Augmented Deep Contexts for Spatially Embedded Video Coding

Anonymous CVPR submission

Paper ID 10325

Abstract

Most Neural Video Codecs (NVCs) only employ temporal references to generate temporal-only contexts and latent prior. These temporal-only NVCs fail to handle large motions or emerging objects due to limited contexts and misaligned latent prior. To relieve the limitations, we propose a Spatially Embedded Video Codec (SEVC), in which the low-resolution video is compressed for spatial references. Firstly, our SEVC leverages both spatial and temporal references to generate augmented motion vectors and hybrid spatial-temporal contexts. Secondly, to address the misalignment issue in latent prior and enrich the prior information, we introduce a spatial-guided latent prior augmented by multiple temporal latent representations. At last, we design a joint spatial-temporal optimization to learn quality-adaptive bit allocation for spatial references, further boosting rate-distortion performance. Experimental results show that our SEVC effectively alleviates the limitations in handling large motions or emerging objects, and also reduces 11.9% more bitrate than the previous state-of-the-art NVC while providing an additional low-resolution bitstream.

1. Introduction

The fundamental problem in video coding lies in determining rich references and effectively utilizing them to obtain an accurate prediction. The richer the references, the better the prediction, and the more bitrate saved.

Over the years, advancements in traditional video codecs [10, 16, 41, 49, 56] have heavily relied on more efficient utilization of temporal references. For example, more complex motion models [27, 54, 55, 57], are proposed to improve the quality of prediction. Inspired by them, recent developments in neural video codecs (NVCs) [18, 25, 26, 45, 46, 51] have also concentrated on improving the utilization efficiency of temporal references.

However, for most recent NVCs, employing only temporal references fails to handle large motions or emerging objects. One reason is that the motion models in existing NVCs are not strong enough to accurately estimate motion

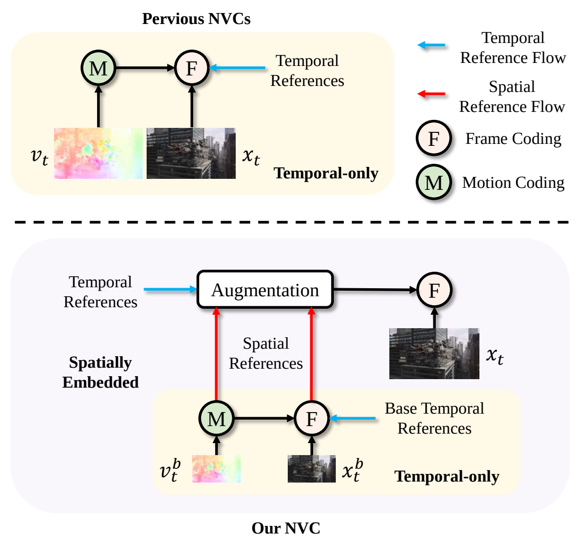
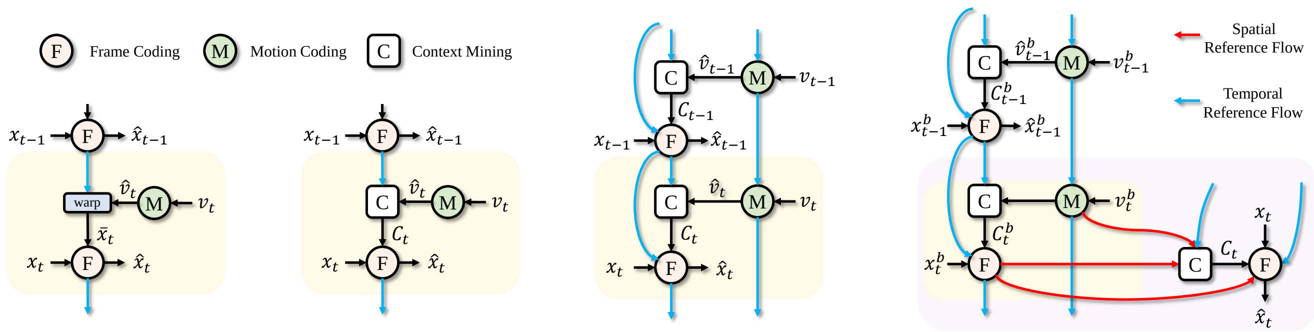


Figure 1. Compared with previous Neural Video Codecs (NVCs), our proposed codec embeds a base temporal-only codec to provide additional spatial references. Spatial and temporal references then augment deep contexts for frame coding.

vectors (MVs) [36], and inevitable motion coding errors exacerbate this inaccuracy. Inaccurate MVs disrupt the context mining, resulting in poor temporal contexts. Another reason is that these temporal-only contexts rely on limited temporal references, such as a single previously decoded feature and the misaligned latent prior [24], challenging to describe emerging objects. Therefore, mitigating the MV errors and generating richer contexts still represent the limitations encountered by NVCs.

Deviating from previous temporal-only NVCs, we draw inspiration from Video Super-Resolution (VSR) [11, 12, 20, 47] by incorporating additional low-resolution spatial references to reconstruct the high-resolution frame. To alleviate the limitations caused by insufficient references, we propose our **S**patially **E**mbedded **V**ideo **C**odec (SEVC). As shown in Figure 1, in previous NVCs, only temporal references are utilized for frame coding. By contrast, our SEVC embeds a base temporal-only codec [25] to provide additional spatial references. Then spatial references and temporal references are augmented for frame coding.



(a) DVC (b) DCVC (c) DCVC-DC (d) Our SEVC

Figure 2. Comparison of our Spatially Embedded Video Codec (SEVC) and previous NVCs. (a) compresses the residual between the prediction frame $\bar{x}_t$ and the input frame x_t. (b) extracts deep contexts C_t from the previously reconstructed frame to serve as the coding condition. (c) further improves the utilization of temporal references in both motion coding and frame coding. (d) conducts a base coding on the low-resolution input frame x_t^b and utilizes both spatial and temporal references for frame coding.

A more detailed comparison is shown in Figure 2. DVC [33] considers the aligned reconstructed frame as the prediction and the residual between the prediction and input frame is compressed. DCVC [23] proposes replacing the prediction frame with deep contexts, which serves as the condition in frame coding. DCVC-DC [25] further improves the utilization of temporal references in both motion coding and frame coding. However, these NVCs still rely on temporal-limited references.

Regarding DCVC-DC [25] as our base codec, we mainly focus on designing the augmentative codec. Specifically, our SEVC first compresses a 4x downsampling low-resolution frame, resulting in three low-resolution spatial references [7]: the base MVs, the spatial feature, and the spatial latent representation. Then a Motion and Feature Co-Augmentation (MFCA) module is proposed to augment base MVs and the spatial feature. The MFCA leverages the temporal feature to progressively amplify the quality of base MVs and the spatial feature. With the augmented base MVs and the spatial feature, hybrid spatial-temporal deep contexts are obtained for better frame coding. In addition to generating augmented deep contexts, introducing spatial references also benefits the entropy model. Given spatial latent representation as the guidance, multiple temporal latent representations could be aligned implicitly via Transformers [32, 47]. To further boost rate-distortion (RD) performance, we introduce a joint spatial-temporal optimization strategy to help our SEVC learn better bit allocation for spatial references.

Experiments show that our proposed SEVC effectively alleviates the coding limitations encountered by previous NVCs in handling large motions or emerging objects. When evaluated on commonly used test sets, our SEVC achieves state-of-the-art (SOTA) performance and far surpasses previous NVCs [17, 24–26, 45, 46] in sequences with aforementioned conditions. Our contributions are summarized as follows:

- We introduce embedding a base codec for additional spatial references to alleviate limitations in handling large motions and emerging objects. Spatial references benefit the generation of contexts and latent prior.
- For augmenting contexts, we propose a Motion and Feature Co-Augmentation module that progressively generates hybrid spatial-temporal contexts. For augmenting the latent prior, we introduce a spatial-guided latent prior that merges multiple temporal latent representations.
- We develop a joint spatial-temporal optimization strategy that helps the spatially embedded codec learn better bit allocation, further boosting RD performance.
- Our proposed codec surpasses the previous SOTA NVC by 11.9% more bitrate saving while providing a low-resolution bitstream.

2. Related Work

2.1. Neural Video Coding

Recent advancements in Neural Video Codecs (NVCs) have demonstrated the the superior potential of neural video coding. Existing studies on NVCs can be broadly categorized into two types: residual coding and conditional coding. Residual coding [3, 14, 18, 19, 31, 33, 34, 42, 43] addresses the frame differences to reduce redundancy, while conditional coding [17, 21–26, 45] exploits correlation based on contextual information from temporal references. Our SEVC also adopts conditional coding due to its flexible use of conditions, but exploring spatial references to generate richer contextual information for frame coding.

2.2. Deep Contexts Augmentation

Conditional coding exploits contextual information as the prediction, where the richness of contexts determines how good the prediction is and the performance of the frame coding. Most existing NVCs augment contexts [4, 23, 25, 38, 45, 46] through better utilization of temporal references. DCVC-TCM [45] incorporates a Temporal Context Mining

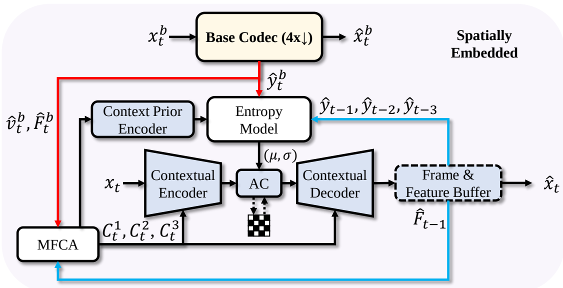


Figure 3. Overview of our SEVC. our SEVC embeds a base temporal-only codec for obtaining spatial references: base MVs $\hat{v}_t^b$, spatial feature $\hat{F}_t^b$, and spatial latent representation $\hat{y}_t^b$. Both the spatial and temporal references are utilized for the coding of the input frame x_t.

module that extracts multi-scale contexts from propagated features instead of decoded frames. DCVC-DC [25] proposes the Group-Based Offset Diversity Alignment to estimate offset maps for different feature groups, augmenting the capacity of temporal contexts to describe diverse scenarios. In the entropy model, Mentzer *et al.* use Transformers [52] to directly predict the contexts in latent space. To distinguish contexts in feature space and latent space, the term "contexts" refers to contexts in feature space, while "prior" denotes contexts in latent space. Based on this description, DCVC-HEM [24] introduces the temporal latent prior to further exploite temporal correlation.

However, temporal contexts are inadequate for representing newly appearing objects and are vulnerable to the accuracy of MVs [36], thereby failing to describe large motions or emerging objects. Meanwhile, there are works [4, 59] that leverages spatial references. Yang *et al.* [59] compress the space-time down-sampled video and reconstruct it through a space-time super-resolution network. Alexandre *et al.* [4] interpolate the prediction of the input low-resolution frame with the properties of random-access scenarios, and a super-resolution network is used to get a high-resolution prediction frame. These methods are not oriented to low-delay scenarios, where the interpolated MVs are not available. In addition, they consider prediction in pixel space which is inferior to contexts in feature space [18, 23]. By contrast, our SEVC augments the base MVs and the spatial feature synergistically to generate hybrid contexts in feature space.

3. Method

3.1. Overview

We consider downsampling the input video x to the low-resolution (LR) base video x^b. From a perspective of information theory [13], the mutual information between the source of x^b and x is

$$I(X^b;X) = H(X^b) - H(X^b|X), \tag{1}$$

where X^b and X denote the source of the base video and original video. Given that x^b is fully derived from x through a fixed downsampling algorithm, the conditional entropy $H(X^b|X)$ is constant to 0. Therefore, we can follow

$$I(X^b;X) = H(X^b). \tag{2}$$

Furthermore, the mutual information can be expressed equivalently as

$$I(X^b;X) = I(X;X^b) = H(X) - H(X|X^b). \tag{3}$$

Equation (2) and (3) follow directly from there with

$$H(X) = H(X^b) + H(X|X^b), \tag{4}$$

which shows that the coding process for the original video can be divided into two parts: one for the base video and another for the original video conditioned on the base video.

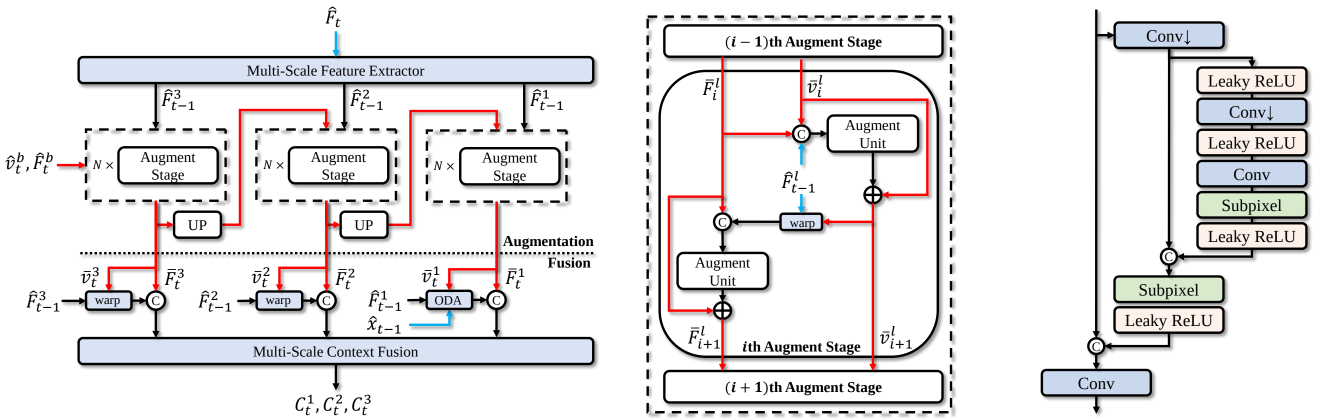
Fig 3 shows the diagram of our spatially embedded video codec. The base codec first compresses the 4x downsampling base frame x_t^b, deriving three spatial references: the base MVs $\hat{v}_t^b$, the spatial feature $\hat{F}_t^b$, and the spatial latent representation $\hat{y}_t^b$. These spatial references are subsequently augmented by the augmentative codec across two branches. One is the progressive augmentation of $\hat{v}_t^b$ and $\hat{F}_t^b$ through our proposed MFCA module to generate hybrid spatial-temporal contexts C_t^1, C_t^2, C_t^3. The other branch is the augmentation of $\hat{y}_t^b$ using multiple temporal latent representations $\hat{y}_{t-1}, \hat{y}_{t-2}, \hat{y}_{t-3}$ to generate the latent prior $\bar{y}_t$. We will introduce them in Section 3.2 and Section 3.3. In addition, a joint spatial-temporal optimization strategy is proposed in Section 3.4 to further boost RD performance.

3.2. Motion and Feature Co-Augmentation

Most previous NVCs estimate and compress full-resolution MVs [23–26, 30, 33, 34, 45], and reconstructed MVs are used for the alignment of decoded temporal frames or features. However, these NVCs only depend on temporal references, which can lead to suboptimal performance in complex scenarios involving large motions or emerging objects because of epistemic uncertainty [36]. To alleviate this, [18, 19, 25] estimate offset map instead of optical flow to represent motion. However, they still ignore motion coding errors. Another aspect is that temporal-only references are not enough to describe emerging objects in the input frame that are not available in the previous frames.

Inspired by Video Super-Resolution [11, 12, 20, 47], we consider whether we can introduce LR base MVs and an LR spatial feature to provide basic information for reconstructing the input frame, and then the temporal feature can be used to augment them progressively. To this end, we propose the Motion and Feature Co-Augmentation (MFCA) module to progressively generate hybrid spatial-temporal deep contexts.

As shown in Figure 4 (a), the MFCA module follows a bottom-up structure, progressively augmenting the base MVs and the spatial feature with incremental qualities and



(a) Illustration of our proposed Motion and Feature Co-Augmentation module.

(b) Illustration of multi-stage augmentation.

(c) Architecture of Augment Unit. "↓" indicates the stride is 2.

Figure 4. Diagrams of the Motion and Feature Co-Augmentation module. "UP" represents upsampling using subpixel convolution [48] with a factor of 2. "ODA" represents Offset Diversity Alignment proposed in [25]. © indicates channel dimension concatenation. ⊕ indicates element-wise addition.

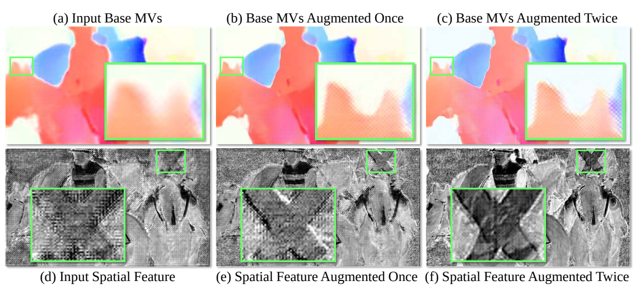


Figure 5. Visualization of base MVs and the spatial feature within the multi-stage augmentation in the middle scale.

resolutions. The temporal feature $\hat{F}_{t-1}$ are first extracted into multi-scale features [45]. Then starting from the smallest scale, the base MVs $\hat{v}_t^b$ and spatial feature $\hat{F}_t^b$ are augmented by the corresponding temporal feature through a multi-stage augmentation. The multi-stage augmentation in each scale consists of several Augment Stages, and the operation within one Augment Stage is illustrated in Figure 4 (b). In ith Augment Stage of lth scale, the base MVs $\bar{v}_i^l$ are first augmented by adding the residual predicted through one Augment Unit:

$$\bar{v}_{i+1}^l = \bar{v}_i^l + f_{AU}(cat(\hat{F}_{t-1}^l, \bar{F}_i^l, \bar{v}_i^l)), l = 1, 2, 3, \quad (5)$$

where cat represents channel dimension concatenation and $f_{AU}(\cdot)$ represents the Augment Unit. The augmented base MVs $\bar{v}_{i+1}^l$ are then used for a better alignment of multi-scale features $\hat{F}_{t-1}^l$. This better-aligned temporal feature could provide more accurate information to augment the spatial feature, and the augmented spatial feature $\bar{F}_{i+1}^l$ is also obtained by adding the residual predicted through another Augment Unit:

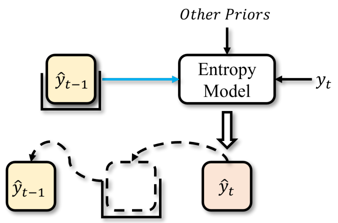
$$\bar{F}_{i+1}^l = \bar{F}_i^l + f_{AU}(cat(warp(\hat{F}_{t-1}^l, \bar{v}_{i+1}^l), \bar{F}_i^l)), l = 1, 2, 3. \quad (6)$$

The augmented base MVs $\bar{v}_{i+1}^l$ and spatial feature $\bar{F}_{i+1}^l$ are treated as low-quality inputs for the next Augment Stage and further augmented. Since the resolutions differ across scales, subpixel convolution [48] with a factor of 2 is used to ensure that the augmentation loop works properly. The last augmented base MVs in each scale are utilized to align the temporal feature by warp operation or Offset-Diversity Alignment [25]. The eventually aligned multi-scale temporal features are then concatenated with augmented spatial features, fed into the context fusion module to generate hybrid spatial-temporal contexts C_t^1, C_t^2, C_t^3. Although multi-stage augmentation may introduce considerable computational overhead, our Augment Unit, as illustrated in Figure 4 (c), operates at a declining resolution to mitigate this issue.

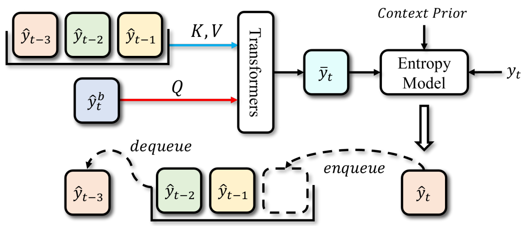
Figure 5 presents a visualization of the multi-stage augmentation in the middle scale. Before the augmentation, the base MVs and the spatial feature exhibited significant blurring at edges and regions with complex textures. After two-stage augmentation, it can be observed that the blurring in the MVs is effectively mitigated. The augmented MVs lead to a more accurate alignment of the temporal feature and thus a higher quality spatial feature.

3.3. Spatial-Guided Latent Prior Augmentation

Entropy models in NVCs rely on the prior information to estimate the distribution of the quantized latent representation $\hat{y}_t$, and subsequently, an entropy coding algorithm [15] can be used to compress $\hat{y}_t$ into a bitstream. The accuracy and richness of the priors significantly affect the precision of the estimated distribution. In addition to the commonly used context prior [23] and the hyperprior [6], as shown in Figure 6 (a), previous NVCs [24–26, 46] consider the previously decoded latent representation $\hat{y}_{t-1}$ as the latent prior. This off-the-shelf latent prior is not optimal because of the



(a) Latent prior $\hat{y}_{t-1}$ used in previous NVCs.



(b) Latent prior $\bar{y}_t$ used in our SEVC.

Figure 6. Comparison of different latent priors. (a) is the temporally misaligned latent representation $\hat{y}_{t-1}$. Given the spatial latent representation $\hat{y}_t^b$ as the query, (b) is augmented by multiple temporal latent representations $\hat{y}_{t-1}, \hat{y}_{t-2}, \hat{y}_{t-3}$. Dashed lines indicate updates for the latent representations.

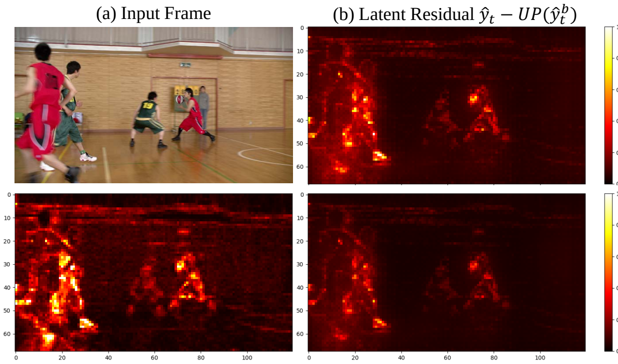
misalignment with the current index. As show in Figure 7 (c), a large residual occurs between $\hat{y}_t$ and the latent prior $\hat{y}_{t-1}$, adversely affecting distribution estimation.

In our SEVC, in addition to the temporal latent representation $\hat{y}_{t-1}$, there is a spatial latent representation $\hat{y}_t^b$. Shi *et al.* [47] gives a perspective that the attention mechanism in Transformers [5, 32, 52] can directly capture the temporal correlation between multiple frames. Unlike existing NVCs, which lack a suitable low-quality representation of the input frame for querying, our spatial latent representation y_t^b fulfills this requirement well.

The spatial latent representation $\hat{y}_t^b$ not only guides the alignment of the temporal latent representation $\hat{y}_t$, but also could complement richer prior information. To further extend prior information, we also query for multiple temporal latent representations. As shown in 6 (b), given $\hat{y}_t^b$ (4x upsampled by subpixel convolution to align the resolution) as the query Q and three temporal latent representations $\hat{y}_{t-3}, \hat{y}_{t-2}, \hat{y}_{t-1}$ as the keys K and values V, the Transformers [29, 47] align and fuse multiple temporal latent representations as the latent prior $\bar{y}_t$. The temporal latent representations are organized in a queue structure.

As shown in Figure 7, our latent prior $\bar{y}_t$ has a less residual compared to the latent prior $\hat{y}_{t-1}$ used in previous NVCs and the upsampled spatial latent representation $UP(\hat{y}_t^b)$. The latent prior and context prior extracted from hybrid contexts, will be fed together into the entropy model to estimate the distribution. Due to similar characteristics of hyper encoder/decoder and our base codec, we discard the hyperprior to avoid posterior collapse [35].

(a) Input Frame (b) Latent Residual $\hat{y}_t - UP(\hat{y}_t^b)$



(c) Latent Residual $\hat{y}_t - \hat{y}_{t-1}$ in DCVC-DC (Frame bpp 0.075, Rec. PSNR 36.54 dB)

(d) Latent Residual $\hat{y}_t - \bar{y}_t$ in SEVC (Frame bpp 0.052, Rec. PSNR 36.57 dB)

Figure 7. Illustration of latent residuals for different priors. The residual of our latent prior $\bar{y}_t$ and latent representation $\hat{y}_t$ is much less than that in DCVC-DC. Compared to the upsampled spatial latent representation $UP(\hat{y}_t^b)$, our prior is augmented by temporal latent representations and also performs a less residual.

3.4. Joint Spatial-Temporal Optimization

In traditional spatially scalable video coding [9, 37, 44], when two videos with different resolutions are compressed, the LR video is commonly given a higher quality to provide better spatial references. Better spatial references benefit the coding for the full-resolution video and improve the performance. Such a hierarchical quality structure is usually accomplished by assigning a smaller Quantization Parameter (QP) to the base layer, thereby allocating more bits to it [50].

However, our SEVC differs from those traditional codecs in two aspects. One is that, the performance of the base codec can be sacrificed for the purpose of obtaining a better full-resolution RD performance. Another aspect is that the bit allocation strategy for spatial references could be optimized end-to-end without handcrafted QP design.

To stabilize the training, we first independently optimize the two parts like previous NVCs [17–19, 23, 33, 34, 45].

$$L = \begin{cases} \dfrac{1}{T}\displaystyle\sum_{t}^{T}(w_t \cdot \lambda \cdot D_t^b + R_t^b), & \text{if base codec,} \\ \dfrac{1}{T}\displaystyle\sum_{t}^{T}(w_t \cdot \lambda \cdot D_t + R_t), & \text{otherwise,} \end{cases} \tag{7}$$

where w_t denotes the frame-level hierarchical quality weight [25] and t indicates the frame index. R_t^b is the estimated bpp (bits per pixel) for the base codec and R_t for the whole codec. D_t^b measures the distortion between LR input frame x_t^b and LR reconstructed frame $\hat{x}_t^b$, while D_t measures it between input frame x_t and reconstructed frame $\hat{x}_t$. The base codec is trained first, followed by the augmentative part. When one part is being trained, the parameters of the other are frozen. After the independent optimization, we

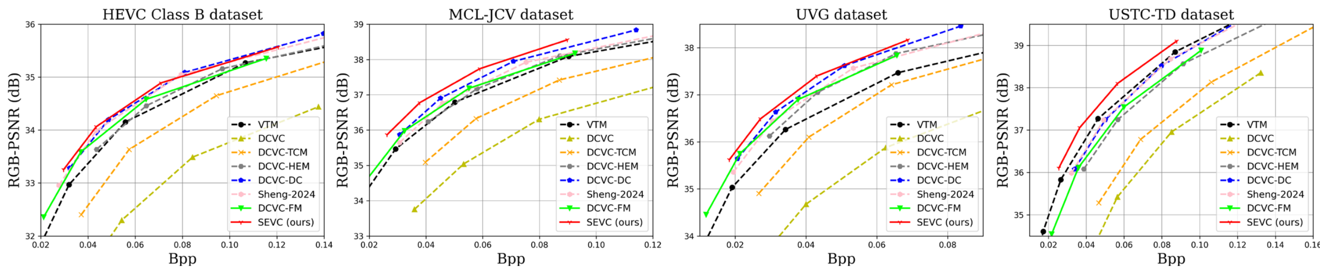
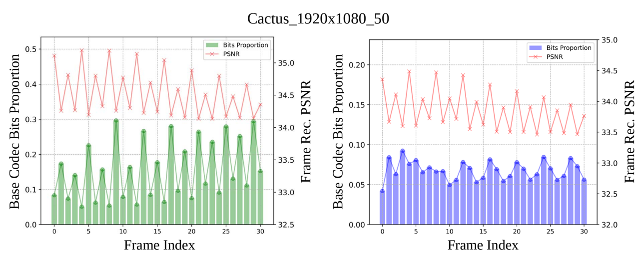


Figure 8. Rate and distortion curves on four 1080p datasets. The Intra Period is 32 with 96 frames.

Table 1. BD-Rate (%) comparison for PSNR. The Intra Period is 32 with 96 frames. The anchor is VTM-13.2 LDB.

	HEVC B	HEVC C	HEVC D	HEVC E	MCL-JCV	UVG	USTC-TD	Average
DCVC [23]	119.6	152.5	110.9	274.8	106.6	133.9	130.4	147.0
DCVC-TCM [45]	32.8	62.1	29.0	75.7	38.2	23.1	73.5	47.8
DCVC-HEM [24]	–0.7	16.1	–7.1	20.7	–1.6	–17.2	21.5	4.6
DCVC-DC [25]	–13.9	–8.8	–27.7	–19.2	–14.4	–25.9	7.6	–14.5
Sheng-2024 [46]	–13.7	–2.3	–24.9	–8.4	–7.1	–19.7	6.6	–9.9
DCVC-FM [26]	–10.3	–8.4	–25.8	–21.9	–8.1	–20.4	19.6	–10.8
SEVC (ours)	**–16.4**	**–15.8**	**–30.0**	**–28.5**	**–23.3**	**–30.2**	**–12.8**	**–22.4**

[1] Since the YUV MSE weight is higher than the RGB MSE in the loss function of DCVC-FM, its performance is inferior to DCVC-DC in this setting.



(a) w/ joint spatial-temporal optimization

(b) w/o joint spatial-temporal optimization

Figure 9. Illustration of base codec bits proportion. The joint spatial-temporal optimization leads our codec to learn a quality-adaptive bits allocation. Given the temporally hierarchical quality structure, the low-quality reconstruction frame implies more base bits while the high-quality one favors less.

conduct the following joint spatial-temporal optimization.

$$L = \frac{1}{T}\sum_{t}^{T}(w_t \cdot \lambda \cdot (D_t + w_l \cdot D_t^b) + R_t), \tag{8}$$

where w_l is a regular constraint of the base reconstruction. w_l is set to a small value to ensure the base reconstruction while barely affecting the full-resolution RD performance. Experiments in Sec. 4.3 show that this constraint is necessary to guarantee proper spatial references.

As shown in Figure. 9, without joint optimization, the bits proportion shows a relatively flat trend. However, after conducting joint optimization, our SEVC learned a more reasonable bit allocation. Given the temporally hierarchical quality structure, the low-quality reconstruction implies more base bits while the high-quality one favors less. This suggests that our codec learned to adaptively allocate bits for spatial references based on the required quality.

4. Experimental Results

4.1. Experimental Settings

Training Setup Our proposed SEVC is trained on Vimeo-90k [58] using the multi-stage training strategy [45]. The independent optimization for the two parts is conducted on the official train split, while the joint optimization is conducted on a selected subset of 9000 sequences from the original videos of the Vimeo-90k dataset. The sequences are randomly cropped into 256×256 in the independent optimization training stage and 384×256 in the joint optimization finetuning stage. A group of 6 frames is used for independent optimization and 32 frames for joint optimization. The λ in Equation (7) and (8) is set to {50, 95, 200, 400} for different bitrates, and the batch size is set to 8. The LR input frame is generated by bicubic downsampling [2].

Testing Setup UVG [39], MCL-JCV [53], HEVC Class B∼E [8], and USTC-TD [28] are used as the test sets. The distortion is calculated on full-resolution RGB colorspace and the BT.601 coefficient is used for the conversion between YUV420 and RGB colorspace. To demonstrate the superiority of our proposed SEVC, we compare our codec with the traditional codec–VTM-13.2 LDB [1] and several advanced conditional NVCs designed for low-delay scenarios: DCVC [23], DCVC-TCM [45], DCVC-HEM [24], DCVC-DC [25], Sheng-2024 [46], and DCVC-FM [26].

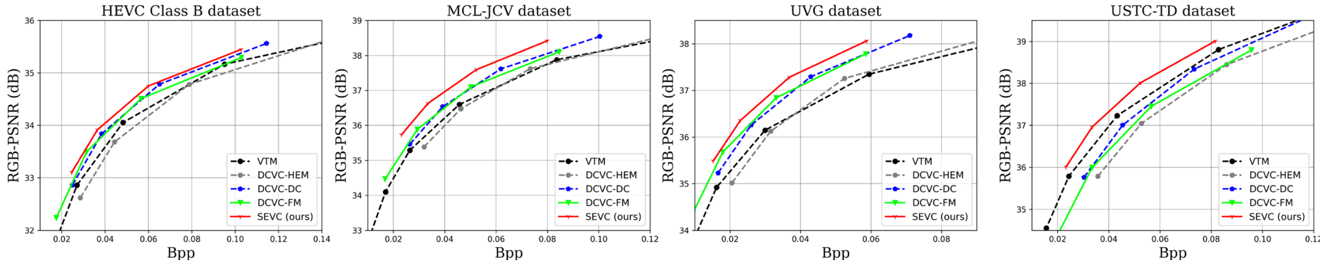


Figure 10. Rate and distortion curves on four 1080p datasets. The Intra Period is –1 with 96 frames.

Table 2. BD-Rate (%) comparison for PSNR. The Intra Period is –1 with 96 frames. The anchor is VTM-13.2 LDB.

	HEVC B	HEVC C	HEVC D	HEVC E	MCL-JCV	UVG	USTC-TD	Average
DCVC-HEM [24]	10.0	30.0	–1.1	68.6	4.9	1.2	27.2	20.1
DCVC-DC [25]	–10.8	–0.1	–24.2	–7.7	–13.0	–21.2	11.9	–9.3
DCVC-FM [26]	–11.7	–8.2	–28.5	–26.6	–12.5	–24.3	23.9	–12.6
SEVC (ours)	**–17.5**	**–15.1**	**–31.6**	**–34.0**	**–27.7**	**–33.2**	**–12.5**	**–24.5**

Table 3. BD-Rate (%) comparison for PSNR. The Intra Period is 32 with 96 frames. The anchor is DCVC-HEM.

Sequence Name	DCVC-DC	SEVC
* *USTC_BicycleDriving* [28]	–17.9	–55.9
* *videoSRC21_1920x1080_24* [53]	–14.5	–31.3
† *BasketballDrive_1920x1080_50* [8]	–8.5	–24.7
† *BQMall_832x480_60* [8]	–28.7	–35.9

1 * indicates sequences including large motions.
2 † indicates sequences including significant emerging objects.

We follow [24–26, 45, 46] to test 96 frames for each video. The Intra Period (IP) is set to 32 when SEVC is compared to all codecs and IP is set to –1 when compared to DCVC-DC and DCVC-FM.

4.2. Compared to Previous NVCs

Figure 8 and Figure 10 shows the RD curves of our SEVC and other codecs on four 1080p datasets, and Table 1 and Table 2 shows the BD-Rate results on all datasets. When the IP is set to 32, DCVC-DC achieves the highest bitrate saving compared to other previous NVCs, averaging 14.5% over the anchor VTM. However, our SEVC achieves 7.9% more bitrate saving compared to DCVC-DC. Since DCVC-FM is mainly designed for IP –1, DCVC-FM achieves the highest bitrate saving under IP –1, averaging 12.6% over the anchor VTM. In this case, our SEVC still achieves 11.9% more bitrate saving. In addition to the superior performance, our SEVC provides an additional base bitstream which can be independently decoded into a low-resolution video.

Furthermore, as demonstrated in Table 3, our SEVC far surpasses the previous SOTA codec in sequences with large motions or significant emerging objects. Through the visualization of intermediate MVs and contexts in Figure 11, we can observe that our augmented MVs have an even higher quality compared to MVs estimated by Spynet [40]. Although better MVs obtained, certain areas that newly appeared, such as the Thor hammer marked by the red box, still exhibit a large residual. This indicates that temporal references are not rich enough to generate a good prediction for the emerging object. However, our hybrid spatial-temporal contexts incorporate spatial references and complement the additional description for the hammer.

4.3. Ablation Studies

To validate the effectiveness of each technique used to improve the performance of our SEVC, we conduct comprehensive ablation studies. The average BD-Rate, calculated in terms of RGB PSNR on the HEVC datasets, is used here for comparison. Our method is marked in bold.

Motion and Feature Co-Augmentation. Table 4 shows the ablation study on the effectiveness of our MFCA. "w/o augmentation" represents that the base MVs and the spatial feature are directly sent to the context fusion part without any augmentation. "N stages" represents N Augment Stages in each scale. A positive correlation can be observed between performance and the number of Augment Stages. However, compromising complexity and performance, we adopt two Augment Stages in each scale.

Different Latent Priors. To demonstrate the advantages of our spatial-guided latent prior, we split and reassemble the prior generation components into four methods in Table 5. The comparison between M_1 and M_2 indicates benefits brought by multiple temporal latent representations. M_3 uses convolution networks instead of Transformers in M_2, resulting in a 5% increase in bitrate, which demonstrates the effectiveness of implicit alignment of Transformers. The last M_4 method further discards spatial latent representation in prior generation, and also performs bitrate increase.

CVPR
#10325

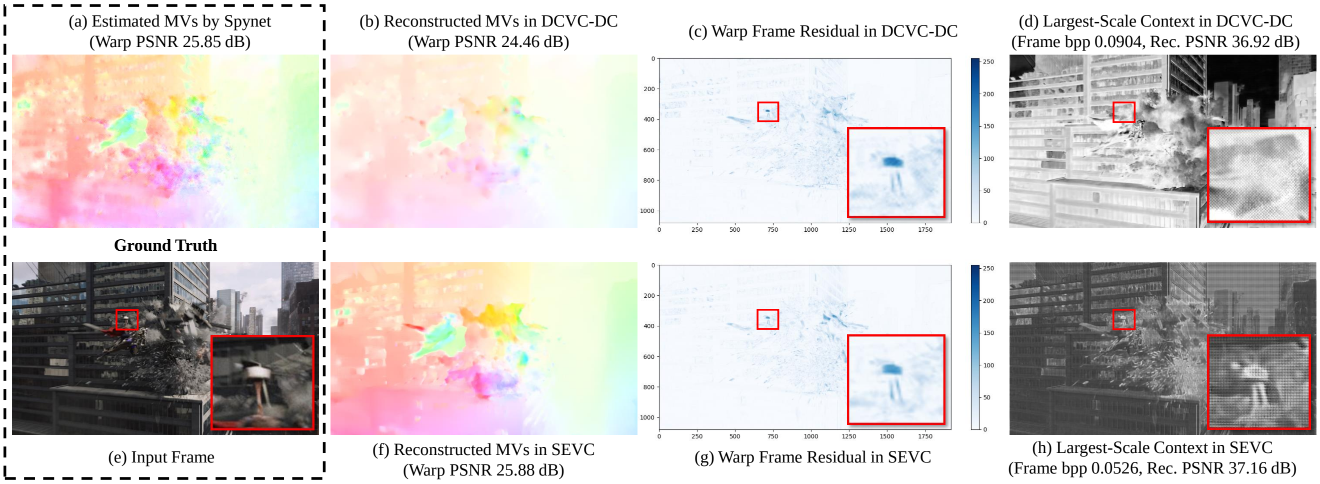


Figure 11. Visualization of the MVs and contexts in DCVC-DC and our SEVC.

Table 4. Ablation Study on MFCA.

	w/o augmentation	1 stages	**2 stages**	3 stages
MACs	2616G	2938G	**3263G**	3588G
BD-Rate (%)	17.5	4.7	**0.0**	–1.9

Table 5. Ablation Study on Different Latent Priors.

	$\boldsymbol{M_1}$	M_2	M_3	M_4
$\hat{y}_t^b$	•	•	•	
$\hat{y}_{t-1}$ [24]		•	•	•
$\hat{y}_{t-1}, \hat{y}_{t-2}, \hat{y}_{t-3}$	•			
Transformers [47]	•	•		
BD-Rate (%)	**0.0**	1.5	6.5	8.4

Table 6. Ablation Study on Joint Spatial-Temporal Optimization.

	w/o joint optimization	w_l				
		0	0.01	$\mathbf{0.05}$	0.1	1.0
BL bitrate proportion	13%	8%	12 %	**23%**	31%	54%
BD-Rate (%)	5.1	11.6	5.3	**0.0**	1.2	5.0

Table 7. Complexity Comparison.

	DCVC-HEM	DCVC-DC	Sheng-2024	DCVC-FM	**SEVC**
MACs	3435G	2764G	3830G	2334G	**3263G**
ET	548ms	663ms	968ms	587ms	**775ms**
DT	213ms	557ms	775ms	495ms	**734ms**

[1] Tested on a NVIDA RTX 3090 GPU with 1080p sequences as inputs.

Joint Spatial-Temporal Optimization. To better understand how the joint spatial-temporal optimization affects the RD performance, we conduct experiments under different strategies as shown in Table 6. If the joint optimization is not conducted, i.e., the base codec parameter is fixed during finetuning, the base codec occupies 13% bitrate. This is almost the same as setting w_l in Equation (8) to 0.01. Furthermore, a modest increment in w_l to 0.05 and 0.1 permits additional bits allocated to the base codec, facilitating a hierarchical quality structure akin to that achieved in traditional spatially scalable video coding. However, unconstrained setting with $w_l = 0$ or excessive bits allocated to the base codec has detrimental effects.

4.4. Comlexity

Table 7 shows the MACs, encoding time (ET), and decoding time (DT) comparison. The MACs of our SEVC are about the same as DCVC-HEM but reduce 44.6% more bitrate compared to VTM. Although compressing two videos, our MACs are still reduced by 15.4% compared to Sheng-2024 [46]. However, our SEVC is inferior to DCVC-DC and DCVC-FM, which made excellent optimizations for complexity and coding time. Considering the higher compression ratio and additional functionality, the increase in complexity is worthwhile.

5. Conclusion

In this paper, we present our solution to augmenting deep contexts and the latent prior for frame coding. For augmenting contexts, we synergistically augment the MVs and the spatial feature to generate hybrid spatial-temporal contexts. For augmenting the latent prior, we employ the spatial latent representation to merge multiple temporal latent representations. Finally, a joint spatial-temporal optimization is conducted to adjust the bit allocation for spatial references. Our SEVC well alleviates the limitations in handling large motions or emerging objects, achieving an 11.9% more bitrate saving compared to the previous SOTA codec.

However, the base codec used in our SEVC is limited to an existing NVC. Although spatial references can be obtained directly from the reconstruction of low-resolution video, the reconstruction limits the characteristics of spatial references. In the future, we will investigate more efficient ways to model and extract spatial references.

References

[1] VTM-17.0. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. Accessed July 28, 2024. 6

[2] bicubic-pytorch. https://github.com/sanghyun-son/bicubic_pytorch. Accessed July 28, 2024. 6

[3] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8503–8512, 2020. 2

[4] David Alexandre, Hsueh-Ming Hang, and Wen-Hsiao Peng. Hierarchical B-frame video coding using two-layer CANF without motion coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10249–10258, 2023. 2, 3

[5] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 4

[7] Yifan Bian, Xihua Sheng, Li Li, and Dong Liu. LSSVC: A Learned Spatially Scalable Video Coding Scheme. *IEEE Transactions on Image Processing: a publication of the IEEE Signal Processing Society*, 33:3314–3327, 2024. 2

[8] Frank Bossen et al. Common test conditions and software reference configurations. *JCTVC-L1100*, 12(7):1, 2013. 6, 7

[9] Jill M Boyce, Yan Ye, Jianle Chen, and Adarsh K Ramasubramonian. Overview of SHVC: Scalable extensions of the high efficiency video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(1):20–34, 2015. 5

[10] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 1

[11] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4947–4956, 2021. 1, 3

[12] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5972–5981, 2022. 1, 3

[13] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 3

[14] Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6421–6429, 2019. 2

[15] Jarek Duda. Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013. 4

[16] Bernd Girod, Eckehard G Steinbach, and Niko Faerber. Comparison of the H. 263 and H. 261 video compression standards. In *Standards and Common Interfaces for Video Information Systems: A Critical Review*, pages 230–248. SPIE, 1995. 1

[17] Yung-Han Ho, Chih-Peng Chang, Peng-Yu Chen, Alessandro Gnutti, and Wen-Hsiao Peng. Canf-vc: Conditional augmented normalizing flows for video compression. In *European Conference on Computer Vision (ECCV)*, pages 207–223. Springer, 2022. 2, 5

[18] Zhihao Hu, Guo Lu, and Dong Xu. FVC: A new framework towards deep video compression in feature space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1502–1511, 2021. 1, 2, 3

[19] Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, and Dong Xu. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5921–5930, 2022. 2, 3, 5

[20] Kappeler, Armin and Yoo, Seunghwan and Dai, Qiqin and Katsaggelos, Aggelos K. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016. 1, 3

[21] Théo Ladune, Pierrick Philippe, Wassim Hamidouche, Lu Zhang, and Olivier Déforges. Optical flow and mode selection for learning-based video coding. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020. 2

[22] Théo Ladune, Pierrick Philippe, Wassim Hamidouche, Lu Zhang, and Olivier Déforges. Conditional coding and variable bitrate for practical learned video coding. *arXiv preprint arXiv:2104.09103*, 2021.

[23] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:18114–18125, 2021. 2, 3, 4, 5, 6

[24] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1503–1511, 2022. 1, 2, 3, 4, 6, 7, 8

[25] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22616–22626, 2023. 1, 2, 3, 4, 5, 6, 7

[26] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26099–26108, 2024. 1, 2, 3, 4, 6, 7

[27] Zhuoyuan Li, Yao Li, Chuanbo Tang, Li Li, Dong Liu, and Feng Wu. Uniformly Accelerated Motion Model for Inter Prediction. *arXiv preprint arXiv:2407.11541*, 2024. 1

[28] Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li,

Changsheng Gao, et al. USTC-TD: A Test Dataset and Benchmark for Image and Video Coding in 2020s. *arXiv preprint arXiv:2409.08481*, 2024. 6, 7

[29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1833–1844, 2021. 5

[30] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-LVC: Multiple frames prediction for learned video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3546–3554, 2020. 3

[31] Haojie Liu, Ming Lu, Zhan Ma, Fan Wang, Zhihuang Xie, Xun Cao, and Yao Wang. Neural video coding using multiscale motion compensation and spatiotemporal context model. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3182–3196, 2020. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2, 5

[33] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11015, 2019. 2, 3, 5

[34] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Li Chen, Zhiyong Gao, and Dong Xu. An end-to-end learning framework for video compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3292–3308, 2020. 2, 3, 5

[35] James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 5

[36] Wufei Ma, Jiahao Li, Bin Li, and Yan Lu. Uncertainty-Aware Deep Video Compression With Ensembles. *IEEE Transactions on Multimedia*, 26:7863–7872, 2024. 1, 3

[37] Gwenaelle Marquant, Charles Salmon-Legagneur, Fabrice Urban, and Philippe de Lagrange. Spatial scalability with VVC: coding performance and complexity. In *Applications of Digital Image Processing XLV*, pages 10–15. SPIE, 2022. 5

[38] Fabian Mentzer, George Toderici, David Minnen, Sung-Jin Hwang, Sergi Caelles, Mario Lucic, and Eirikur Agustsson. Vct: A video compression transformer. *arXiv preprint arXiv:2206.07307*, 2022. 2

[39] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference (ACM MMSys)*, pages 297–302, 2020. 6

[40] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4161–4170, 2017. 7

[41] Karel Rijkse. H. 263: Video coding for low-bit-rate communication. *IEEE Communications magazine*, 34(12):42–45, 1996. 1

[42] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G Anderson, and Lubomir Bourdev. Learned video compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3454–3463, 2019. 2

[43] Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14479–14488, 2021. 2

[44] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Overview of the scalable video coding extension of the H. 264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, 2007. 5

[45] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 25:7311–7322, 2022. 1, 2, 3, 4, 5, 6, 7

[46] Xihua Sheng, Li Li, Dong Liu, and Houqiang Li. Spatial Decomposition and Temporal Fusion Based Inter Prediction for Learned Video Compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6460–6473, 2024. 1, 2, 4, 6, 7, 8

[47] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36081–36093, 2022. 1, 2, 3, 5, 8

[48] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 4

[49] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. 1

[50] Gary J Sullivan, Jill M Boyce, Ying Chen, Jens-Rainer Ohm, C Andrew Segall, and Anthony Vetro. Standardized extensions of high efficiency video coding (HEVC). *IEEE Journal of selected topics in Signal Processing*, 7(6):1001–1016, 2013. 5

[51] Chuanbo Tang, Xihua Sheng, Zhuoyuan Li, Haotian Zhang, Li Li, and Dong Liu. Offline and Online Optical Flow Enhancement for Deep Video Compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5118–5126, 2024. 1

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3, 5

[53] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavouni-

dis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513. IEEE, 2016. 6, 7

[54] Yao Wang and O. Lee. Use of two-dimensional deformable mesh structures for video coding .i. the synthesis problem: mesh-based function approximation and mapping. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(6):636–646, 1996. 1

[55] Yao Wang, O. Lee, and A. Vetro. Use of two-dimensional deformable mesh structures for video coding. ii. the analysis problem and a region-based coder employing an active mesh representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(6):647–659, 1996. 1

[56] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003. 1

[57] Thomas Wiegand, Eckehard Steinbach, and Bernd Girod. Affine multipicture motion-compensated prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):197–209, 2005. 1

[58] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 6

[59] Jiayu Yang, Chunhui Yang, Fei Xiong, Feng Wang, and Ronggang Wang. Learned low bitrate video compression with space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1786–1790, 2022. 3