LSTFE-Net:Long Short-Term Feature Enhancement Network for Video Small Object Detection

Jinsheng Xiao[1], Yuanxu Wu[1], Yunhua Chen[2]*, Shurui Wang[1], Zhongyuan Wang[1], Jiayi Ma[1]
1. Wuhan University, China
2. Guangdong University of Technology,China
{xiaojs,whuwuyuanxu,shuruiwang}@whu.edu.cn, yhchen@gdut.edu.cn
wzy_hope@163.com, jyma2010@gmail.com

Abstract

Video small object detection is a difficult task due to the lack of object information. Recent methods focus on adding more temporal information to obtain more potent high-level features, which often fail to specify the most vital information for small objects, resulting in insufficient or inappropriate features. Since information from frames at different positions contributes differently to small objects, it is not ideal to assume that using one universal method will extract proper features. We find that context information from the long-term frame and temporal information from the short-term frame are two useful cues for video small object detection. To fully utilize these two cues, we propose a long short-term feature enhancement network (LSTFE-Net) for video small object detection. First, we develop a plug-and-play spatio-temporal feature alignment module to create temporal correspondences between the short-term and current frames. Then, we propose a frame selection module to select the long-term frame that can provide the most additional context information. Finally, we propose a long short-term feature aggregation module to fuse long short-term features. Compared to other state-of-the-art methods, our LSTFE-Net achieves 4.4% absolute boosts in AP on the FL-Drones dataset. More details can be found at https://github.com/xiaojs18/LSTFE-Net.

1. Introduction

Video small object detection plays an important role in many fields such as automatic driving, remote sensing, medical image, and industrial defect detection [26]. However, it is still a difficult task due to the lack of pixel information and the difficulty of feature extraction. Therefore, the topic of how to enhance the features of small objects has attracted great attention [1, 7, 16].

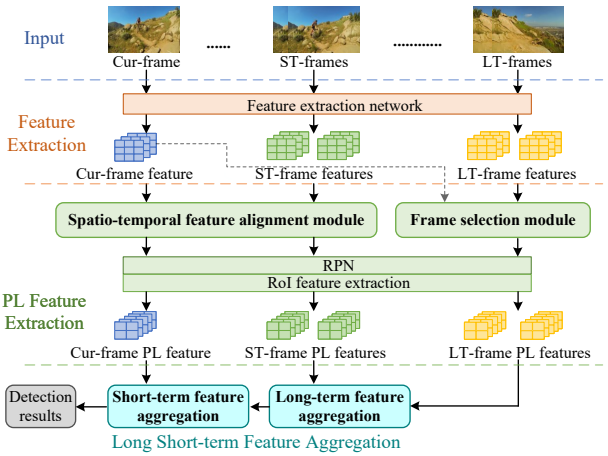

Figure 1. The architecture of the proposed LSTFE-Net. The current frame (Cur-frame), short-term frames (ST-frames) near the Cur-frame, and long-term frames (LT-frames) sampled from the whole video first go through the feature extraction network. Then the Cur-frame feature and ST-frame features are connected through the spatio-temporal feature alignment module, and the frame selection module searches the background context of LT-frame features. After getting the Proposal-Level (PL) features, the long short-term feature aggregation module finally integrates the long short-term features into the Cur-frame to make feature enhancement. Best viewed in color and zoomed in.

Some recent works have proved that the improvement of video small object detection performance requires full utilization of information in the temporal dimension. While detecting small objects in the current frame may suffer from many problems such as motion blur, low resolution, and too small size, effective modeling of information from other frames can help address these problems [2, 4, 9, 17]. There is a high similarity in nearby frames because of the strong time continuity between them, so it is natural to emphasize the importance of short-term frames, which are near the current frame. According to FGFA [28] and STSN [2],

*The corresponding author

short-term frame features can be aligned and aggregated to provide more useful information for small objects.

Context information is important for small object detection [16,25]. Because a large number of images are sampled from the same video for the same object, the background context information of the object is single. Additionally, only part of the frames in the video are sampled as current frames (such as 15 frames) while training, which results in the lack of real background context information and reduces the robustness of the training model. Compared to features in a short range, features from the whole video level can be more discriminative and robust [21, 23]. And it is noticed in prior works [9, 23] that more contextual information will be provided when using long-term frames sampled from the whole video.

The impacts of short-term and long-term frames in detection have been studied in recent methods. However, these methods have obvious disadvantages in both efficiency and accuracy, especially for small objects in videos. Some methods [22, 27, 28] extract information from the short-term frame and exploit the flow model to propagate features across adjacent frames, however, this is expensive because the flow model is hard to construct and transplant. Other methods [21, 23] focus on semantic information from long-term frames and incorporate randomly sampled long-term frames in detection, which causes uncertainty of detection performance and the loss of valuable information. Besides, these methods above cannot figure out the specific information that matters most for small objects from frames. Some methods [21–23, 27, 28] think the information in the video is single and miss considering distinct information from different frames, getting inadequate features. Other methods [3, 5] focus on extracting high-level features from the video which are not suitable for small objects due to their special properties.

To better mine information from both short-term frames and long-term frames, we propose a long short-term feature enhancement network (LSTFE-Net) for video small object detection. Specifically, the features of short-term frames are expected to correspond to the current frame in a low-cost and effective way, so a spatio-temporal feature alignment module is designed to propagate features across nearby frames. Further, in order to increase the benefit of aligned features while not increasing too much complexity of the model, a spatio-temporal feature aggregation method is also added. The context information is expected to be highlighted from the whole video, prompting a frame selection module to select the long-term frame feature. The goal is to make effective feature enhancement after the features are collected, and the establishment of connections between different features is enforced. A long short-term feature aggregation module is devised to aggregate features from the current frame, the short-term frames, and the long-term frames by stages. The performance of the proposed method is evaluated on the open dataset, and experiment results demonstrate that our method has obvious advantages in video small object detection. The architecture of the network is shown in Fig. 1.

Our main contributions are summarized as follows:

(1) An LSTFE-Net is proposed to effectively enhance small object features and improve detection performance.

(2) A plug-and-play spatio-temporal feature alignment module is designed for aligning features across adjacent frames. A flexible way to make Pixel-Level feature enhancement for small objects using aligned features is also explored. Combined with the Proposal-Level feature enhancement, this module achieves multi-level enhancement to improve the feature expressiveness. The whole module is easy to transplant and proved to be effective, which reveals its potential ability to benefit most of the works.

(3) A frame selection module is proposed to ensure the utilization of high-value input data, and it selects the long-term frame feature with the most context information. This module reinforces the network to automatically look for useful information for small objects, improving its stability and performance in video small object detection.

(4) To effectively integrate the long-term frame features and short-term frame features into the current frame, a long short-term feature aggregation module is proposed to aggregate different features in different stages. This enables the relations between Proposal-Level features to be built adaptively based on an attention mechanism, which also means our feature enhancement for small objects can be accomplished in a general and limitless way.

2. Related Works

Object detection from images has achieved considerable success and introduced some leading detectors in recent years. Based on the early proposed detectors, video object detection is intensively studied as a more challenging task. It is proved that exploiting information from other frames can significantly enhance the detection of the current frame, so temporal feature aggregation is gaining attention and introduced into video small object detection.

Object detection in videos: The appropriate use of the temporal information in the video to increase detection accuracy is a fundamental difficulty in video object detection. Adapting the image object detection algorithm to the video domain is challenging because of the complex spatial and temporal changes in videos. SeqNMS [10], TCN [15], and T-CNN [14] apply post-processing to video object detection to handle the complexities of videos. SeqNMS [10] links bounding boxes from different frames iff their Intersection over Union (IoU) is above some certain threshold and re-ranks those linked boxes; TCN [15] introduces the tubelet components and proposes a temporal convolu-

tional neural network to incorporate temporal information, which improves detection performance across frames. T-CNN [14] first uses the still-image object detection algorithm to get single frame detection results and associates the results using optical flows. However, none of these methods based on post-processing can be trained end-to-end, and their performance is yet to be improved. Our method, in contrast, directly uses temporal information in videos without post-processing and can be trained end-to-end, which brings great convenience.

Temporal feature aggregation: Recent works have also focused on ways to aggregate temporal features with different distances from the current frame, including short-term frame and long-term frame features, to enhance the feature of the object. DFF [27] uses a flow field predicted by FlowNet [13] to make the key frame feature align to short-term frames, thus reducing the redundant computation and accelerating the network. Unlike DFF, FGFA [28] mainly employs flow motion to make short-term frame features align to the current frame and then fuses these two kinds of features to improve detection accuracy. MANet [22] achieves the feature calibration module and the feature aggregation module on both pixel-level and instance-level based on FGFA and uses the motion pattern reasoning module to aggregate features from two different levels. Unlike short-term frames, long-term frames are usually sampled from the whole training video. Since there is little time continuity between the long-term frame and the current frame, the fusion of high-level semantic information is paid more attention to. To make more effective feature aggregation, SELSA [23] performs one innovative work by calculating the semantic similarity between current frames and long-term frames. STMN [24] uses the recurrent computation unit as the spatio-temporal memory module to pass semantic information between different frames. While using both local and global features to enhance the features of the current frames, MEGA [3] introduces a memory module to store more temporal features.

Video small object detection:A few researchers have proposed to improve the detection accuracy of small objects by leveraging temporal information in videos. Motion R-CNN [7] provides additional auxiliary information for small objects based on the frame differencing method and improves small object detection performance in optical remote sensing videos. DogFight [1] proposes to use a two-stage segmentation-based approach employing spatio-temporal attention cues, where objects are localized during the first stage and the object location is tracked and filtered using temporal information during the second stage.

Although both short-term frame features and long-term frame features have been used to improve detection performance, the information mined from these features is still not enough for them to build strong connections with the small object. We focus on several major problems and aim to get more powerful small object features through enhancement. To be specific, while DFF [27] and FGFA [28] require FlowNet which is difficult to construct and transplant, our method uses a plug-and-play spatio-temporal feature alignment module to make feature alignment and Pixel-Level enhancement. Instead of being simply randomly sampled from the video like MEGA [3], the long-term frame in this paper is selected by the frame selection module to enrich context information for small object detection. Unlike the above methods exclusively focusing on exploiting one kind of information from either the short-term frame or the long-term frame [22, 23, 27, 28], we fully consider the various information from both the short-term frame and long-term frame. The high-level features in this paper can work better for small objects than those in [3, 5] because of these purposeful designs.

3. Methods

3.1. Framework overview

The architecture of the LSTFE-Net is shown in Fig. 1. Besides the current frame, multiple short-term frames and long-term frames are also sampled from the video and input to the network. The feature extraction network first extracts the frame features of the input. Then different frame features are processed differently according to their characteristics. Because the short-term frames are sampled adjacent to the current frame, there is extensive temporal information between them. A plug-and-play spatio-temporal feature alignment module is designed to make alignment of the current frame feature and short-term frame features using cascaded deformable convolution blocks. This alignment can be described in the form of offsets in convolution, which helps the convolutional neural network sample more features from object regions to precisely enhance the small object of the current frame. The aligned features are further fused into the current frame feature on a low level.

It is inappropriate to make feature alignment between the current frame and long-term frames because there is no evident time continuity between them and their object features have great displacement in space. Considering that background context information is crucial and single to small object detection, long-term frames are employed to provide more background context information. It is natural to expect the long-term frames with the most context information to be further used in the network, so a frame selection module is devised. After feature embedding, it selects the long-term frame feature with the lowest similarity to the current frame feature. After the above modules, the Region Proposal Network (RPN) is used to generate proposals and Region of Interest (RoI) Align is applied to make feature extraction, getting the Proposal-Level features of different

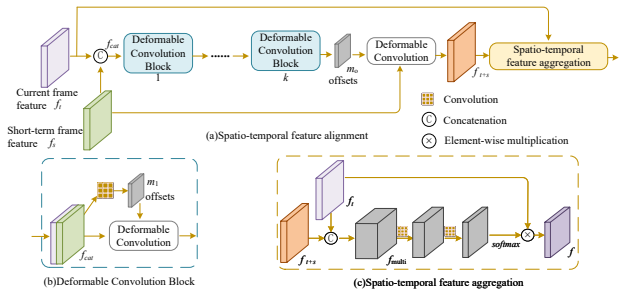

Figure 2. Spatio-temporal feature alignment module. The short-term frame feature f_s is first aligned based on deformable convolution, and then integrated into the current frame feature f_t using spatio-temporal feature aggregation. Best viewed in color and zoomed in.

frames.

Effectively integrating long-term frame features and short-term frame features into the current frame feature is also a key part of feature enhancement. To better integrate different kinds of frame features, we divide the long short-term feature aggregation into two stages: long-term feature aggregation and short-term feature aggregation. In the first stage, the feature aggregation between long-term frames and short-term frames is conducted, which provides additional context information for small objects. Next, the feature aggregation between short-term frames and the current frame is carried out in the second stage, and also the location differences between the frames are considered. This module models the relations between Proposal-Level features adaptively and universally, ensuring feature enhancement for small objects. The above algorithms will be introduced in detail in the following sections.

3.2. Spatio-temporal feature alignment module

There is a strong similarity between the current frame and the short-term frames. Features of the same object are typically not spatially aligned between frames due to the movement in videos. Hence direct fusion of features between frames will result in feature interference, ghost effects, and even lower detection performance. Deformable convolution [8] learns multiple offsets of the spatial location during convolution. We design a spatio-temporal feature alignment module based on deformable convolution to learn offsets and make feature alignment between frames, and a spatio-temporal feature aggregation module is proposed to conduct Pixel-Level feature enhancement of the current frame. The specific architecture is shown in Fig. 2.

Given a frame I_t at time t and a nearby short-term frame I_s, let f_t and f_s indicate spatial features through the feature extraction network from frame I_t and I_s, as shown in Fig. 2 (a). Suppose the shape of f_t or f_s is $[C, H, W]$, where C stands for the channels, H for the height of tensors, and W for the width of tensors. To fuse the features, f_t and f_s are concatenated to get f_{cat}. Then f_{cat} is fed to a deformable convolutional block, as shown in Fig. 2 (b). The block generates offsets $m_1 \in [2 \times 9, H, W]$ with 3×3 convolution, where 9 denotes kernel size 3×3, 2 denotes offsets in two directions: x and y. Finally, m_1 and f_{cat} are fed to a deformable convolutional layer to get the aligned feature. Our deformable convolutional block is plug-and-play and cascaded, which means multiple blocks can be used in series to make multiple times feature alignments. The final block outputs the final offsets m_o, and m_o and short-term frame feature f_s are fed to a deformable convolutional layer to get the aligned short-term feature f_{t+s}. In general, the temporal information between the current frame and the short-term frame is used to estimate the spatial offset between object features, which is further used to make the short-term frame feature align with the current frame feature.

To effectively integrate the aligned short-term frame feature into the current frame feature, a spatio-temporal feature aggregation module is designed to make the adaptive information fuse. The adaptive weight can be expressed by formula 1.

$$\omega(f_t, f_{t+s}) = \rho(l(f_t, f_{t+s})) \tag{1}$$

where l is a spatio-temporal function that describes spatio-temporal connections between f_t and f_{t+s}, and ρ is the mask function used to calculate the adaptive weight. To fully use temporal information between frames, the frame differencing method is introduced into l, which concatenates $f_t - f_{t+s}$,$f_{t+s} - f_t$, f_t, and f_{t+s} into f_{multi}. f_{multi} is then fed to mask function ρ. It goes through two convolutional layers to squeeze the number of channels and fully fuse the information. To improve the generalization ability of the model, the *softmax* function is then used to generate the final adaptive weight $\omega(f_t, f_{t+s})$. as shown in Fig. 2 (c). Finally, the enhanced current frame feature f is computed by

$$f = \sum_{f_j \in (f_t)} \{\omega(f_t, f_j) \otimes f_t\} \tag{2}$$

where (f_t) denotes a collection of aligned short-term frame features near f_t, $\otimes$ denotes element-wise multiplication. In this paper, multiple short-term frames are sampled, aligned to the current frame, and adaptively integrated into the current frame to enhance the current frame feature on Pixel-Level.

3.3. Frame selection module

Because a large number of images are sampled in the same video for the same object, the background context information of the object is very similar. Object detection algorithms based on deep learning tend to use this similarity to make the feature enhancement for the current object.

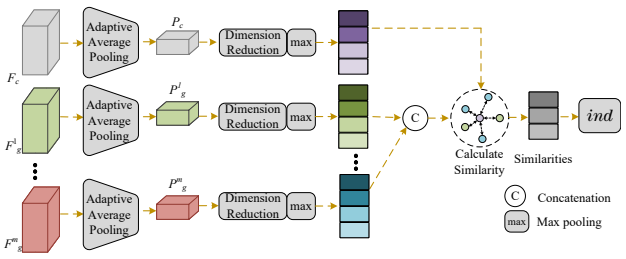

Figure 3. Frame selection module. The current frame feature F_c and m long-term frame features $\{F_g^i\}_{i=1}^m$ first get the feature embedding. Then similarities between them are calculated and finally, the index of the long-term frame with the lowest similarity is output. Best viewed in color and zoomed in.

However, this strategy results in the poor generalization performance of the training model, especially for training data with a high degree of similarity. Considering the significance of background context information for small object detection, our network samples long-term frames from the whole video and tries to acquire training frames with more background differences. The frame selection module can select the long-term frame with the lowest similarity to the current frame, which is further used to enrich background context information. The architecture of the frame selection module is shown in Fig. 3.

Given the current frame feature F_c and the sampled long-term frame feature F_g^i, this module tries to select the long-term frame that is most dissimilar to F_c from m candidate long-term frames to be used for Proposal-Level feature aggregation. As a result, this module outputs the index of the long-term frame with the lowest similarity, which can be interpreted as

$$ind = \underset{i\in[1,\mathrm{m}]}{\operatorname{argmin}} \left\{ Sim(\, g(F_c)\,, g(F_g^i)\,) \right\} \tag{3}$$

where ind is the index of the selected long-term frame, Sim is the similarity function used to calculate the similarity between different input features, g represents the feature embedding function, $i \in [1,\mathrm{m}]$ indicates the i^{th} candidate long-term frame feature, and argmin is used to calculate the index of the long-term frame with the lowest similarity. Similarity function Sim can be formulated as

$$Sim(X,Y) = \frac{X \cdot Y}{\sqrt{dim}} \tag{4}$$

According to formula 4, cosine distance is used to calculate the similarity between the current frame feature $g(F_c)$ and long-term frame feature $g(F_g^i)$, and dim represents the dimension of the input feature. In formula 3, feature embedding function g incorporates adaptive pooling, dimension reduction, and max pooling, as shown in Fig. 3. Suppose the shape of the input feature F_c is $[N, C, H, W]$, where

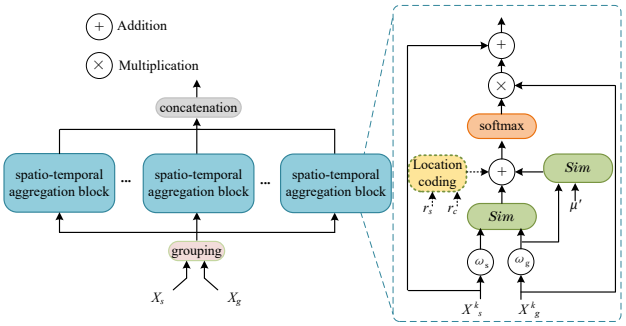

Figure 4. Details of long short-term feature aggregation. Input features are first grouped, then make spatio-temporal aggregation, and finally concatenated to get the enhanced feature. The "Location coding" is used in the second stage: short-term feature aggregation. Best viewed in color and zoomed in.

N, C, H, W stands for batch size, channels, height, and width, respectively. The feature is processed by adaptive average pooling on spatial domains and reshaped to a 4D tensor $[N, C, 1, 1]$ to get the global spatial feature of the image. To calculate the similarity more conveniently, the feature then reduces the dimension and is reshaped to the 2D tensor $[N, C]$, and the most representative feature is selected by max pooling. Finally, the 4D feature $[N, C, H, W]$ is reshaped to 2D $[1, C]$.

For selected index ind, the long-term frame feature F_g^{ind} has the lowest background context similarity with the current frame feature. In training, it provides more diverse real context information for small object detection.

3.4. Long short-term feature aggregation module

Short-term frames in the local temporal range are employed in several recent works to enhance the current frame [2, 4, 5]. Nevertheless, only fusing short-term frames can not fully use the information from the whole video and has significant limitations. The long-term frame is therefore introduced to supplement context information. Motivated by the multi-head attention mechanism [12], a long short-term feature aggregation module that conducts feature aggregation on Proposal-Level is proposed. It effectively integrates the long-term frame features and short-term frame features into the current frame features. The long short-term feature aggregation includes two stages: long-term feature aggregation and short-term feature aggregation, the details of long short-term feature aggregation are shown in Fig. 4.

Long-term feature aggregation: The input of the long-term feature aggregation module is the short-term frame Proposal-Level features $X_s \in [N_1, C]$ and the long-term frame Proposal-Level features $X_g \in [N_2, C]$, where N_1 and N_2 represent the number of features reserved and C represents the channels of features. To combine information from different channels and subspaces, input features

are first divided into K groups. The features after grouping can be formulated as

$$X^k = X[:, (k-1)\frac{C}{K} : k\frac{C}{K}] \tag{5}$$

where X could be long-term frame Proposal-Level features X_g or short-term frame Proposal-Level features X_s, C represents the channels of features, K is the number of groups, k means the k^{th} group, and X^k denotes the segment of the k^{th} group on the channel dimension. Every group makes spatio-temporal aggregation and is concatenated to each other.

$$X_{g+s} = concat(\varphi(X_s^k, X_g^k)) \tag{6}$$

where X_{g+s} denotes the enhanced short-term frame features, $concat$ denotes concatenation, and the spatio-temporal aggregation function φ effectively aggregates long-term frame features and short-term frame features. φ can be formulated as

$$\begin{cases} \varphi(X_s^k, X_g^k) = X_s^k + \mathrm{softmax}(\omega) \cdot X_g^k \\ \omega = Sim(\omega_s \cdot X_s^k, \omega_g \cdot X_g^k) + Sim(\mu', \omega_g \cdot X_g^k) \end{cases} \tag{7}$$

To obtain a plug-and-play module, the spatio-temporal aggregation function φ utilizes residual connection (as shown in the dotted box in Fig. 4). X_g^k and X_s^k are fused in residual function, attention ω serves as a correlation weight that consists of two parts: the cosine similarity between short-term frame features and long-term frame features to associate short-term frame and long-term frame, and the cosine similarity between long-term frame features and a group of learnable weights μ' to associate different channels of the long-term frame, ω_s and ω_g are linear transformation matrixes and also fully connected layers. Definition of Sim is same as formula 4.

Short-term feature aggregation: Only high-level semantic information of the long-term frame is integrated into the short-term frame during long-term feature aggregation. Considering the strong time continuity between the short-term frame and the current frame, the location information of the short-term frame is also fused into the current frame, as shown in Fig. 4 "location coding". Let $r_s = \{x_s, y_s, h_s, w_s\}$ be the location information of Regions of Interest (RoIs) in the short-term frame which includes the center point x_s, y_s and height and width h_s, w_s of each RoI, and let r_c be the location information of RoIs in the current frame. The location correlation weight can be represented as

$$\begin{cases} \omega_r = \mathrm{relu}(\omega_{cs} \cdot \psi(r_c, r_s)) \\ \psi(r_c, r_s) = \left\{ \log(\frac{|x_c - x_s|}{w_c}), \log(\frac{|y_c - y_s|}{h_c}), \log(\frac{w_c}{w_s}), \log(\frac{h_c}{h_s}) \right\} \end{cases} \tag{8}$$

where ω_r denotes location correlation weight, the nonlinear function relu is used to reduce redundant information, ω_{cs} is the linear transformation matrix and also a fully connected layer in the network, and ψ is utilized to code the location information of the current frame and the short-term frame to make the module translation invariant. The correlation weight of short-term feature aggregation can be represented as

$$\begin{aligned} \omega = & Sim(\omega_c \cdot X_c^k, \omega_{g+s} \cdot X_{g+s}^k) \\ & + Sim(\mu'', \omega_{g+s} \cdot X_{g+s}^k) + \omega_r^k \end{aligned} \tag{9}$$

where X_c^k stands for the current frame Proposal-Level features after grouping, ω_{g+s} , ω_c are fully connected layers used to reshape the feature, ω is the adaptive weight including three parts: the cosine similarity between enhanced short-term frame features and current frame features, the cosine similarity between enhanced short-term frame features and a group of learnable weights μ'', and the location correlation weight between the current frame and short-term frame after grouping. The other steps of short-term feature aggregation are the same as long-term feature aggregation and needless to be described again.

4. Experiments

4.1. Experiment Setup

Datasets and Evaluation Setup: We conduct experiments on ImageNet-VID [20] and FL-Drones [19]. ImageNet-VID is a widely used large-scale benchmark for video object detection, which contains 3862 videos in the training set, 555 videos in the validation set, and a total of 30 categories. mAP is used for evaluation on ImageNet-VID. To evaluate the performance of the proposed method on video small object detection, we also make use of the FL-Drones dataset consisting of 14 videos and 38948 frames. The flying drones in the dataset can be regarded as small objects because the average size of annotated drones is 25×16 and the frame resolutions are 640×480 and 752×480. This dataset is quite challenging due to the extreme illumination and relatively small size. According to the author [19], half of the data is used for training, and the other half is used for testing. And AP is used as the evaluation metric for FL-Drones since it is a video dataset for small objects.

Implementation Details: The proposed algorithm is trained and tested using 4 16GB NVIDIA Tesla V100. The baseline used in this paper is Faster R-CNN. Following the multi-task loss in baseline, the whole network is optimized with both classification and regression losses. SGD is selected as the optimizer and the model is trained for 120000 iterations on ImageNet-VID and 66000 iterations on FL-Drones. The learning rate is set to 0.001 and updated using WarmupMultiStepLR [18], which means the model first has 500 warm-up iterations to improve stability.

4.2. Comparison with state-of-the-art

ImageNet-VID dataset: Comparison results on this dataset are shown in Table 1. All algorithms adopt the same feature extraction network to be compared fairly, and the results of the state-of-the-art works are collected from published papers.

Table 1. Quantitative results on the ImageNet-VID dataset (%).

Method	Backbone	mAP
FGFA [28]	ResNet101	76.3
STMN [24]	ResNet101	80.5
LongRange [21]	ResNet101	81.0
D&T [6]	ResNet101	75.8
MANet [22]	ResNet101	78.1
STSN [2]	ResNet101	78.9
RDN [5]	ResNet101	81.8
SELSA [23]	ResNet101	80.3
MEGA [3]	ResNet101	82.9
LSTFE-Net	ResNet101	**83.4**

It is obvious that our method outperforms other state-of-the-art methods and achieves the highest mAP of 83.4%. Results show that the LSTFE-Net can better collect the information from long-term and short-term frames and enhance the feature.

FL-Drones dataset: As a video dataset for small objects, few results of other classical methods can be found on this dataset which means we need to conduct experiments using these methods ourselves. For a more comprehensive comparison, ResNet50 and ResNet101 are respectively used as the feature extraction network for analysis. The results are shown in Table 2.

According to Table 2, our method achieves the best small object detection performance no matter what the feature extraction network is. Our LSTFE-Net obtains the highest AP of 37.8% and outperforms the strongest competitor TransVOD by 0.9% AP when ResNet50 is used. For ResNet101, our LSTFE-Net achieves an AP of 46.8%, 4.4% higher than MEGA. Fig. 5 shows some qualitative comparison results of our LSTFE-Net versus other state-of-the-art works.

4.3. Ablation Studies

We conduct ablation studies to validate the effectiveness of each component in our model on the FL-Drones dataset.

Spatio-temporal feature alignment module: k cascaded deformable convolutional blocks are applied in this module, and experiments are conducted in this section to study the influence of the value of k on spatio-temporal modeling. The short-term frame and the long-term frame

Table 2. Quantitative results on the FL-Drones dataset (%).

Method	AP	
	ResNet50	ResNet101
DFF [27]	19.5	22.4
FGFA [28]	21.2	24.5
TransVOD [11]	36.9	35.2
RDN [5]	35.9	42.1
MEGA [3]	36.5	42.4
LSTFE-Net	**37.8**	**46.8**

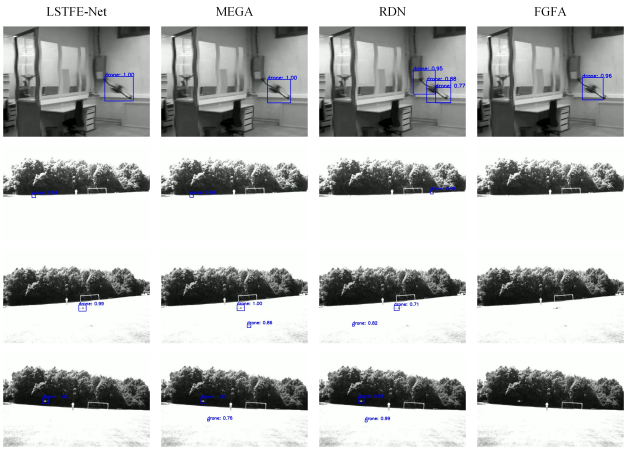

Figure 5. Qualitative comparison of our LSTFE-Net versus other state-of-the-art works. Blue boxes represent the detection. Best viewed in color and zoomed in.

Table 3. Ablation study of the spatio-temporal feature alignment module(%).

	Backbone	AP
$k=0$	ResNet101	42.1
$k=1$	ResNet101	**45.4**
$k=2$	ResNet101	39.4
$k=3$	ResNet101	39.3

are first added to the baseline, and then the value of k is adjusted. Results are shown in Table 3:

Table 3 shows that the network achieves an AP of 42.1% if the spatio-temporal feature alignment module is not used. When $k=1$, AP increases by 3.3% to 45.4%. As k increases from 1 to 3, AP decreases gradually. To explore the reasons for the decreasing detection performance, the offset matrixes with different values of k are visualized, as shown in Fig. 6.

The offset matrix with $k=1$ is visualized in Fig. 6(b), where it is obvious that offsets are generally larger at the

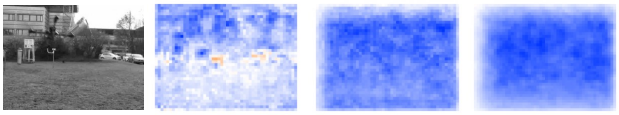

Figure 6. Visualization of offset matrix. The deeper the color, the larger the offset.

Table 4. Ablation study of frame selection module (%).

Number of long-term frames	Backbone	AP	
		Long-term frames	Frame selection
1	ResNet101	29.5	29.9
2	ResNet101	35.1	**35.9**
3	ResNet101	33.9	34.8
4	ResNet101	29.6	30.7

moving target and smaller in the stationary background area. This indicates that the network can better model the correlation between the same object of different frames. With the increase of k, temporal information is mixed up, which makes the offset learned by the spatio-temporal feature alignment module gradually disordered. When $k = 3$, the offset matrix can hardly represent the relationship between frames. Adding the offset in the convolution will hinder the normal extraction of features and make the detection result worse.

Frame selection module: The detection results before and after using the frame selection module are compared. To emphasize the impact of the frame selection module, the number of short-term frames is set to 0. After m long-term frames are sampled, additional $2m$ frames are then sampled and m frames are selected by the frame selection module. The results are shown in Table 4:

Table 4 demonstrates how the number of long-term frames will affect the detection performance of small objects. The promotion effects of the frame selection module vary for different numbers of sampled long-term frames. When there are 4 sampled frames, AP improves most (1.1%). And when the number of sampled frames is 2, AP improves by 0.8% to 35.9% (highest).

Long short-term feature aggregation module: In this section, concatenation is used as a simple feature aggregation method to be compared. The only difference between short-term feature aggregation and long-term feature aggregation is the utilization of location coding, so the contribution of location coding is verified. Short-term frames and long-term frames are first added to the baseline, and then simple feature aggregation, long short-term feature aggregation without location coding, and long short-term feature aggregation are used for experiments separately. The results are shown in Table 5.

Table 5. Ablation study of long short-term feature aggregation module (%).

Method	Backbone	AP
simple feature aggregation	ResNet101	39.9
long short-term feature aggregation without location coding	ResNet101	41.2
long short-term feature aggregation	ResNet101	**42.1**

Table 5 shows that the simple feature aggregation method obtains an AP of 39.9%, while the proposed long short-term feature aggregation method can improve AP to 42.1%. Additionally, the AP drops to 41.2% when the location coding is removed, which means that the location coding can bring a gain of 0.9% AP.

5. Conclusion

For small objects, we propose an LSTFE-Net to enhance the feature with two important sources of information from the video. A spatio-temporal feature alignment module is proposed to model the temporal information between the short-term frame and the current frame. It uses deformable convolution to make connections between objects of different frames, and conducts feature enhancement on Pixel-Level. Considering that the background context information of the object is single, the frame selection module is used to select the long-term frame with the most distinct feature from the current frame, which is further used to enrich the vital context information for small object detection. To effectively fuse the features of different types of frames, we develop a long short-term feature aggregation module, where context information from the long-term frames and the temporal information from the short-term frames are incorporated into the current frame by stages. Experiments are carried out on ImageNet-VID and FL-Drones to compare our method with other state-of-the-art methods and conduct ablation studies. The results show that the proposed method achieves superior small object detection performance in videos.

Acknowledgement

This work is supported by National Natural Science Foundation of China (U1903214), Major Science and Technology Projects of Jilin Province (20210301030GX), Natural Science Foundation of Guangdong Province (2021A1515012233)

References

[1] Muhammad Waseem Ashraf, Waqas Sultani, and Mubarak Shah. Dogfight: Detecting drones from drones videos. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, pages 7067–7076, 2021. 1, 3

[2] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018. 1, 5, 7

[3] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10337–10346, 2020. 2, 3, 7

[4] Yiming Cui, Liqi Yan, Zhiwen Cao, and Dongfang Liu. Tf-blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8138–8147, 2021. 1, 5

[5] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7032, 2019. 2, 3, 5, 7

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE international conference on computer vision*, pages 3038–3046, 2017. 7

[7] Jie Feng, Yuping Liang, Zhanwei Ye, Xiande Wu, Dening Zeng, Xiangrong Zhang, and Xu Tang. Small object detection in optical remote sensing video with motion guided r-cnn. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 272–275. IEEE, 2020. 1, 3

[8] Hang Gao, Xizhou Zhu, Steve Lin, and Jifeng Dai. Deformable kernels: Adapting effective receptive fields for object deformation. *arXiv preprint arXiv:1910.02940*, 2019. 4

[9] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *European conference on computer vision*, pages 431–446. Springer, 2020. 1, 2

[10] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016. 2

[11] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021. 7

[12] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018. 5

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3

[14] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017. 2, 3

[15] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 817–825, 2016. 2

[16] Jeong-Seon Lim, Marcella Astrid, Hyun-Jin Yoon, and Seung-Ik Lee. Small object detection using context and attention. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 181–186. IEEE, 2021. 1, 2

[17] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 1

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[19] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Detecting flying objects using a single moving camera. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):879–892, 2016. 6

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[21] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9756–9764, 2019. 2, 7

[22] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 542–557, 2018. 2, 3, 7

[23] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9217–9225, 2019. 2, 3, 7

[24] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 485–501, 2018. 3, 7

[25] Jinsheng Xiao, Haowen Guo, Jian Zhou, Tao Zhao, Qiuze Yu, and Yunhua Chen. Tiny object detection with context enhancement and feature purification. *Expert Systems with Applications*, 211:118665, 2023. 2

[26] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2022. 1

[27] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7218, 2018. 2, 3, 7

[28] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017. 1, 2, 3, 7