Video Tiny-Object Detection Guided by the Spatial-Temporal Motion Information

Xin Yang†, Gang Wang†*
Academy of Military Sciences, Beijing, China
dlmuwanggang@163.com

Weiming Hu, Jin Gao, Shubo Lin
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Liang Li, Kai Gao, Yizheng Wang
Academy of Military Sciences, Beijing, China

Abstract

Detecting tiny/small objects (e.g., drone targets) in videos is highly desired in many realistic scenarios. Nevertheless, current object detection algorithms can hardly recognize tiny targets against extremely complex backgrounds. To address this problem, we propose a motion-guided video tiny-object detection method (MG-VTOD), in which the spatial-temporal motion strength maps play an important role in object searching and locating. Inspired by the biological retinal structure, we compute the motion strength using a sequential frame cube that has been aligned and registered. Subsequently, the motion strength maps are employed to enhance the potential areas of the moving targets, thereby facilitating the target detection procedure. Experimental results obtained on the Anti-UAV-2021 dataset validate that the proposed MG-VTOD method significantly outperforms the competing object detection methods.

1. Introduction

Tiny/small object detection is an important task in many realistic scenarios, *e.g.*, locating unauthorized flying targets around airports through cameras. Nevertheless, current object detection algorithms can hardly recognize tiny targets against extremely complex backgrounds. It is highly desired to develop intelligent techniques that can locate and recognize tiny object with low miss rate and low false alarm rate. In the computer vision field, video tiny-object detection (VTOD) is a particular case of visual object detection. Since tiny objects have limited appearance features, VTOD is among the most challenging tasks of visual object detection, as illustrated in Fig. 1.

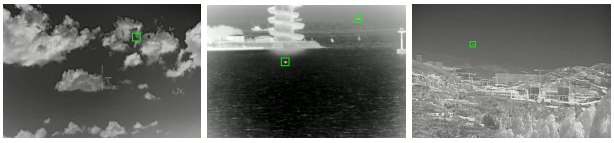

Figure 1. Example video frames containing tiny objects against complex backgrounds.

As a fundamental task, visual object detection has been intensively studied in the past decades [47]. With the big success of deep learning, many advanced and popular object detection algorithms are developed based on deep neural networks [17]. These methods can be roughly divided into two categories: two-stage and one-stage algorithms. In two-stage algorithms, visual object detection is implemented by a target positioning step and a subsequent target recognition step. Representative methods include the R-CNN [11] and Faster R-CNN [32] methods. The other type of object detection, namely one-stage algorithms, predicts the target categories and location information simultaneously. Among them, the YOLO [29] serial methods have been widely acknowledged.

Compared to object detection in static images [39], video object detection is more difficult in practical scenarios [18]. The accuracy of video object detection suffers from deformable object appearances and degraded imaging quality, *e.g.*, motion blur, object occlusion, video defocus, rare pos-

*Corresponding author
† Equal contributions

tures [7]. Therefore, increasing attention has been paid to object detection in videos. Existing such algorithms can be mainly categorized into feature-flow-based, tracking-based, and post-processing-based algorithms.

Although the aforementioned video object detection methods are able to achieve considerable detection performance on some large-scale datasets, they underperform when detecting tiny objects in complex backgrounds. The reason is that the size of targets is tiny and sometimes the appearance information is missing [24]. There are also many fictitious artifacts within the images that have similar appearances to the desired small targets, leading to serious false detection [38]. As a result, it is very challenging to identify the tiny targets for deep convolutional neural networks that only utilize static image features [44].

Comparatively, biological visual systems can recognize tiny objects in complex backgrounds easily. The retina converts light signals into electrical signals and integrates the signals into the cortex through a series of preprocessing operations [1]. The retina is not only a photoelectric converter but also contains two information processing channels, namely a motion-processing pathway and a appearance-processing pathway. The motion-processing pathway is responsible for extracting movement information, while the appearance-processing pathway takes charge of processing detailed visual information [28]. Retinal signals are then transmitted into the lateral geniculate nucleus or the visual cortex, *e.g.*, the primary visual cortex (V1). Correspondingly, there are also two visual pathways in the cortex: the dorsal pathway, and the ventral pathway [6]. The ventral pathway, *a.k.a.* object recognition pathway, generally starts from the V1 area and extends to the superior parietal lobule through the V2, V4, inferior temporal cortex, *etc*. The dorsal pathway starts from the V1 area and reaches the superior parietal lobule area through the V2, V3, middle temporal(MT), *etc*. The dorsal pathway is associated with the location and movement of the object. The two channels mutually modulate each other and jointly process visual signals [9].

Aiming at the challenge of VODT, we propose to both use the static image features and the visual motion features inspired by the mechanism of biological visual systems. The visual motion features are extracted based on the registered sequential frames, which is employed as an enhancement module for the object detector. We implement the backbone of the object detection architecture based on the widely validated deep convolutional neural networks. Consequently, the motion strength computation model can guide the conventional object detector to better locate the tiny objects in video frames. The proposed method as well as the competing methods are extensively evaluated on the publicly available Anti-UAV-2021 Challenge dataset [46].

2. Related work

2.1. Visual object detection

Visual object detection methods can be roughly divided into two categories: two-stage and one-stage algorithms.

In two-stage algorithms, visual object detection is implemented by a target positioning step and a subsequent target recognition step. Representative methods include the R-CNN [11], Fast R-CNN [10], Faster R-CNN [32] and Mask R-CNN [14] methods. The R-CNN method [11] selects thousands of region proposals from the input image by a selective search module and then scales each region proposal into a feature extraction network. The feature extraction network model [20] obtains a high-dimensional feature vector and then trains a support vector machine [43] classifier to determine if the region contains a target or not. As an improved version, the Fast R-CNN method [10] trains object classification and detection box regression in the same network framework, considerably reducing the training computational workload and prediction time. Furthermore, the Faster R-CNN method [32] uses a region proposal network to replace the selective search module, in which the region proposal network shares features with the whole detection network. And the anchor box was introduced to adapt to the change of the target shape, which improved the detection accuracy and speed. The advanced Mask R-CNN method [14] was built on the Faster R-CNN method, adding a mask prediction branch based on a fully convolutional network to each region of interest. In this way, this method can obtain the image segmentation result on pixel level as well as the object detection result simultaneously.

The other type of object detection, namely one-stage algorithms, predicts the target categories and location information simultaneously. Among them, the YOLOv1 [29], YOLOv2 [30], YOLOv3 [31], YOLOv4 [3] and YOLOv5 [1] methods have been widely acknowledged. The YOLOv1 method [29] avoids generating a series of proposals and directly performs regression and classification on the entire input image, thereby improving the object detection speed considerably. The YOLOv2 method [30] further employs batch normalization, passthrough layers, multi-scale resolution training, and other strategies to increase the detection accuracy. Then, the YOLOv3 method [31] uses a number of residual blocks and feature pyramid networks in the detection architecture, significantly improving the detection performance of small targets. In the YOLOv4 method [3], the authors used multi-anchors to recognize a single object, easing the imbalance problem between positive and negative samples. Besides, the YOLOv4 method employs the complete intersection-over-union loss to compute the cost function, which can better describe the difference between the detection result and the ground truth. The latest YOLOv5

[1] https://github.com/ultralytics/yolov5

method uses adaptive learning bounding box anchors to predict the areas of the potential targets. Also, the YOLOv5 method further develops the cross-stage partial networks to upgrade the whole model. In addition, the YOLOv5 series includes several versions that have different characteristics and advantages, depending on the application scenarios.

2.2. Video object detection

Video object detection has gained increasing attention from researchers over the past several years. Compared to object detection in static images [39], video object detection is more challenging and more important in practical scenarios [18]. As aforementioned, popular video object detection algorithms can be roughly grouped into feature-flow-based, tracking-based, and post-processing-based algorithms.

The feature-flow-based methods use either a feature propagation or a feature aggregation scheme to enhance the feature of blurred frames. The deep feature flow (DFF) method [49] only extracts features on keyframes. Optical flow computing is utilized to establish a temporal correspondence between objects. The keyframe features are then propagated by optical flow information, which embodies temporal correspondence between objects, through a bilinear interpolation process to obtain the features of non-keyframes. The FGFA method proposed in [48] aligns the feature maps appropriately through optical flow, and then aggregates the feature maps from adjacent frames to obtain features of the current frame. Consequently, the problem of motion blur can be eased. Another method named MEGA [4] weights aggregated features through cosine similarity of target features in video frames. In addition, it uses global semantic information to enhance the feature representation. The method proposed in [8] gradually enhances the feature of the aggregated proposal through the reference frames. However, such methods underperform when the optical flow is inaccurately obtained.

The tracking-based methods use a tracker to accompany the object detection progress, combining both the tracking and detection results. For example, the Detect or Track method [25] determines to choose the results of object detection or object tracking by a scheduler network. The CaT-Det method [26] consists of two deep convolutional models that form a cascaded detector, and an additional tracker to predict regions of interests based on historic detections, thereby leveraging the temporal correlation in videos to accelerate the detection efficiency.

As for post-processing-based methods, they optimize the detection result through the spatio-temporal consistency of the video target. When the image quality is degraded, the obtained target confidence score is usually low. The Seq-NMS method [13] links the box of the adjacent frame and uses the box with the high confidence score from the adjacent frame to correct the box of the current frame. In addition, the Seq-Bbox Matching method [2] matches the boxes to a tubule. The bounding box of the same tubule is adjusted by the average score. This scheme of bounding box linking helps to reduce missed detection and to improve detection recall rate.

3. MG-VODT: Model design and implementation

Inspired by biological visual systems, it is highly believed that motion information play an important role in tiny object searching and recognition. In this section, we design an tiny video-object detection method guided by the visual motion features. The visual motion information can be obtained by the sequential frame cuboid elaborated in the previous section. The motion features are subsequently used to guide tiny/small object detection in infrared videos. The whole detector is implemented as follows.

3.1. Visual motion strength computational module

For both artificial or biological vision systems, visual motion features play an important role in vision-related tasks, *e.g.*, target detection [48], object tracking [50] and video denoising [45]. Conventional popular motion feature extraction methods include optical flow [16] and frame difference [21]. Nevertheless, optical flow extraction methods, which are computationally heavy, are sensitive to illumination variation and random noise. Moreover, the frame difference can hardly represent motion features accurately, especially in scenarios of dynamic backgrounds and degraded frames. In this paper, we propose to use the bio-inspired model to extract visual motion strength.

In biological visual systems, the retina area morphologically includes photoreceptor cells, horizontal cells, bipolar cells, amacrine cells and ganglion cells [40]. When processing the visual information, photoreceptor cells receive light stimuli and convert light signals into electrical signals. The horizontal cells are laterally connected with photoreceptor cells, receiving signals from photoreceptors while feeding inhibitive control signals back [12]. Subsequently, the electrical signals are transmitted to ganglion cells through bipolar cells [15]. The ganglion cells work together with bipolar cells and amacrine cells [27] to process the visual information, and then project the coded signals to the lateral geniculate nucleus or the primary visual cortex [34].

Investigators in the neuroscience field have found that in biological visual systems, the visual information is processed in two paralleled pathways, *i.e.*, the motion-processing pathway and appearance-processing pathway. The two pathways process visual motion information and static appearance visual information, respectively. Specifically, in the motion-processing pathway of the inner plexiform, the amacrine cell has one end connected with the

bipolar cell, and the other end connected with the ganglion cell. Amacrine cells act as a temporal high-pass filter bank [41] that enhances the temporal and spatial variation. Its output depends on both the current input and the previous output, like the solution a difference equation. The ganglion cells participate in both the motion-processing pathway and appearance-processing pathways [33]. On the one hand, they act as a spatial low-pass filter, retaining the temporal and spatial variation information yield by the amacrine cells. On the other hand, they subsequently boost the contrast information, amplifying the visual motion signals. We implement the motion-processing pathway computation model based on the adjacent frames that have gone by. These adjacent frames can be reshaped into a sequential frame cuboid. Besides, to address the problem of camera movement (including camera rotation, pitching and rolling), the images in the frame cuboid are aligned and registered according to the backgrounds. It is worth noting that since the motion strength are obtained based on multiple frames, the designed method above can effectively suppress random noise.

3.2. Motion-guided networks

Tiny object sometimes occupy only a few pixels in each frame, showing limited visual appearance features. Although many solutions to tiny object recognition have been developed, *e.g.*, blob detection methods [37] and tiny objection detectors for static images [5], tiny object detection yet remain a very difficult task, especially in complex backgrounds. It is widely known that visual motion information can facilitate attention guidance and object searching, for both artificial intelligent systems and biological vision systems.

Since target searching and tracking are crucial for animal survival, visual motion information processing is very important for biological visual systems. Once the target moves, it arouses temporal and spatial variations, leading to visual attention responses to the changing areas. Many visual areas in the brain, formed by various types of neurons, take part in this critical task. Besides, these visual areas integrate visual motion information together with spatial appearance information hierarchically [19]. In a biological retina, the motion-processing pathway and the appearance-processing pathway focus on temporal and spatial visual features, respectively. It is intuitive to utilize the motion-processing response to enhance the potential target areas and suppress the undesired backgrounds.

To this end, we propose to integrate the motion strength into deep convolutional neural networks, thereby building a motion-guided visual object detection model. In this way, we can employ the visual motion information as an attention guidance module, while retaining the spatial appearance information that represented by the deep neural networks.

3.3. MG-VTOD model

Aiming at object detection in infrared videos, we have proposed a scheme that combines the Magno model with deep convolutional neural networks. We next implement such an algorithm based on the popular YOLOv5 method, which is famous for its high executive efficiency and high accuracy. The original YOLOv5 method includes four versions that have different model capacities, namely YOLOv5-s, YOLOv5-m, YOLOv5-l, and YOLOv5-x. We employ the YOLOv5-s as the backbone, and utilize the motion strength to guide the object detection process, thereby proposing a MG-VTOD object detector that is illustrated in Fig. 2.

The motion computational module outputs the visual motion strength, yielding a one-channel grayscale map, in which the motion areas have large intensity (*e.g.*, 1), while the static backgrounds have small intensity values (*e.g.*, 0). Meanwhile, the original video frames are preprocessed to improve the visual quality. Following the working mechanism of the YOLOv5 method, all the input channels are sliced and sent to the convolutional layers. Subsequently, the convolutional responses of the motion strength maps and the preprocessed video frames are concatenated together. As a result, in the subsequent processing procedures, the areas that have larger motion strength response values are more likely to be activated.

Within the MG-VTOD model, we employ the cross-stage partial (CSP) network [36] that can mitigate the problem of heavy inference computations. On the one hand, the visual features are learned and represented through a series of sequential convolution operations. On the other hand, the CSP network integrates the feature maps from the beginning and the end of a specified network module. In addition, two types of CSP networks are used. The difference comes from the manner of intensive convolution operations. One type contains residual blocks and the other one has continuous convolution operations. The CSP network scheme improves the efficiency of the whole model because it processes feature maps through two paths, and consequently reduces gradient repetition. The concatenated feature maps contain visual features at different levels.

On object detection tasks, the target size varies significantly, depending on the sensor parameters and observation distances. To cope with the problem of size heterogeneity, the MG-VTOD model employs both the feature pyramid network (FPN) [22] and the path aggregation network (PAN) [23] schemes. The FPN has a down-sampling feature extraction process, like many classic deep convolutional networks. It also has an upsampling process that concatenates the upsampled maps with the corresponding maps in the down-sampling branch. Since the concatenated maps contain visual features at different levels, they are fed to different detection heads separately to detect objects with

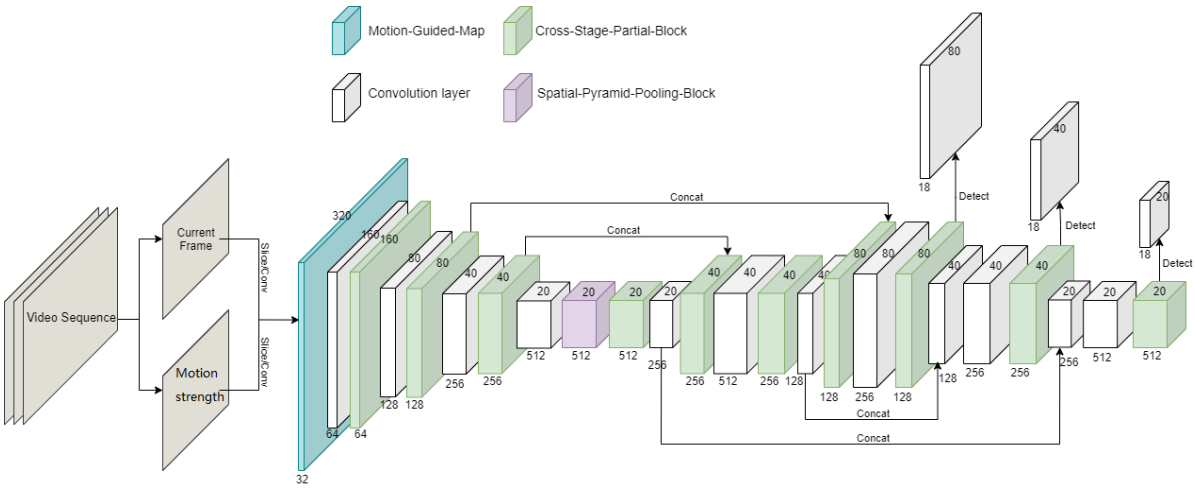

Figure 2. Structural illustration of the proposed MG-VTOD model.

various scales. The PAN takes an additional bottom-up path that uses the upsampling maps yielded by the FPN. The PAN also performs Region-of-Interest align operations on each feature map to extract the features for the object. In this way, the whole designed model can pool features from all the levels, and particularly, can shorten the distance between the lowermost and the top layers. Visual features are therefore substantially enriched for each level with the help of the augmented visual processing paths. Furthermore, a multi-scale prediction scheme and pre-defined anchor boxes are used to improve the detection ability for targets with different sizes. In our method, the prediction module is performed on the feature maps that are down-sampled 8, 16, and 32 times, respectively.

In order to improve the generalization ability of the designed method, we use the mosaic data augmentation scheme by splicing four images onto one image. In the yielded sample images, the target may appear in different positions within the four image patches. Therefore, this scheme helps the model improve the ability of target positioning and background adaption.

4. Experimental results and discussions

Towards an accurate and robust object detection in infrared videos, we have designed the MG-VTOD method that combines deep convolutional neural networks with the motion strength computation model. In this section, we evaluate the performance of the MG-VTOD method on a publicly available large-scale dataset. Firstly, we test different YOLOv5 versions to determine the backbone network of our method. Secondly, we apply the MG-VTOD and the competing methods to the large-scale dataset. Thirdly, we investigate the specific advantages of the proposed method over the competing methods. The experimental results along with discussions are reported as follows.

4.1. Experimental setup

4.1.1 UAV video target detection dataset

The dataset used in this paper comes from the Anti-UAV-2021 Challenge dataset[2] that contains 304451 frames in total. The resolutions of the video frames are mostly 640×512. The training dataset and test dataset is composed of 140 high-quality full HD thermal infrared video sequences, respectively. All the targets appearing in the frames are manually annotated. Note that these annotation files are employed to evaluate the performance of the proposed method as well as the competing methods, but should not be used in the Anti-UAV-2023 Challenge. It is challenging to detect the object targets in these videos because the backgrounds are various and complex. These backgrounds include clouds, buildings, trees, mountains, and other complex backgrounds, reflecting realistic scenarios in UAV surveillance. Moreover, the object targets span multi-scale occurrences in terms of size, *i.e.*, large, medium, small and tiny. To be more precise, the target sizes in the Anti-UAV-2021 dataset are mostly less than 2500 pixels, according to the statistical data. Compared to the field-of-view of the frame (327680 pixels), most of the targets only occupy a small region.

4.1.2 Experimental parameter settings

The experiment is based on a Ubuntu operating system. The algorithm is performed on a PC configured with AMD EPYC 7502 32-Core Processor, A100-PCI-E-40GB GPU. The network input size is 640×640. As for the hyper-parameter settings, the batch size is set as 64, the initial

[2] https://anti-uav.github.io/dataset/

learning rate is set as 0.01. In addition, we select the Adam as the optimizer.

4.1.3 Quantitative evaluation metric

When evaluating the performance of the object detectors, we adopt the widely used Average Precision (AP) value (when Intersection-over-Union is greater than 0.5) as the quantitative metric, which is calculated based on the Precision-Recall pairs. Precision stands for the correct proportion of all detection results, while Recall represents the proportion of all objects that are detected correctly:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \end{aligned} \tag{1}$$

where TP represents the number of objects detected correctly, FP denotes the number of non-object targets detected as object targets, and FN stands for the number of missed object targets.

4.2. Experimental results

4.2.1 Comparison to the advanced competing methods

To verify the effectiveness of our method, we compared it with the state-of-the-art YOLOv5-s, FCOS [35], FGFA [48], and SELSA methods [42]. The FCOS method is an representative one-stage and anchor-free object detector that completely avoids the complicated computation related to anchor boxes [35]. The FGFA method is a video object detector guided by optical flow information [48]. The SELSA method aggregates visual features in the full-sequence level for video object detection [42].

The quantitative evaluation results are reported in Table 1. In terms of detection accuracy, the AP, precision, and recall values, the proposed MG-VTOD method are all the highest, with an AP value of 90.7%. Our method improves the recall rate more significantly, which indicates that our method can detect more true objects in complex backgrounds. The YOLOv5 method performs on a single frame, thereby failing to use motion information. The FGFA and SELSA methods underperform since they are mainly designed to detect significant moving objects in videos. They are more likely to miss small and tiny targets in complex backgrounds. The FCOS method yields the worst target detection results.

Example visual experimental results are shown in Fig. 3. Within the target detection results, the green boxes indicate the ground truth annotation, while the red boxes stand for the bounding boxes of detection results.

As can be seen, video #111 reflects a mountain background in which the visual textures are very complex. When the object appears in the complex background, the YOLOv5-s and SELSA methods fail to yield detection results, neglecting the tiny object targets. Moreover, The FGFA method outputs false detection results while missing the true object targets. Comparatively, our method can accurately detect all the object targets. This is mainly because the proposed MG-VTOD method is able to utilize the motion information through the motion strength computation model. One can hardly pinpoint the target in the original frames. With the help of temporal-spatial contextual information, the motion strength computation model yields response maps in which the target areas have been significantly enhanced. The motion features can subsequently facilitate the object detection process.

Table 1. Comparison of experimental results.

Methods	AP(%)	Precision(%)	Recall(%)	FPS
YOLOv5-s	86.1	93.1	79.9	95
FCOS	80.1	90.6	74.6	21
FGFA	83.3	92.0	75.4	9
SELSA	84.1	93.0	75.6	4
MG-VTOD	**90.7**	**94.7**	**86.3**	28

As for the runtime, the YOLOv5-s method is the most time-efficient method. The FGFA, SELSA and FCOS methods are computationally heavy, thereby failing to detect target in a fast manner. Our method can process 28 frames per second, achieving a real-time object target detection.

4.2.2 Superiority validation of the proposed method

To further analyze the advantages of the MG-VTOD method over the competing method, we obtain the quantitative evaluation results of the tested methods when they detect targets with different sizes and with different types of backgrounds.

According to the target size in the dataset, we divided all the targets to be detected into seven groups, as reported in Table 2. One can see that the YOLOv5-s and MG-VTOD methods obtain very close performance when the targets are large, *i.e.*, the target area is larger than 1600 pixels. This demonstrates that the two methods are both good at detecting large targets. It is worth noting that the proposed MG-VTOD has significant superiority over the competing method when detecting small and tiny targets. In particular, the MG-VTOD has advantages in detecting small-size targets, which are challenging for traditional methods. For example, when detecting object targets smaller than 100 pixels, the AP value obtained by the MG-VTOD is 15.4% higher than that of the YOLOv5-s method. Since the Mago-DCNN and the YOLOv5-s have similar backbone networks, it can be inferred that the motion strength computation model substantially contributes to the improvement of the

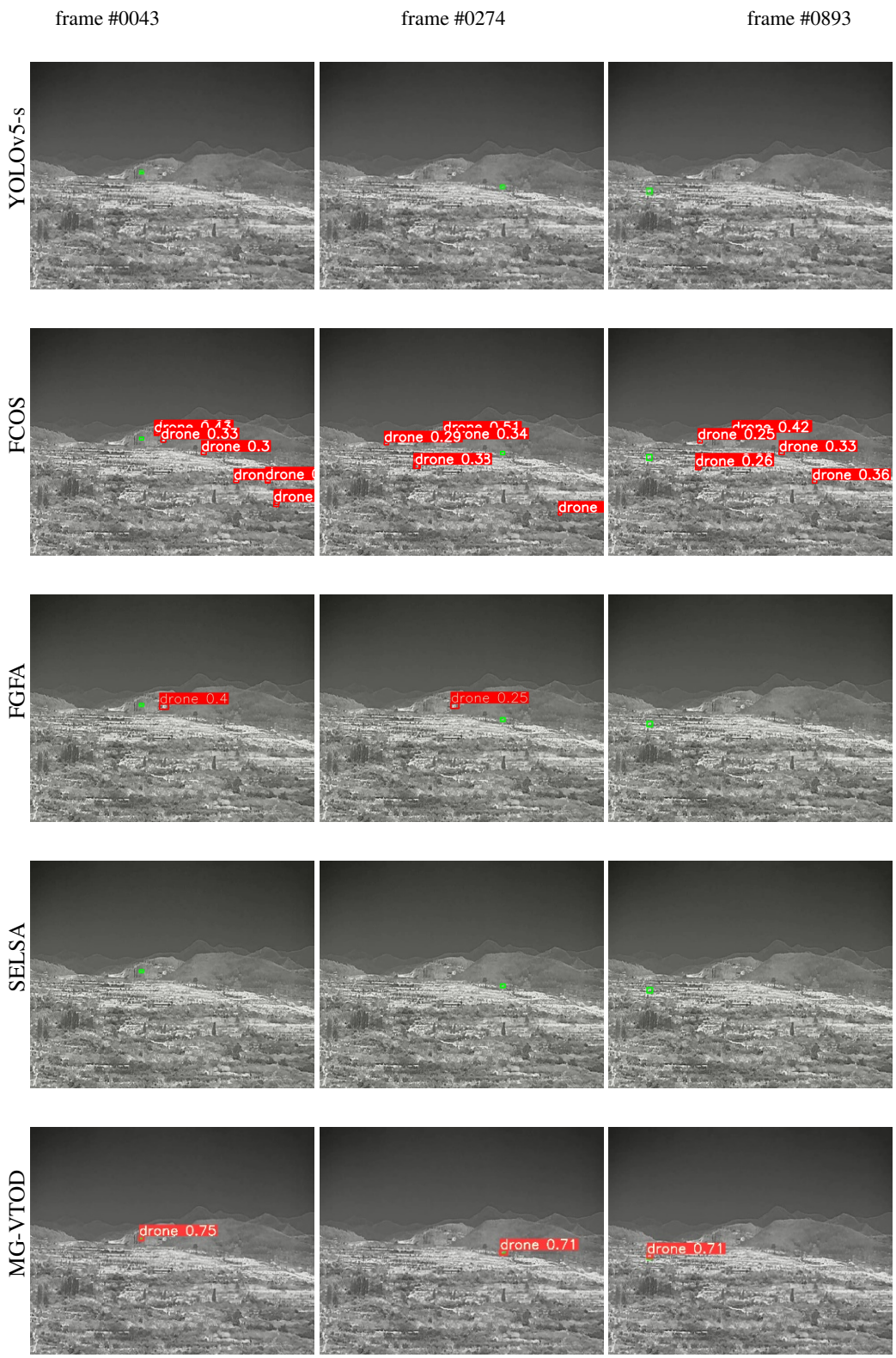

Figure 3. Visual comparison among detection results obtained by tested methods on video #111. Green rectangles stand for the ground truth of object targets; Red rectangles represent detection results obtained by the test methods; The marked numbers are the confidence scores for the corresponding results.

detection accuracy.

Table 2. Comparison of AP values (%) in terms of target sizes (*i.e.* number of pixels).

Size	(0,100]	(100,400]	(400,900]	(900,1600]
YOLOv5-s	61.3	78.0	92.9	87.7
MG-VTOD	**76.7**	**89.0**	**94.6**	**88.3**
Difference	↑15.4	↑11.0	↑1.5	↑0.6

To investigate the advantages of our method over the state-of-the-art competing method in different types of background, we further classify the data into five categories that reflect backgrounds of clean, cloud, building, wood and mountain, respectively. The experimental results are reported in Tab. 3. It can be seen that when detecting object in clean backgrounds, both the MG-VTOD and YOLOv5-s yield reliable detection results, achieving an AP value of 98.5%. Nevertheless, in most scenarios, the MG-VTOD obtained higher AP values than the YOLOv5-s method. In particular, when detecting objects in wood and mountain backgrounds, which usually have high-degree visual complexity, the MG-VTOD shows overwhelming superiority compared to the YOLOv5-s method. The target detection results illustrated in Fig. 3 have also verified the efficacy of the MG-VTOD method.

Table 3. Comparison of AP values (%) in terms of background types.

Background	Clean	Cloud	Building	Wood	Mountain
YOLOv5-s	98.5	89.3	85.7	62.3	72.0
MG-VTOD	**98.5**	**96.0**	**86.9**	**79.4**	**85.7**
Difference	0	↑6.7	↑1.2	↑17.1	↑13.7

In order to further evaluate the robustness to random noise, we apply the MG-VTOD and YOLOv5-s methods to videos that have been degraded by zero-mean, Gaussian white noise with a variance of 0.001. The obtained AP values of the two tested methods are shown in Table 4. Since we did not use noisy video as training samples during the model optimization procedure, the performance of the two methods decrease in artificially noisy data. Comparing Table 4 with Table 1, we can find that the MG-VTOD are more noise-robust than the YOLOv5-s method. This is because the motion strength computation model embedded in the MG-VTOD method can utilize multiple frames to compute the motion information. During this procedure, the random noise can be substantially suppressed.

Table 4. Comparison of AP values (%) obtained on videos degraded by random noise.

Methods	AP(%)	Precision(%)	Recall(%)
YOLOv5-s	73.0	**91.1**	67.2
MG-VTOD	**82.2**	90.8	**77.2**

5. Conclusions

It is widely acknowledged that monitoring remote unauthorized objects is of great importance for public security. Nevertheless, tiny-object detection in videos remain a challenging task in the computer vision filed. In order to detect tiny objects with complex backgrounds in realistic videos, we have proposed a novel real-time object detection method that jointly utilizes static and visual motion information. Inspired by the biological retina mechanism, we have designed a motion strength computation model to extract the motion responses of moving targets. The motion responses are subsequently employed to enhance the potential areas of the moving targets. The whole method has been implemented by the widely adopted deep convolutional neural networks that are guided by the motion strength maps. To verify the efficacy of the proposed method, we have evaluated the MG-VTOD method as well as the advanced competing methods on a publicly available large-scale dataset. Experimental results have validated that the MG-VTOD method significantly outperforms the competing methods, including the YOLOv5, FCOS, FGFA and SELSA methods, especially when detecting tiny-size targets against extremely complex backgrounds, *e.g.*, woods and mountains. Among future work, the proposed MG-VTOD method will be further employed to contribute to multi-target tracking systems.

Acknowledgement

This work is funded by National Natural Science Foundation of China (NSFC) (62102443), and is also sponsored by Beijing Nova Program (2022038).

References

[1] Martin S Banks, Jenny CA Read, Robert S Allison, and Simon J Watt. Stereoscopy and the human visual system. *SMPTE Motion Imaging Journal*, 121(4):24–43, 2012. 2

[2] Hatem Belhassen, Heng Zhang, Virginie Fresse, and El-Bay Bourennane. Improving video object detection by Seq-Bbox matching. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 226–233, 2019. 3

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[4] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020. 3

[5] Keyang Cheng, Honggang Cui, Humaira Abdul Ghafoor, Hao Wan, Qirong Mao, and Yongzhao Zhan. Tiny object detection via regional cross self-attention network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 4

[6] Sang-Han Choi, Gangwon Jeong, Young-Bo Kim, and Zang-Hee Cho. Proposal for human visual pathway in the extrastriate cortex by fiber tracking method using diffusion-weighted mri. *Neuroimage*, 220:117145, 2020. 2

[7] Yiming Cui, Liqi Yan, Zhiwen Cao, and Dongfang Liu. TF-Blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8138–8147, 2021. 2

[8] 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 International Conference on Computer Vision*, pages 7023–7032, 2019. 3

[9] Fang Fang and Sheng He. Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, 8(10):1380–1385, 2005. 2

[10] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 2

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1, 2

[12] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010. 3

[13] 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. 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[15] Chase B Hellmer, Leo M Hall, Jeremy M Bohl, Zachary J Sharpe, Robert G Smith, and Tomomi Ichinose. Cholinergic feedback to bipolar cells contributes to motion detection in the mouse retina. *Cell Reports*, 37(11):110106, 2021. 3

[16] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981. 3

[17] 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. 1

[18] Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2021. 1, 3

[19] Michael I Jordan, Michael J Kearns, and Sara A Solla. *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, volume 10. MIT Press, 1998. 4

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012. 2

[21] Yi Li, ZX Sun, Bo Yuan, and Yan Zhang. An improved method for motion detection by frame difference and background subtraction. *Journal of Image and Graphics*, 14(6):1162–1168, 2009. 3

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 4

[23] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 4

[24] Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602, 2021. 2

[25] Hao Luo, Wenxuan Xie, Xinggang Wang, and Wenjun Zeng. Detect or track: Towards cost-effective video object detection/tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8803–8810, 2019. 3

[26] Huizi Mao, Taeyoung Kong, and William J Dally. Catdet: Cascaded tracked detector for efficient object detection from video. In *Proceedings of Machine Learning and Systems*, volume 1, pages 201–211, 2019. 3

[27] Richard H Masland. The tasks of amacrine cells. *Visual Neuroscience*, 29(1):3–9, 2012. 3

[28] IJa Murray and Sb Plainis. Contrast coding and magno/parvo segregation revealed in reaction time studies. *Vision Research*, 43(25):2707–2719, 2003. 2

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 1, 2

[30] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017. 2

[31] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, pages 91–99, 2015. 1, 2

[33] Stelios M Smirnakis, Michael J Berry, David K Warland, William Bialek, and Markus Meister. Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386(6620):69–73, 1997. 4

[34] Paul T Sowden and Philippe G Schyns. Channel surfing in the visual brain. *Trends in Cognitive Sciences*, 10(12):538–545, 2006. 3

[35] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 6

[36] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 390–391, 2020. 4

[37] Gang Wang, Carlos Lopez-Molina, and Bernard De Baets. Blob reconstruction using unilateral second order Gaussian kernels with application to high-ISO long-exposure image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4817–4825, 2017. 4

[38] Gang Wang, Carlos Lopez-Molina, and Bernard De Baets. High-iso long-exposure image denoising based on quantitative blob characterization. *IEEE Transactions on Image Processing*, 29:5993–6005, 2020. 2

[39] Gang Wang, Gilbert Van Stappen, and Bernard De Baets. Automated detection and counting of artemia using u-shaped fully convolutional networks and deep convolutional networks. *Expert Systems with Applications*, 171:114562, 2021. 1, 3

[40] Wei Wei. Neural mechanisms of motion processing in the mammalian retina. *Annual Review of Vision Science*, 4:165–192, 2018. 3

[41] Frank Werblin, Greg Maguire, Peter Lukasiewicz, Scott Eliasof, and Samuel M Wu. Neural interactions mediating the detection of motion in the retina of the tiger salamander. *Visual Neuroscience*, 1(3):317–329, 1988. 4

[42] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9217–9225, 2019. 6

[43] Yuan Xiaofang and Wang Yaonan. Parameter selection of support vector machine for function approximation based on chaos optimization. *Journal of Systems Engineering and Electronics*, 19(1):191–197, 2008. 2

[44] Chang Xu, Jinwang Wang, Wen Yang, and Lei Yu. Dot distance for tiny object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1192–1201, 2021. 2

[45] Xiangyu Xu, Muchen Li, Wenxiu Sun, and Ming-Hsuan Yang. Learning spatial and spatio-temporal pixel aggregations for image and video denoising. *IEEE Transactions on Image Processing*, 29:7153–7165, 2020. 3

[46] Jian Zhao, Gang Wang, Jianan Li, Lei Jin, Nana Fan, Min Wang, Xiaojuan Wang, Ting Yong, Yafeng Deng, Yandong Guo, et al. The 2nd Anti-UAV workshop & challenge: Methods and results. *arXiv preprint arXiv:2108.09909*, 2021. 2

[47] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019. 1

[48] 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. 3, 6

[49] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2017. 3

[50] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–557, 2018. 3