training Resolution	Bit-rate (Mbit/s)	FPS	Video Length
Game	LoL Game: `https://www.youtube.com/watch?v=BQG92HATfvE`	3840×2160	1920×1080	10.04	30	15s-5min
Vlog	Make-up tutorial: `https://www.youtube.com/watch?v=MYGZ2_X5L3E`	3840×2160	1920×1080	10.10	30	15s-5min
Inter	Blackpink interview: `https://www.youtube.com/watch?v=6FBCVpU3XG4`	3840×2160	1920×1080	10.15	30	15s-5min
Sport	Extreme sports: `https://www.youtube.com/watch?v=M0jmSsQ5ptw`	3840×2160	1920×1080	10.04	30	15s-5min
Dance	izone performance: `https://www.youtube.com/watch?v=hBlLaEt1VjI`	3840×2160	1920×1080	10.03	30	15s-5min
City	London city drive: `https://www.youtube.com/watch?v=QI4_dGvZ5yE`	3840×2160	1920×1080	9.91	30	15s-5min

Table 1. Details of VSD4K datasets.

Method	Model	Dataset	Vid4 Calender	City	Foliage	Walk	Average
External learning	EDVR-M[6]	REDS	21.82	25.91	24.67	28.83	25.31
	EDVR-L[6]	REDS	21.89	25.68	24.77	29.17	25.38
	EDVR-L[6]	Vimeo-90K	22.18	26.30	25.00	29.55	25.76
Content-aware learning	EDSR-M	Vid4	25.23	30.56	26.48	31.00	28.32
Content-aware learning	EDSR-L	Vid4	27.19	32.19	27.66	32.47	29.88

Method	Model	Dataset	REDS 000	011	015	020	Average
External learning	EDVR-M[6]	REDS	27.72	31.26	33.42	29.57	30.49
	EDVR-L[6]	REDS	28.01	32.17	34.06	30.09	31.09
	EDVR-L[6]	Vimeo-90K	27.80	31.03	33.45	29.50	30.45
Content-aware learning	EDSR-M	REDS	27.27	31.31	34.02	29.07	30.42
Content-aware learning	EDSR-L	REDS	27.63	32.38	34.94	29.86	31.20

Table 2. Comparisons of content-aware learning versus external learning. EDVR-M, EDVR-L, EDSR-M, EDSR-L has 10, 40, 16, 32 resblocks respectively. Red indicates the best results.

Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[6] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[7] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

	game-45s			dance-45s			inter-45s			vlog-45s		
Method	x2	x3	x4	x2	x3	x4	x2	x3	x4	x2	x3	x4
H.264	37.72	33.43	30.63	29.12	24.51	21.86	36.54	33.26	31.02	42.44	39.79	37.65
H.265	38.32	34.56	32.28	30.90	27.09	24.86	36.94	33.92	31.85	43.39	41.04	39.13
Ours(ESPCN)	36.09	31.06	29.05	43.56	36.89	35.30	38.88	32.22	28.75	46.19	41.72	39.52
Storage(MB)	14.46	6.48	3.90	14.08	6.39	3.80	13.97	6.38	3.79	14.00	6.37	3.78

Table 3. Quantitative comparisons with H.264/H.265. We use a lightweight model (ESPCN) in these comparisons. Red and blue indicate the best and the second best results.

	90027457			72549854		
Method	x2	x3	x4	x2	x3	x4
M0	49.86	45.60	43.55	40.22	35.21	32.50
S_{1-n}	50.01	46.00	44.07	40.32	35.47	32.84
Ours	50.14	46.33	44.15	40.41	35.40	32.73
Margin	+0.13	+0.33	+0.08	+0.09	-0.07	-0.11
H.264	41.84	40.33	39.18	33.10	32.05	31.06
H.265	42.02	40.81	39.29	33.22	32.55	31.95
Size(MB)	24.10	14.41	10.97	13.45	10.02	8.38

Table 4. PSNR results on public Vimeo-90K dataset. Red and blue indicate the best and the second best results among our method, H.264, and H.265.

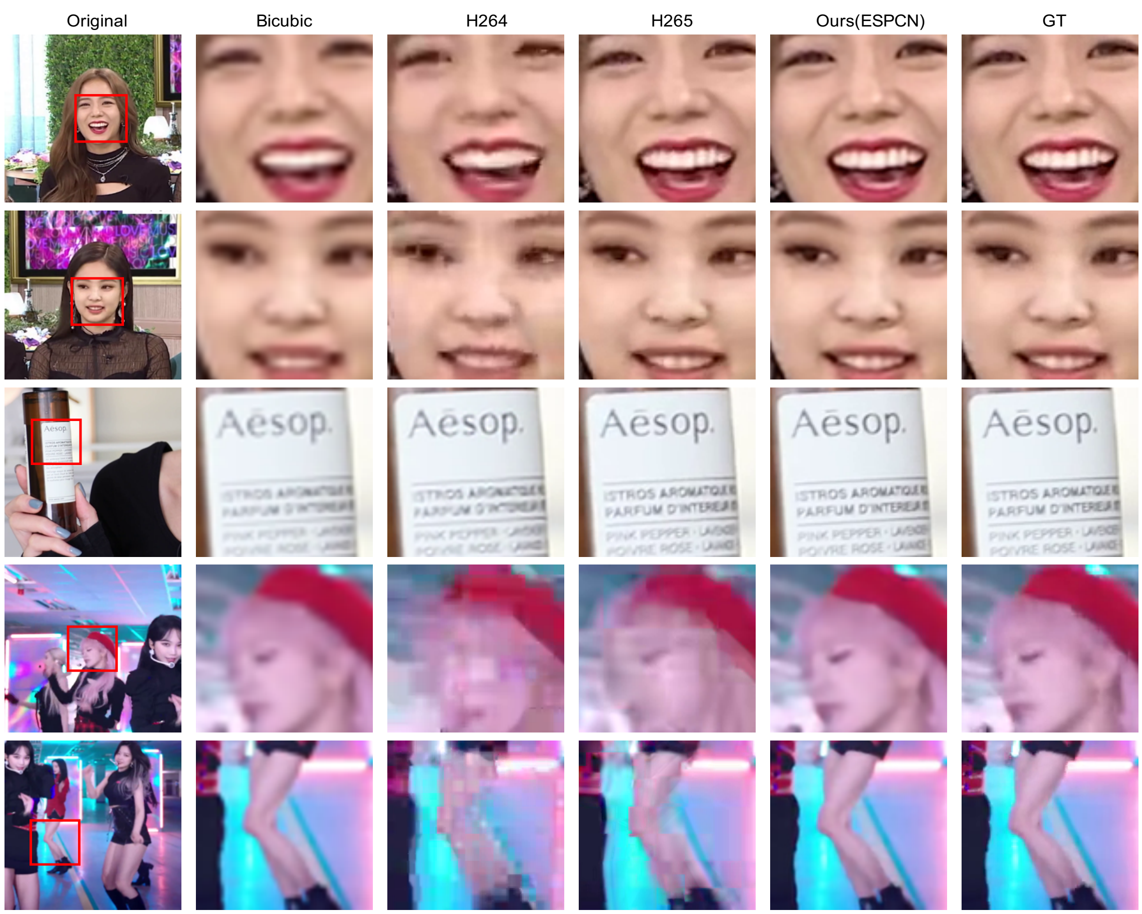

Figure 1. Qualitative comparisons with H.264/H.265. We use a lightweight model (ESPCN) in these comparisons. Best viewed by zooming x4.