Efficient Meta-Tuning for Content-aware Neural Video Delivery

Xiaoqi Li[1*], Jiaming Liu[1,2*], Shizun Wang[3*], Cheng Lyu[3], Ming Lu[4†], Yurong Chen[4], Anbang Yao[4], Yandong Guo[2], and Shanghang Zhang[1†]

[1] Peking University
[2] OPPO Research Institute
[3] Beijing University of Posts and Telecommunications
[4] Intel Labs China
clorisleef0313@gmail.com, shzhang.pku@gmail.com, yandong.guo@live.com

In this appendix, we present more detailed comparisons with other neural video delivery method. In addition, we further apply EMT on long videos with more settings.

1 Supplementary Comparisons with Neural Video Delivery Methods

We report the supplementary comparisons with H.264/H.265, CaFM [2], SRVC [1], and DVC [3] on more video sequences of VSD4K dataset. For the two commercial codec standards H.264 and H.265, we use ffmpeg with libx264 codec and libx265 codec to compress the HR videos at lower bit-rate while maintain the resolution. We obtain these compressed videos of the same size as our method (LR video and models). We also compare our method with DVC at four different bitrate-distortion trade-off operating points $\lambda \in \{256, 512, 1024, 2048\}$ (DVC1, DVC2, DVC3, DVC4). As shown in Fig. 1, our method outperforms these methods under same or less storage size in most cases. In Tab. 1, EMT achieves promising results compared with other methods with less training time in most circumstances, demonstrating the advantage of our method.

2 Extension to Long Videos

In this section, we report more results on long videos of VSD4K [2]. For long videos, previous neural video delivery methods take too much computational cost. Therefore, we only compare with commercial codec standards. We evaluate our method under two settings, which are denoted as M' and M separately. M' uniformly divides the long video into 5-second chunks and sequentially delivers the content-aware SR models. M first divides the long video into groups and applies EMT to each group. To be specific, we extract all the I-frames from the input video and make each group contain 30 I-frames. As shown in Tab. 2, both

*Equal contribution
†Corresponding Author

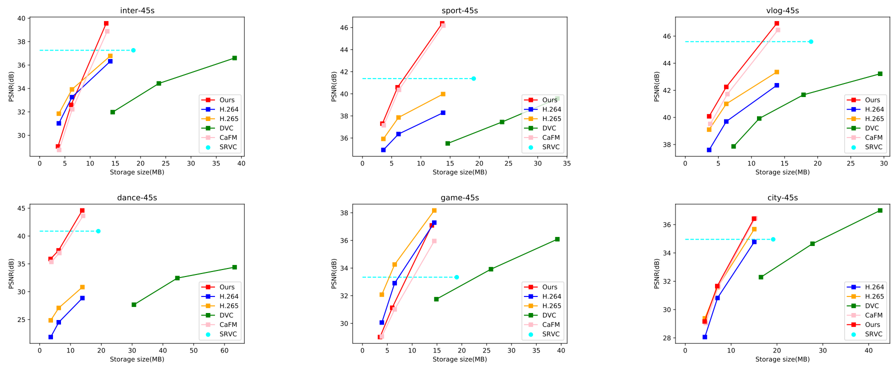

Fig. 1: Comparisons with neural video delivery methods in terms of PSNR and storage.

of our methods outperform commercial codec standards on long videos, showing the great potential of our approach. In addition, the margins of M surpasses the margins of M' since the temporal consistency between neighboring chunks is not always true for long videos. Dividing the long video into groups further improve the results of EMT.

Table 1: Comparisons with neural video delivery in terms of PSNR and training time. Red and blue indicate the best and the second best results among all methods.

	inter-45s		**sport-45s**		**vlog-45s**		**dance-45s**		**game-45s**		**city-45s**	
	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time
C1-n	38.95	11.2h	46.03	11.2h	46.20	11.2h	43.47	11.2h	35.61	11.2h	36.44	11.2h
CaFM	38.90	10.2h	46.12	10.2h	46.45	10.2h	43.63	10.2h	35.96	10.2h	36.43	10.2h
SRVC	37.26	12.1m	41.38	12.1m	45.59	12.1m	40.87	12.1m	33.34	12.1m	34.97	12.1m
DVC1	31.98	35.6m	35.52	33.6m	37.86	31.5m	27.67	38.4m	31.76	35.8m	32.30	36.2m
DVC2	34.44	36.1m	37.45	34.3m	39.92	30.8m	32.46	37.9m	33.93	36.0m	34.65	35.9m
DVC3	36.60	37.1m	39.58	34.8m	41.67	31.3m	34.40	38.5m	36.10	36.5m	37.00	36.6m
DVC4	38.70	38.0m	41.28	34.8m	43.22	33.6m	36.33	39.1m	38.10	36.2m	39.03	38.0m
Ours(M)	39.18	7.6m	46.25	7.6m	46.71	7.6m	44.24	7.6m	36.51	7.6m	36.42	7.6m

3 Further ablation study

We further evaluate the effectiveness of meta-tuning. Shown in Tab. 3, P_{1-n} removes meta-tuning but still reserves the pretrained model on DIV2K while MP_{1-n} eliminates the pretrain model on DIV2K and adopts meta-tuning strategy to train from scratch. As can be seen, pretrained model on DIV2K along with the meta-tuning strategy contribute jointly to the overall performance of EMT.

Table 2: Comparisons with H.264 and H.265 on long videos. We show the results of our M' and M methods using 3 epochs for fine-tuning. The storage is measured in megabytes and Acc is measured in PSNR. Margin indicates the difference of our method and H.265.

	vlog-5min		**vlog-10min**		**vlog-20min**		**vlog-30min**	
	Acc	Storage	Acc	Storage	Acc	Storage	Acc	Storage
H.264	34.45	18.58	35.08	35.41	35.05	70.75	34.88	144.41
H.265	36.67	18.58	37.08	35.41	37.11	70.75	37.00	144.41
Ours(M')	37.44	18.58	37.99	35.41	38.17	70.75	38.21	144.41
Margin	+0.77	-	+0.91	-	+1.06	-	+1.21	-
H.264	34.68	18.62	35.78	35.64	35.29	71.19	35.02	145.12
H.265	36.75	18.62	37.15	35.64	37.18	71.19	37.07	145.12
Ours(M)	37.67	18.62	38.33	35.64	38.31	71.19	38.41	145.12
Margin	+0.92	-	+1.18	-	+1.13	-	+1.34	-

Table 3: Effectiveness of meta-tuning.

Method	**PreD**	**MT**	**CPS**	**inter-45s** PSNR	Time	**sport-45s** PSNR	Time
P_{1-n}	✓	-	✓	39.08	18.4m	46.11	4.2m
MP_{1-n}	-	✓	✓	39.08	2.2m	46.11	11.3m
Ours(S)	✓	✓	✓	39.08	1.2m	46.11	1.2m

4 Further comparisons with baseline

In this section, we demonstrate the advantages of EMT compared with baseline. Shown in Tab. 4, We set the default training epoch of C_{1-n} to 300 and add a baseline method C^*_{1-n} under 1000 epochs. Both C_{1-n} and C^*_{1-n} are trained from scratch. We further adopt a pretrained model on DIV2K for the baseline method, denoted as C^{*p}_{1-n}. As can be seen, training with more epochs along with adopting pretrained model can further improve the baseline result. Both C^*_{1-n} and C^{*p}_{1-n} reach the highest PSNR at about 800 epochs. Nevertheless, our result is still competitive with C^{*p}_{1-n} and significantly faster.

Table 4: Comparisons with baseline [29].

	inter-45s		**sport-45s**		**dance-45s**	
	PSNR	Time	PSNR	Time	PSNR	Time
C_{1-n}	38.95	11.2h	46.03	11.2h	43.47	11.2h
C^*_{1-n}	39.23	29.8h	46.44	29.8h	43.79	29.8h
C^{*p}_{1-n}	39.48	29.8h	46.54	29.8h	44.09	29.8h
Ours(L)	39.56	55.5m	46.41	1.76h	44.59	1.53h

References

1. Khani, M., Sivaraman, V., Alizadeh, M.: Efficient video compression via content-adaptive super-resolution. arXiv preprint arXiv:2104.02322 (2021)
2. Liu, J., Lu, M., Chen, K., Li, X., Wang, S., Wang, Z., Wu, E., Chen, Y., Zhang, C., Wu, M.: Overfitting the data: Compact neural video delivery via content-aware feature modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4631–4640 (2021)
3. Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., Gao, Z.: Dvc: An end-to-end deep video compression framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11006–11015 (2019)