Learning to Fuse Residual and Conditional Information for Video Compression and Reconstruction

Ran Wang[1], Zhuang Qi[1], Xiangxu Meng[1], and Lei Meng[1,2] *

[1] Shandong University, Jinan, Shandong, China
[2] Shandong Research Institute of Industrial Technol, Jinan, China
{wr0702,z_qi}@mail.sdu.edu.cn
{mxx,lmeng}@sdu.edu.cn

Abstract. With the rapid development of the Internet, video compression and reconstruction have attracted more and more attention as the use and transmission frequency of video data have increased dramatically. Traditional methods rely on hand-crafted modules for inter-frame and intra-frame coding, but they often fail to fully exploit the redundant information of video frames. To address this problem, this paper proposes a deep learning video compression method which combines conditional context information and residual information to fully compress intra-frame and inter-frame redundancy. Specifically, the proposed algorithm uses conditional coding to provide rich context information for residual methods. At the same time, residual coding supports conditional coding in dealing with redundant information. By fusing the video frames generated by the two methods, information complementarity is achieved. Experimental results from two benchmark datasets show that our method can effectively remove redundancy between video frames and reconstruct video frames with low distortion to achieve better than state-of-the-art (SOTA) performance.

Keywords: video compression · residual coding · conditional coding.

1 Introduction

With the rapid development of digital media technology, a large amount of video content has been generated and contributes about 80% of the Internet traffic. However, the large-scale and high-redundancy properties prevent a large number of videos from being widely available. Therefore, it is very meaningful and critical to design an efficient video compression method to reduce the bandwidth required for video transmission and the storage space on the terminal device. Meanwhile, this may bring benefits for other vision tasks, such as video object detection [8, 23] or tracking [9, 40].

Traditional video compression methods [28, 36] are implemented through hand-crafted modules, however, it has been observed that they cannot achieve end-to-end optimization and have limited compression efficiency. Recently, deep learning has found applications in areas such as recommendation [19–22, 24],

* Corresponding author

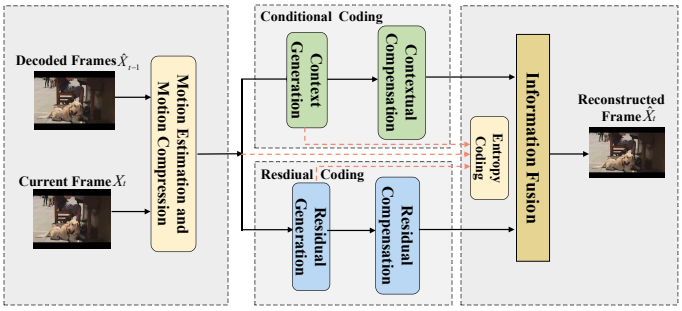


Fig. 1. Illustration of RCVC for Video compression with residual coding and conditional coding based on motion information.

classification [3, 13, 14, 16, 32, 34], image generation [11, 12, 29, 33] and federated learning [17, 27]. Recently, deep learning has demonstrated superior capabilities in computer vision tasks and has also received increasing attention in video compression tasks [10,15,18,38]. Existing deep learning-based video compression frameworks mainly include: residual coding framework and conditional coding framework. Residual coding method [5,18,38] uses residual information to reduce redundancy between video frames, but the subtraction operation it uses is too simple to obtain enough effective information. Compared with residual coding, conditional coding [10] can obtain rich context information, which can help the model to learn high-frequency information in video frames and supplement the residual information obtained in residual coding, but conditional coding methods prone to artifacts.

To address these problems, this paper proposes a dual-channel video compression framework, named RCVC, which combines residual coding and conditional coding. Figure 1 shows the entire framework, which can achieve information complementation and mitigate the interference of redundant information. Specifically, RCVC first uses the motion estimation module to obtain the original optical flow information, and uses the relevant encoder to obtain the encoded optical flow information to mine the temporal redundancy between video frames. Second, RCVC uses the residual coding module to obtain residual information between video frames to reduce the interference of redundant information. Meanwhile, RCVC uses the conditional coding module to generate rich conditional information. This enables the compression of residual information and context information. Finally, RCVC fuses the residual reconstruction and conditionally reconstructed video frames based on the fusion module to complete the integration of video frame information. Throughout the process, the entropy encoding module compresses the potential representation losslessly to create a bitstream.

Experiments are conducted on the Vimeo90K and UVG datasets in terms of performance comparison and ablation study for the effectiveness of the proposed. The experimental results show that the proposed method is superior to the traditional methods and existing state-of-the-art method in PSNR and MS-SSIM evaluation indexes. In summary, the main contributions of this paper are as follows:

- We propose a new video compression framework called RCVC that can fuse residual and conditional information from sequential video frames for improved video reconstruction.
- We propose a feature fusion module based on residual reconstruction frame and conditional reconstruction frame, which combines context information and residual prediction information to achieve information complementarity and squeeze redundancy.
- Our framework adds information to the reconstruction of video frames, reducing information redundancy and preserving high-frequency information, achieving better performance than traditional and state-of-the-art methods.

2 Related works

2.1 Image compression

In the past few years, image compression based on deep learning [1, 2, 4, 7, 26, 31] has developed rapidly. Image compression technology based on deep learning has achieved significant performance, surpassing traditional hand-designed lossy image encoders [30], which has greatly promoted the development of video compression technology. For example, the superprior model proposed by Balle et al. [1], which helps transform the marginal probability model of encoded symbols into a joint model by introducing additional latent variables as priors. This reduces redundancy and lowers the bit rate. It is also widely used in motion codecs and residual codecs for video compression. He et al. [4] proposed to use checkerboard convolution as a parallel alternative to the serial autoregressive context model, which has a better degree of parallelism under the same complexity. This provides an idea for us to further speed up the coding speed in conditional coding video compression framework.

2.2 Video Compression

Traditional video compression algorithms [28, 36] mainly follow the prediction coding structure and rely on manually designed modules, such as discrete cosine transform (DCT) and block-based motion estimation, to reduce the spatiotemporal redundancy between video frames. Manual methods provide effective compression results under standards such as H.264 [36] and H.265 [28], however, due to splitting between modules, they are difficult to achieve overall joint optimization, and there are compression artifacts and other problems. Therefore, video compression technology based on deep learning [5, 6, 10, 15, 18, 38, 39] has received more and more attention. Lu et al. [18] proposed the first end-to-end rate-distortion optimization video compression framework, which replaced the key modules in traditional hybrid video codecs with deep neural networks. Hu et al. [5] proposed a resolution adaptive optical flow compression method, which considered rate-distortion optimization when encoding motion vectors (MV). Researchers at Microsoft Research [10] proposed a new coding paradigm called conditional coding framework, which uses motion estimation and motion compensation modules to generate contexts as conditional guidance for codecs and

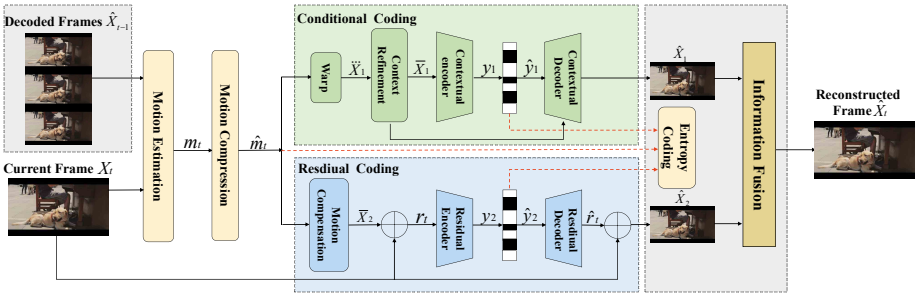


Fig. 2. Overview of our proposed video compression scheme. Motion Estimation and Motion Compression obtain motion information $\hat{m}_t$ from the current frame X_t and the previous reconstructed frame X_{t-1}, Conditional Coding extracts video contextual features based on the motion information $\hat{m}_t$, and Residual Coding evaluates the information quality of the reconstructed frame $\hat{X}_t$ to remove redundant video information.

entropy models. In this paper, we propose a video compression method (RCVC) which integrates the feature information obtained by the conditional encoding method and the residual encoding method, so as to achieve better compression effect.

3 Method

In this section, we will detail the motion estimation and motion compression module, residual coding module, conditional coding module, information fusion module and strategies to train the framework. An overview of our scheme is depicted in Fig 2.

3.1 Motion Estimation and Compression Module

The process of video compression is as follows, first we feed the current frame X_t and the previously reconstructed frame $\hat{X}_{t-1}$ into a neural network-based motion estimation module to estimate the optical flow. Then we can get two-dimensional optical flow information, that is, the displacement deviation in adjacent video frames. In this paper, the motion estimation module is based on the pre-trained Spynet. The obtained motion vector $\mathbf{m}_t$ is compressed lossily by the motion encoder, and then the optical flow information is reconstructed at the decoding end to mine the time redundancy between the adjacent two frames. The reconstructed motion vector is denoted as $\hat{\mathbf{m}}_t$. The motion estimation and motion compression process can be expressed as:

$$m_t = ME\left(X_t, \hat{X}_{t-1}\right) \tag{1}$$

$$\hat{m}_t = f_{Dn}\left(Q\left(f_{En}\left(m_t\right)\right)\right) \tag{2}$$

Where X_t is the current frame. $\hat{X}_{t-1}$ refers to the previously decoded frame. $ME(\cdot)$ represents the function of generating the motion vector. $\mathbf{m}_t$ refers to the motion vector. $Q(\cdot)$ is the quantization operation. $f_{En}(\cdot)$ and $f_{Dn}(\cdot)$ are motion encoder and decoder. $\hat{\mathbf{m}}_t$ represents the reconstructed motion vector.

3.2 Residual Coding Module

The decoded motion vector $\hat{\mathbf{m}}_t$ and the previously reconstructed frame $\hat{X}_{t-1}$ are input into the motion compensation module, which can obtain the predicted frame $\bar{X}_2$ of the current frame. Subtracting the current frame and the predicted frame can obtain residual information r_t. The residual information enters the residual encoder, and the quantization operation is realized by adding random noise, and the quantized information is put into the entropy model to obtain the estimated potential rate. Finally, the quantified residual information $\hat{y}_2$ is entered into the residual decoder to obtain the reconstructed residual $\hat{r}_t$. The residual coding module can be expressed as:

$$\hat{X}_2 = f_{RD}\left(Q\left(f_{RE}\left(X_t - \bar{X}_2\right)\right)\right) + \bar{X}_2 \tag{3}$$

Where $\bar{X}_2$ refers to the predicted frame. $Q(\cdot)$ is the quantization operation. $f_{RE}(.)$ and $f_{RD}(\cdot)$ are residual encoder and decoder. $\hat{X}_2$ represents the reconstructed frames obtained by the residual method.

3.3 Conditional Coding Module

The purpose of the conditional coding module is to obtain context information and encode and decode context information. By distorting motion vector and previously decoded frames, conditional context information contains more dimensions than residual information, allowing information in video frames to be more fully exploited. The context information is compressed into its latent representation by the conditional encoder, and then the same quantization operation is performed, the context information knows the guidance of the conditional decoder decoding, and finally the reconstructed context information is obtained. The conditional coding module can be expressed as:

$$\hat{X}_1 = f_{CD}\left(Q\left(f_{CE}\left(X_t \mid \bar{X}_1\right), \bar{X}_1\right)\right) \tag{4}$$

Where $\bar{X}_1$ refers to the context information. $Q(\cdot)$ is the quantization operation. $f_{CE}(\cdot)$ and $f_{CD}(\cdot)$ are conditional encoder and decoder. $\hat{X}_1$ represents the reconstructed frames obtained by the conditional method.

3.4 Information Fusion Module

To better eliminate redundancy, we combine residual information and confition information, which complement each other. With the residual coding framework, we can get a preliminary reconstruction of the original frame, however, due to the simple subtraction, there will be artifacts and noise. At this time, the context information provided by conditional coding can help us further refine the reconstruction frame, so the fusion module we propose is to input the context information and two reconstruction frames at the same time, and combine the reference features of multiple frames through the CNN network to achieve better frame reconstruction. The fusion module can be formulated as:

$$\hat{X}_t = FN\left(\hat{X}_1, \hat{X}_2\right) \tag{5}$$

Where $\hat{X}_t$ refers to the decoded frame. $\hat{X}_1$ represents the reconstructed frames obtained by the conditional method. $\hat{X}_2$ represents the reconstructed frames obtained by the residual method. $FN(\cdot)$ represents the function for fusing the information from the two video frames.

3.5 Training strategy

In our proposed framework, we optimize the following Rate Distortion (RD) trade-off realizing using least bitrate to get the best reconstruction quality:

$$L_t = \lambda \cdot D\left(X_t, \hat{X}_t\right) + R_t^{\hat{m}} + R_t^{\hat{y}_1} + R_t^{\hat{y}_2} \tag{6}$$

L_t is the loss function for the current time step t. λ controls the trade-off between the distortion D and the bitrate cost R. $D(X_t, \hat{X}_t)$ refers to the distortion between the input frame X_t and the reconstructed frame $\hat{X}_t$, where $D(\cdot)$ denotes MSE (mean squared error) or MS-SSIM (multiscale structural similarity) for different targets. In this paper we select MSE, where D consists of three parts. R is calculated as the cross-entropy between the true probability and the estimated probabilities of the latent code. $R_t^{\hat{m}}$ represents the bit rate used for encoding the quantized motion vector latent representation and the associated hyper prior. $R_t^{\hat{y}_1}$ represents the bit rate used for encoding the quantized residual latent representation and the associated hyper prior. $R_t^{\hat{y}_2}$ represents the bit rate used for encoding the quantized contextual latent representation and the associated hyper prior.

4 Experiments

4.1 Experimental Setup

Training dataset. The training dataset is selected from the Vimeo90K dataset [37], which is 82G in size and contains 89,800 independent video clips with different contents downloaded from vimeo.com, which cover various scenes and actions, each with a sequence of 7 video frames and a resolution of 448 x 256 for the training images.

Test datasets. To evaluate the performance of the proposed method, UVG [25] and HEVC [28] datasets are used for the evaluation. UVG dataset contains seven high frame rate videos. The resolution is 1920×1080, where the difference between adjacent frames is small. The HEVC dataset contains 16 Class B, C, D, and E videos ranging in resolution from 416×240 to 1920×1080.

Evaluation Metrics. To measure the distortion of reconstructed frames, two evaluation metrics are used in this paper: PSNR and MS-SSIM [35]. MS-SSIM can reflect the perception of distortion better than PSNR.

Implementation Details. Our learning rate is set to 1e-4 at the beginning and then decayed to 1e-5. The batch size was set to 16. For the λ, we set it to 256,512,1024, and 2048, respectively.

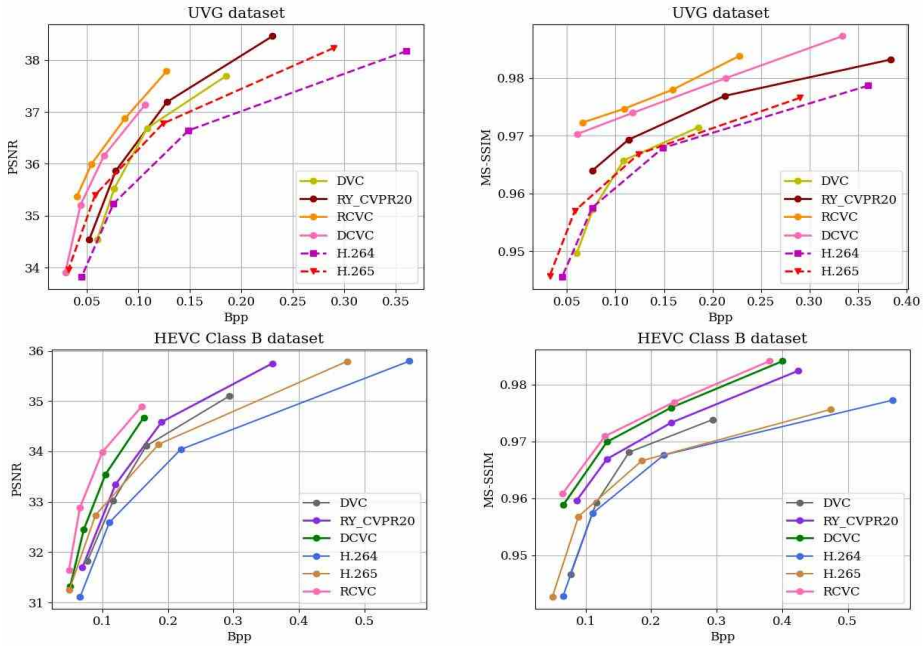


Fig. 3. Comparsion between our proposed method with traditional video compression methods [28,36] and the deep learning-based video compression methods [10,18,38] on the UVG and HEVC ClassB datasets.

4.2 Performance Comparison

Figure 3 shows the rate-distortion performance of the traditional video compression methods [28, 36] and the deep learning-based video compression methods [10, 18, 38] on the UVG and HEVC ClassB datasets. The horizontal coordinate is the bit rate. The higher the bit rate, the larger the volume occupied by the compressed video, the vertical axis is the compressed mass PSNR, and the larger the PSNR, the higher the reconstructed video quality. Fixed the horizontal coordinates, look at the vertical coordinates, the curve means the compression quality of the image at the same bit rate, and the upper curve above means that we get better compression quality at the same bit rate. Fixed vertical coordinates, which means that in the bit rate case with the same compression mass, the curve on the left can get a higher compression rate and a smaller bit rate. We have made the following findings:

- The video compression method based on deep learning outperforms the traditional video compression method in both the PSNR and MS-SSIM metrics. Deep learning can realize the end-to-end joint optimization, which is more effective than the manually designed traditional video compression.
- From the PSNR evaluation index, the proposed video compression method (RCVC) for fusion residual encoding and conditional encoding is about 1 dB higher than the DVC, and about 0.2dB higher than the DCVC. Our

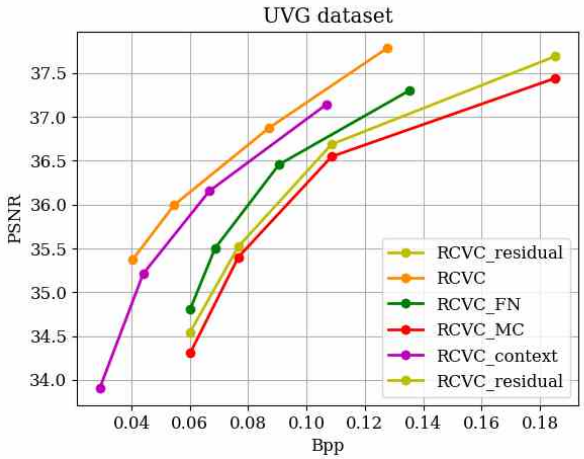


Fig. 4. Ablation Experiments. RCVC_residual verify the role of the residual coding module. RCVC_FN verify the role of the Feature Fusion Module. RCVC_MC verify the importance of motion compensation modules in video compression frameworks. RCVC_context validates the effect of context conditions on reconstructing video frames.

method can obtain the sample compressed reconstruction quality while saving the code rate. It shows that by integrating the information of the two modules, we can greatly reduce the redundancy between frames and reduce the information loss in the compression reconstruction process.

– From the MS-SSIN evaluation metric, at the same bpp level, our RCVC method was 0.1 dB higher than the DVC method, slightly better compared with the DCVC method. It shows that we can get more video frames in line with our subjective visual perception by integrating information.

Table 1. The BD-Bitrate comparison

Method	RCVC	DCVC	DVC	X265	X264
UVG	-28.8%	-25.3%	17.2%	0.0%	30.3%
HEVC ClassB	-29.3%	-26%	7.9%	0.0%	35%

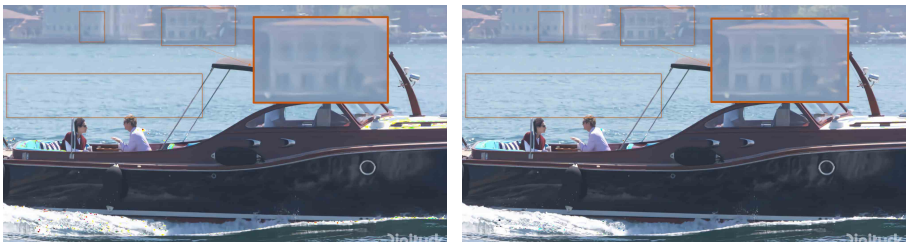
Table 1 shows the corresponding BD-Bitrate results. Our proposed RCVC method saves 28.8% and 29.3% bitrate in UVG dataset and HEVC ClassB dataset, respectively, which is better than DVC and DCVC, indicating that we have obtained better bitrate savings through information complementarity.

4.3 Ablation Study

In this section, ablation experiments were performed to validate the role of each module in our proposed RCVC framework.

The experimental results show that the PSNR at the same bpp level drops by about 1 dB, indicating that the motion compression module is crucial for us to obtain more accurate motion information. After the removal of the residual

(a) Original frame bpp / MS-SSIM (b) DVC 0.16bpp / 0.936



(c) DCVC 0.13bpp / 0.965 (d) RCVC 0.12bpp / 0.972

Fig. 5. Qualitative comparison.The reconstructed frames are from DVC, DCVC, and our RCVC method. Our method either achieves better visual quality or uses fewer bits.

coding module, the PSNR decreased at the same bpp level by approximately 1 dB, indicating that the residual information can remove the temporal redundancy and thus put attention to the critical information. The PSNR decreased by about 2dB, indicating that conditional context information can be learned to high-frequency information in video compression. After removing the fusion module, the performance of video compression decreased by 1dB, indicating that our fusion information can assist us in video reconstruction.

4.4 Case Study

Figure 5 shows the reconstructed video frames of different video compression frameworks, which are from the residual coding framework DVC [18], the conditional coding framework DCVC [10], and the RCVC deep learning video compression framework we propose. The initial frame picks up one of the frames in the UVG dataset for a high frame rate video ReadySteadyGo.

From the figure, it can be found that the background color of the video frame reconstructed by the DVC method is quite different from the original picture, and the blur degree of the object is higher, indicating that its distortion degree is high. Moreover, from the details of the amplification, the DVC method has obvious compression artifacts, indicating that the image quality obtained by using only a simple subtraction operation is relatively rough and has noise interference. Overall, the image quality reconstructed by DCVC and RCVC methods is significantly better than that of DVC methods. However, in the recovery of high-frequency details, the RCVC method is better than the DCVC method. Specifically, we observed that the contours of distant houses in the images reconstructed by the RCVC method were more clearly visible, closer to the original

picture, and their BPP was smaller. This indicates that RCVC obtains both the rebuild quality and a smaller proportion in memory. It shows that we can not only effectively remove noise by stitching context information and RGB prediction, but also achieve better detail restoration through information complementarity.

5 Conclusion

The goal of video compression is to obtain the best reconstruction quality at the cost of minimal bit rate. Traditional video compression method and compression method based on a priori deep learning mostly adopt residual coding framework, theoretically, the current to code pixels may be associated with all the previously reconstructed pixels, for the traditional encoder, due to the huge search space, it is difficult to use artificial rules to show the correlation between the pixels. Thus, the deep learning-based video compression method first generates the prediction frame from the prior decoded frame, and then calculates the residual difference between the current frame and the predicted frame. The residue is encoded into a stream, and the decoder decods the stream to obtain the reconstructed residual, and finally adds with the predicted frame to obtain the decoded frame. Given the prediction frame residual frame is a good way to denoising, but it is simple and effective, but not the optimal solution, because to find the residual operation is a simple manual design subtraction operation, can not completely remove the amount of redundancy of the whole frame.

In this paper, we combine residual coding and conditional coding, by obtaining the correlation between high frequency information to better redundancy. Experiments show that video compression methods integrating residual coding and conditional coding can achieve better performance. In the future, our work will combine advanced causal inference [39] technology to infer the invariant factors that affect video quality. Second, expand it to more challenging settings, such as federated learning [23].

Acknowledgment

This work is supported by the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073), the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048), the TaiShan Scholars Program (Grant no. tsqn202211289).

References

1. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. arXiv preprint arXiv:1611.01704 (2016)
2. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436 (2018)
3. Guan, Q.L., Zheng, Y., Meng, L., Dong, L.Q., Hao, Q.: Improving the generalization of visual classification models across iot cameras via cross-modal inference and fusion. IEEE Internet of Things Journal (2023)
4. He, D., Zheng, Y., Sun, B., Wang, Y., Qin, H.: Checkerboard context model for efficient learned image compression. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14766–14775 (2021). https://doi.org/10.1109/CVPR46437.2021.01453

5. Hu, Z., Chen, Z., Xu, D., Lu, G., Ouyang, W., Gu, S.: Improving deep video compression by resolution-adaptive flow coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 193–209. Springer (2020)
6. Hu, Z., Lu, G., Xu, D.: Fvc: A new framework towards deep video compression in feature space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1502–1511 (2021)
7. Johnston, N., Vincent, D., Minnen, D., Covell, M., Singh, S., Chinen, T., Hwang, S.J., Shor, J., Toderici, G.: Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4385–4393 (2018)
8. Lee, S.J., Lee, S., Cho, S.I., Kang, S.J.: Object detection-based video retargeting with spatial–temporal consistency. IEEE Transactions on Circuits and Systems for Video Technology **30**(12), 4434–4439 (2020)
9. Li, C., Liu, X., Zhang, X., Qin, B.: Design of uav single object tracking algorithm based on feature fusion. In: 2021 40th Chinese Control Conference (CCC). pp. 3088–3092. IEEE (2021)
10. Li, J., Li, B., Lu, Y.: Deep contextual video compression. Advances in Neural Information Processing Systems **34**, 18114–18125 (2021)
11. Li, X., Wu, L., Chen, X., Meng, L., Meng, X.: Dse-net: Artistic font image synthesis via disentangled style encoding. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2022)
12. Li, X., Wu, L., Wang, C., Meng, L., Meng, X.: Compositional zero-shot artistic font synthesis. Proceedings of IJCAI (2023)
13. Li, X., Ma, H., Meng, L., Meng, X.: Comparative study of adversarial training methods for long-tailed classification. In: Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia. pp. 1–7 (2021)
14. Li, X., Zheng, Y., Ma, H., Qi, Z., Meng, X., Meng, L.: Cross-modal learning using privileged information for long-tailed image classification. CVM (2023)
15. Lin, J., Liu, D., Li, H., Wu, F.: M-lvc: Multiple frames prediction for learned video compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3546–3554 (2020)
16. Liu, J., Xiao, J., Ma, H., Li, X., Qi, Z., Meng, X., Meng, L.: Prompt learning with cross-modal feature alignment for visual domain adaptation. In: CAAI (2022)
17. Liu, T., Qi, Z., Chen, Z., Meng, X., Meng, L.: Cross-training with prototypical distillation for improving the generalization of federated learning. ICME (2023)
18. 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)
19. Ma, H., Li, X., Meng, L., Meng, X.: Comparative study of adversarial training methods for cold-start recommendation. In: Proceedings of ADVM (2021)
20. Ma, H., Qi, Z., Dong, X., Li, X., Zheng, Y., Meng, X.M.L.: Cross-modal content inference and feature enrichment for cold-start recommendation. IJCNN (2023)
21. Ma, H., Xie, R., Meng, L., Chen, X., Zhang, X., Lin, L., Zhou, J.: Exploring false hard negative sample in cross-domain recommendation. In: Recsys (2023)
22. Ma, H., Xie, R., Meng, L., Chen, X., Zhang, X., Lin, L., Zhou, J.: Triple sequence learning for cross-domain recommendation. arXiv preprint arXiv:2304.05027 (2023)
23. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)

24. Meng, L., Feng, F., He, X., Gao, X., Chua, T.S.: Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In: Proceedings of MM (2020)
25. Mercat, A., Viitanen, M., Vanne, J.: Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. In: Proceedings of the 11th ACM Multimedia Systems Conference. pp. 297–302 (2020)
26. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. Advances in neural information processing systems **31** (2018)
27. Qi, Z., Wang, Y., Chen, Z., Wang, R., Meng, X., Meng, L.: Clustering-based curriculum construction for sample-balanced federated learning. In: CAAI International Conference on Artificial Intelligence. pp. 155–166. Springer (2022)
28. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on circuits and systems for video technology **22**(12), 1649–1668 (2012)
29. Sun, W., Li, X., Li, M., Wang, Y., Zheng, Y., Meng, X., Meng, L.: Sequential fusion of multi-view video frames for 3d scene generation. In: CAAI International Conference on Artificial Intelligence. pp. 597–608. Springer (2022)
30. Taubman, D., Marcellin, M.: Jpeg2000: standard for interactive imaging. Proceedings of the IEEE **90**(8), 1336–1357 (2002). https://doi.org/10.1109/JPROC.2002.800725
31. Toderici, G., O'Malley, S.M., Hwang, S.J., Vincent, D., Minnen, D., Baluja, S., Covell, M., Sukthankar, R.: Variable rate image compression with recurrent neural networks. arXiv preprint arXiv:1511.06085 (2015)
32. Wang, Y., Li, X., Ma, H., Qi, Z., Meng, X., Meng, L.: Causal inference with sample balancing for out-of-distribution detection in visual classification. In: CAAI International Conference on Artificial Intelligence. pp. 572–583. Springer (2022)
33. Wang, Y., Li, X., Qi, Z., Li, J., Li, X., Meng, X., Meng, L.: Meta-causal feature learning for out-of-distribution generalization. In: European Conference on Computer Vision. pp. 530–545. Springer (2022)
34. Wang, Y., Qi, Z., Li, X., Liu, J., Meng, X., Meng, L.: Multi-channel attentive weighting of visual frames for multimodal video classification. IJCNN (2023)
35. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
36. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the h. 264/avc video coding standard. IEEE Transactions on circuits and systems for video technology **13**(7), 560–576 (2003)
37. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**, 1106–1125 (2019)
38. Yang, R., Mentzer, F., Gool, L.V., Timofte, R.: Learning for video compression with hierarchical quality and recurrent enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6628–6637 (2020)
39. Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A.: A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD) **15**(5), 1–46 (2021)
40. Yao, R., Lin, G., Xia, S., Zhao, J., Zhou, Y.: Video object segmentation and tracking: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(4), 1–47 (2020)