A Flexible Algorithm Design of Spatial Scalability for Real-time Surveillance Applications

Zhe Zheng
AI Department
Beijing Smart-chip Microelectronics Technology Co., China
Zheng31zhe@outlook.com

Jinghua Liu
Equipment Management Department
State Grid Corporation of China
China
jinghualiua@126.com

Darui Sun
General Research and Development Department, Beijing Vimicro AI Chip Technology Corporation, China
daruisun052@outlook.com

Jinghui Lu
General Research and Development Department, Beijing Vimicro AI Chip Technology Corporation, China
hua17988@163.com

Song Qiu
Product Research and Development Center, Beijing Vimicro AI Chip Technology Corporation, China
ma2646154488@163.com

Yanwei Xiong
AI Department Beijing Smart-chip Microelectronics Technology Co., China
Yanwei0x@outlook.com

Rui Liu
AI Department Beijing Smart-chip Microelectronics Technology Co., China
zhigvjoairg@163.com

Wenpeng Cu
AI Department Beijing Smart-chip Microelectronics Technology Co., Ltd, China
wu1564466@163.com

Abstract: *The Surveillance Video and Audio Coding （SVAC） working group is currently developing the third-generation video compression standard, SVAC 3.0. To extend this standard, this paper proposes a spatial scalable coding Scalable Surveillance Video Coding （SSVC） framework, so that the SVAC3.0 video stream can gracefully adapt to different transmission bandwidth limitations and the requirements of decoding hardware, while keep the quality of the reconstructed image without degradation. In order to achieve the scalability of hardware implementation, SSVC designs a flexible reference frame marking and usage scheme, so that the enhanced layer coding does not directly depend on the basic layer, and effectively reduces the coding coupling between layers. SSVC improves the motion vector prediction method, effectively utilizes the encoding information of the base layer, and is mainly compatible with SVAC3.0 syntax structures. In addition, SSVC provides several different operational modes to adapt to various application scenarios and achieve an optimal trade-off between coding efficiency and complexity. The performance comparisons among simulcast stream and SSVC, single-layer stream and SSVC enhancement layer, as well as experimental data for different operational modes are provided. The coding efficiency and computational complexity are also analyzed.*

Keywords: *Video compression, video transmission, scalable coding, inter-layer reference, image rescaling, reference picture management.*

Received June 17, 2024; accepted October 14, 2024
https://doi.org/10.34028/iajit/21/6/11

1. Introduction

As digital video evolves towards higher spatial and temporal resolutions, wider color gamut, and broader dynamic range, it not only triggers upgrades and advancements in various segments of the content production, chip manufacturing, and network transmission industries but also drives the transformation of industries centered around video, such as broadcast television, security surveillance, intelligent transportation, video sharing, and video-on-demand services. The exponentially growing data volume poses significant challenges to the efficient transmission and storage of ultra-high-definition videos, especially in the field of video surveillance. There is an unprecedented pressure on storage, forwarding, analysis, and browsing of videos. With the widespread adoption of mobile devices, people are increasingly using smartphones, tablets, and laptops to connect to monitoring devices, browse, and even share videos. There has been an explosive growth in both public domain and home surveillance videos. Therefore, effective video compression and convenient encoding/decoding adaptability will provide a better service experience for a wide range of users [4, 8, 12].

To address the challenges of high bandwidth requirements and storage difficulties in ultra-high-definition video, the Surveillance Video and Audio Coding (SVAC) working group has taken the lead in developing a proprietary video encoding standard specifically designed for the surveillance industry. This standard aims to provide solutions with independent intellectual property rights for encoding ultra-high-definition videos. Since 2008, the SVAC working group has been closely following the development of video compression technologies worldwide. They have successively proposed surveillance video compression standards such as SVAC 1.0 and SVAC 2.0 [13]. Currently, the working group is in the process of developing SVAC 3.0.

However, the challenge lies in the fact that most personal monitoring devices have low-resolution screens, limited computational power, and battery life. Additionally, network video sharing can also lead to issues with network connectivity quality. To address these challenges and provide a suitable video quality for each user based on their receiving device and network connectivity, it is possible to encode and output multiple streams of the same video content with different spatial resolutions and qualities [18]. In the past few decades, with the advancement of video compression technology, scalable coding has been an active research area. To adapt to a wider range of application scenarios, there is a demand for the extension of scalable coding in the SVAC standard.

The organization of this paper is as follows. In s ection 2, the target and current advancements in scalable coding will be presented. Section 3 compares the Scalable Surveillance Video Coding (SSVC) solution with existing ones. Section 4 discusses inter-layer reference schemes and introduces the design considerations of SSVC. Section 5 provides a detailed explanation of the technical aspects of the SSVC solution based on the SVAC standard. Section 6 suggests the operational modes of the SVAC-based solution. Section 7 presents experimental data and analysis, while section 8 concludes the paper with a summary.

2. The Target of Scalable Coding

According to the initial concept, the scalability of an encoder refers to the ability to extract a portion of the bitstream and compose a new decodable bitstream to adapt to a new decoding terminal. The reconstructed quality of the partial bitstream is lower than that of the complete bitstream but higher than that of a single encoded stream with the same bitrate [10]. Common modes of scalable coding include temporal, spatial, and quality-SNR scalability. With the development of encoding technology, the application solutions for scalable coding have also evolved. Essentially, scalable coding and layered coding are synonymous concepts. Scalable coding enables natural hierarchy in decoding, display, user interaction, storage, forwarding, and other aspects, better matching application scenarios.

Spatial scalable coding involves generating multiple encoding substreams with different resolutions for a single image. Within one bitstream, it supports multiple representations with varying resolutions and bitrates. For the decoding terminal, the decoder can selectively decode partial substreams based on its capabilities. In terms of transmission, the network layer can discard some substreams based on the terminal's capabilities and network conditions, without causing decoding errors or interruptions due to packet loss.

Early video compression standards such as H.263 [15] and MPEG-4 Visual [6] defined their own scalable coding schemes. In particular, MPEG-4 introduced the design of the complex Fine Granularity Scalable (FGS) coding and Progressive Fine Granularity Scalable (PFGS) coding [17]. H.264/AVC introduced a scalable coding version called Scalable Video Coding (SVC) [10]. Its purpose is to encode video bitstreams that contain multiple substreams, with decoding complexity and image quality comparable to H.264/AVC single-stream encoding. However, the adoption of scalable coding solutions has been relatively limited. The reasons for this are primarily related to algorithm design, as scalable coding solutions often have higher encoding and decoding complexity, resulting in a loss of coding efficiency. In terms of application, compared to simulcast, hardware devices may not be compatible, leading to limited flexibility. Additionally, compared to transcoding, scalable coding imposes higher computational requirements on terminal devices, while transcoding mainly requires software and hardware updates on the server side. In light of these issues, H.265/HEVC extension Scalable High-Efficiency Video Coding (SHVC) [10] focused on simplifying the approach and achieving compatibility with single-layer encoders to the greatest extent possible. It became the first scalable coding standard built on a higher-level syntax framework only. By leveraging efficient inter-layer reference image processing modules, SHVC achieved high scalability coding gains without requiring changes to block-level coding logic. The recent introduction of Versatile Video Coding H.266/(VVC) [16] inherits the simplified approach of SHVC and further supports resolution switching. However, it does not explicitly define scalable coding extensions. Instead, it incorporates scalable coding and multi-view coding into the coding standard using the Video Parameter Set (VPS).

We can observe that the efficiency of scalable coding is heavily influenced by the performance of single-layer encoding. Improvements in single-layer compression performance inevitably erode some of the gains achieved through scalable coding. In the process of balancing algorithm complexity and coding efficiency, the complexity of algorithms has gone through a cycle of being initially simple, then becoming complex, and finally simplifying again. Initially, enhancement layer coding, as the name suggests, focused on enhancing the lower-layer signals, which implied improving image quality and increasing image resolution. However, with the development of video compression standards and the diversity of applications, a lower layer does not necessarily mean lower quality or always inferior to the enhancement layer. The requirements for enhancement layers can be relatively straightforward in many cases, such as pursuing smoothness, emphasizing improvements in image quality, or adapting to the display plane's dimensions. With the provision of more efficient encoding tools, the potential gain from inter-layer information diminishes accordingly. This necessitates controlling the inter-layer dependencies to a certain extent, which becomes more important in order to reduce hardware costs.

In the field of video surveillance, video images need to be displayed on terminals with different computing capabilities and resolutions. Furthermore, it is necessary to store the video data for future reference. Continuous video recording over a long period generates a significant amount of data, leading to high storage costs. Regularly deleting older videos can also result in unpredictable information loss. Scalable coding can effectively address this issue. Low-resolution bitstreams can be stored for long periods, while high-resolution videos are only stored for a short period of time. This can significantly reduce storage costs, which is particularly meaningful in the field of personal or home surveillance. Since temporal scalability can be achieved through the design of reference frames and is inherently supported in current standards, this paper will primarily focus on spatial scalability.

3. Standardization of Scalable Coding

3.1. Scalability in Early Standard

Enabling inherent scalability in video bitstreams to adapt to different applications and network environments has been one of the pursued goals since the standardization of video compression. After the release of the basic version of H.263 [15], scalable coding tools were quickly provided in the form of appendices, supporting temporal, spatial, and SNR scalability. Three image types were defined for scalable coding, namely B, EI, and EP frames. Each type has an Enhancement Layer Number (ELNUM) to indicate its corresponding layer, as well as a Reference Layer Number (RLNUM) to indicate the layer used for prediction. B frames are not used as reference frames for other frames, so they can be discarded without affecting the decoding of other frames, thus providing temporal scalability in scalable coding. Due to the differences between the reconstructed and original images caused by compression, this difference can be used as the encoding for the enhancement layer. The Peak Signal-to-Noise Ratio (PSNR) of the enhancement layer will be higher than that of the base layer, resulting in SNR scalability coding. If interpolated reference images are required for the enhancement layer, it is referred to as spatial scalability coding.

In MPEG4 VISUAL [6], temporal scalability supports both rectangular Video Object Planes (VOP) and VOPs of arbitrary shapes. However, in terms of spatial scalability, only rectangular VOPs are supported. The reference frames can be selected from the following four frames: the most recently decoded enhancement layer VOP, the closest decoded VOP in display order from the reference layer, the next VOP in display order from the reference layer, and the temporally corresponding VOP from the reference layer. The prediction for the enhancement layer is formed by combining temporal prediction based on motion compensation within the same layer and upsampling from the lower layer. The lower layer only requires upsampling and does not need motion compensation, so the motion vector (mv) is set to 0. Due to limitations in hardware performance, early layered coding schemes such as H.263 and MPEG4 VISUAL only utilized reference layer samples and did not fully exploit other inter-layer information.

3.2. Scalable Video Coding (SVC)

SVC [10], as an extension of H.264/AVC [1], is the first scalable coding standard that was meticulously planned and designed. It not only inherits all the well-established main coding tools of H.264/AVC but also introduces a comprehensive set of new tools aimed at enhancing performance. To fully utilize inter-layer information and improve compression performance, the SVC encoder employs a mechanism called inter-layer prediction. This mechanism consists of three parts: inter-layer motion prediction, inter-layer residual prediction, and inter-layer intra prediction. The SVC encoder can selectively use or not use these tools based on local signal characteristics. Except for inter-layer intra prediction, SVC does not directly use reference layer samples.

To distinguish it from traditional macroblock types, SVC introduces a new macroblock type indicated by the base_mode_flag. This flag indicates that the current macroblock utilizes inter-layer prediction. When both reference layer and enhancement layer macroblocks are inter-frame coded, the splitting of enhancement layer macroblocks, as well as the associated reference indices and motion vectors, are derived from corresponding data in the reference layer's 8x8 blocks at the same position. This process is referred to as inter-layer motion prediction. If a Macroblock (MB) utilizes inter-layer motion prediction, it does not need to encode reference frame indices. Instead, it directly uses the reference indices from the corresponding positions in the reference layer. The predicted values for its motion vectors are also obtained by scaling the motion vectors of the corresponding blocks in the reference layer at the same position. This allows efficient utilization of inter-layer information for motion prediction in SVC. When a reference layer MB is intra-coded, the prediction signal for the enhancement layer can be obtained through inter-layer intra prediction. Before upsampling the reconstructed intra signals in the reference layer, it is necessary to determine whether the neighboring block is intra-coded. If it is also intra-coded, the reconstruction needs to be decoded. If it is inter-coded, there is no need for decoding, and only the current reconstructed block needs to be extended at the boundaries before upsampling to generate the inter-layer intra prediction signal. It is evident that the decoding of the enhancement layer does not require complete decoding of the base layer, which is a distinctive feature of SVC's single-loop design.

In SVC, inter-layer residual prediction can be used for all Macroblocks (MBs) in the enhancement layer. In this case, the corresponding 8x8 sub-block residuals in the reference layer are upsampled and used as the predicted enhancement layer MB residual signal. In the enhancement layer bitstream, only the difference signal of the residuals is encoded. The upsampling of the reference layer residuals needs to be performed based on transform blocks to ensure that filtering is not performed at the boundaries of the transform blocks.

3.3. Scalable High Efficiency Video Coding (SHVC)

In fact, while the architecture of SVC is well-designed in terms of algorithmic aspects, it is not compatible with H.264/AVC single-layer encoders in practical applications. Supporting SVC in existing hardware designs is not an easy task. In view of this, the SHVC [3] of H.265/HEVC [7, 11] completely abandons the design approach of SVC. It ensures consistency between the base layer and single-layer encoders while restricting modifications in the enhancement layer that affect the lower layers.

The method used by SHVC to utilize inter-layer information is called Inter Layer Prediction (ILP). Although the name is similar to SVC, its essence has shifted from focusing on utilizing encoding information in SVC to primarily relying on traditional methods of referencing and reconstructing pixels. To achieve efficient ILP, the reconstructed reference layer images are obtained from the decoding process of the reference layer Decoder Picture Buffer (DPB). These reconstructed reference layer images, after undergoing interlayer processing, are then placed into the enhancement layer's DPB. They are used as Inter Layer Reference images (ILR) for predicting and encoding the enhancement layer images. This indicates that SHVC adopts a multiloop design, where the reference layer must be fully decoded before it can be utilized by the enhancement layer. For spatial scalable coding, SHVC has the following two characteristics:

1. In SHVC, arbitrary ratio resampling is adopted. The reference layer is upsampled to match the size of the enhancement layer and serves as a reference frame in the encoding process. During encoding, the original image can be downsampled to obtain lower-layer images at any scale ratio. Additionally, SHVC supports flexible phase adjustment. In cases where the phase offset during encoder downsampling does not match the decoder, the sample phase offset during decoder upsampling can be adjusted to match the encoder’s downsampling in order to reduce coding artifacts.
2. In terms of usage, the ILR can be considered as a long-term reference frame for the enhancement layer (although it has the same POC as other frames in the same AU). Additionally, due to the inherent design of HEVC, any reference frame can be used as a collocated frame for TMVP. In this case, the ILR can potentially be used as a collocated frame (although it can be avoided in practice). By utilizing the prediction modes of corresponding blocks in the lower-layer reference images, the reference frame index is directly used for TMVP generation. However, the lower-layer reference images need to possess the same information as other reference frames, such as POC, prediction information, reference frame lists, etc. If there are differences in inter-layer resolutions, MVs may need to be appropriately scaled.

3.4. Scalability in H.266/VVC

The new generation video compression standard, H.266/VVC [14, 18], does not have a specific definition for scalable coding extension. Instead, it incorporates multi-layer coding and single-layer coding together. In previous scalable coding standards, the spatial resolution of the image could not be freely changed unless it was re-encoded as a new Compressed Video Sequence (CVS). In the design of H.266/VVC, resolution updates can be achieved without updating the CVS or sending Sequence Parameter Set (SPS), intra frames, etc. This means that inter-frame prediction can be performed between images of different resolutions without the need for additional signaling or re-encoding. To achieve this, it is necessary to perform real-time resampling of reference samples to match the size of the current sample block. This technique is known as Reference Pictures Resampling (RPR). RPR ensures that the reference samples are appropriately rescaled to align with the resolution of the current sample block for efficient inter-frame prediction in H.266/VVC. By using RPR, VVC (H.266) gains the ability to perform inter-frame prediction from reference frames of different sizes. This allows VVC to easily support multi-layer bitstreams with different resolutions. The flexibility provided by RPR enables efficient compression and transmission of video content at various resolutions within the same VVC framework.

The design of scalable coding in H.266/VVC takes into consideration the compatibility with single-layer coding. For instance, parameters related to decoding capability and the definition of DPB size are independent of the number of layers in the bitstream. The key operations involved in RPR, such as sample resampling and motion vector mapping, are performed at the block level and seamlessly integrated with the single-layer encoder. To support multiple resolutions within a single VVC bitstream, significant efforts have been made in the higher-level syntax of H.266/VVC. Essentially, a single-layer decoder can decode multi-layer bitstreams with minimal modifications.

3.5. SSVC Scheme

The Scalable Surveillance Video Coding (SSVC) scheme presented in this paper is designed based on the surveillance video compression standard SVAC 3.0. In the design process, it takes into consideration the issues encountered in the design and promotion of SVC and SHVC, while also optimizing algorithms based on the characteristics of SVAC 3.0. The primary objective of the SSVC design is to provide a scalable coding solution with controllable complexity and ease of use for surveillance video applications. This differs from the goal of H.266/VVC, which aims to encompass a wide range of application scenarios to the maximum extent possible. An SSVC bitstream consists of a base layer and several enhancement layers. The base layer is fully compatible with the SVAC 3.0 single-layer encoder, while most of the basic lower-layer modules in the enhancement layer can be interchangeably used with the single-layer encoder to reduce hardware design costs. However, for compression efficiency, necessary algorithm updates have been made in the enhancement layer to fully utilize inter-layer information from the reference layer. These updates are implemented to enhance the coding efficiency of the SSVC scheme. SSVC still employs a multiloop decoding strategy, which means that the base layer image must be fully decoded before the decoding of the enhancement layers, which depend on the base layer image, can proceed. SSVC requires that the resolution of the lower-layer images is smaller than the resolution of the higher-layer images. The lower-layer images serve as inter-layer reference images for the higher-layer images and need to be upsampled to match the resolution of the higher-layer images. SSVC does not impose constraints on the upsampling algorithm but provides an encoding scheme for upsampling filters. To differentiate between inter-layer reference images and enhancement layer reconstructed images, an identifier is defined to indicate their source. Inter-layer reference images can be treated as regular reference frames and are given equal treatment as other reference frames within the coding framework. With this identifier in frame inter-prediction, it is possible to strategically adjust the frame inter-prediction algorithm based on the actual situation. This allows for leveraging the inter-layer information effectively. By considering the source of the reference frames, the frame inter-prediction algorithm can be slightly modified or optimized to make better use of the inter-layer information and improve coding efficiency. Unlike SHVC, SSVC imposes necessary limitations on TMVP. However, in the generation of spatial neighboring Motion Vector Predictor (MVP) candidates, SSVC considers the influence of inter-layer motion information. This means that while SSVC places restrictions on TMVP, it still leverages inter-layer motion information when generating MVP candidates in the spatial neighborhood.

SSVC provides several different operational modes for the encoder to adapt to various application scenarios. For example, in certain scenes with simple and static backgrounds, the inter-layer reference for P frames can be disabled to reduce design complexity while maintaining compression efficiency. These operational modes allow flexibility in adjusting SSVC encoding settings according to specific requirements and trade-offs between complexity and compression efficiency. Table 1 presents the technical characteristics of different standards' scalable coding extensions, allowing observation of the focal points of each approach.

Table 1. The characteristics of existing scalable coding extensions.

	H.263+	**MPEG4 Visual**	**SVC**	**SHVC**	**H.266/VVC**	**SSVC**
Inter-layer pixel reference	YES	YES	NO	YES	NO	YES
Inter-layer reference frame type	N/A	N/A	Short-term	Long-term	Long-term	Short-term
Inter-layer MVP	N/A	N/A	inter-layer motion information inherited	TMVP	TMVP	TMVP/ AMVP
Constrained intra prediction	NO	NO	YES	NO	NO	NO
Residual prediction	NO	NO	YES	NO	NO	NO
Scaling ratio	Fixed, up	Fixed, up	Fixed, up	Arbitrary, up	Arbitrary, down/up	Arbitrary, up
Decoding mode	Multi-loop	Multi-loop	Single-loop	Multi-loop	Single-loop	Multi-loop

4. Inter-Frame and Inter-Layer Prediction

Video scenes exhibit high information redundancy, combined with variations in human visual sensitivity to brightness and colors, providing ample opportunities for video compression. The pursuit of high-speed, high-frame-rate, and high-resolution videos, which is currently market-driven, further increases the redundancy in the information. Efficient utilization of inter-frame information redundancy has always been a major focus of successive video coding standards. From the initial use of a single forward reference frame with no phase offset prediction to the utilization of multiple bidirectional reference frames with multiple phase offset predictions, inter-frame information has been more effectively applied, resulting in significant improvements in compression efficiency. However, the increase in the number of reference frames and the finer granularity of phase offsets correspondingly lead to a manifold increase in complexity. Another consideration is the attempt to utilize prior information to mark reference frames, replacing the previous practice of treating reference frames indiscriminately. H.264/AVC introduced the concept of long-term reference frames. By combining short-term and long-term reference frames, the compression efficiency for approximately periodically repetitive videos can be effectively improved. It is generally believed that long-term reference frames lack temporal information and can only be used as pixel references, while short-term reference frames are traditional reference frames that can participate in complex motion prediction.

Regarding the effective utilization of long-term reference frames, there have been many beneficial efforts in the industry. In AV1 [5], there is an option to selectively encode a frame with high quality and designate it as the "GOLDEN PICTURE," which essentially serves as a long-term reference frame. The long-term reference frame is still a genuine frame that exists within the video sequence. However, in certain standards like Audio Video coding Standard (AVS) [19], the concept of background frames has been introduced [20]. Background frames are used in a similar manner to long-term reference frames. However, unlike long-term reference frames, background frames do not correspond to real frames in the video sequence. Instead, they are generated by learning non-motion regions of the video content and represent an image. Background frames are encoded using smaller quantization parameters. They are particularly effective in suppressing noise and achieving excellent prediction performance for videos with significant background components. However, the drawback is that they can lead to high instantaneous bitrate, making it challenging to set random access points within the video stream. To address the aforementioned challenges, SVAC 3.0 further introduces the concept of a library picture [21], which breaks the limitations of Random Access Points (RAP). The library picture consists of representative images that can reference each other but do not reference non-library pictures. Library pictures can be transmitted through out-of-band signaling or within slices as part of the bitstream. This concept allows for more flexible and efficient coding, enabling improved random access and reduced dependence on specific frames as reference points.

Due to the similarity between the reconstructed frames of the base layer and the current enhanced layer in terms of texture and motion information, SSVC considers it appropriate to manage them as short-term reference frames. Subsequent experimental analyses have also shown that both texture and motion information play important roles in scalable coding performance. When the resolution ratio between the base layer and the enhancement layer is large, texture prediction dominates. However, when the resolution ratio is small, motion information becomes more influential. The similarity between the reconstructed frames of the base layer and the texture of the enhancement layer complements the focus on background regions in long-term reference frames mentioned earlier. Experimental results indicate that scalable coding performs well for video sequences with significant content variations. In contrast, the library picture mode excels in scenes with simple backgrounds. Thus, the choice of encoding mode depends on the characteristics of the video content, with scalable coding demonstrating good performance for dynamic sequences and the library picture mode being effective for scenes with simple backgrounds.

5. The Algorithm Design of SSVC

Video surveillance is an important application area for scalable coding. One of the fundamental requirements is the access of terminals with different resolutions. Additionally, managing a large amount of recorded video data poses a dilemma between storage and deletion. Scalable coding can address both of these challenges effectively. When viewing videos on low-resolution terminals, only decoding the base layer images is sufficient, without compromising the viewing experience. On the other hand, high-resolution terminals such as television screens can display the enhanced layer images to examine necessary details. This way, scalable coding allows for flexible access and optimal utilization of video content based on the capabilities of different devices or terminals. Regarding storage, for older recordings, it is possible to store only the base layer bitstream to save disk space. Alternatively, image content analysis can be utilized to determine whether to retain the enhanced layer bitstream. However, even with these considerations, when promoting scalable coding in video surveillance, the SVAC standard faces similar challenges as SVC, SHVC, etc., The convenience of application and cost-effectiveness remain key issues that scalable coding needs to address. During the development of the new generation surveillance video compression standard, SVAC 3.0, efforts were made to investigate and address the implementation issues of its scalable coding scheme, SSVC. This process involved thorough examination and analysis to ensure the practical feasibility and effectiveness of the SSVC approach within the SVAC 3.0 framework.

Among the commonly used modes of scalable coding, temporal scalability is the simplest. Although there are some differences in various standards, the approach of utilizing B-frames for temporal scalability has remained unchanged since H.263. However, the trade-off between dropping B-frames and the resulting video stuttering is often considered less acceptable than sacrificing image quality. Signal-to-Noise Ratio (SNR) scalable coding was initially designed for network video transmission with fluctuating bandwidth. However, its application has been limited in practice. Nevertheless, the algorithm itself can be considered as a subset of spatial scalable coding techniques. Color gamut scalable coding is defined in SHVC, but its specific application is rarely seen in the field of surveillance video. Therefore, the design of spatial scalable coding becomes particularly important. The SSVC scheme primarily focuses on addressing spatial scalability issues.

5.1. Reference Frame Identification

To align with the reference frame management mode of the enhancement layer, the reconstructed images of the base layer are upsampled and included in the reference frame sequence of the enhancement layer without altering the reference frame management of the enhancement layer. In the encoding process of the enhancement layer, a reference frame is defined with an identifier and layer number to indicate whether the frame originates from a lower layer and its corresponding layer number. This allows for proper referencing between layers, ensuring efficient inter-layer prediction and coding.

In addition to the identifier and layer number, the base layer frame, as an inter-layer reference frame, needs to retain its temporal reference along with Coding Unit (CU) information such as motion vectors and reference frame information. These pieces of information are essential for inter-frame prediction among the enhancement layer frames. By incorporating temporal references and CU information from the base layer, efficient prediction and coding can be achieved across the frames of the enhancement layer. It is important to note that if the number of layers is defined as 2 in the sequence parameter set, the inter-layer reference frames mentioned earlier only need to be identified without explicitly specifying the layer number, which can be inferred naturally as layer 0 in this case.

5.2. Reference Frame Management

The management of reference frames for the base layer and enhancement layer is done independently, each having its own set of reference frame lists. This aspect can be defined at the layer level in the sequence parameter set. After upsampling, the reconstructed frames of the base layer are inserted into the reference frame list of the enhancement layer as inter-layer reference frames. There are no I-frames in the enhancement layer. Specifically, when the base layer is an I-frame, the enhancement layer uses P-frames and only references upsampled reference frames from the base layer. When the base layer is a P-frame, the enhancement layer can encode using both intra-layer reference frames and inter-layer reference frames derived from the lower layer. This allows for more flexibility in inter-layer prediction within the enhancement layer encoding process. In reference frame management, the enhancement layer encoder provides the position of inter-layer reference frames within the reference frame list of the enhancement layer and labels them accordingly. When the base layer is a B-frame, the enhancement layer also uses B-frames but does not utilize inter-layer prediction (no referencing of upsampled reference frames from the base layer). This behavior is similar to the case of dual-stream coding.

During the encoding process, the enhancement layer encoder first acquires the identifier of the current reference frame and determines whether it is an inherent reference frame of the enhancement layer or a frame from the base layer. It then proceeds with different processing steps based on this determination.

The base layer frame represents a lower-resolution image of the same scene as the current frame, but with some loss in video texture information. Therefore, both the content and motion information have region-to-region reference ability. This is fundamentally different from long-term reference frames, which consider certain parts of the image to have repetitive patterns for encoding gains. Long-term reference frames primarily utilize pixel information, but due to the significant variations in scenes, their motion information is not suitable for referencing in the current frame. Hence, in the SSVC scheme, the base layer frames are treated as regular short-term reference frames. Therefore, in the parts involving MVP, adaptation based on the actual Picture Order Count (POC) difference of the base layer is necessary.

5.3. Temporal Motion Vector Prediction

SHVC allows the encoder to specify collocated frames, which means that temporal motion vector prediction can potentially use inter-layer reference frames. When performing temporal motion prediction using inter-layer reference frames, the coding information of the layer containing the inter-layer reference frame is required. This information is prepared when constructing the reference frame information, so no additional operations are needed at the block level.

In SVAC 3.0, collocated frames are defined as the immediately preceding frame in the decoding order. This means that the reference frames used in TMVP are fixed and cannot be specified by the encoder. In the encoding process of the enhancement layer in SSVC, except for the first encoded frame, neither forward nor backward references can use inter-layer reference frames as the first reference frame. In other words, the reference index for inter-layer reference frames cannot be 0. This implies that, apart from the first frame in the enhancement layer, inter-layer reference frames are not utilized for Temporal Motion Vector Prediction (TMVP). Considering that two-layer coding is common in surveillance video scalable coding applications, with a base layer and an enhancement layer, when encoding the first frame of the enhancement layer, the base layer is an intra-frame, and temporal motion prediction does not come into play.

Since inter-layer reference frames are rarely involved in temporal motion prediction, and experimental results have shown that their impact on coding gains in the enhancement layer is minimal, SSVC simplifies the process by setting the temporal motion vector prediction values to 0 when the reference frame is an inter-layer reference frame. Another scenario is when the enhancement layer reference frame is not an inter-layer reference frame but an in-layer reference frame, and the temporal reference block within this reference frame refers to an inter-layer reference frame. In such cases, SSVC defines the TMVP value as 0.

5.4. Spatial Motion Vector Prediction

The operation of the SKIP/direct mode in the surveillance video standard SVAC 3.0 is different from H.266/VVC. In SVAC 3.0, it involves sequentially examining the prediction reference modes of the adjacent luminance prediction blocks F, G, C, A, B, D, Figure 1 for the current prediction unit. The goal is to obtain the first available motion information that matches the current prediction mode and use it as the current spatial prediction information. Specifically, the process begins by determining the availability of the prediction information from adjacent luminance prediction blocks. If the adjacent luminance prediction block has been encoded, is not at the image or patch boundary, and is not intra-coded, it can be considered available. Next, following the checking order mentioned earlier, the algorithm searches for the first adjacent luminance prediction block with the same prediction mode as the current one. If the current prediction mode is forward prediction, the algorithm searches for the first forward-predicted block. If it is backward prediction, it looks for the first backward-predicted block. And if it is bidirectional prediction, it locates the first bidirectional-predicted block. The spatial prediction information obtained from this search, along with the temporal prediction information, is used together for predicting the block coding information in the SKIP/direct mode.

When there is redundancy between spatial neighborhood prediction information and temporal prediction information, it can lead to a decrease in prediction accuracy. In the encoding process of the enhancement layer, this issue can be addressed by utilizing the reference information from the base layer to improve prediction accuracy. In the encoding of the enhancement layer, the utilization of base layer information is only employed when necessary. The aim is not to significantly alter the generation method of the candidate MVP list but rather to maintain consistency with single-layer encoders and reduce design complexity and cost. In the encoding process of the enhancement layer, the encoder first compares the motion vectors of the spatial prediction information with the temporal prediction information. If they are equal, it retrieves the motion information of the collocated block in the inter-layer reference frame. It then performs a scaling operation based on the first frame in the forward reference frame list, which serves as the current reference frame. This scaling operation generates a scaled MVP, which replaces the MVP of the aforementioned spatial prediction information. Additionally, the forward reference frame is set to the first frame in the forward reference frame list. If there is backward prediction, the same approach is used: utilizing the inter-layer reference frame information and performing corresponding scaling to update the backward MVP. The backward reference frame is then set to the first frame in the backward reference frame list.

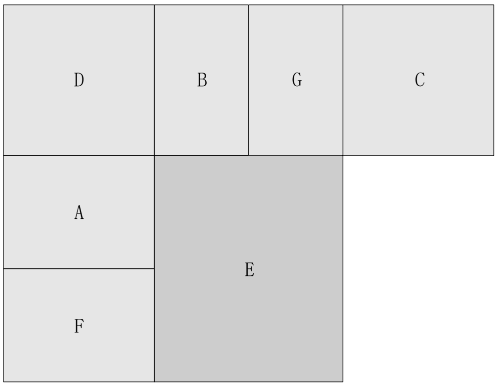

Figure 1. Spatial domain block position relationships, where E represents the current block and others represent neighboring blocks.

In the Advanced Motion Vector Prediction (AMVP) mode, MVPs are generated from the motion vectors of neighboring blocks. The generation method of MVPs depends on the availability of neighboring blocks, the availability of motion information in those blocks, and the source of motion information from neighboring blocks. In SVAC 3.0, MVPs are defined to originate from three neighboring blocks: A, B, and C. If block C is unavailable, block D is used as a substitute for MVP generation. In typical single-layer prediction, if a neighboring block is an intra block, its corresponding MVP is set to 0. If the neighboring block is an inter block, the motion vector of that block needs to be scaled based on the difference between its reference frame's Picture Order Count (POC) and the POC of the current reference frame. This scaling operation maps the motion vector of the neighboring block to the current reference frame.

In scalable coding, it is necessary to handle inter-layer reference frames differently because the POC of the base layer is the same as that of the current enhancement layer. Additionally, in terms of motion continuity, the motion vectors based on inter-layer reference images lack the same level of coherence as those based on in-layer reference frames. Therefore, special considerations are required for inter-layer reference frames. This part only involves improvements to algorithmic logic and does not introduce additional modes that would impact the Rate-Distortion Optimization (RDO) process, which is similar to H.264's Scalable Video Coding (SVC). As a result, the encoding complexity is almost unaffected and does not significantly increase.

If the reference frame of the current block in the enhancement layer is not an inter-layer reference frame, and the reference frame of the neighboring blocks being checked is also not an inter-layer reference frame, the MVP generation method remains unchanged. If the reference frame of the current block in the enhancement layer is not an inter-layer reference frame, but the reference frame of the neighboring block being checked is an inter-layer reference frame, a scaling operation is performed on the motion vector based on the inter-layer reference frame. This scaling operation takes into account the temporal information, such as POC, of both the inter-layer reference frame's reference block and the reference frame of that block. Additionally, the motion vector based on the inter-layer reference frame is added to the scaled motion vector to obtain the MVP for the current block. If the reference frame of the current block in the enhancement layer is an inter-layer reference frame, and the reference frame of the neighboring block being checked is also an inter-layer reference frame, and they both refer to the same inter-layer reference frame, then the motion vectors of the neighboring block can be directly used as the predicted MVP. If the reference frame of the current block in the enhancement layer is an inter-layer reference frame, but the reference frame of the neighboring block being checked is not an inter-layer reference frame, it indicates that the motion information of that neighboring block has no reference value for predicting the motion of the current block in the enhancement layer. In this case, the MVP is defined as 0.

5.5. Reference Frame Upsampling

Although the utilization of coding information from base layer images determines the gain in inter-layer coding performance in scalable coding, upsampling of base layer images is not mandatory. In SVC [10], the design principle of single-loop decoding allows for the decoding of the enhancement layer without fully decoding the base layer, eliminating the need for upsampling of base layer images. However, SHVC [3] reintroduces the multi-loop decoding mode, where the base layer images need to be upsampled before being used as reference frames for the enhancement layer. In H.266 [4], the spatial scalability coding mode and spatial resolution switching are integrated, and the reference frame resolutions are not necessarily the same. Therefore, inter-layer references are performed at a block level and share operations with fractional pixel interpolation. As a standard focused on video compression for surveillance applications, SVAC 3.0 does not require the support for diverse high-level syntax designs like H.266/VVC, which are intended to cater to various application scenarios. The algorithm design discussed in this paper adopts a mode that directly utilizes reference frames for inter-layer prediction and employs the upsampling scheme from SHVC. SVAC3.0 has reserved interfaces to support user-defined upsampling operations.

6. SSVC Operating Mode

The complexity of the SVC scheme, its incompatibility with single-layer encoders, and the relatively high hardware costs are among the reasons why it is difficult to promote. In the scalable coding schemes of SHVC and H.266/VVC, the base layer is treated as a regular long-term reference frame, requiring minimal modifications to the underlying logic. The benefit of this approach is a simpler algorithm architecture. However, it also results in the loss of some performance and necessary flexibility, leaving little room for improvement in encoder efficiency. SSVC continues to adhere to the principle of basic compatibility with single-layer encoders and introduces a flexible inter-layer coding scheme. The SSVC encoder can select different operational modes based on the actual application requirements and design needs. These operational modes primarily consider various requirements such as performance improvement, simplified control, and resource consumption.

6.1. Inter-Layer Reference Constrained Mode (IRCM)

SSVC not only provides scalable coding to adapt to different devices, network environments, and application needs, but also has a potential advantage of essentially eliminating the problem of high frame codewords at random access points. Random access points typically consist of intra frames, which often have codewords that are 10 times or more than those of inter frames to ensure consistent image quality. This high instantaneous bitrate during network transmission can cause significant impacts on real-time transmission capabilities. In SSVC, only the base layer is encoded as intra frames, which have a smaller image resolution. The enhancement layer does not encode any intra frames, thereby avoiding the phenomenon of bitrate overshoot. If an SSVC application does not consider inter-layer dependency and only intends to use the simulcast mode without wanting the bitrate fluctuation caused by intra frames, the IRCM of SSVC can be employed. In this mode, inter-layer reference is only allowed when the base layer is an INTRA frame. In such cases, the motion vectors are set to 0. Subsequent P frames still undergo independent encoding. The only difference from other simulcast modes lies in the coding of frames at random access points.

6.2. Inter-Layer MVP Switchable Mode (IMSM)

The available information for inter-layer reference mainly includes pixel information and motion information. The SSVC encoder can utilize a flag transmitted at the Sequence Parameter Set (SPS) level to control the usage of motion information references, such as skip/direct mode and AMVP, during the inter-frame prediction process. The enhancement layer only utilizes pixel information from the base layer during encoding. Single-layer encoders can support SSVC scalable functionality by modifying the higher-level architecture. If compression efficiency is a priority, the MVP open mode can be selected. However, if hardware reuse is more important and a partial loss in performance is acceptable, the MVP close mode can be chosen.

6.3. Forced MV Mode

During the design process of the SSVC scalable coding scheme, we have observed that when using inter-layer reference frames for inter-frame predictive coding in the enhancement layer, the effectiveness of motion prediction obtained through motion estimation may not necessarily be better than direct prediction. This observation holds true for different video scenes and various ratios of image resolutions. After analysis, this is likely due to the deviation between the cost calculation method used and the actual cost. However, due to unavoidable issues such as upsampling phase offset and uneven image encoding quality (e.g., variations in QP per region, ROI, etc.,), motion estimation remains indispensable.

Therefore, SSVC provides a forced MV zero mode based on inter-frame prediction from the base layer. In the field of video surveillance, many application scenarios are predictable, allowing the SSVC encoder to choose whether or not to utilize this mode based on actual circumstances.

6.4. Inter-layer Reference Constrained Mode (IRCM)

The Library picture tool has demonstrated excellent compression capabilities for repetitive scenes [21]. For fixed cameras, the background can be considered to have certain invariance or repetitiveness, allowing the Library picture tool to leverage its advantages. However, the process of creating a Library picture unavoidably introduces latency. Additionally, the Library pictures require high encoding quality and larger codewords, which impose significant pressure on network transmission. These drawbacks are generally intolerable for real-time video surveillance applications. As a scalable coding scheme, SSVC is more suitable for dynamic scenes and does not introduce latency. It outputs multiple layers of bitstreams, making it better suited to meet the application requirements in the field of video surveillance.

Furthermore, from an algorithmic perspective, SSVC utilizes inter-layer prediction to capture information that cannot be effectively expressed through inter-frame prediction alone. On the other hand, the Library picture tool, similar to regular long-term reference frames, primarily relies on inter-frame redundancy. It is essentially limited in its ability to handle temporal variations in scenes, as evidenced by the experimental data.

To better support the Library picture tool in SSVC and ensure ease of application design, SSVC recommends a compatibility mode where the Library picture tool can be optionally enabled in the enhancement layer. In this mode, SSVC disables inter-layer prediction, effectively degrading to simulcast mode. Alternatively, the Library picture tool can be disabled, allowing SSVC to maintain its full functionality. The choice between these modes can be made by the encoder based on the specific application scenario.

7. SSVC Coding Performance

It is essential to select representative test sequences for a comprehensive evaluation of SSVC's performance. Since the SSVC scalable coding scheme is evaluated using the reference code platform of the surveillance video compression standard SVAC 3.0, it is natural to choose the commonly used test sequences from SVAC 3.0 Table 2, which cover typical surveillance scenes.

Table 2. The characteristics of video sequence for SSVC.

Video name	Bitdepth	Resolution	fps	Description	Frames
Huochezhan	8	3840x2160	25	Fixed camera, crowd, trains moving in opposite direction	250
Lijiaoqiao	8	3840x2160	50	Fixed camera, trees，traffic on crossroads	500
Beihaihumian	8	1920x1080	50	Fixed camera, water surface with ripples	500
Qiaoxialuduan1	8	1920x1080	25	Fixed camera，pedestrian，traffic	250
Qiaoxialuduan2	8	1920x1080	25	Fixed camera，bike, traffic	250
Tingchechang	8	1920x1080	25	Cruising camera, pedestrian，traffic	250
MarketPlace	10	1920x1080	60	Vibrated camera，crowd，plants	600
NightTraffic3	10	1920x1080	50	Fixed camera，night，street light，traffic on crossroads	500
DaylightRoad	10	3840x2160	60	Car camera，zooming，building	600
Cactus	8	1920x1080	50	Fixed camera, scenery indoors, rotation, swing, character	500

Before conducting comparative analysis, it is necessary to clarify several terms. “Single bitstream” refers to the SVAC 3.0 single-layer encoder that outputs a bitstream of a single resolution. “Dual bitstream” refers to the SVAC 3.0 simulcast mode, which outputs multi-layers of bitstreams with different resolutions. These bitstreams are not entirely independent as they share the syntax of higher-level parameter sets. “SSVC bitstream” refers to the scalable coding bitstream of SSVC, which utilizes inter-layer prediction and outputs a mixed stream of base and enhancement layers. The enhancement layer in the dual bitstream only computes the codewords for the high-resolution component of the dual bitstream. In SSVC, the enhancement layer only calculates the codewords for the SSVC enhancement layer.

The experiments utilized three different encoding configurations: Low Delay P-frame (LDP) mode, Low Delay B-frame (LDB) mode, and Inter-Bi-Prediction (IBP) mode. LDP mode is characterized by low latency and utilizes single forward prediction for P-frames. LDB mode is also low latency but employs dual forward prediction for B-frames. IBP mode involves bidirectional prediction.

The experiments first compared the BDrate gains [2] (negative values indicating performance improvement, positive values indicating performance degradation) of dual bitstream and SSVC across different encoding configurations. Tables 3, 4, and 5 present the results for LDP, LDB, and IBP under four different resolution ratios. From these tables, it can be observed that as the resolution ratio between the base and enhancement layers decreases, the effectiveness of inter-layer prediction deteriorates. This is because the upsampled base layer image loses more texture details, which are insufficient to aid in prediction and coding of the enhancement layer.

Table 3. Comparison between dual bitstream and SSVC with LDP configuration.

	2:3			**1:2**			**2:5**			**1:4**		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Beihaihumian	-31.38%	-32.43%	-32.34%	-7.52%	6.30%	4.81%	-3.74%	8.87%	6.53%	-2.58%	-0.10%	-1.36%
Qiaoxialuduan1	-20.52%	-11.21%	-14.86%	-7.76%	-1.34%	1.34%	-5.25%	-0.79%	2.10%	-3.73%	-4.01%	-6.41%
Tingchechang	-31.53%	-31.04%	-30.24%	-13.16%	-9.51%	-7.86%	-7.70%	-4.69%	-3.22%	-5.35%	-5.80%	-5.86%
Cactus	-37.47%	-35.57%	-33.16%	-17.09%	-8.68%	-6.62%	-10.74%	-1.89%	-1.05%	-6.47%	-1.93%	-2.97%
Qiaoxialuduan2	-27.78%	-29.51%	-27.66%	-8.81%	-3.84%	-4.40%	-4.28%	-1.54%	-2.45%	-3.01%	-3.82%	-6.22%
MarketPlace	-37.86%	-32.91%	-31.65%	-15.93%	-7.90%	-7.82%	-9.70%	-1.08%	-1.19%	-4.38%	-1.01%	-0.81%
NightTraffic3	-23.06%	-19.77%	-21.65%	-11.51%	-6.24%	-8.16%	-9.11%	-3.13%	-5.59%	-7.57%	-4.77%	-5.83%
Huochezhan	-28.82%	-25.98%	-26.48%	-10.42%	-1.72%	-4.04%	-6.48%	-1.37%	-2.12%	-4.79%	-4.10%	-5.78%
Lijiaoqiao	-30.42%	-28.01%	-29.37%	-12.00%	-0.41%	-4.40%	-8.26%	4.47%	-0.10%	-6.12%	-1.03%	-4.43%
DaylightRoad2	-38.58%	-36.65%	-36.61%	-16.22%	-10.71%	-12.39%	-9.44%	-3.12%	-5.33%	-5.25%	-1.89%	-3.43%
Overall	-30.74%	-28.31%	-28.40%	-12.04%	-4.41%	-4.95%	-7.47%	-0.43%	-1.24%	-4.92%	-2.85%	-4.31%

Table 4. Comparison between dual bitstream and SSVC with LDB configuration.

	2:3			**1:2**			**2:5**			**1:4**		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Beihaihumian	-31.56%	-33.34%	-33.90%	-7.46%	4.64%	1.39%	-3.90%	8.95%	5.00%	-2.38%	1.03%	-1.37%
Qiaoxialuduan1	-22.98%	-14.02%	-14.67%	-7.99%	0.24%	2.61%	-5.39%	-0.88%	3.22%	-3.66%	-4.08%	-5.67%
Tingchechang	-32.44%	-31.10%	-30.60%	-13.02%	-9.04%	-8.19%	-8.36%	-5.19%	-4.46%	-5.58%	-5.52%	-6.34%
Cactus	-37.46%	-35.80%	-33.22%	-16.23%	-8.25%	-5.88%	-10.74%	-2.30%	-1.23%	-6.25%	-2.32%	-2.49%
Qiaoxialuduan2	-28.69%	-28.59%	-25.93%	-9.42%	-5.68%	-4.67%	-4.97%	-2.21%	-2.49%	-2.94%	-3.81%	-6.18%
MarketPlace	-38.12%	-33.84%	-32.57%	-15.02%	-7.53%	-7.05%	-9.82%	-1.64%	-1.80%	-4.30%	-1.74%	-2.20%
NightTraffic3	-23.11%	-20.33%	-21.91%	-10.54%	-5.28%	-7.29%	-8.83%	-3.63%	-5.87%	-7.25%	-4.49%	-5.78%
Huochezhan	-29.27%	-24.75%	-26.31%	-10.72%	0.13%	-1.79%	-6.57%	0.83%	-1.36%	-4.61%	-4.58%	-6.17%
Lijiaoqiao	-30.61%	-26.93%	-28.30%	-12.31%	2.02%	-2.41%	-8.27%	5.51%	0.85%	-5.80%	-2.78%	-4.79%
DaylightRoad2	-38.43%	-36.64%	-36.37%	-16.55%	-11.48%	-12.63%	-9.76%	-3.14%	-5.31%	-5.23%	-3.37%	-3.62%
Overall	-31.27%	-28.53%	-28.38%	-11.93%	-4.02%	-4.59%	-7.66%	-0.37%	-1.34%	-4.80%	-3.17%	-4.46%

Table 5. Comparison between dual bitstream and SSVC with IBP configuration.

	2:3			**1:2**			**2:5**			**1:4**		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Beihaihumian	-27.46%	-28.42%	-29.71%	-8.28%	8.63%	5.91%	-4.05%	11.85%	9.66%	-2.50%	1.30%	-1.95%
Qiaoxialuduan1	-19.90%	-11.19%	-15.06%	-8.67%	0.24%	3.68%	-6.04%	-0.06%	3.98%	-4.47%	-4.99%	-7.42%
Tingchechang	-25.18%	-24.10%	-24.09%	-12.53%	-7.75%	-6.76%	-8.52%	-4.12%	-3.75%	-6.19%	-5.71%	-7.70%
Cactus	-30.00%	-28.01%	-24.97%	-16.05%	-7.44%	-5.01%	-11.26%	-2.04%	-0.04%	-6.96%	-3.32%	-3.45%
Qiaoxialuduan2	-22.30%	-22.45%	-20.87%	-10.17%	-5.15%	-5.02%	-5.79%	-0.43%	-2.87%	-3.76%	-3.70%	-6.97%
MarketPlace	-30.98%	-26.74%	-25.19%	-14.24%	-4.54%	-4.69%	-9.71%	0.12%	-0.07%	-4.46%	-0.56%	-1.69%
NightTraffic3	-23.02%	-19.82%	-21.10%	-11.33%	-5.26%	-7.40%	-9.31%	-3.22%	-6.04%	-7.75%	-4.89%	-6.35%
Huochezhan	-20.62%	-16.93%	-17.32%	-9.41%	0.54%	-1.11%	-6.91%	0.76%	-1.33%	-5.50%	-4.73%	-6.84%
Lijiaoqiao	-22.85%	-20.38%	-21.93%	-11.38%	2.01%	-2.64%	-8.60%	5.86%	0.70%	-6.82%	-2.58%	-4.79%
DaylightRoad2	-30.02%	-27.97%	-28.05%	-14.32%	-8.10%	-9.90%	-9.21%	-1.80%	-4.49%	-5.65%	-3.48%	-4.06%
Overall	-25.23%	-22.60%	-22.83%	-11.64%	-2.68%	-3.29%	-7.94%	0.69%	-0.43%	-5.41%	-3.27%	-5.12%

SSVC and the dual bitstream mode share the same base layer, with the difference lying in the enhancement layer. In SSVC, the enhancement layer utilizes inter-layer reference to reduce the bitrate, while in the dual bitstream mode, the enhancement layer is independently encoded. Table 6 provides a performance comparison between the SSVC enhancement layer and the single bitstream. It can be observed that as the resolution of the base layer decreases, the bitrate contribution of the base layer decreases as well. Consequently, the performance gain of the SSVC enhancement layer gradually approaches the gain achieved by SSVC compared to the dual bitstream. At a resolution ratio of 2:3, where the base layer has a larger proportion, the performance gain of the enhancement layer is significantly better compared to Table 3. However, at a ratio of 1:4, where the base layer has a smaller proportion, the performance gain is comparable to that shown in Table 3.

Table 6. Comparison between SSVC enhanced layer and single bitstream with LDP configuration.

	2:3			1:2			2:5			1:4		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Beihaihumian	-61.90%	-63.10%	-62.91%	-9.66%	4.15%	2.49%	-3.65%	9.02%	6.12%	-2.39%	0.55%	-1.02%
Qiaoxialuduan1	-36.91%	-27.92%	-31.85%	-9.63%	-2.89%	-0.94%	-5.54%	-0.75%	2.92%	-3.75%	-3.98%	-6.01%
Tingchechang	-60.61%	-59.26%	-59.06%	-19.99%	-16.08%	-14.48%	-9.10%	-5.96%	-4.66%	-5.28%	-5.81%	-6.08%
Cactus	-72.29%	-71.19%	-69.76%	-25.93%	-18.37%	-16.45%	-12.91%	-4.44%	-3.11%	-6.34%	-2.13%	-3.00%
Qiaoxialuduan2	-54.16%	-54.42%	-53.04%	-12.37%	-7.75%	-8.18%	-4.68%	-1.70%	-1.94%	-2.98%	-4.19%	-6.19%
MarketPlace	-75.67%	-73.56%	-73.04%	-24.60%	-17.47%	-17.15%	-12.14%	-3.75%	-4.19%	-4.38%	-0.42%	-1.03%
NightTraffic3	-41.57%	-37.83%	-39.54%	-14.88%	-9.41%	-11.62%	-9.59%	-3.57%	-6.16%	-7.35%	-4.53%	-5.86%
Huochezhan	-57.04%	-54.89%	-55.40%	-14.97%	-6.50%	-8.23%	-7.25%	-1.78%	-2.83%	-4.69%	-3.93%	-5.45%
lijiaoqiao	-59.57%	-58.31%	-58.99%	-16.63%	-5.23%	-9.36%	-9.27%	4.23%	-0.92%	-6.16%	-0.69%	-4.47%
DaylightRoad2	-75.31%	-73.86%	-73.81%	-25.90%	-21.10%	-22.49%	-11.89%	-5.96%	-7.51%	-4.98%	-1.65%	-2.99%
Overall	-59.50%	-57.44%	-57.74%	-17.46%	-10.07%	-10.64%	-8.60%	-1.47%	-2.23%	-4.83%	-2.68%	-4.21%

SSVC supports flexible encoding settings, allowing the encoder to choose the trade-offs for certain tools based on the requirements of the application. This is discussed in section 6. Specifically, experiments were conducted regarding the Inter-layer MVP switchable mode IMSM. Table 7 presents the experimental data for the LDP configuration. It can be observed that IMSM has clear gains at commonly used layer ratios such as 1:2 and 2:5. However, the performance gain decreases when the ratio becomes too large or too small. This is because, in cases of larger ratios, pixel prediction provides the main gain, and inter-layer motion vectors are not significantly better than in-layer motion vectors. Similarly, when the ratio is smaller, the reference value of inter-layer motion vectors becomes less significant.

Table 7. Experiment results of turning off and on the IMSM.

	2:3			1:2			2:5			1:4		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Beihaihumian	-0.11%	-0.77%	-0.21%	-0.46%	-0.05%	0.30%	-0.21%	1.08%	-0.49%	0.06%	-1.55%	-1.99%
Qiaoxialuduan1	-0.13%	-0.02%	-0.56%	-0.18%	-0.25%	0.61%	-0.08%	0.26%	-0.03%	0.00%	0.01%	0.25%
Tingchechang	-0.23%	-0.51%	-0.30%	-0.71%	-0.41%	0.41%	-0.59%	-0.40%	-0.02%	0.01%	-0.46%	-0.08%
Cactus	-0.03%	0.04%	-0.13%	-0.31%	0.22%	-0.20%	-0.39%	-0.17%	-0.18%	-0.15%	0.13%	-0.04%
Qiaoxialuduan2	-0.31%	-0.27%	-0.49%	-0.40%	0.71%	-0.46%	-0.26%	0.92%	0.56%	-0.04%	-0.33%	-0.66%
MarketPlace	0.41%	-0.04%	0.08%	-0.80%	-0.03%	-1.38%	-0.81%	-0.33%	0.34%	-0.26%	0.88%	0.39%
NightTraffic3	0.04%	-0.22%	0.03%	-0.42%	-0.38%	-0.53%	-0.12%	-0.43%	-0.39%	0.02%	0.13%	-0.19%
Huochezhan	-0.27%	-0.28%	0.04%	-0.44%	-0.48%	-0.09%	-0.26%	0.08%	1.49%	-0.03%	0.16%	0.43%
Lijiaoqiao	-0.17%	-0.36%	-0.45%	-0.31%	0.39%	-0.09%	-0.31%	0.50%	1.36%	-0.07%	-0.95%	0.40%
DaylightRoad2	0.06%	-0.23%	-0.09%	-0.76%	-0.50%	-0.35%	-0.84%	-0.51%	-0.48%	-0.11%	1.52%	-0.87%
Overall	-0.07%	-0.26%	-0.21%	-0.48%	-0.08%	-0.18%	-0.39%	0.10%	0.22%	-0.06%	-0.05%	-0.24%

Table 8 presents the effect of the Inter-layer reference Constrained Mode (IRCM) in the LDP configuration. When this mode is enabled, only the images at randomly accessed points utilize inter-layer prediction. This mode is designed to mitigate the impact of intra-frame coding on the bitstream, and as a result, subsequent P-frames do not employ inter-layer prediction. Therefore, the table only includes statistics for the data at randomly accessed points. In this scenario, both layers of the dual bitstream encode I-frames, while the SSVC enhancement layer encodes P-frames. It can be observed that the individual gain for I-frames is higher than the average gain, especially for cases with larger inter-layer ratios. This effectively suppresses the occurrence of bitrate overshoot. Similar experimental results were observed in LDB and IBP configurations.

Table 9 presents the results of the Forced MV mode. According to the initial assumption, inter-layer motion estimation is unnecessary because the base layer and enhancement layer have almost corresponding texture information. There is no logical reason to not select reference blocks in the same positions. Therefore, disabling inter-layer motion search was considered a practical option, and this was the motivation behind offering this option in SSVC. However, from Table 9, it can be seen that the reality is different from the assumption. After performing motion estimation, some sequences indeed show performance degradation, which conforms with the initial assumption. However, there are also sequences that demonstrate gains, such as NightTraffic3 with a gain of 0.22%. Therefore, SSVC supports the switch mode for inter-layer motion search to accommodate various needs. Ultimately, in hardware design, the motion estimation module is fully reusable.

Table 8. Experiment results of turning off and on the IRCM.

	2:3			**1:2**			**2:5**			**1:4**		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Beihaihumian	-40.75%	-30.44%	-35.19%	-26.64%	4.24%	-6.65%	-17.30%	20.65%	7.44%	-6.08%	13.63%	9.67%
Qiaoxialuduan1	-37.75%	-31.74%	-38.04%	-24.12%	-10.33%	-16.18%	-14.10%	1.79%	-0.82%	-3.77%	3.47%	2.48%
Tingchechang	-38.99%	-39.35%	-38.34%	-24.56%	-20.54%	-18.18%	-14.96%	-9.59%	-7.06%	-5.07%	-2.06%	-1.01%
Cactus	-40.76%	-38.87%	-35.94%	-26.51%	-16.62%	-14.13%	-17.29%	-4.90%	-4.04%	-6.51%	1.66%	0.38%
Qiaoxialuduan2	-38.02%	-34.23%	-37.97%	-24.15%	-12.37%	-18.76%	-14.60%	-0.32%	-8.78%	-4.67%	5.15%	-0.71%
MarketPlace	-45.62%	-44.18%	-42.60%	-33.79%	-25.10%	-24.76%	-24.71%	-10.18%	-10.88%	-10.20%	2.84%	2.44%
NightTraffic3	-39.21%	-39.52%	-39.92%	-25.13%	-20.23%	-21.73%	-15.63%	-8.23%	-10.74%	-5.59%	-0.62%	-2.18%
Huochezhan	-40.65%	-42.33%	-44.49%	-25.17%	-18.24%	-22.05%	-15.39%	-5.77%	-9.65%	-5.57%	2.22%	-1.68%
Lijiaoqiao	-43.59%	-43.74%	-48.82%	-30.01%	-18.99%	-28.37%	-19.99%	0.46%	-13.47%	-7.29%	9.52%	-1.23%
DaylightRoad2	-44.33%	-45.88%	-46.52%	-30.79%	-25.47%	-27.83%	-20.59%	-11.44%	-15.80%	-8.01%	0.99%	-2.76%
Overall	-40.97%	-39.03%	-40.78%	-27.09%	-16.36%	-19.86%	-17.46%	-2.75%	-7.38%	-6.28%	3.68%	0.54%

Table 9. Performance variation between turning off and on the forced MV mode with LDP configuration, where the resolution ratio between the base layer and enhancement layer is 1:2.

	Y	**U**	**V**
Beihaihumian	0.09%	-0.21%	-0.59%
Qiaoxialuduan1	0.06%	0.02%	-0.01%
Tingchechang	0.00%	-0.25%	-0.55%
Cactus	-0.21%	-0.36%	-0.32%
Qiaoxialuduan2	-0.16%	0.12%	-0.48%
MarketPlace	-0.18%	-0.50%	-0.17%
NightTraffic3	0.22%	0.26%	-0.03%
Huochezhan	-0.14%	-0.14%	-0.49%
Lijiaoqiao	-0.23%	-0.34%	-0.43%
DaylightRoad2	-0.22%	-0.03%	-0.44%
Overall	-0.08%	-0.14%	-0.35%

Table 10 provides the gains achieved by encoding the library pictures in SVAC 3.0. It can be observed that the library pictures exhibit a certain dependency on the image content. For periodic images like "cactus" or images with a stable background like "huochenzhan," there is a significant performance improvement, which is highly attractive. However, for scenes with camera shake such as "MarketPlace" or car-mounted camera footage like "DaylightRoad2," the performance is poor, which is disproportionate to the coding resources allocated. In such cases, it becomes challenging to achieve bitrate reduction significantly. However, referring to Table 3, these scenes with camera movements are precisely where SSVC excels. Therefore, SSVC provides a compatible mode for library pictures during scalable coding, effectively alleviating the burden on the encoder. This allows for a more flexible balance between encoding gains and costs, achieving a trade-off that suits the specific requirements.

Table 10. Performance of enabling library pictures based on SVAC3, which can show the difference of preference of video content from SSVC that is indicated in Table 3.

	Y	**U**	**V**
Beihaihumian	-5.23%	-13.38%	-19.19%
Qiaoxialuduan1	-18.65%	-24.51%	-40.62%
Tingchechang	-8.70%	-12.68%	-12.81%
Cactus	-23.35%	-24.35%	-26.82%
Qiaoxialuduan2	-15.42%	-26.31%	-31.18%
MarketPlace	-1.08%	-2.71%	-2.11%
NightTraffic3	-23.39%	-25.19%	-21.71%
huochezhan	-27.33%	-38.00%	-38.28%
Lijiaoqiao	-24.93%	-41.86%	-38.17%
DaylightRoad2	-1.89%	-4.67%	-3.17%

8. Conclusions

The achievement of video coding scalability has long been a desire of video professionals. However, due to practical limitations, there are relatively few real-world examples of scalable coding. With the continuous expansion of short videos and video-on-demand markets on the internet, scalable coding applications based on paid services have gained widespread popularity. Additionally, in the field of video surveillance, there is a significant concern regarding the cost of hard disk storage. While cloud storage can be utilized to reduce costs, there are often requirements for video browsing that must be combined with permissions and privacy considerations. Therefore, layered scalable coding meets practical demands in such scenarios.

This paper proposes a flexible and configurable spatial scalable video coding scheme called SSVC, based on the video surveillance standard SVAC 3.0. It aims to improve encoding efficiency while considering practical application convenience. SSVC is not a fixed and unchangeable solution. Its flexibility lies not only in the encoder's ability to configure different tools according to specific needs but also in the provision of extensibility for future developments.

Taking inspiration from intelligent signal processing methods, the next step in SSVC would involve utilizing the restoration of enhancement layer texture using the base layer image as a reference, or extraction of the spatio-temporal features to predict the future frame [9]. As video standards become increasingly complex, with a growing number of encoding tools available, it becomes important to examine which encoding tools' performance overlaps with SSVC in the context of scalable coding. Another aspect that SSVC needs to consider is selectively disabling these overlapping tools to optimize its performance. It is worth noting that the switches for most tools can be configured in the high-level syntax of the encoder, without affecting the consistency of the standard bitstream. Therefore, the configuration of these switches does not bring about any issues related to the consistency of the standard bitstream.

Acknowledgment

This study has been supported by the academician expert Open Fund of Beijing Smart-chip Microelectronics Technology Co., Ltd. SGITZXDTNXQT2201061

References

[1] Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC, 2007.

[2] Bjontegaard G., "Calculation of Average PSNR Differences between Rd-Curves," *in Proceedings of the 13thVCEG-M33, VCEG Meeting,* Austin, 2001.

[3] Boyce J., Ye Y., Chen J., and Ramasubramonian A., "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20-34, 2016. DOI:10.1109/TCSVT.2015.2461951

[4] Bross B., Wang Y., Ye Y., Liu S., Chen J., Sullivan G., and Ohm J., "Overview of the Versatile Video Coding (VVC) Standard and its Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736-3764, 2021. DOI:10.1109/TCSVT.2021.3101953

[5] Chen Y., Murherjee D., Han J., Grange A., Xu Y., and Liu Z., et al., "An Overview of Core Coding Tools in the AV1 Video Codec," *in Proceedings of the Picture Coding Symposium (PCS),* San Francisco, 2008. DOI:10.1109/PCS.2018.8456249

[6] Coding of Audio-Visual Objects-Part 2: Visual, ISO/IEC 14492-2 (MPEG-4 Visual), ISO/IEC JTC 1, Version 3: 2004.

[7] High Efficiency Video Coding, Document Rec. ITU-T H.265 and ISO/IEC 23008-2, 2014.

[8] Lee H., Kang J., Lee J., Choi J., Kim J., and Sim D., "Scalable Extension of HEVC for Flexible High-Quality Digital Video Content Services," *ETRI Journal*, vol. 35, no. 6, pp. 990-1000, 2013. https://doi.org/10.4218/etrij.13.2013.0040

[9] Mokhtar Z. and Dawwd S., "3D VAE Video Prediction Model with Kullback Leibler Loss Enhancement," *The International Arab Journal of Information Technology*, vol. 21, no. 5, pp. 879-888, 2024. https://doi.org/10.34028/iajit/21/5/9

[10] Schwarz H., Marpe D., and Wiegand T., "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, 2007. DOI:10.1109/TCSVT.2007.905532

[11] Sjoberg R., Chen Y., Fujibayashi A., Hannuksela M., Samuelsson J., Tan T., Wang Y., and Wenger S., "Overview of HEVC High-Level Syntax and Reference Picture Management," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1858-1870, 2012. DOI:10.1109/TCSVT.2012.2223052

[12] Sullivan G., Ohm J., Han W., and Wiegand T., "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, 2012. DOI:10.1109/TCSVT.2012.2221191

[13] Technical Specifications for Surveillance Video and Audio Coding (SVAC), Chinese National Standard: GB/T 25724.

[14] Versatile Video Coding, Standard ITU-T H.266, ISO/IEC 23090-3, 2020.

[15] Video Coding for Low Bit Rate Communication, ITU-T, ITU-T Recommendation H.263 Version 2, 2005.

[16] Wang Y., Skupin R., Hannuksela M., Deshpande S., Hendry., Drugeon V., Sjoberg R., Choi B., Seregin V., Sanchez Y., Boyce J., Wan W., and Sullivan J., "The High-Level Syntax of the Versatile Video Coding (VVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3779-3800, 2021. DOI: 10.1109/TCSVT.2021.3070860

[17] Wu F., Li S., and Zhang Y., "A Framework for Efficient Progressive Fine Granularity Scalable Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 11, no. 3, pp. 332-344, 2001. DOI: 10.1109/76.911159

[18] Ye Y., He Y., Wang Y., and Hendry., "SHVC, the Scalable Extensions of HEVC, and its Applications," *Zte Communications*, vol. 14, no. 1, pp. 2016.

[19] Zhang J., Jia C., Lei M., Wang S., Ma S., and Gao W., "Recent Development of AVS Video Coding," *in Proceedings of the Picture Coding Symposium*, Ningbo, 2019. DOI: 10.1109/PCS48520.2019.8954503

[20] Zhang X., Huang T., Tian Y., and Gao W., "Background-Modeling-Based Adaptive Prediction for Surveillance Video Coding," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 769-784, 2014. DOI: 10.1109/TIP.2013.2294549

[21] Zuo X., Yu L., Yu H., Mao J., and Zhao Y., "Scene-Library-Based Video Coding Scheme Exploiting Long-Term Temporal Correlation," *Journal of Electronic Imaging,* vol. 26, no. 4, pp. 043026, 2017.

Zhe Zheng is employed by Beijing Smart-chip Microelectronics Technology Co., Ltd, AI Department. He graduated from the University of Southern California in the United States in 2009, and specializes in the research of chip technology in the fields of Artificial Intelligence and Communication.

Jinghua Liu works for the equipment management Department of State Grid Corporation of China, mainly engaged in research on new technologies in the field of Transmission.

Darui Sun (memeber, IEEE) received the Ph.D. degree from Southeast University, Nanjing, China, in 2003. He is working at Beijing Vimicro AI Chip Technology Corporation.as an advanced video engineer.

Jinghui Lu received the B.E and ME degrees in electronic engineering from Tsinghua University, China, in 1997 and 2000, respectively. He is working at Beijing Vimicro AI Chip Technology Corporation as Algorithm director, engaged in research on Digital Signal Processing and video Codec Related fields.

Song Qiu received the M.E Degree in Electronic Engineering from Tsinghua University, China, in 1999. He is working at Beijing Vimicro AI Chip Technology Corporation, and is the vice chairman of Beijing Security Video and Audio Codec Technology Industry Alliance, and a member of the National Security Alarm system standardization Technical Committee SAC/TC100. He mainly engaged in the Key Technology Research and Product Development of Intelligent Internet of Things Data Processing.

Yanwei Xiong is employed by Beijing Smart-chip Microelectronics Technology Co., Ltd, AI Department, mainly engaged in Artificial Intelligence Chip Technology research.

Rui Liu is employed by Beijing Smart-chip Microelectronics Technology Co., Ltd, AI Department, mainly engaged in research on Artificial Intelligence Technology and Battery Sampling Technology.

Wenpeng Cui is employed by Beijing Smart-chip Microelectronics Technology Co., Ltd, AI Department, mainly engaged in the research of Chip Technology in the fields of Artificial Intelligence, Communication and other fields.