Enhanced Surveillance Video Compression With Dual Reference Frames Generation

Lei Zhao, Shiqi Wang, *Member, IEEE*, Shanshe Wang, *Member, IEEE*, Yan Ye, *Senior Member, IEEE*, Siwei Ma, *Senior Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*— In this paper, we improve the inter coding performance of surveillance videos by simultaneously investigating the distinct characteristics of background and foreground redundancy, and introduce two novel reference frames in a complementary manner. On one hand, a block level background reference frame (BRF) is proposed to reduce the background redundancy. The proposed scheme incorporates semantic information into the compression process, and makes use of instance segmentation to facilitate the background block decision, making the generated BRF free from foreground pollution. On the other hand, in order to handle foreground redundancy, a foreground reference frame (FRF) is generated based on Surveillance Prediction Generative Adversarial Network (SP-GAN), which utilizes previous reconstructed frames, optical flow based prediction, as well as BRF to infer the foreground objects of the to-be-coded frame. We integrate the proposed scheme into HM-16.6 software and append BRF and FRF to the reference pictures list (RPS). Simulation results demonstrate considerable superiority of the proposed scheme. In particular, by adding the proposed BRF to RPS, 3% coding gains are observed compared with the state-of-the-art BRF method. When both BRF and FRF are incorporated into RPS, 5.8% gains are achieved for surveillance video coding.**

***Index Terms*— Surveillance video, video compression, background reference frame, foreground reference frame.**

I. Introduction

THE dramatic increasing of surveillance video big data poses critical challenges to the surveillance system especially in terms of efficient storage and transmission. The IDC report [1] claims that surveillance data are expected to make up a significant share of the big data (more than 40%) in 2020, which puts forward an urgent demand for efficient compression methodologies towards video surveillance.

Inter coding aims to remove the temporal redundancy between adjacent frames. In particular, the specified blocks in the reference frames are employed as predictions, thus only residual signals need to be transmitted in the bitstream. Inter coding uses reference frames to predict the current to-be-coded frame. Employing already-coded and reconstructed pictures as the reference is the most classical way in the literature. Motivated by the high correlations between the adjacent frames, multiple already decoded frames are assigned as reference to the to-be-coded frame, which are able to produce high similarity predictions by multi-hypothesis. Although reconstructed frames have been widely employed for promising performance and ease of implementation, the limited diversity within them spurs researchers to explore other types of reference frames, such as the synthesized reference frame (SRF). In particular, Zhao *et al.* [2] introduced a virtual reference frame (VRF) by making use of frame interpolation. As the proposed VRF is registered with the to-be-coded frame, motion data can be implicitly conveyed by the VRF, and hence lightens the bitstream burden. In [4], a clip level reference buffer organizing scheme was investigated, which manages the reference in a more economical manner. Apart from these generic approaches, there are also some SRFs designed towards specific scenarios. In [5]–[11], background reference frame (BRF) was investigated to facilitate surveillance video coding. Besides, external information from Google Earth was utilized to aid satellite video coding in [12], where Google Earth pictures serve for predictions of I frames.

The High Efficiency Video Coding (HEVC) standard [13], [14] achieves remarkable compression performance improvement compared with previous ones. However, HEVC and other popular video coding standards target at dealing with generic video content, whereas the distinct characteristics in surveillance video, e.g. strong background redundancy and relative sparse foreground objects, have not been fully exploited. It motivates researchers to explore specific coding algorithms for such videos. Generally speaking, existing surveillance video coding algorithms can be divided into two categories: background based approaches [5]–[11] and foreground based approaches [16]–[18]. In particular, background based coding methods aim to remove the background redundancy in the surveillance video, which is typically realized by introducing the BRF. Zhang *et al.* [7], [8] proposed a segment-and-weight

Manuscript received October 26, 2020; revised February 9, 2021; accepted March 25, 2021. Date of publication April 14, 2021; date of current version March 9, 2022. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010133001, in part by PKU-Alibaba Fund, and in part by the High-Performance Computing Platform of Peking University. This article was recommended by Associate Editor V. Stankovic. *(Corresponding author: Shanshe Wang.)*

Lei Zhao, Shanshe Wang, Siwei Ma, and Wen Gao are with the School of Electronic Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing 100871, China, and also with the National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China (e-mail: lei_zhao@pku.edu.cn; sswang@pku.edu.cn; swma@pku.edu.cn; wgao@pku.edu.cn).

Shiqi Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: shiqwang@cityu.edu.hk).

Yan Ye is with Alibaba Group U.S., Sunnyvale, CA 94085 USA (e-mail: yan.ye@alibaba-inc.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3073114.

Digital Object Identifier 10.1109/TCSVT.2021.3073114

1051-8215 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

based running average method, which utilizes original frames to obtain the BRF. Besides, considering that BRF is normally compressed in high quality that leads to bandwidth fluctuation, Chen *et al.* [11] proposed to generate BRF in block level, achieving better coding performance while avoiding the bit-rate burst. Although the aforementioned works have made great progress in surveillance coding, there is still a major drawback to be addressed. Specifically, existing BRF methods are typically based on sparse foreground hypothesis, and modeling the BRF relies on first identifying the potential foreground pixels. This strategy, however, may not be applicable to those videos with dense or low-speed foreground objects.

In contrast with background based coding, foreground coding methods exploit global object redundancy and optimize foreground coding performance. In essence, these approaches leverage internal correlations and prior knowledge to efficiently compress the foreground objects. In [16], an object based surveillance coding framework was proposed, which online learns the appearance model of each object by using incremental principal component analysis (IPCA), and represents the objects through the coefficients of the principal components. However, for the objects with dramatic appearance variations, this scheme lacks efficiency since it is difficult to accurately characterize the following status using only the previous observations. Moreover, in order to overcome the limitation of short term references, Ma *et al.* [17] proposed to build a vehicle library with high quality copies of the similar vehicles. During encoding, similar vehicles are retrieved from this library and serve as additional reference. Similar work can also be found in [18], where a knowledge based coding scheme is presented by modeling the vehicles as the constant part and variational part.

Recently, the rapid development of deep learning methodology has brought many computer vision tasks to higher levels, which also provides a new horizon for video coding optimization. Previous works have verified the effectiveness of employing deep learning based computer vision applications for higher compression efficiency [2], [19]–[25]. In this paper, we take advantage of the recent advances in deep learning and computer vision to improve the inter coding performance of surveillance videos. Concretely, we present two complementary reference frames to respectively deal with the background and foreground redundancy. The novelty and contributions of this paper are as follows:

- We present a hybrid surveillance processing scheme towards efficient compression and intelligent analysis. The extracted semantic information not only can be utilized in analysis tasks, but also contributes to the efficient compression by facilitating the BRF generation.
- We propose to remove background redundancy by introducing the high quality BRF based on instance segmentation. Different from existing approaches that exploit only the statistical and motion information, we improve the quality of BRF from the semantic perspective, and adopt instance segmentation to avoid foreground pollution. As such, our method is more robust in complex scenes with dense or low-speed foreground objects.
- We propose to remove foreground redundancy based on Generative Adversarial Network (GAN). In particular, a GAN based surveillance video future frame prediction architecture, termed as Surveillance Prediction GAN (SP-GAN), is put forward to generate a foreground reference frame (FRF), which is able to infer the foreground objects of the to-be-coded frame.

II. Related Works

The related works are reviewed in this Section. Firstly, existing works on synthesized reference frame (SRF) generation are introduced. Subsequently, the recent developments on deep learning based future frame prediction are reviewed.

A. SRF Generation

Directly employing reconstructed frames as reference does not fully exploit the temporal and spatial characteristics, leading to the inefficiency in complex scenes. To tackle this problem, many researchers further investigate the underlying knowledge among reconstructed frames, and synthesize additional high performance reference frames by inferring the potential status of the to-be-coded frame.

Recent works [2]–[4] have demonstrated the effectiveness of employing SRF for generic videos. Inspired by the intuitive relationship between frame interpolation and hierarchical B coding structure, Zhao *et al.* [2] presented a SRF based on deep neural network. Since the proposed SRF and the to-be-coded picture are registered, motion data are no longer needed to be signalled and thus brings considerable bitrate savings. Similar work can also be found in [3], where a SRF is generated for quality scalable high efficiency video coding (SHVC). Besides the picture level SRFs, there are also strategies that organize the SRF in clip level. In particularly, Ma *et al.* [4] divide the reconstructed content into singular clips and repetitive clips, and place only the singular clips into reference buffer to maintain the content diversity.

In order to remove background redundancy, many pioneer works have explored the BRF in surveillance coding. In [5], a BRF is generated based on Gaussian Mixture Model (GMM) [15] which outperforms the reconstructed frames in terms of rate-distortion (RD) performance on H.264/AVC. Though significant coding performance improvement is achieved by GMM, the huge memory cost and large number of floating operations make it barely acceptable for surveillance system. In view of this, Zhang *et al.* [7] investigated a compression friendly modeling method by calculating the weighted average. Based on this method, an adaptive prediction scheme is further developed, which adopts different coding strategies for background blocks and foreground blocks respectively [8]. Recently, Chen *et al.* [11] improved the BRF by developing a block level building approach. In contrast to [7], [8] that maintain individual background models for each pixel, Chen *et al.* explicitly use the reconstructed content as the background, achieving higher efficiency in terms of both RD performance and computational complexity. Although numerous explorations have been conducted, the quality of the

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

(a) Zhang *et al*.'s method [7][8] (b) Chen *et al*.'s method [11] (c) The proposed method

Fig. 1. The BRFs generated by (a) Zhang *et al*'s method, (b) Chen *et al*.'s method and the proposed method, respectively. As can be observed, (a) and (b) suffer from severe foreground pollution when the objects are in slow motion.

generated BRF is still far from perfect especially in complicated scenes with dense or low-speed objects. Fig. 1 shows the BRFs generated by the methods in [7], [8] and [11] respectively. As can be observed, both methods suffer from severe foreground pollution owing to the misclassification of the background pixels. In view of this, this paper improves the BRF by incorporating semantic cues, and takes advantage of instance segmentation to exclude foreground pixels. The details of our BRF generation scheme will be presented in Section IV.

B. Future Frame Prediction

Predicting future frames is a challenging task due to the rich diversity of the potential variations over time. Recently, the promising generation ability of deep learning has inspired researchers to resolve this problem by resorting to the neural networks. In general, existing works on future frame prediction can be roughly divided into two categories: directly synthesizing future frames or deducing the future pixels with motion cues.

Earlier approaches intend to predict pixel values without explicitly involving the motion information. In [26], a multi-layer long short-term memory (LSTM) Encoder-Decoder framework was designed to learn a fixed length representation, which is then decoded using single or multiple decoder LSTMs to perform future frame prediction. In order to deal with the blurry predictions obtained from the Mean Squared Error (MSE) loss function, Mathieu *et al.* [27] proposed a new loss function based on image gradients, and trained the network with generative adversarial strategy. In [28], a predictive coding inspired RNN architecture with both bottom-up and top-down connections is developed to continually predict the appearance of future video frames. Besides, Amersfoort *et al.* [29] regard future frame prediction as the transform process, and learn to predict local affine transforms to the future frame based on the previous observations. Nevertheless, these methods lead to blurry predictions owing to the intrinsic ambiguity of the motion itself, which motivates researchers to leverage motion cues for sharper predictions.

Many improved works explicitly model the motion trajectory to facilitate frame prediction [30]–[35]. Liu *et al.* [30] predicted video frames by flowing pixel values from existing ones, leading to sharp and realistic results. In [31], a hybrid multi-tasking architecture is devised to simultaneously optimize optical flow estimation and future frame prediction. Similar work is also presented in [32], where a dual motion GAN is developed to enforce the generated predictions to be consistent with the predict motion flows. In [33], Lu *et al.* developed a flexible spatio-temporal network (FSTN) based on convolutional LSTM autoencoder framework, which incorporates an optical flow module to capture temporal changes. Besides, Liu *et al.* [34] tackled the anomaly detection problem within a future frame prediction framework, and introduce the optical flow loss to constrain the predictions. In [35], a shallow network with auto-encoder architecture is presented, which deduces the optical flow to the next frame on the basis of the previous optical flow frames.

In this paper, we propose to generate a FRF based on future frame prediction, which aims to provide high quality reference for foreground objects in surveillance videos. In contrast to existing methods that mainly focus on the prediction quality of moving foregrounds, we incorporate background information into the generation process to make the produced FRF more suitable for surveillance video coding task.

III. Hybrid Surveillance Processing Scheme

Efficient video compression and intelligent analysis are two critical parts concerning the advanced surveillance system. Traditional surveillance systems regard them as individual tasks and deal each problem with independent solutions. In this paper, we propose to handle these two tasks in a hybrid way, in which the compression efficiency is improved by taking advantage of the interactions between each other.

The proposed hybrid surveillance processing scheme is depicted in Fig. 2, which can be separated to two flows including intelligent analysis flow and video compression flow. Specifically, intelligent analysis comprises of semantic information extraction module and intelligent analysis module. The former module extracts important semantic information such as foreground objects mask in the local server, which is then fed to analysis module to perform analytical tasks. The advantages of the proposed intelligent analysis flow are twofold. First, the analysis module is deployed in the local server such that the delay caused by compression and transmission can be eliminated. Secondly, the analysis tasks are performed based upon original frames, hence avoiding the negative effect brought by compression artifacts. Regarding

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

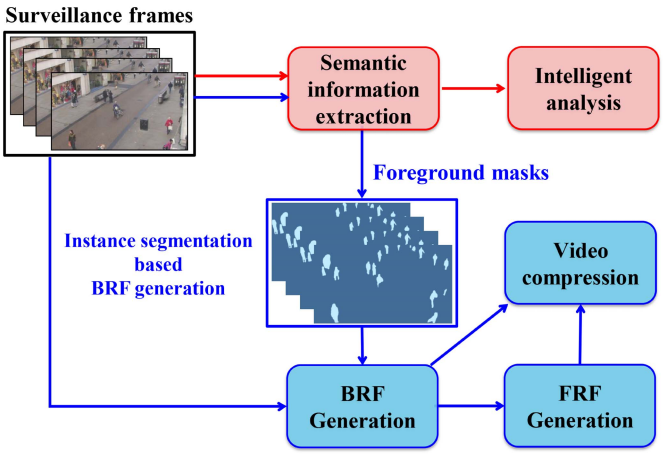


Fig. 2. Hybrid surveillance processing framework towards video compression and intelligent analysis.

the video compression, two complementary reference frames are introduced to comprehensively remove the redundancy in surveillance videos. In particular, an instance segmentation based BRF is obtained using the original video frames and foreground masks produced in the semantic extraction module. Based on the BRF, a FRF is subsequently generated, which is added to the reference pictures list (RPS) together with the BRF. Compared with the existing reference frames in RPS, the proposed BRF and FRF are able to provide more accurate predictions in background and foreground regions respectively, leading to less residues to be transmitted. In general, the semantic extraction is a crucial module in the proposed scheme which links the two independent flows. In particular, the extracted semantic information not only can be utilized in analysis process, but also contributes to the efficient compression by facilitating the BRF generation. In the following sections, we will elaborate how to produce the BRF with foreground masks, and make further use of BRF to facilitate better FRF generation.

IV. Instance Segmentation-Based BRF Generation

In order to build a robust BRF that is capable of dealing with complicated situations such as dense or slow-motion objects, we propose to impose semantic constraints in the BRF generation process. Inspired by the recent work in [11], we generate the BRF in block level to avoid bitrate burst. The framework of the proposed BRF generation is illustrated in Fig. 3. In particular, the to-be-coded frame F_T is firstly fed to semantic extraction module to produce a foreground mask M_T based on instance segmentation, which serves as an important part in determining the background blocks. Subsequently, for each coding tree unit (CTU) in F_T, the background block detection process is performed to identify the potential background blocks. Concretely, motion metric and semantic metric are designed to accurately detect the background CTUs, and a CTU level flag is signalled in the bitstream to indicate whether it is a background block or not. After all the CTUs in F_T finish the detection process, the selected background CTUs are compressed in high quality, while the non-background

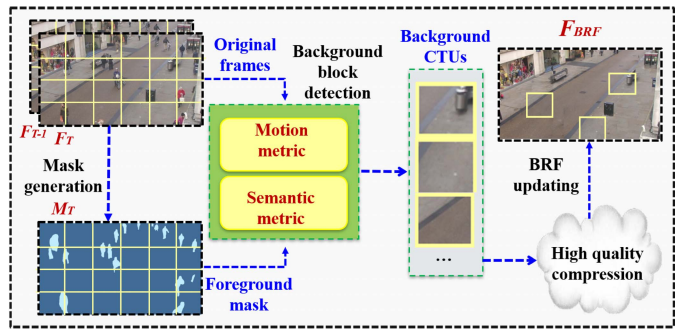


Fig. 3. The proposed BRF generation framework, where motion metric and semantic metric are complementarily utilized to constrain the background blocks detection.

CTUs are compressed in normal quality. At the decoder side, the instantaneous decoding refresh (IDR) frame is employed as the initial BRF F_{BRF}, and the decoded background CTUs are not only used to reconstruct the current frame, but also utilized to update the co-located blocks in the F_{BRF}. The key components of the proposed BRF generation scheme are elaborated as follows.

A. Foreground Mask Generation

The derivation of the foreground masks is depicted in this subsection. Before compressing the to-be-coded frame, the instance segmentation method Mask R-CNN [36] is firstly employed as a pre-processing operation to generate a pixel level foreground mask. These foreground masks contain precise position information of the specific objects, which can be utilized by both intelligent analysis and BRF generation tasks. Regarding BRF generation task, the foreground masks provide extra cues for encoder to identify the background blocks in the following detection process. A major issue here is determining the foreground objects that are of great importance. In this paper, several typical types of movable objects are studied, which can be basically classified to human relevant objects (HRO) and vehicle relevant objects (VRO). Specifically, HRO comprises of human and some related objects such as backpack, umbrella, whilst VRO contains the popular transportation tools including cars, trucks and bicycles. Let Φ_{SO} denote the HRO and VRO sets, the derivation of the M_T is expressed as follow:

$$M_T(i) = \begin{cases} 1, & \text{if } F_T(i) \in \Phi_{SO}, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $F_T(i)$ and $M_T(i)$ represent the ith pixel values of F_T and M_T, respectively. When M_T is derived, each CTU in F_T can be classified into three categories: semantic CTUs (Θ_S), background CTUs (Θ_{BG}) and hybrid CTUs (Θ_{HB}), where Θ_S only contain the pixels that belong to semantic objects, Θ_{BG} contain only background pixels, and Θ_{HB} contain both semantic objects and background pixels.

B. Background Block Detection

After deriving the foreground mask, the background block detection is subsequently initiated to choose the most stable

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

blocks as background reference. In order to ensure the robustness of the generated BRF, two complementary metrics are developed to mutually constrain the detection process by using the motion and semantic information. The details of these metrics are given as follows.

1) Motion Metric: Typically, background blocks correspond to the stable areas that stay unchanged for a certain duration. The proposed motion metric aims to detect those stable contents by calculating the variations between the co-located areas within successive frames. Let V_T^K represent the variation measurement of the Kth CTU in frame F_T, which is calculated as:

$$V_T^K = \frac{\sum_{i=1}^{N} \left| F_T^K(i) - F_{T-1}^K(i) \right|}{N}, \tag{2}$$

where $F_T^K(i)$ and $F_{T-1}^K(i)$ denote the ith pixel values in Kth CTU of F_T and F_{T-1} respectively, and N is the number of pixels within CTU. In ideal conditions, V_T^K depends only on the motion states within the block. However, due to the varying camera types and deployment conditions in practical applications, typical surveillance video would be inevitably polluted by the environmental noise. Let N_T be the expectation of the noise in F_T, which can be estimated as:

$$N_T = \frac{\sum_{i \in (L_T \cap L_{T-1})} \left| F_T^K(i) - F_{T-1}^K(i) \right|}{card(L_T \cap L_{T-1})}, \tag{3}$$

where $L_T = \{i \mid M_T(i) = 0\}$ denote the set of locations that do not belong to Φ_{SO}, and $card(\cdot)$ is the operation to get cardinal number. To overcome the negative effects caused by the noise, the proposed motion metric measurement MM_T^K is formulated as:

$$MM_T^K = V_T^K - N_T, \tag{4}$$

Intuitively, the blocks with lower motion metric value normally correspond to the stable area that should be determined as background blocks. However, as motion metric is barely capable of distinguishing the real background and low-speed foreground objects, employing only motion metric always lead to the severe foreground pollution in BRF. In view of this, we further propose the semantic metric to avoid wrong background block decision.

2) Semantic Metric: The semantic metric is designed to re-weight the motion metric based on the distinct foreground mask features in each block. A key insight here is that some static foregrounds, e.g. parking cars and sitting people, are supposed to be identified as background in the detection process. On the contrary, those objects with slow motion should be excluded from the BRF. Toward this end, we assign different weighting factor according to the block type. Let SM_T^K denote the semantic weighting factor of F_T^K, which can be formulated as:

$$SM_T^K = \begin{cases} 1, & \text{if } F_T^K \in \Theta_{BG}, \\ \mathrm{W_M}, & \text{if } F_T^K \in \Theta_{HB} \ and \ \widetilde{MM}_T^K / \widehat{MM}_T^K > \sigma, \\ \mathrm{W_S}, & \text{otherwise} \end{cases} \tag{5}$$

where $\widetilde{MM}_T^K$ denotes the motion metric measurement MM_T^K calculated within Φ_{SO} area in F_T^K, and $\widehat{MM}_T^K$ denotes the MM_T^K in non-Φ_{SO} area, W_M and W_S represent the weighting parameters for blocks with semantic foreground pixels, and σ is the threshold for local motion detection. As such, it implies that the blocks without foreground pixels are favored in our algorithm, and thus with the highest priority to serve as background blocks. For those blocks containing foreground objects pixels, however, larger weighting factors are imposed to avoid misclassification. In particular, different weighting factors are assigned according to the respective foreground property. In the proposed method, the factor W_S is set larger than 1 to mitigate the wrong background decision for Θ_S blocks. Regarding Θ_{HB}, we further calculate local motion metric $\widetilde{MM}_T^K$ and $\widehat{MM}_T^K$ to better capture local motion. When $\widetilde{MM}_T^K / \widehat{MM}_T^K$ lagers than σ, it implies local motion exists within the block, hence an extreme large factor is assigned to these blocks to avoid wrong background decision. In our implementations, W_M, W_S and σ are respectively set to 1000, 9 and 1.5.

Subsequently, the background confidence (BC) of each block is eventually calculated as the product of these two metrics, i.e.

$$BC_T^K = -SM_T^K \cdot MM_T^K, \tag{6}$$

where BC_T^K denotes the BC value of the Kth CTU of F_T. The CTU will be identified as background candidate only if the BC_T^K for both luma and chroma components is larger than the pre-defined threshold T_{BGY} and T_{BGUV}. Besides, at most N_{NUM} background candidates will be determined as ultimate background blocks in order to avoid bandwidth fluctuation, where N_{NUM} is set to 1/10 of the total CTU number within one frame. If the number of background candidates is larger than N_{NUM}, only the top-N_{NUM} candidates with the largest BC will be determined as the ultimate background blocks.

C. Background Block Compression

The determined background blocks are compressed in lower quantization parameter (QP) to provide high quality reference, and the coding parameters are set in accordance with the method in [11]. Analogous to the strategy in [11], the reconstructed IDR frame is used as the initial background frame, which serves as an additional reference frame during coding process. Moreover, a flag is assigned to each CTU to indicate whether it is a background block or not. During compression, the reconstructed background blocks are not only used to reconstruct the current frame, but also used to update the corresponding blocks in the F_{BRF}.

V. SP-GAN-Based FRF Generation

The generation of the proposed FRF is elaborated in this Section. In contrast to BRF that provides high quality reference for background regions, FRF aims to forecast the current status of foreground objects by the previously reconstructed frames, and thus mitigates the foreground redundancy. The diagram of the FRF generation scheme is illustrated in Fig. 4. In particular, a GAN based surveillance video prediction net, termed as SP-GAN, is developed to forecast the to-be-coded frame with three dedicated inputs, which are previously

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

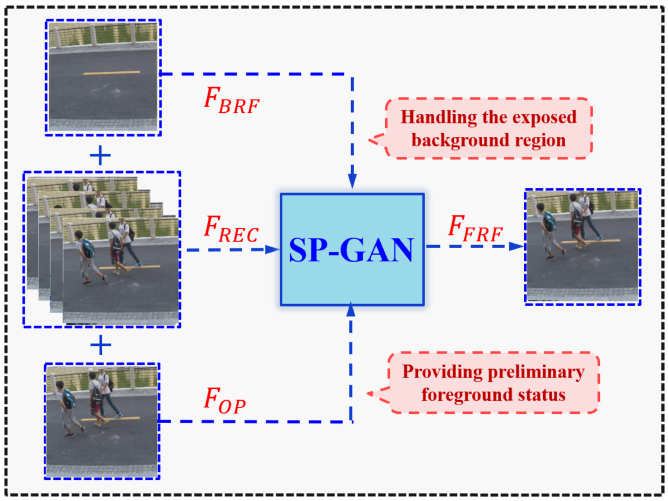


Fig. 4. The diagram of the proposed FRF generation scheme. Besides previous reconstructed frames, the F_{OP} and F_{BRF} are also employed to facilitate better FRF generation.

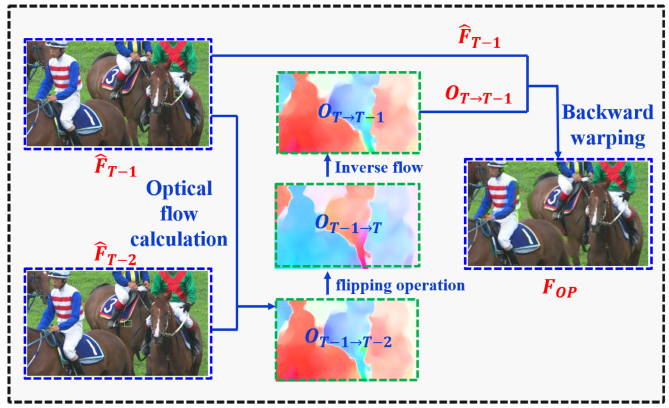


Fig. 5. The flow chart of the F_{OP} generation process.

reconstructed frames F_{REC}, optical flow based preliminary prediction F_{OP} and BRF F_{BRF}. Let F_T denote the to-be-coded frame, F_{REC} comprises of S successive reconstructed frames $\{\hat{F}_{T-1}, \hat{F}_{T-2}, \ldots, \hat{F}_{T-S}\}$. In addition to F_{REC}, two synthesized pictures, i.e. F_{OP} and F_{BRF}, are also employed to facilitate the future frame prediction. The SP-GAN fuses these $S+2$ input frames to output the F_{FRF}, which subsequently serves as an additional reference frame when compressing the F_T. Compared with previous methods, the proposed FRF is applicable to generic foreground contents, and does not require any objects segmentation or extraction. The details of the FRF generation are given as follows.

A. Insights on FRF Generation

Existing future frame prediction methods make use of previous observations to predict the next frame, which normally leverage motion information in an explicit or implicit manner. However, these methods tend to simultaneously resolve the motion flow prediction and future frame generation tasks in a single network, leading to the ambiguous of the learning target. In view of this, we incorporate an optical flow based preliminary future prediction F_{OP} as an additional input, which explicitly conveys both motion information and primary prediction, hence easing the learning process.

As analyzed in [8], how to efficiently compress the co-existing blocks (CBs) of background and foreground pixels is a key issue in surveillance video coding. As a matter of fact, the background pixels in CBs always correspond to the exposed regions that are covered by the foreground objects in the previous frames, which are however difficult to be synthesized using only the previous reconstructed frames. In light of this, besides F_{REC} and F_{OP}, the F_{BRF} is also utilized as an additional network input to handle the exposed background regions.

B. F_{OP} Generation Based on Backward Warping

The generation process of F_{OP} is depicted in Fig. 5. In particular, assuming F_T is the to-be-coded frame, the F_{OP} acts as a preliminary prediction of F_T, which is produced by backward warping of the previous reconstructed frame $\hat{F}_{T-1}$ based on the predicted optical flow $O_{T\to T-1}$ from F_T to F_{T-1}. In order to predict the $O_{T\to T-1}$, the optical flow from $\hat{F}_{T-1}$ to $\hat{F}_{T-2}$, which termed as $O_{T-1\to T-2}$, is firstly calculated by using the LiteFlowNet [37]. Then the future flow $O_{T-1\to T}$ is estimated as the flip of the $O_{T-1\to T-2}$,

$$O_{T-1\to T} = -O_{T-1\to T-2}. \tag{7}$$

An intuitive strategy to generate F_{OP} is directly forward warping the $\hat{F}_{T-1}$ based on $O_{T-1\to T}$. However, forward warping cannot project original pixels to integer position, leading to the poor quality of the F_{OP}. Accordingly, we choose to backward warp the $\hat{F}_{T-1}$ to F_{OP} based on the optical flow $O_{T\to T-1}$. A critical issue here is deriving the inverse flow $O_{T\to T-1}$ based on $O_{T-1\to T}$. In this work, we use the method in [38] to obtain the $O_{T\to T-1}$. Let $IVS(\cdot)$ denote the inverse operation, the derivation of $O_{T\to T-1}$ can be expressed as follows:

$$O_{T\to T-1} = IVS(O_{T-1\to T}). \tag{8}$$

The F_{OP} is finally derived by backward warping the $\hat{F}_{T-1}$ based on $O_{T\to T-1}$.

C. Network Architecture for SP-GAN

GAN has emerged as a powerful deep learning tool in many computer vision tasks especially in the field of image and video generation. In this work, GAN architecture is adopted to predict future frame in view of its strong generation ability. A typical GAN comprises of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$, in which $\mathcal{G}$ aims to produce realistic examples whereas $\mathcal{D}$ struggles to distinguish fakes examples from the genuine ones. Generally speaking, these two modules update alternatively until the generation ability of $\mathcal{G}$ and the discrimination ability of $\mathcal{D}$ achieve balance. As shown in Fig. 6, the U-net [34], [39] structure is adopted as the $\mathcal{G}$ in the SP-GAN. Compared with the former encoder-decoder network, U-net adds skip connections to concatenate the encoded and decoded feature maps that are in the same size, thus the network becomes more capable of preserving the high-frequency details in each scale. Regarding $\mathcal{D}$, we employ the Markovian discriminator that appeared in the PatchGAN [40], which runs

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

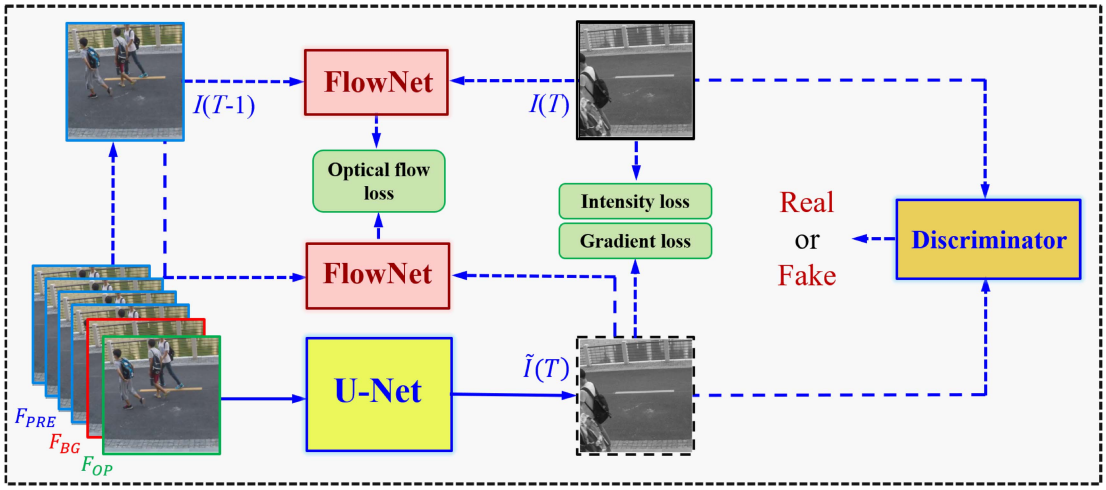


Fig. 6. The pipeline of the SP-GAN, where the $\mathcal{G}$ aims to produce realistic images whereas $\mathcal{D}$ struggles to identify the fake ones. Compared with the existing methods that input only the previous frames F_{PRE}, we leverage the F_{BG} and F_{OP} to make the SP-GAN adopt to the surveillance video future frame prediction task.

convolution operations across image patches and averages all the responses to obtain the final discrimination score of $\mathcal{D}$.

D. Training Strategy

In the training process, three constraints have been imposed to guide the generation of the next frame, which are intensity constraint, gradient constraint [27] and optical flow constraint [34]. Intensity and gradient constraints have been widely employed in the image/video generation tasks. In particular, intensity constraint ensures the general similarity between the generated image and ground truth, while gradient penalty helps produce sharp results. Let $\mathcal{L}_{L_p}$ and $\mathcal{L}_{gdt}$ denote the intensity and gradient constraints, which can be expressed in Eq. 9 and Eq. 10, respectively.

$$\begin{aligned}&\mathcal{L}_{L_p}(\widetilde{I}(T), I(T))\\&\quad= \left\|\widetilde{I}(T) - I(T)\right\|_p^p,\end{aligned} \tag{9}$$

$$\begin{aligned}&\mathcal{L}_{gdt}(\widetilde{I}(T), I(T))\\&\quad= \sum_{i,j}\left|\left|I(T)_{i,j} - I(T)_{i-1,j}\right| - \left|\widetilde{I}(T)_{i,j} - \widetilde{I}(T)_{i-1,j}\right|\right|\\&\qquad+\left|\left|I(T)_{i,j}-I(T)_{i,j-1}\right|-\left|\widetilde{I}(T)_{i,j}-\widetilde{I}(T)_{i,j-1}\right|\right|,\end{aligned} \tag{10}$$

where i, j is the pixel index and $|\cdot|$ represents the absolute value function, $\widetilde{I}(T)$ and $I(T)$ denote the predicted image and ground truth image in T moment respectively, and p is set to 2 in our method. In addition to $\mathcal{L}_p$ and $\mathcal{L}_{gdt}$, we also leverage the optical flow penalty $\mathcal{L}_{op}$ provided in [34] to guarantee the coherence of motion prediction,

$$\begin{aligned}&\mathcal{L}_{op}(\widetilde{I}(T), I(T))\\&\quad= \left\|OF(I(T), I(T-1)) - OF(\widetilde{I}(T), I(T-1))\right\|_1,\end{aligned} \tag{11}$$

where $OF(\cdot)$ denotes the optical flow calculation function, which is realized through a pre-trained FlowNet [41] in this work.

We follow the similar strategy in LSGAN [34], [42] to train the SP-GAN. In particular, the $\mathcal{G}$ and $\mathcal{D}$ are alternatively updated during training process, and the parameters of the other module remain the same when the current module is updating. The training target of $\mathcal{D}$ is classifying ground truth to 1 whilst predictions to 0. When training the $\mathcal{D}$, the loss function $\mathcal{L}^D$ is designed as:

$$\begin{aligned}\mathcal{L}^D &= \mathcal{L}^D_{adv}(\widetilde{I}(T), I(T))\\&= \sum_{m,n}\frac{1}{2}(\mathcal{D}(I(T)_{m,n}) - 1)^2 + \sum_{m,n}\frac{1}{2}\mathcal{D}(\widetilde{I}(T)_{m,n})^2,\end{aligned} \tag{12}$$

where m, n indicates the patch index. When training the $\mathcal{G}$, we combine the three above mentioned penalties together with adversary loss to form the ultimate loss function $\mathcal{L}^G$,

$$\begin{aligned}\mathcal{L}^G &= \lambda_{L_p}\mathcal{L}_{L_p}(\widetilde{I}(T), I(T))\\&\quad+\lambda_{gdt}\mathcal{L}_{gdt}(\widetilde{I}(T), I(T))\\&\quad+\lambda_{op}\mathcal{L}_{op}(\widetilde{I}(T), I(T))\\&\quad+\lambda_{adv}\sum_{m,n}\frac{1}{2}\mathcal{D}(\widetilde{I}(T)_{m,n} - 1)^2.\end{aligned} \tag{13}$$

Here λ_{L_p}, λ_{gdt}, λ_{op} and λ_{adv} are weighting factors to control the trade-off between each penalty.

Note that the SP-GAN is trained upon original videos since the generative ability rather than artifact removal is desired in our scheme. Whereas the reconstructed frames are used to infer the to-be-coded frame when the model is integrated into video codec. To train the SP-GAN, we collect 58 surveillance sequences from AVS test dataset [43], VFC-1M [44] dataset and ShanghaiTech dataset [34]. The selected training sequences are captured in traffic and campus scenes hence the majority of the foreground objects in these sequences are pedestrians and vehicles. To ensure the diversity of training data, we discard useless parts in each sequence and preserve only the clips with abundant foregrounds. The ultimate training set contains in total 14702 frames, which are then resized to 256×256 and scaled to $[-1, 1]$. For the video coding, the SP-GAN models for Y and UV components are trained separately. More specifically, the history frames number S

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

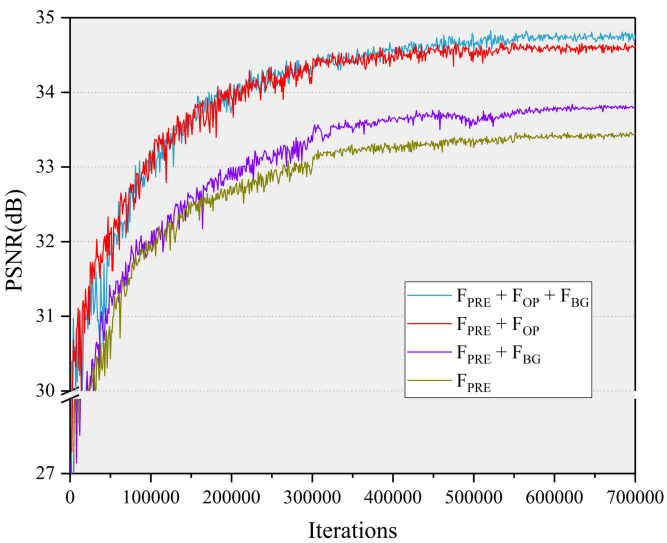


Fig. 7. The training curves of SP-GAN with different inputs. Compared with the baseline scheme that only has F_{PRE} as the inputs, the additional inputs of F_{BG} and F_{OP} bring substantial PSNR improvement.

is set to 4, and we randomly select 5 consecutive frames $\{F_{T-4}, F_{T-3}, F_{T-2}, F_{T-1}, F_T\}$ as a training set, in which $\{F_{T-4}, F_{T-3}, F_{T-2}, F_{T-1}\}$ act as input F_{PRE} and F_T is the ground truth. Besides, the optical flow based prediction F_{OP} is produced by taking advantage of F_{T-2} and F_{T-1}. Additionally, for each sequence, we manually select one clean frame which contains no foregrounds as the background frame F_{BG}. Finally, F_{PRE}, F_{OP} and F_{BG} mutually comprise the ultimate training inputs of SP-GAN, as shown in Fig. 6.

Fig. 7 provides the training curves of SP-GAN with ablative inputs, which depicts the PSNR variations on the test set in the training process. As we can see, compared with the baseline scheme that employs only F_{PRE} as input, the additional inputs of F_{BG} and F_{OP} have greatly improved the generation performance. In particular, when the model is trained with F_{PRE} and F_{BG}, approximately 0.5dB improvement is achieved when the model has converged. Similarly, when the model is trained with F_{PRE} and F_{OP}, 1.2dB PSNR improvement is observed compared with the baseline. Moreover, when the model is trained with all of F_{PRE}, F_{BG} and F_{OP}, the best result is obtained, which outperforms the baseline scheme for more than 1.4dB. We also provide the subjective and objective comparisons of SP-GAN predictions with different inputs in Fig. 8, where best subjective and objective performance is obtained with all of the three inputs. The ablation results in Fig. 7 and Fig. 8 have spoken to the effectiveness of the F_{BG} and F_{OP} in the surveillance future frame prediction task, which motivates us to impose all of the three inputs in the compression process.

E. Implementation

Before compressing the to-be-coded frame, 4 previous reconstructed frames F_{REC} are firstly obtained, and the F_{OP} is subsequently produced based on F_{REC}. Since the ideal background image F_{BG} is not available in the coding process, the F_{BRF} is employed to provide background information for the exposed region. These three inputs are fed to the trained SP-GAN model to output the prediction for the to-be-coded frame. An intuitive strategy to make use of this prediction is directly applying it as the final reconstruction. However, considering that the SP-GAN may fail in some complicated scenes, the rate-distortion optimization (RDO) is preferred to help maintain the compression quality. Towards this end, similar to F_{BRF}, we add the output of SP-GAN as another long-term reference termed as F_{FRF}. Since F_{FRF} is produced by reconstructed frames, no additional information need to be transmitted in the bitstream, and the decoder conducts the identical process to obtain the F_{FRF}.

In HEVC, luma component has overwhelming importance in determining the unified inter coding information that is shared by both luma and chroma. In view of this, we slightly modify the rate-distortion optimization (RDO) process in F_{BRF} & F_{FRF} generation, and assign larger weight to chroma component to make up for the loss caused by F_{FRF}. Concretely, in HEVC, the overall distortion D in inter prediction are calculated as $D = D_{luma} + WD_{chroma}$, where D_{luma} and D_{chroma} are luma and chroma distortions respectively, and W is the weighting factor that balances them. In the modified scheme, in addition to W, another weighting factor ε is introduced and the D is then adjusted as $D = D_{luma} + \varepsilon WD_{chroma}$, where ε is experimentally set to 1.1.

VI. Experimental Results

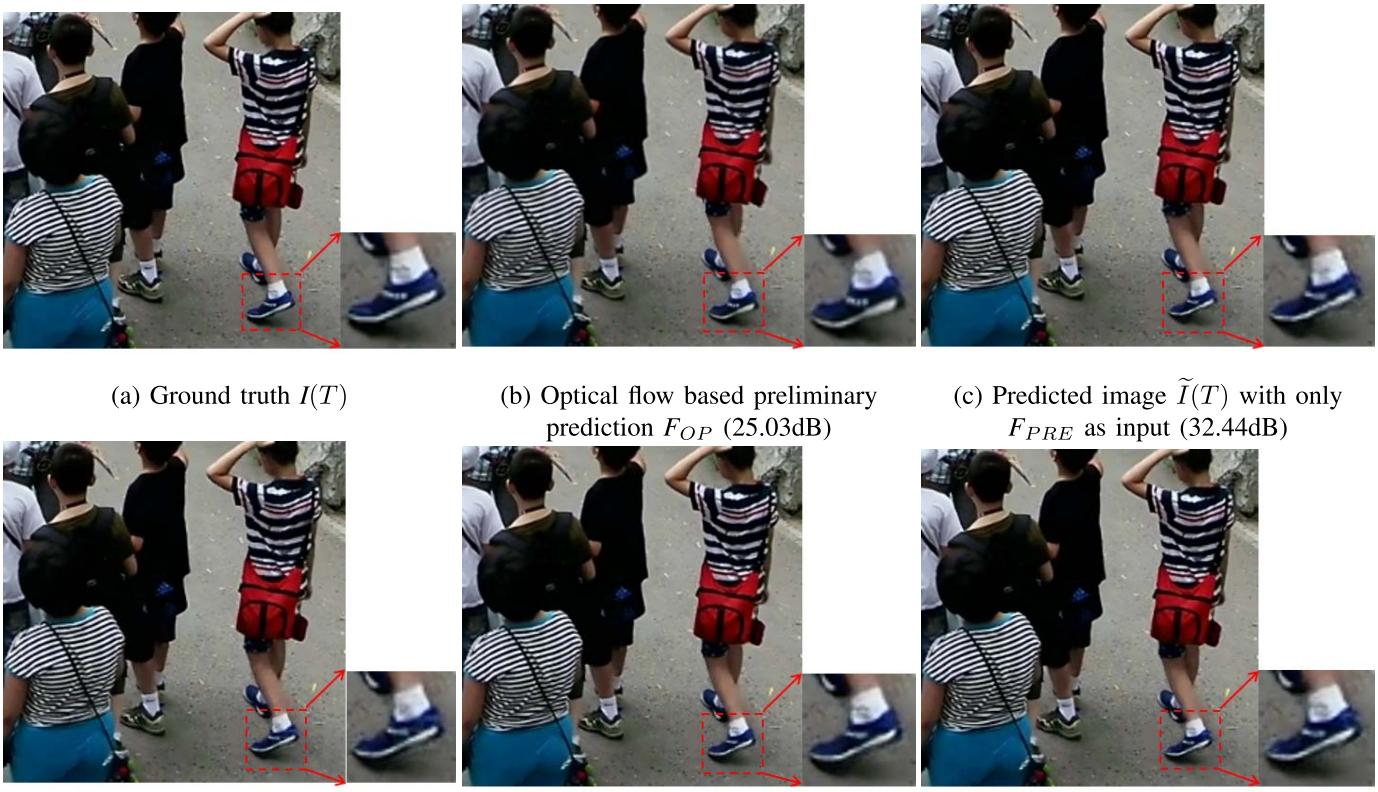
The performance of the proposed methods are provided in this Section. The encoder configurations and training details of SP-GAN are discussed in Section VI-A, and Section VI-B provides the overall performance of the BRF method. Besides, the performance of the proposed FRF method and the integration of BRF and FRF are provided in Section VI-C, and the complexity analysis is provided in Section VI-D.

A. Encoding Configurations and Training Details

1) Encoding Configuration and Evaluations: The proposed methodologies are implemented on HEVC software HM-16.6, and the encoder configurations are in compliance with HEVC common test conditions (CTC) [45]. The experiments are conducted upon CPU and GPU platforms, which are Intel(R) Core(TM) i7-6900K and GeForce GTX TITAN X, respectively. Eleven surveillance sequences with relatively dense foregrounds are tested in the experiment, which are collected from [44] and [46] and have no overlap with the training sequences for SP-GAN. Four standard QP points, i.e. {22, 27, 32, 37}, are tested to verify the performance of the proposed methods. Additionally, we use the widely employed BD-rate [47] metric to evaluate the effectiveness of the proposed methods.

2) Training Details of SP-GAN: The SP-GAN is trained on top of GPU platform with Tensorflow [48]. The SP-GAN model is trained with 700000 iterations, and the batch size is set to 4 during the training process. Besides, we use Adam [49] as the optimizer to update the model parameters. The initial learning rate for $\mathcal{G}$ is set to 10^{-4}, and it decays to 5×10^{-5} and

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.



(a) Ground truth $I(T)$ (b) Optical flow based preliminary prediction F_{OP} (25.03dB) (c) Predicted image $\widetilde{I}(T)$ with only F_{PRE} as input (32.44dB)

(d) Predicted image $\widetilde{I}(T)$ with F_{PRE} and F_{BG} as input (33.05dB) (e) Predicted image $\widetilde{I}(T)$ with F_{PRE} and F_{OP} as input (33.49dB) (f) Predicted image $\widetilde{I}(T)$ with F_{PRE}, F_{PRE} and F_{BG} as input (33.79dB)

Fig. 8. Subjective and objective comparisons of the SP-GAN predictions with different inputs.

TABLE I
THE OVERALL RD PERFORMANCE OF THE PROPOSED F_{BRF} METHOD COMPARED WITH HM-16.6

Sequences	Low-delay B			Low-delay P		
	Y	U	V	Y	U	V
Hongsilouwest1	-38.5%	-33.2%	-27.8%	-37.8%	-30.9%	-26.7%
Hongsilouwest2	-34.1%	-31.6%	-28.3%	-32.8%	-30.1%	-27.6%
Hongsilouwest3	-25.8%	-27.8%	-23.4%	-24.7%	-27.1%	-20.7%
Downtown1	-35.5%	-41.3%	-41.8%	-33.4%	-39.3%	-41.1%
Downtown2	-32.4%	-32.5%	-35.8%	-30.4%	-31.5%	-34.3%
Dongcemen1	-28.3%	-20.9%	-22.2%	-26.3%	-20.6%	-22.3%
Dongcemen2	-19.6%	-17.6%	-17.7%	-17.8%	-18.3%	-16.8%
Yaogannan1	-19.6%	-23.2%	-21.5%	-17.2%	-21.0%	-20.8%
Yaogannan2	-19.9%	-19.9%	-19.8%	-17.6%	-18.9%	-19.2%
Weiminghudong1	-33.6%	-37.0%	-23.9%	-32.5%	-36.6%	-21.4%
Weiminghudong2	-29.6%	-32.5%	-32.6%	-27.7%	-27.7%	-31.9%
Overall	**-28.8%**	**-28.9%**	**-26.8%**	**-27.1%**	**-27.5%**	**-25.7%**

10^{-5} in the 300000th and 550000th iterations, respectively. As for $\mathcal{D}$, the learning rate decays from 10^{-5} to 10^{-6} during the training process. Additionally, the weighting factors λ_{adv}, λ_{L_p}, λ_{gdt} and λ_{op} are experimentally set to 0.25, 50, 1 and 0.2, respectively.

B. Performance of F_{BRF}

We first compare the proposed F_{BRF} with HM-16.6, and the experimental results are provided in Table I. In the experiment, the proposed instance segmentation based BRF is appended to RPS and serves as an additional reference frame. As can be observed, the proposed F_{BRF} achieves remarkable BD-rate gain on HM-16.6. More specifically, in the LDB configuration, on average 28.8%, 28.9% and 26.8% bitrate savings are achieved on Y, Cb and Cr components respectively, and up to 38.5% luma coding gain is observed on *Hongsilouwest1* sequence. Regarding the LDP configuration, overall 27.1% luma bitrate saving is achieved, which also demonstrates the significant superiority of the proposed F_{BRF}.

We also compare our method with the state-of-the-art BRF scheme in [11], which adopts the block level generation strategy as well. Intuitively, ideal BRF generation methods are supposed to well portray the real background hence the generated BRFs typically have fine subjective quality. To verify the effectiveness of the proposed F_{BRF} from subjective quality perspective, we extract several typical BRFs in the compression process which are depicted in Fig. 9. The first row represents the BRFs generated by the method in [11], while the corresponding BRFs generated by our method are shown in the second row. As can be observed, the BRFs produced by the method in [11] suffer from severe foreground pollution, whereas in our method the foreground pollution is effectively mitigated by introducing the semantic constraint. For example, in Fig. 9 (b), the fragments of slow moving pedestrians incorrectly appear in the BRF produced by the method in [11], which is mainly due to the wrong classification

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

(a) DongCeMen1 (b) DownTown1 (c) Hongsilouwest1

(d) Hongsilouwest2 (e) Weiminghudong1 (f) Weiminghudong2

Fig. 9. Subjective comparisons between the proposed BRFs and that produced by the method in [11]. Six typical pairs are shown here, where the first row represents the BRFs generated by the method in [11], and the corresponding BRFs generated by our method are shown in the second row.

of the foreground and background pixels. However, such undesired pollution is well excluded from our BRF. Specifically, the extensive pollution disappears and only scattered foreground pixels left in the BRFs, which results from the inaccurate recognition of the instance segmentation algorithm utilized in our method. It is worth noting that static foregrounds are supposed to be identified as background pixels in the coding process, and our method will not exclude these instances from the BRF. As shown in Fig. 9 (a), the women circled in red remains still during the modeling process, and our scheme successfully distinguishes the static foregrounds and updates the BRF using the foreground pixels. By utilizing motion metric and semantic metric in a combined manner, our mechanism is much more robust in dealing with complex foreground conditions.

In addition to the subjective quality comparison, BD-rate comparison between the proposed BRF and the method in [11] is also provided. As can be observed in Table II, the proposed F_{BRF} comprehensively outperforms the scheme in [11], which achieves 3.5% and 3.3% luma bitrate saving in LDB and LDP configurations respectively, and up to 8.0% gain is observed on *Hongsilouwest3* sequence. These simulation results have provided useful evidences in indicating the advantage of the proposed BRF.

C. Performance of F_{FRF}

The performance of the proposed F_{FRF} is provided in Tables III, IV and V. In our experiment, instead of using F_{FRF} as an independent reference, we append both F_{BRF} and F_{FRF} to RPS since F_{BRF} constitutes the precondition to

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

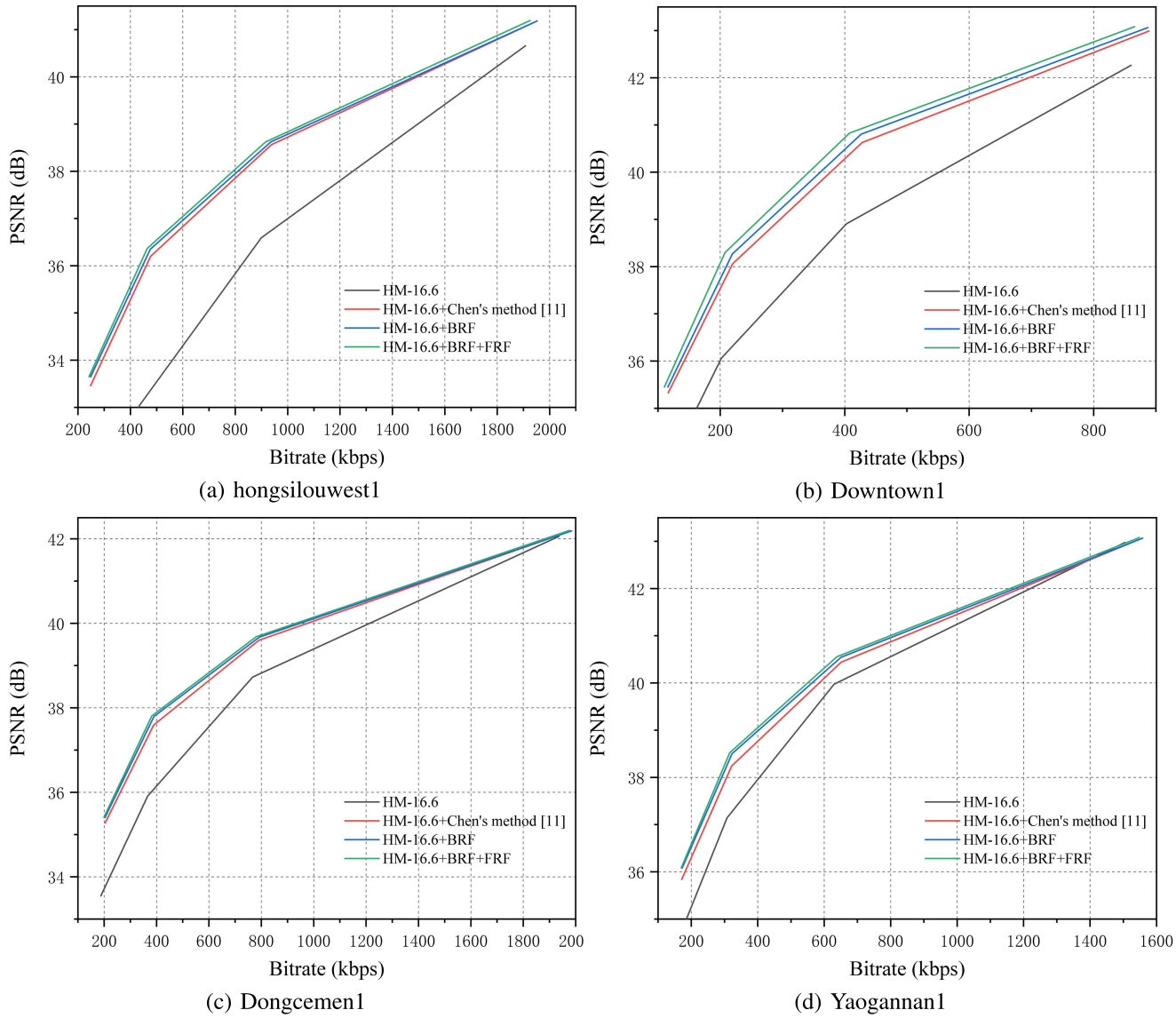


Fig. 10. RD performance of the F_{BRF} and F_{FRF} methods based on Y component.

the derivation of F_{FRF}. The simulation results of the F_{BRF} & F_{FRF} method compared with HM-16.6 are shown in Table III. As we can see, when F_{BRF} and F_{FRF} are both enabled, 30.3% and 29.0% average luma gains are respectively observed under LDB and LDP configurations. Besides, in order to demonstrate the effectiveness of F_{FRF} itself, we also compare the F_{BRF} & F_{FRF} with F_{BRF}, and the results are shown in Table IV. As can be seen, comprehensive BD-rate gains are achieved by adding F_{FRF} as an additional reference, where 2.1% and 2.6% luma bitrate savings are obtained in LDB and LDP configurations, and up to 4.9% and 5.2% performance improvements are observed on *Downtown1* sequence. Moreover, comparisons between the proposed F_{BRF} & F_{FRF} methods and the method in [11] are also provided in Table V, which indicates our method outperforms the state-of-the-art surveillance coding method with 5.8% bit rate savings, providing substantial evidences in validating the effectiveness of F_{BRF} & F_{FRF} method. To illustrate the advantages of the proposed methods in a more intuitive way, we provide some typical RD curves as shown in Fig. 10, from which we can clearly observe the effectiveness of the presented references used in either isolate or combined ways.

We also provide several typical coding frames as shown in Fig. 11, where red boxes represent the PUs that are predicted by F_{BRF}, and blue boxes represent that predicted by F_{FRF}. As can be observed, both F_{BRF} and F_{FRF} are frequently utilized as predictions during coding process, showing great superiority of F_{BRF} and F_{FRF} compared with normal reference frames. More specifically, encoder tends to choose F_{FRF} in foreground regions, indicating considerable forecasting accuracy of the proposed SP-GAN. We also provide the hit ratio of F_{FRF} in different QPs, as shown in Fig. 12. Here, the hit ratio R of F_{FRF} is calculated as follows:

$$R = \frac{N_{C_{FRF}}}{N_{C_{FG}}} \times 100\%, \tag{14}$$

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

(a) Dongcemen2 (b) Downtown1

(c) Hongsilouwest2 (d) Yaogannan2

Fig. 11. Some typical coding frames that predicted with F_{BRF} and F_{FRF}.

TABLE II

THE OVERALL RD PERFORMANCE OF THE PROPOSED F_{BRF} METHOD COMPARED WITH CHEN *et al.*'S METHOD [11]

Sequences	Low-delay B			Low-delay P		
	Y	U	V	Y	U	V
Hongsilouwest1	-3.1%	-1.2%	-0.4%	-2.6%	-0.6%	-0.1%
Hongsilouwest2	-4.3%	-2.1%	-1.4%	-3.8%	-1.6%	-1.2%
Hongsilouwest3	-8.0%	-4.0%	-5.0%	-8.0%	-4.4%	-3.1%
Downtown1	-4.5%	-1.7%	-4.1%	-4.3%	-2.8%	-3.8%
Downtown2	-3.4%	-1.9%	-2.7%	-3.1%	-2.0%	-1.9%
Dongcemen1	-3.7%	-1.0%	-0.7%	-3.7%	-1.0%	-0.4%
Dongcemen2	-0.8%	-0.8%	0.9%	-0.8%	-0.6%	0.9%
Yaogannan1	-4.2%	-1.2%	0.2%	-4.1%	-1.4%	-1.2%
Yaogannan2	-2.8%	-1.6%	-0.5%	-2.4%	-3.1%	0.0%
Weiminghudong1	-2.3%	-1.5%	-0.9%	-2.6%	-1.4%	0.2%
Weiminghudong2	-1.2%	0.1%	0.1%	-1.2%	-1.3%	-0.9%
Overall	**-3.5%**	**-1.5%**	**-1.3%**	**-3.3%**	**-1.8%**	**-1.1%**

TABLE III

THE OVERALL RD PERFORMANCE OF THE PROPOSED F_{BRF}& F_{FRF} METHOD COMPARED WITH HM-16.6

Sequences	Low-delay B			Low-delay P		
	Y	U	V	Y	U	V
Hongsilouwest1	-39.6%	-34.8%	-29.7%	-39.3%	-32.4%	-28.5%
Hongsilouwest2	-35.6%	-33.9%	-30.8%	-34.4%	-32.3%	-29.6%
Hongsilouwest3	-26.3%	-29.9%	-25.9%	-25.6%	-30.1%	-22.8%
Downtown1	-38.8%	-39.1%	-42.6%	-37.0%	-37.1%	-41.1%
Downtown2	-35.7%	-30.4%	-34.6%	-34.0%	-30.3%	-33.8%
Dongcemen1	-29.3%	-18.1%	-19.9%	-27.7%	-18.8%	-21.8%
Dongcemen2	-21.1%	-17.4%	-18.6%	-19.6%	-20.1%	-17.8%
Yaogannan1	-20.7%	-20.1%	-21.2%	-19.0%	-18.3%	-21.6%
Yaogannan2	-21.1%	-16.9%	-18.7%	-19.2%	-16.9%	-18.2%
Weiminghudong1	-34.4%	-38.1%	-25.6%	-33.8%	-37.5%	-21.9%
Weiminghudong2	-30.7%	-34.6%	-33.0%	-29.3%	-30.5%	-33.3%
Overall	**-30.3%**	**-28.5%**	**-27.3%**	**-29.0%**	**-27.5%**	**-26.3%**

where C_{FG} denotes the CTUs that contain foreground pixels, $N_{C_{FG}}$ is the amount of C_{FG}, and $N_{C_{FRF}}$ is the amount of C_{FG} that are predicted with F_{FRF}. It can be observed in Fig. 12 that more than 69% foreground CTUs select F_{FRF} when QP is 22. Considering that the selection of reference frame is realized by means of RDO, such results provide powerful evidences in showing the effectiveness of F_{FRF}. Besides, we also notice that the usage of F_{FRF} decreases as QP gets larger, which indicates that FRF tends to provide finer predictions in higher bitrate conditions. Regarding F_{BRF}, we notice that encoder tends to select F_{BRF} in the surrounding areas of foreground objects, which correspond to the newly

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

THE OVERALL RD PERFORMANCE OF THE PROPOSED F_{BRF} & F_{FRF} METHOD COMPARED WITH F_{BRF}

Sequences	Low-delay B			Low-delay P		
	Y	U	V	Y	U	V
Hongsilouwest1	-2.0%	-2.6%	-3.0%	-2.4%	-2.1%	-2.9%
Hongsilouwest2	-2.1%	-2.4%	-3.2%	-2.3%	-2.3%	-2.8%
Hongsilouwest3	-1.2%	-2.6%	-3.5%	-1.7%	-3.9%	-3.0%
Downtown1	-4.9%	3.0%	-1.7%	-5.2%	2.9%	-0.5%
Downtown2	-4.6%	3.0%	-1.4%	-4.8%	2.1%	1.0%
Dongcemen1	-1.4%	4.9%	4.0%	-1.8%	3.7%	2.2%
Dongcemen2	-1.8%	0.4%	-1.2%	-2.6%	-0.5%	-0.7%
Yaogannan1	-1.6%	4.9%	0.6%	-2.2%	4.6%	-0.4%
Yaogannan2	-1.7%	4.5%	1.6%	-2.2%	3.1%	1.5%
Weiminghudong1	-0.7%	-1.8%	-2.3%	-1.4%	-1.7%	-1.2%
Weiminghudong2	-1.5%	-3.4%	-0.9%	-2.0%	-4.4%	-2.4%
Overall	**-2.1%**	**0.7%**	**-0.7%**	**-2.6%**	**0.1%**	**-0.8%**

TABLE V

THE OVERALL RD PERFORMANCE OF THE PROPOSED F_{BRF} & F_{FRF} METHOD COMPARED WITH CHEN *et al.*'S METHOD [11]

Sequences	Low-delay B			Low-delay P		
	Y	U	V	Y	U	V
Hongsilouwest1	-5.1%	-3.8%	-3.4%	-4.9%	-2.8%	-3.0%
Hongsilouwest2	-6.3%	-4.4%	-4.7%	-6.0%	-3.8%	-4.1%
Hongsilouwest3	-9.1%	-6.5%	-8.5%	-9.6%	-8.3%	-6.2%
Downtown1	-9.4%	1.3%	-5.7%	-9.5%	0.1%	-4.3%
Downtown2	-7.9%	1.1%	-4.1%	-7.9%	0.0%	-1.0%
Dongcemen1	-5.1%	3.8%	3.2%	-5.4%	2.6%	1.8%
Dongcemen2	-2.5%	-0.4%	-0.3%	-3.3%	-1.1%	0.2%
Yaogannan1	-5.7%	3.8%	0.7%	-6.3%	3.2%	-1.5%
Yaogannan2	-4.5%	2.9%	1.1%	-4.6%	0.1%	1.5%
Weiminghudong1	-2.9%	-3.4%	-3.1%	-3.9%	-3.1%	-1.1%
Weiminghudong2	-2.7%	-3.2%	-0.8%	-3.2%	-5.6%	-3.2%
Overall	**-5.6%**	**-0.8%**	**-2.3%**	**-5.8%**	**-1.7%**	**-1.9%**

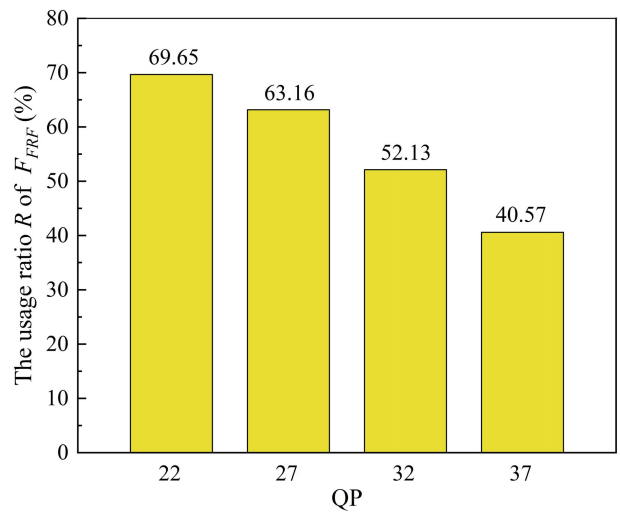


Fig. 12. The usage ratio of F_{FRF} in different QPs.

TABLE VI

THE COMPLEXITY INCREMENTS COMPARISONS BETWEEN THE PROPOSED F_{BRF} AND THAT PRESENTED IN [11]

	F_{BRF} in [11]	Proposed F_{BRF} (CPU)	Proposed F_{BRF} (CPU+GPU)
ΔT_{Enc}	16%	34%	19%
ΔT_{Dec}	2%	3%	3%

TABLE VII

THE COMPLEXITY INCREMENTS OF F_{BRF} & F_{FRF} METHOD

	F_{BRF} & F_{FRF} (CPU)	F_{BRF} & F_{FRF} (CPU+GPU)
ΔT_{Enc}	144%	70%
ΔT_{Dec}	97367%	28107%

exposed backgrounds that can barely obtain matched contents in short term references frames.

D. Complexity Analysis

The computational complexity of the proposed F_{BRF} and F_{FRF} is detailed in this subsection. More specifically, we provide two versions of the computational complexity evaluation including CPU only and the hybrid utilization of CPU and GPU. The complexity overhead is measured by $\Delta T_{Enc/Dec}$, which is computed as follows,

$$\Delta T_{Enc/Dec} = \frac{T'_{Enc/Dec} - T_{Enc/Dec}}{T_{Enc/Dec}} \times 100\%, \qquad (15)$$

where $T_{Enc/Dec}$ denotes the complexity of HEVC anchor in encoder or decoder side, and $T'_{Enc/Dec}$ denotes the corresponding complexity of the proposed method. We first compare the complexity of the proposed F_{BRF} with that proposed in [11], as presented in Table VI. Compared with the F_{BRF} in [11], our method incurs additional 18% encoder complexity on CPU platform, which mainly comes from the generation of foreground masks in our method. In the meanwhile, only 3% additional encoder complexity increments are observed on GPU+GPU platform, indicating considerable application prospects of the proposed BRF.

The major issue of the complexity burden lies in F_{FRF}. When F_{BRF} and F_{FRF} are both integrated to RPS, the complexity of encoder and decoder has respectively increased to 144% and 97367% on CPU platform, as shown in Table VII. At the same time, considerable complexity relief is observed when GPU is utilized. When F_{BRF} and F_{FRF} are both integrated, the complexity overheads on CPU+GPU platform decrease to 70% and 28107%, respectively. In the future, we will make further studies for lighter neural network architectures so as to make our method applicable in real-time scenarios. Besides, with the fast development of acceleration technologies and deep learning oriented hardware architecture, we are fully optimistic for the utilization of the proposed methods in the near future.

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

VII. Conclusion

We present a dual reference frame generation scheme for surveillance videos compression, in which BRF and FRF are complementarily utilized to mitigate the background and foreground redundancy respectively. Different from the previous BRF approaches that exploit only statistical and motion information, the proposed BRF incorporates semantic knowledge, leading to more robust predictions in those scenes with complicated foregrounds. To deal with foreground redundancy, the FRF is generated based on SP-GAN, which fuses previous reconstructed frames, optical based prediction and BRF to forecast the status of the to-be-coded frame. Extensive simulation results have demonstrated the remarkable superiority of the proposed scheme. Compared with the state-of-the-art surveillance coding method, the proposed BRF achieves more than 3% bitrate saving, and 5.8% bitrate savings are obtained when BRF and FRF are both enabled.

References

[1] *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, EMC, Framingham, MA, USA, 2012.

[2] L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Enhanced motion-compensated video coding with deep virtual reference frame generation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4832–4844, Oct. 2019.

[3] Q. Ding, L. Shen, H. Yang, X. Dong, and M. Xu, "VRFCNN: Virtual reference frame generation network for quality SHVC," *IEEE Signal Process. Lett.*, vol. 27, pp. 2049–2053, 2020.

[4] C. Ma, D. Liu, X. Peng, F. Wu, H. Li, and T. Wang, "Reference clip for inter prediction in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 130–143, Jan. 2019.

[5] M. Paul, W. Lin, C. T. Lau, and B.-S. Lee, "Video coding using the most common frame in scene," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 734–737.

[6] S. Chakraborty, M. Paul, M. Murshed, and M. Ali, "A novel video coding scheme using a scene adaptive non-parametric background model," in *Proc. IEEE 16th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2014, pp. 1–6.

[7] X. Zhang, Y. Tian, T. Huang, and W. Gao, "Low-complexity and high-efficiency background modeling for surveillance video coding," in *Proc. IEEE VCIP*, Nov. 2012, pp. 1–6.

[8] X. Zhang, T. Huang, Y. Tian, and W. Gao, "Background-modeling-based adaptive prediction for surveillance video coding," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 769–784, Feb. 2014.

[9] X. Zhang, Y. Tian, T. Huang, S. Dong, and W. Gao, "Optimizing the hierarchical prediction and coding in HEVC for surveillance and conference videos with background modeling," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4511–4526, Oct. 2014.

[10] M. Paul, W. Lin, C.-T. Lau, and B.-S. Lee, "Explore and model better I-frames for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1242–1254, Sep. 2011.

[11] F. Chen, H. Li, L. Li, D. Liu, and F. Wu, "Block-composed background reference for high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2639–2651, Dec. 2017.

[12] X. Wang, R. Hu, Z. Wang, and J. Xiao, "Virtual background reference frame based satellite video coding," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1445–1449, Oct. 2018.

[13] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[14] *High Efficiency Video Coding*, documentITU-T Rec. H.265 and ISO/IEC 23008-2 (HEVC), ITU-T and ISO/IEC, 2013.

[15] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 246–252.

[16] A. Hakeem, K. Shafique, and M. Shah, "An object-based video coding framework for video sequences obtained from static cameras," in *Proc. 13th Annu. ACM Int. Conf. Multimedia (MULTIMEDIA)*, 2005, pp. 608–617.

[17] C. Ma, D. Liu, X. Peng, and F. Wu, "Surveillance video coding with vehicle library," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 270–274.

[18] J. Xiao, R. Hu, L. Liao, Y. Chen, Z. Wang, and Z. Xiong, "Knowledge-based coding of objects for multisource surveillance video data," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1691–1706, Sep. 2016.

[19] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, and J. Yang, "CNN-based bi-directional motion compensation for high efficiency video coding," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2018, pp. 1–4.

[20] N. Yan, D. Liu, H. Li, B. Li, L. Li, and F. Wu, "Convolutional neural network-based fractional-pixel motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 840–853, Mar. 2019.

[21] J. Liu, S. Xia, W. Yang, M. Li, and D. Liu, "One-for-all: Grouped variation network-based fractional interpolation in video coding," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2140–2151, May 2019.

[22] Y. Li, H. Li, L. Li, F. Wu, H. Zhang, and H. Yang, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2316–2330, Sep. 2018.

[23] T. Wang, M. Chen, and H. Chao, "A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 410–419.

[24] C. Jia, S. Wang, X. Zhang, S. Wang, and S. Ma, "Spatial-temporal residue network based in-loop filter for video coding," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[25] L. Zhou, X. Song, and J. Yao, *Convolutional Neural Network Filter (CNNF) for Intra Frame*, Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, document JVET-I0022, Jan. 2018.

[26] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. ICML*, 2015, pp. 843–852.

[27] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2015, *arXiv:1511.05440*. [Online]. Available: http://arxiv.org/abs/1511.05440

[28] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," 2016, *arXiv:1605.08104*. [Online]. Available: http://arxiv.org/abs/1605.08104

[29] J. van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," 2017, *arXiv:1701.08435*. [Online]. Available: http://arxiv.org/abs/1701.08435

[30] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4473–4481.

[31] N. Sedaghat, M. Zolfaghari, and T. Brox, "Hybrid learning of optical flow and next frame prediction to boost optical flow in the wild," 2016, *arXiv:1612.03777*. [Online]. Available: http://arxiv.org/abs/1612.03777

[32] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future-flow embedded video prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1762–1770.

[33] C. Lu, M. Hirsch, and B. Scholkopf, "Flexible spatio-temporal networks for video prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2137–2145.

[34] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proc. IEEE CVPR*, Jun. 2018, pp. 6536–6545.

[35] W. Liu, A. Sharma, O. Camps, and M. Sznaier, "DYAN: A dynamical atoms-based network for video prediction," in *Proc. ECCV*, 2018, pp. 170–185.

[36] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[37] T.-W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8981–8989.

[38] J. Sánchez, A. Salgado, and N. Monzón, "Computing inverse optical flow," *Pattern Recognit. Lett.*, vol. 52, pp. 32–39, Jan. 2015.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

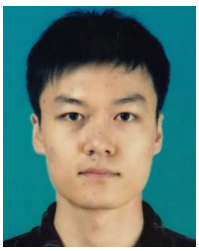
[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[41] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[42] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," 2016, *arXiv:1611.04076*. [Online]. Available: http://arxiv.org/abs/1611.04076

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.

[43] S. Ma, S. Wang, and W. Gao, "Overview of IEEE 1857 video coding standard," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 1500–1504.
[44] L. Ding, Y. Tian, H. Fan, Y. Wang, and T. Huang, "Rate-performance-loss optimization for inter-frame deep feature coding from videos," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5743–5757, Dec. 2017.
[45] F. Bossen, *Common Test Conditions and Software Reference Configurations*, Joint Collaborative Team on Video Coding (JCT-VC), document JCTVC-F900, Torino, Turin, Italy, 2011.
[46] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. CVPR*, Jun. 2011, pp. 3457–3464.
[47] G. Bjontegaard, "Calcuation of average PSNR differences between RDcurves," VCEG-M33, presented at the 13th meeting ITU-T Video Coding Experts Group (VCEG), Apr. 2001.
[48] M. Abadi *et al.* (2015). *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*. [Online]. Available: http://download.tensorflow.org/paper/whitepaper2015.pdf
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980



Lei Zhao received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2014. He is currently pursuing the Ph.D. degree with Peking University, Beijing, China. His research interests include video compression and machine learning.

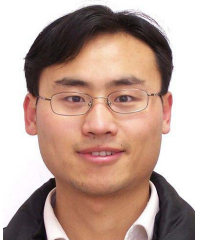
Shiqi Wang (Member, IEEE) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008, and the Ph.D. degree in computer application technology from Peking University in 2014. From March 2014 to March 2016, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. From April 2016 to April 2017, he was with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore, as a Research Fellow. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong. He has proposed over 40 technical proposals to ISO/MPEG, ITU-T, and AVS standards, and authored more than 150 refereed journal/conference papers. His research interests include video compression, image/video quality assessment, and image/video search and analysis. He received the Best Paper Award of IEEE MULTIMEDIA 2018, the Best Paper Award at the 2017 Pacific-Rim Conference on Multimedia (PCM), and is the coauthor of a paper that received the Best Student Paper Award in the IEEE International Conference on Image Processing 2018.

Shanshe Wang (Member, IEEE) received the B.S. degree from the Department of Mathematics, Heilongjiang University, Harbin, China, in 2004, the M.S. degree in computer software and theory from Northeast Petroleum University, Daqing, China, in 2010, and the Ph.D. degree in computer science from the Harbin Institute of Technology. He held a post-doctoral position with Peking University, Beijing, from 2016 to 2018. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, where he is currently a Research Assistant Professor. His current research interests include video compression and image and video quality assessment.

Yan Ye (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China in 1994 and 1997, respectively, and the Ph.D. degree in electrical engineering from the University of California at San Diego in 2002.

She leads the Video Standards and Implementation group at Alibaba Video Cloud, Alibaba Group U.S., where she oversees video coding standard development at the various standards organizations, software and hardware video codec implementations, and learning-based video processing and enhancement technology development. Prior to Alibaba, she was with the R&D Labs, InterDigital Communications; Image Technology Research at Dolby Laboratories; and Multimedia R&D and Standards at Qualcomm. She has been involved in the development of various video coding and streaming standards, including H.266/VVC, H.265/HEVC, scalable extension of H.264, MPEG DASH, and MPEG OMAF. She is an inventor of more than 200 granted and published U.S. patents and patent applications, and has published more than 50 papers in peer-reviewed journals and conferences. Her research interests include machine-learning-based video coding, processing and content analysis, virtual and augmented reality, video coding, processing and streaming, and real-time and immersive video communications.

Dr. Ye is an Editor of the VVC test model and the MPEG LCEVC standard, and was previously a Co-Editor of the scalable extension and the screen content coding extension of the HEVC standard.



Siwei Ma (Senior Member, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He held a post-doctoral position with the University of Southern California, Los Angeles, CA, USA, from 2005 to 2007. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, where he is currently a Professor. He has authored over 200 technical articles in refereed journals and proceedings in image and video coding, video processing, video streaming, and transmission. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Journal of Visual Communication and Image Representation*.

Wen Gao (Fellow, IEEE) received the Ph.D. degree in electronic engineering from The University of Tokyo, Japan, in 1991. He was a Professor of computer science with the Harbin Institute of Technology from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, from 1996 to 2006. He is currently a Professor of computer science with Peking University, Beijing, China. He has authored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ISCAS, ICME, and the ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. He served or serves on the Editorial Board for several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, the *EURASIP Journal of Image Communications*, and the *Journal of Visual Communication and Image Representation*.

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on August 31,2023 at 11:21:49 UTC from IEEE Xplore. Restrictions apply.