# Explore Spatio-temporal Aggregation for Insubstantial Object Detection: Benchmark Dataset and Baseline

Kailai Zhou, Yibo Wang, Tao Lv, Yunqian Li, Linsen Chen, Qiu Shen*, Xun Cao*
Nanjing University, Nanjing, China
{calayzhou,ybwang,lvtao,lyq,linsen_chen}@smail.nju.edu.cn, {shenqiu,caoxun}@nju.edu.cn

## Abstract

*We endeavor on a rarely explored task named Insubstantial Object Detection (IOD), which aims to localize the object with following characteristics: (1) amorphous shape with indistinct boundary; (2) similarity to surroundings; (3) absence in color. Accordingly, it is far more challenging to distinguish insubstantial objects in a single static frame and the collaborative representation of spatial and temporal information is crucial. Thus, we construct an IOD-Video dataset comprised of 600 videos (141,017 frames) covering various distances, sizes, visibility, and scenes captured by different spectral ranges. In addition, we develop a spatio-temporal aggregation framework for IOD, in which different backbones are deployed and a spatio-temporal aggregation loss (STAloss) is elaborately designed to leverage the consistency along the time axis. Experiments conducted on IOD-Video dataset demonstrate that spatio-temporal aggregation can significantly improve the performance of IOD. We hope our work will attract further researches into this valuable yet challenging task. The code will be available at:* https://github.com/CalayZhou/IOD-Video.

## 1. Introduction

Recently, the emergence of deep learning based approaches [21, 22, 58] has witnessed significant advancements of object detection. Nevertheless, they still face intractable problems on some insubstantial objects captured by multispectral cameras [26] under specific wavelength, e.g. smoke, steam and gas leak. Due to frequent occurrences of smoke poisoning, fire accident, toxic gas leakage and explosion, it is urgent and crucial to realize real-time intelligent monitoring as well as early warning for insubstantial objects. This research topic is fresh and challenging, as insubstantial objects are quite different from conventional objects from several aspects.

Classical paper *what is an object* [1] defined a measure

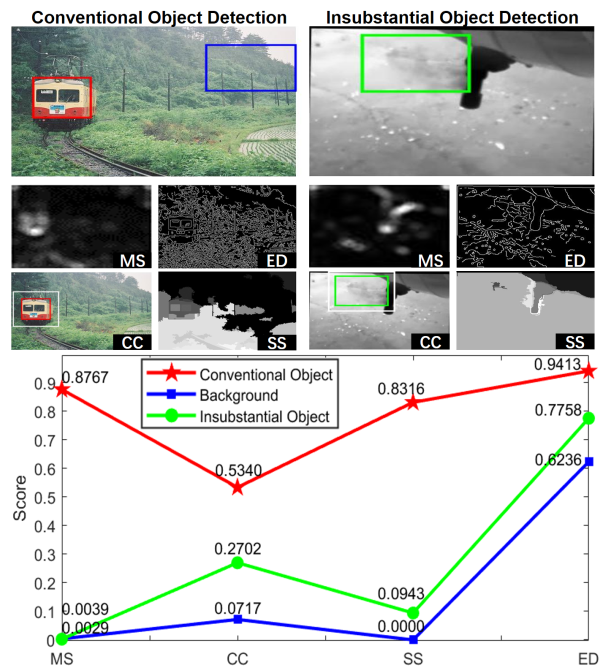



Figure 1. Comparisons between the conventional object and insubstantial object. The insubstantial object is more similar to the background under MS, CC, SS and ED image cues.

of objectness generic over classes which regards objects as standalone things with a well-defined boundary and center. It is considered that any object has at least one of three distinctive characteristics: (1) a well-defined closed boundary in space; (2) a different appearance from their surroundings; (3) sometimes it is unique within the image and stands out as salient. Based on the above observations, Alexe et al. [1] proposed four image cues to distinguish objects: Multi-scale Saliency (MS), Contrast (CC), Edge Density (ED) and Superpixels Straddling (SS). MS [33] indicates that an object is the salient region with a unique appearance; CC reflects the color dissimilarity between the foreground and background; ED measures the average edge magnitude as closed boundary characteristics; SS segments an image into

*Corresponding author.

small regions with uniform color or texture. As illustrated in Fig. 1, The foreground (train, in red box) in the left achieves relatively high scores for the four measurements, while the background (forest, in blue box) is the opposite. Considering the gas leak (in green box) in the right, it is deprived of trichromatic information owing to the monochromaticity of infrared images. Moreover, the shape of gas leak changes over time and there is no fixed and clear boundary. In terms of above image cues, the gas leak is more similar to the background rather than the foreground. Accordingly, mature algorithms for conventional object detection may fail in this special case, so that specific dataset and algorithm for insubstantial object are urgently needed.

To facilitate the study on this challenging problem, we collect a video-level insubstantial object detection (IOD-Video) dataset via multispectral camera under various scenes, including the smoke from chimney, hot steam, gas leak, etc. The object characteristics in IOD-Video dataset are summarized as follows: (1) indistinct boundary and amorphous shape; (2) the similarity to the background surroundings; (3) absence of color information and saliency. Consequently, the collaborative representation of spatial and temporal features is crucial under the limitation of spatial information within a static frame.

Additionally, we develop a spatio-temporal aggregation framework for IOD task. Unlike conventional object detection, IOD should distinguish visual temporal variation more than static semantic appearance. Traditional video background subtraction methods such as Gaussian Mixture Model (GMM) [64] and Visual Background Extractor (Vibe) [3] need static background, while the optical flow methods [6] require the moving target to have prominent feature points. They are not applicable to capture the time-variant pattern of insubstantial objects. Deep learning based methods for exploiting spatio-temporal features mainly focus on the classification field, e.g. action recognition [8,91]. Moreover, video object detection [38] dedicated to propagating the plentiful information from key frame features to non-key frame features and exploiting temporal modeling. The majority paradigm of above approaches first extracts features of a single frame through 2D convolutional network (2D-CNN), and then mines the temporal relation based on the extracted features. In this case, action detection is close to IOD, but it is highly dependent on sufficient spatial information in a single frame and is unsatisfactory for insubstantial object. In this paper, we explore a spatio-temporal aggregation framework from two aspects: First, representative spatio-temporal backbones of action recognition are introduce into our framework to evaluate their accuracy on IOD task. Second, spatio-temporal aggregation loss (STAloss) is designed to impose constraints in the three-dimensional space, as traditional 2D detection losses only concentrate on the static frame and do not take the concordance along time dimension into account. Experimental results reveal that temporal shift models have the best video-level detection performance which preserve the feature-level integrity of spatial dimension, and STAloss can improve the performance significantly.

Our contributions are summarized as follows:

• We propose an IOD-Video dataset for insubstantial object detection to promote research on this challenging task.

• We develop a spatio-temporal aggregation framework in which the video-level detection capability of representative action recognition backbones can be fairly evaluated.

• Based on the temporal shift backbone which achieves best performance, the STAloss is specifically designed to leverage the temporal consistency for further improvement.

## 2. Related Work

**Feature Extraction in Action Recognition.** Two-stream networks, 3D-CNNs and compute-effective 2D-CNNs are explored to extract spatio-temporal information in action recognition. The classic paradigm of two-stream network utilized extra modalities that reflect temporal motion information as the second input pathway, e.g. optical flow [11, 71]. 3D-CNNs are applied to the extraction of spatio-temporal information without pre-computing the input stream which explicitly represents temporal information [7,16,29,67]. Nevertheless, extracting spatio-temporal representation by 3D convolution kernel is of high computational cost, several works [57,68] explored the idea of 3D factorization to reduce the complexity. Another straightforward and effective method is to extract frame-level features with 2D-CNNs and then model temporal correlation [12, 18, 46, 48, 52, 65, 70, 82]. TSM [48], TAM [18] extended the temporal shift operation to video understanding. TIN [61] further fused temporal dependency via interlacing spatial representation. TEA [46] employed motion excitation and multiple temporal aggregation to enhance the motion pattern, which is similar to the channel-wise spatio-temporal module proposed by STM [36]. TDN [70] adopted the low-cost difference operation to establish multi-scale motion temporal representation by utilizing two temporal difference modules. The study of spatio-temporal collaborative representation mostly focuses on the classification task, there are few attempts in the video-level detection task.

**Video Object Detection.** Feature degradation (e.g., motion blur, occlusion, and defocus) is the primary challenge of video object detection (VOD). Early box-level VOD methods tackle this problem in a post-processing way which links bounding boxes predicted by still frames [9,19,28,41]. Feature-level VOD methods aggregate temporal contexts to improve the feature representation. STSN [4] predicted sampling locations directly and adopted deformable convolutions across space and time to leverage temporal information. STMN [77] built the spatio-temporal memory to capture the long-term appearance and motion dynamics. Methods in [39, 72, 88–90] aligned and warped adjacent features under the guidance of optical-flow. Besides optical flow, some approaches [14, 15, 20, 27, 42, 53, 62, 62, 75, 80] enhanced the object-level feature by exploring semantic and spatio-temporal correspondence among the region proposals. PSLA [25] applied self-attention mechanisms in the temporal-spatial domain without relying on extra optical flow. LSTS [37] adaptively learned the offsets of sampling locations across frames. MEGA [10] and LSFA [73] enhanced global-local aggregation by modeling longer-term and short-term dependency. TF-Blender [13] depicted the temporal feature relations and blended valuable neighboring features. TransVOD [31] first introduced temporal Transformer in VOD task to aggregate both the spatial object queries and the feature memories of each frame. Due to the redundant computational cost caused by applying still object detectors to each individual frame, the study of previous VOD methods focuses on propagating the rich information from key frame features to non-key frame features.

**Spatio-temporal Action Detection.** Different from VOD, action detection task pays attention to identifying and localizing human actions in the video. The frame-level detectors [56, 69] generate final action tubes by linking frame level detection results. To take full advantage of temporal information, some clip-level detectors have been proposed [24, 40, 44, 45, 51, 55, 60, 66, 74, 79]. ACT [40] handled a short sequence of frames and output action tubes by regressing from anchor cuboids. ACRN [66] computed actor-scene pair relation information for action classification. Context-Aware RCNN [76] rethought the importance of resolution in actor-centric and MOC [47] treated an action instance as a trajectory of moving points. Sarmiento et al. [60] made introduction of two cross attention blocks to effectively model the spatial relations and capture short range temporal interactions. In recent years, some approaches attempt to recognize action based on 3D convolution features [32,44,51,55,79]. ACDnet [51] intelligently exploited the temporal coherence between successive video frames to approximate their CNN features rather than naively extracting them. TFNet [44] applied attention mechanism to fuse the temporal features extracted by 3D-CNN and the frequency features extracted by 2D-CNN. ACAR-Net [55] built upon a novel high-order relation reasoning operator and an actor-context feature bank to enable indirect relation reasoning. STEP [79] proposed a progressive approach to obtained high-quality proposals gradually by incorporating more related temporal context. Nevertheless, they lack analysis of how different 3D-CNN architecture designs influence the final performance.

In conclusion, previous VOD and action detection methods mostly rely on 2D-CNN to propose bounding boxes, this paradigm may be unsatisfactory on the IOD task which needs collaborative extraction of spatio-temporal features.

## 3. IOD-Video Dataset

### 3.1. Data Collection and Annotation

**Data Collection.** As most Volatile Organic Compound (VOC) gases do not appear in the visible spectrum and hence they can not be seen by human eyes or traditional RGB cameras. The characteristic absorption peaks of many VOC gases are concentrated in the mid infrared spectrum, which are considered as the fingerprint region. So the IOD-Video dataset is captured in a restrained portion of the infrared (IR) domain range in $3 \sim 5\mu m$ and $8 \sim 12\mu m$. Specifically, part of the IOD-Video samples are collected from active deflation experiments of chemical gases by infrared spectral imaging or portable devices. The rest are the real-world samples captured by the alerting or monitoring system deployed in different petrochemical factories, which display various insubstantial objects, including the smoke emission, the water vapor and VOC gas leakage (e.g. olefins, alkanes, carbon monoxide). After continuous collection for nearly three years, we obtained thousands of videos which are manually cleaned up as follows: 1) The original videos are cropped into multiple short clips around 10s, each frame in clips is ensured to contain the insubstantial object. 2) At most two representative clips are reserved for the same scene to ensure the diversity of dataset. 3) Eliminate clips that cannot be identified by human eyes or have severe imaging noise. We finally get 600 video samples, amount to a total of 141,017 frames.

**Dataset annotation.** IOD-Video is carefully labeled by three experienced experts with bounding boxes, which are intuitive and practicable. A specifically developed tool is used to improve the annotation quality, by providing the pseudo color, motion information extracted by background modeling and historical annotated frames across two seconds, since longer video-level sequences can significantly improve the ability of human eyes to identify the insubstantial object. We provide frame-level bounding annotations which are double-checked to avoid inconsistency by following rules: 1) Annotations are temporally continuous without sudden change. 2) Bounding boxes tightens the object boundary well by human's subjective perception. 3) Bounding boxes reacts immediately when diffusion direction varies. The samples are labeled every five frames and middle frames are interpolated due to the slight difference between adjacent frames. All captured insubstantial objects are integrated into one category as foreground.

**Dataset Statistics.** Although insubstantial objects of IOD-Video dataset captured in mid infrared band are extremely difficult to collect, the annotated frames of IOD-Video exceed the quantity of KAIST [34] multispectral pedestrian detection and LNG [5] gas leakage dataset, as shown in

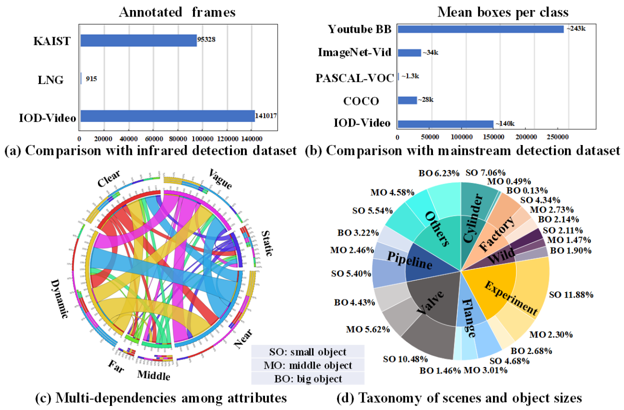



Figure 2. Statistics of the IOD-Video dataset and comparisons with other infrared and main stream detection datasets.

Fig. 2 (a). In terms of average boxes per class, IOD-Video is no less than the RGB detection datasets such like COCO [49], PASCAL-VOC [17], ImageNet-Vid [59], as shown in Fig. 2 (b). Fig. 2 (c) demonstrates multi-dependencies among IOD-Video attributes, which achieves good diversity by providing various distances (0$\sim$100m), sizes, visibility, and scenes captured by different spectral ranges. The lager width of a link between two super-classes indicates a higher probability. For example, the dynamic background samples in the near distance occupy a larger proportion than the static background samples in the far distance. IOD-Video dataset covers a wide range of scenarios (including pipeline, factory, flange, valve, experiments, cylinder, wild and others) and object with different size (e.g., small, middle and big). The distributions of scene categories and object size are displayed in Fig. 2 (d).

### 3.2. Challenges

As shown in Fig. 3, IOD faces multiple challenges caused by its characteristics , photography restrictions and environmental interference. First, the color absence in infrared video samples with indistinct boundary makes it intractable to identify and localize insubstantial object. Second, some IOD-video samples are captured by patrol inspection devices and the camera movement makes it difficult to locate the insubstantial object due to scene switching and camera shaking. Due to temperature drift caused by ambient temperature and infrared radiation absorption, the infrared cameras (e.g., cooled infrared detection arrays and Uncooled infrared Focal Plane Arrays (UFPA) suffer from low signal-to-noise ratio for infrared inhomogeneity and various imaging noises especially in the absence of heat source. Third, insubstantial objects sometimes appear to be invisible when it coincides with the complex background. In addition, the impact of environmental interference will be magnified due to the monochromaticity of infrared images, such as leaves, grasses, dusts blown by the wind.

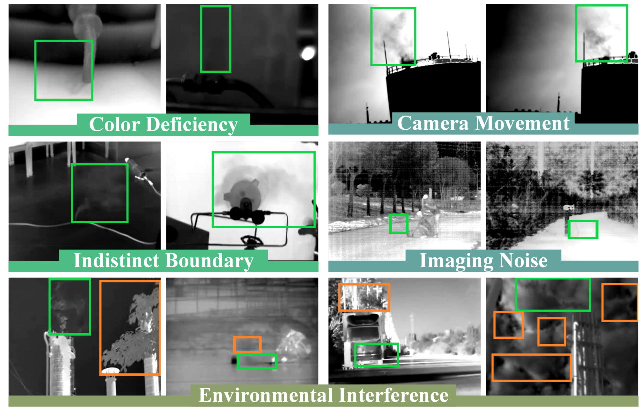



Figure 3. Three main challenges: insubstantial characteristics (color absence, indistinct boundary); photography restrictions (camera movement, imaging noise); environmental interference.

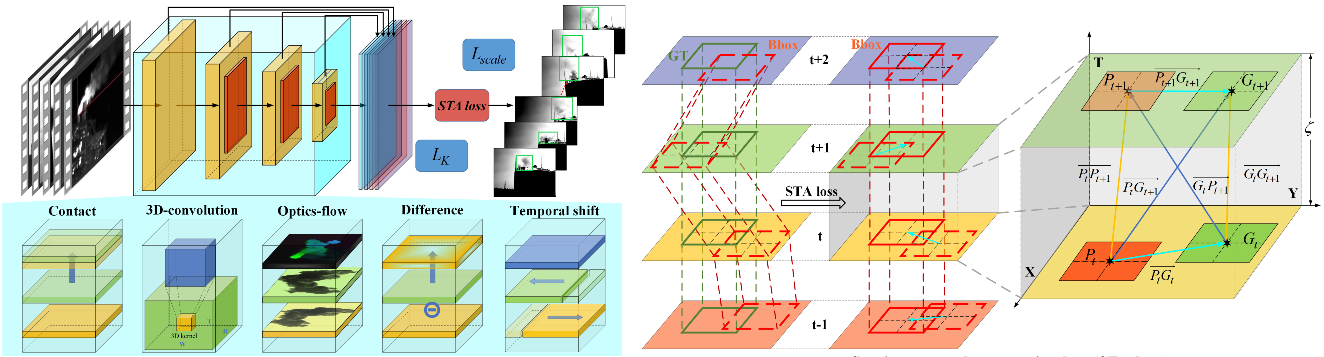
### 3.3. Evaluation Metrics and Protocols

The IOD-Video dataset samples are divided into clear and vague sets according to whether annotator can judge the boundary of the object within a single frame subjectively. We refer the COCO evaluation protocol [49] and report average precision over all IOU thresholds (AP), AP at IOU thresholds 0.5 (@0.5), 0.75 (@0.75), clear set (@clear) and vague set (@vague) in the frame level. In addition, IOD-Video dataset is randomly divided into three split (train/test split at a ratio of 2:1) and K-Fold cross-validation is adopted to report results averaged over three splits following the common setting [40, 47, 56].

## 4. Spatio-temporal Aggregation Framework

### 4.1. Overview

Insubstantial objects can hardly be located in a single static frame, as they differ from convectional objects with unclear contours. Thus, IOD needs spatio-temporal collaborative representation of adjacent several frames which remains unexplored in video-level detection task before. In this section, we design a general video-level detection framework as illustrated in Fig. 4, where representative action recognition models can be adopted as the spatio-temporal backbone. The anchor-free model CenterNet [87] is employed as the basic architecture to make the whole pipeline as simple as possible. All spatio-temporal backbones are built on ResNet-50 [30], hence the detection performance of different action recognition models on IOD can be fairly compared with our framework. To be specific, multi-scale feature maps from ResNet-50 $stage$ $2, 3, 4, 5$ are up-sampling with the deconvolution layer into the same resolution, $T$ images with resolution of $W \times H$ are fed into the spatio-temporal backbone to generate a feature volume $\mathcal{F}^{T\times \frac{W}{R} \times \frac{H}{R} \times 64}$, $R$ is spatial down-sampling ratio. In order to make full use of annotations, the original loss of




(a) Spatio-temporal Aggregation framework

(b) Spatio-temporal aggregation loss (STA loss)

Figure 4. (a) Our framework is inserted with five different designed spatio-temporal backbones to evaluate the detection capability of action recognition models. (b) The STAloss acts as an extra constraint to gather the predicted boxes (red dashed boxes) to the true boxes (green solid boxes) along the temporal dimension. The final results adjusted by STAloss are shown as red solid boxes.

base detector CenterNet is calculated based on the output feature volume $\mathcal{F}$ for each input frame. Since IOD needs spatio-temporal representation in feature extraction stage, naturally, the loss function design should also make corresponding changes. Then aim to leverage the temporal consistency, we design the STAloss that will be presented in section 4.3. Next the technical details of these spatio-temporal backbones will be presented.

## 4.2. Spatio-temporal Backbone

We choose state-of-the-art approaches [7, 18, 43, 46, 48, 61, 70, 78] on Something-Something (Sth-Sth) dataset [23] as spatio-temporal backbones. Analogous to IOD dataset, the static background in Sth-Sth dataset contributes little to the final prediction [91], and strong motion reasoning is required to capture the long-term temporal structure.

**Concat.** Direct concatenation of input $T$ frame features extracted by 2D-CNN is the simplest way to model temporal information. We follow the concatenation implementation of previous MOC [47] methods.

**3D-CNN.** 3D-CNN [7, 78] jointly learns spatial and temporal features by extending standard 2D spatial convolutional networks to the temporal dimension. Nevertheless, it brings higher computational cost and lacks specific consideration in the temporal information.

**Flow-based.** Many action recognition methods utilize the pre-computed dense optical flow as explicit motion representation. We argue it may fail on insubstantial object for the lack of outstanding feature points, and therefore Motion Squeeze Network (MSNet) [43] is adopted to establish flow correspondences implicitly.

**Temporal Difference.** As an approximate motion representation, temporal difference explicitly computes motion information and captures the distinctive properties of the adjacent frames. Numerous methods [70, 71, 84] have shown effectiveness in action recognition by utilizing the temporal difference operator into the network design.

**Temporal Shift.** The temporal dynamics are embedded into spatial representations by shifting the channels along the temporal dimension, which is a simple yet effective design with strong spatio-temporal modeling ability.

## 4.3. Spatio-temporal Aggregation Loss

After getting predicted boxes for each input frame, they may have a good regression within single static frames, but the consistency of predicted results along the time dimension is not subject to constraints. For instance, all four predicted boxes in Fig. 4 (b) has the same Intersection-over-Union (IoU) with the ground truth (GT) in terms of spatial dimension. Nevertheless, they are staggered in the time axis. The STAloss is able to pull the predicted boxes to the true boxes across multi frames.

In view of above observations, STAloss (i.e., $L_{STA}$) is proposed to impose constraints in the three-dimensional space. $L_{STA}$ is consisted of $L_{STA\cos\theta}$ and $L_{STA\sin\beta}$, they can be optimized in a collaborative manner. Let $P_t^{pre}(x_{pc}^t, y_{pc}^t, t)$ be the predicted box center of the $t^{th}$ frame and $G_t(x_{gc}^t, y_{gc}^t, t)$ be the corresponding GT center, an extra STA branch is established based on the output feature volume $\mathcal{F}$ to predict the offset $(\Delta_x^t, \Delta_y^t, t)$, which adjusts the box center of each input frame to a proper location $P_t(x_{pc}^t+\Delta_x^t, y_{pc}^t+\Delta_y^t, t)$ and the adjusted box of next frame is $P_t(x_{pc}^t + \Delta_x^t, y_{pc}^t + \Delta_y^t, t+\zeta)$ . $\zeta$ is the only hyperparameter in $L_{STA}$ which represents temporal interval length between adjacent frames. $L_{STA\cos\theta}^{cross}$ represents the dot multiplication of cross vectors $\overrightarrow{G_tP_{t+1}}$ and $\overrightarrow{P_tG_{t+1}}$, which will pull the prediction boxes towards GT boxes when they are far away. The $L_{STA\cos\theta}^{self}$ restricts the vector direction of the center line of adjacent prediction boxes $\overrightarrow{P_tP_{t+1}}$ to be consistent with the GT boxes $\overrightarrow{G_tG_{t+1}}$. $\theta$ reflects the angle between the cross vectors $\overrightarrow{P_tG_{t+1}}$, $\overrightarrow{G_tP_{t+1}}$ and the angle between self vectors $\overrightarrow{P_tP_{t+1}}$, $\overrightarrow{G_tG_{t+1}}$. The angle $\theta$ is expected to approach $0°$ during training. The offset is opti-

| | Method | Base Backbone | Pretrained Model | Frame AP (%) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | @0.5 | @0.75 | @clear | @vague | 0.5:0.95 |
| Frame-based Detector | Faster RCNN [58] | ResNet-50 | ImageNet | 33.49 | 6.76 | 16.52 | 8.82 | 12.31 |
| | SSD [50] | ResNet-50 | ImageNet | 30.21 | 3.95 | 12.78 | 7.82 | 9.99 |
| | CenterNet [87] | ResNet-50 | ImageNet | 24.80 | 4.32 | 11.21 | 6.65 | 8.50 |
| Video-based Detector | CRCNN [76] | ResNet101&FPN | ImageNet | 36.15 | 7.46 | 18.84 | 9.14 | 13.52 |
| | MOC [47] | DLA-34 | COCO | 36.81 | **8.94** | 19.96 | 9.62 | 10.96 |
| | MOC + Flow [47] | DLA-34 | COCO | 34.50 | 7.28 | 18.18 | 8.49 | 9.41 |
| Ours | TEA [46] | ResNet-50 | ImageNet | **42.19** | 8.66 | **22.97** | **9.95** | **15.69** |

Table 1. Comparisons with previous frame-based detectors and video-based detectors. Our basic spatio-temporal aggregation framework simply replaces the backbone of CenterNet with TEA [46] without any other complex design.

mized by the following items:

$$\begin{cases} L_{STA\cos\theta}^{cross} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\overrightarrow{G_tP_{t+1}}\cdot\overrightarrow{P_tG_{t+1}}}{\left|\overrightarrow{G_tP_{t+1}}\right|\left|\overrightarrow{P_tG_{t+1}}\right|} \\ L_{STA\cos\theta}^{self} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\overrightarrow{P_tP_{t+1}}\cdot\overrightarrow{G_tG_{t+1}}}{\left|\overrightarrow{P_tP_{t+1}}\right|\left|\overrightarrow{G_tG_{t+1}}\right|} \\ L_{STA\sin\beta}^{pre} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\left|\overrightarrow{P_tG_t}\right|}{\left|\overrightarrow{P_tG_{t+1}}\right|} \\ L_{STA\sin\beta}^{next} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\left|\overrightarrow{P_{t+1}G_{t+1}}\right|}{\left|\overrightarrow{P_{t+1}G_t}\right|} \end{cases} \tag{1}$$

while $\cos\theta$ trends to be a smooth curve, this may cause difficulty to further converge. To address this issue, we introduce a $L_{STA\sin\beta}$ term which is the division of $\left|\overrightarrow{P_tG_t}\right|$ and $\left|\overrightarrow{P_tG_{t+1}}\right|$ inspired by Distance-IoU loss [85]. The vector $\overrightarrow{G_tG_{t+1}}$ is almost perpendicular to $XY$ plane due to the tiny changes between adjacent frames in most cases, so $L_{STA\sin\beta}$ term can be approximate equal to $\sin\beta$ where $\beta$ is the angle between the vector $\overrightarrow{P_tG_{t+1}}$ and $\overrightarrow{G_tG_{t+1}}$. When the predicted boxes and GT boxes become closer during the later training process, the $L_{STA\sin\beta}$ term will locate the center more accurately in term of steeper gradient when angle $\beta\to 0°$. The cooperative effect of $L_{STA\cos\theta}$ and $L_{STA\sin\beta}$ leverage the temporal consistency along the time dimension. We average the above terms and final $L_{STA}$ is formulated as:

$$\begin{cases} L_{STA\cos\theta} = \frac{1}{2}L_{STA\cos\theta}^{cross} + \frac{1}{2}L_{STA\cos\theta}^{self} \\ L_{STA\sin\beta} = \frac{1}{2}L_{STA\sin\beta}^{pre} + \frac{1}{2}L_{STA\sin\beta}^{next} \\ L_{STA} = \lambda(1 - L_{STA\cos\theta}) + (1-\lambda)L_{STA\sin\beta} \end{cases} \tag{2}$$

where $\lambda$ is the hyperparameter between the $L_{STA\sin\beta}$ and $L_{STA\cos\theta}$, empirically we set $\lambda$ to 0.5. The overall training objective is:

$$L = L_K + \lambda_{size}L_{size} + L_{STA} \tag{3}$$

where $L_K$ and $L_{size}$ is the original classification loss and scale loss of the base detector CenterNet [87], and $L_{STA}$ acts as an extra loss which can be inserted into any other video-level detection pipeline.

# 5. Experiments

## 5.1. Implementation Details

We apply the same data augmentation with MOC [47] to the whole training video clips: mirroring, distorting, expanding and cropping. Specifically, in the training we crop a patch with the size of [0.3, 1] of the input image and resize it to 288 $\times$ 288, then each image is randomly distorted and horizontal flipped with a probability of 0.5 to increase the diversity. The spatial down-sampling ratio $R$ is set to 4 and the temporal interval length $\zeta$ is set to 4. The whole network is trained by Adam optimizer with a learning rate of 5e-4 and a batch size of 16 on two nvidia 3090 GPUs, we decrease the learning rate by 0.1$\times$ at the 6th and 8th epochs, and stop at the 12th epoch. For video-based detectors, the number of input frames is set to 8 unless otherwise stated.

## 5.2. Performance of Classic Detectors

We first analyze detection performances of classic frame-based and video-based detectors on IOD-Video dataset in Tab. 1. In terms of the frame-based detectors, the video clips are split into frames as training samples. Faster RCNN [58] achieves the best results under the condition of the same backbone. On the one hand, compared with SSD [50], the two-stage architecture design of Faster RCNN proves to be beneficial since there existing a large number of hard negatives in IOD-Video samples. On the other hand, the original deep layer aggregation backbone (DLA-34 [81]) is deprived from CenterNet, and this seems to preclude the benefits of hierarchical feature fusion which is important for anchor-free design. Note that all methods in Tab. 1 utilize ImageNet pretrained model except for MOC [47], which uses the COCO pretrained model to provide stronger spatial representation ability while reducing sen-

| | Spatio-temporal Backbone | Base Backbone | Frame AP (%) | | | | |
|---|---|---|---|---|---|---|---|
| | | | @0.5 | @0.75 | @clear | @vague | 0.5:0.95 |
| Concat | Concat [47] | ResNet-50 | 27.41 | 4.45 | 12.38 | 7.32 | 9.32 |
| 3D-Convolution | S3D [78] | ResNet-50 | 35.82 | 6.73 | 17.72 | 8.81 | 12.72 |
| | I3D [7] | ResNet-50 | 36.83 | 7.39 | 18.78 | 9.29 | 13.43 |
| Flow-based | MSNet [43] | ResNet-50 | 41.19 | 7.91 | 21.35 | 10.01 | 14.90 |
| Difference | TDN [70] | ResNet-50 | 41.69 | 8.48 | 21.42 | 10.46 | 15.40 |
| Temporal Shift | TSM [48] | ResNet-50 | 42.13 | 8.20 | 21.98 | 10.28 | 15.38 |
| | TAM [18] | ResNet-50 | 41.95 | 8.53 | 21.49 | 10.61 | 15.50 |
| | TIN [61] | ResNet-50 | 42.77 | 8.01 | 22.35 | 10.51 | 15.73 |
| | TEA [46] | Res2Net-50 | 42.19 | 8.66 | 22.97 | 9.95 | 15.69 |
| +STAloss | TIN [61] | ResNet-50 | 43.72 | 9.26 | 23.81 | 10.35 | 16.27 |
| | TEA [46] | Res2Net-50 | **45.08** | **9.50** | **24.43** | **10.91** | **16.99** |

Table 2. The representative action recognition models are selected from the state-of-the-arts on Sth-Sth dataset [23], and we present the video-level detection performance of different spatio-temporal backbones. STAloss replaces the original offset loss $L_{off}$ of CenterNet.

| input frames | Frame AP (%) | | | | |
|---|---|---|---|---|---|
| | @0.5 | @0.75 | @clear | @vague | 0.5:0.95 |
| 2 | 33.60 | 5.65 | 13.35 | 7.98 | 11.78 |
| 4 | 40.46 | 8.02 | 14.51 | 9.77 | 14.89 |
| 6 | 44.44 | 9.91 | 24.09 | 11.17 | 16.97 |
| 8 | 45.08 | 9.50 | 24.43 | 10.91 | 16.99 |

Table 3. Evaluation results with different input frames.

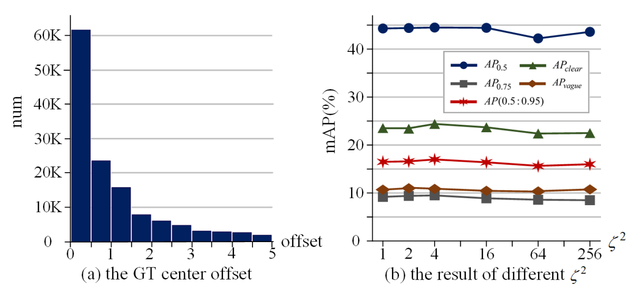



Figure 5. (a) The statistical GT center offset of IOD-Video dataset. (b) Analysis of different hyperparameter $\zeta^2$ value settings.

sitivity to dynamic change. In addition, the concatenation design lacks in-depth mining of temporal information. The dense-flow extracted by external off-the-shelf method [6] usually obtains gains on action recognition tasks but shows minimal improvements on IOD task, due to the shortage of heavily dependent on texture and color features in spatial domain. Our spatio-temporal aggregation framework simply replaces the backbone of CenterNet [87] with TEA [46], and it achieves overall better performances which suggest the spatio-temporal aggregation is crucial.

### 5.3. Spatio-temporal Backbones

Next we explore the spatio-temporal backbones with different architecture designs. As shown in Tab. 2, although simple concatenation operation [47] performs well on the spatio-temporal action detection task, it fails on the IOD task due to the significant difference between two tasks within a single static frame. The action detection datasets, UCF101-24 [63] and JHMDB [35], tend to be scene-focused while our IOD-Video dataset is motion-focused. 3D convolution [7] or spatial and temporal separable 3D convolution [78] provide a feasible paradigm in action recognition task, but they lack explicit mining of spatio-temporal datacube along the temporal dimension, which causes performance degradation compared with compute-effective 2D-CNNs. A trainable neural module proposed in MSNet [43] has a relatively good performance on the IOD task, which establishes correspondences across multi-frames and converts them into motion features (flow-based). The implicit extraction of motion information may alleviate the difficulty of extracting dense optical flow directly on IOD-Video dataset to some extent. TDN [70] presents a two-level temporal modeling framework to generalize the idea of RGB difference for motion modeling. The similar idea can be found in the design of TEA [46] structure, which indicates that direct subtraction is a simple yet effective way to capture the insubstantial feature. Moreover, the temporal shift based methods are capable to achieve superior performance. Among them, TSM [48] and TAM [18] are the initial design which achieve the comparable performance. TIN [61] adopts deformable shift module and has the highest accuracy on AP@[0.5:0.95] without STAloss added. TEA is the combination of subtraction and temporal shift in a light- weight configuration. In our opinion, temporal shift models preserve the feature-level integrity of spatial dimension and can be an effective design for strong temporary motion reasoning as well as spatial semantic representation on video-level detection tasks.

### 5.4. The Number of Input Frames

It is believed that the input range plays an important role to capture the temporal information, we evaluate the performance under different input frames sets. Intuitively, the

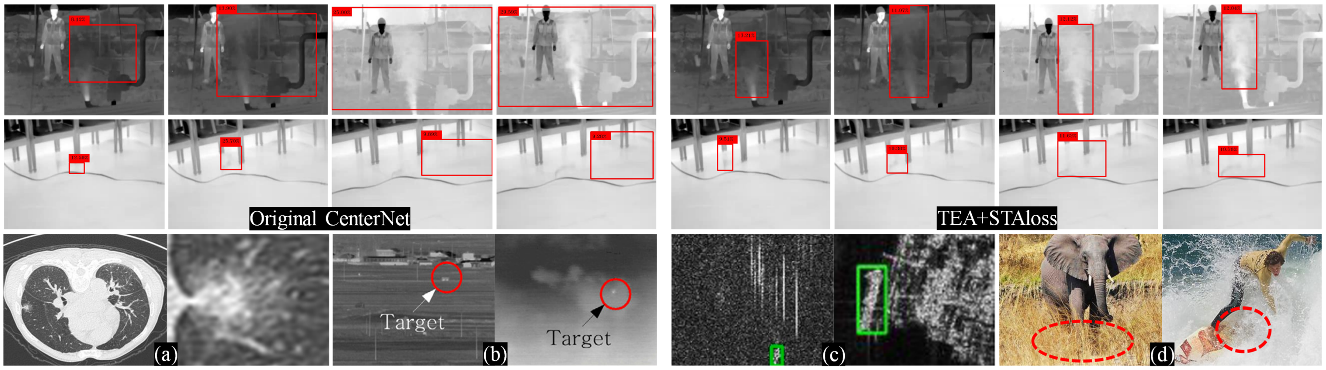



Figure 6. Detection results of CenterNet and our baseline method (TEA+STAloss) for clear and vague samples. Potential applications: (a) Medical diagnosis (b) Infrared dim small objection detection (c) Synthetic Aperture Radar (SAR) detection (d) Partially-occluded targets.

more input frames to be, the higher AP is expected to be obtained. Nevertheless, the input frames 8 shows tiny advantage compared with input frames 6 in in Tab. 3. It is reflected that our baseline method mainly focuses on short-term temporal modeling, the long-term temporal modeling ability with novel insight is needed for further improvement.

### 5.5. STAloss

To further leverage temporal consistency in the loss level, we apply STAloss to the TEA and TIN which performs best among spatio-temporal backbones. The STAloss can bring a substantially better localization accuracy, especially at AP@0.5 and AP@clear sets. The comprehensive results over five sets indicate that it is a feasible way to impose constraints in spatio-temporal space of loss function design for IOD task. Since training a video-level detector with the STAloss as auxiliary objective only involves one hyperparameter $\zeta$ in the internal structure of loss function, we conduct several experiments to investigate the robustness of hyperparameter $\zeta$, which is used to adjust the temporal interval length. We make a statistic of the offset between the center of adjacent GT boxes on IOD-Video dataset. The GT offset of a single sample and hyperparameter $\zeta$ are combined to form the vector $\overrightarrow{G_t G_{t+1}}$. Fig. 5 (a) shows a sharp downward curve which illustrates the slight change for most cases along the time axis. Under this condition, different values of $\zeta^2$ in 1, 2, 4, 16, 64, 256 are set and we observe our baseline is relatively insensitive to the variations of $\zeta^2$ from 1 to 256 in Fig. 5 (b). Combine the statistical curve of GT center offset and $\zeta^2$ settings into consideration, $\zeta$ should be larger than most offset statistics empirically for the convergence at the early stage of training. When the $\zeta$ is set too large, the $L_{STA\cos\theta}$ will always be 1 and $L_{STA\sin\beta}$ tend to be 0, which brings difficulty for STAloss to be optimized. Overall, the only hyperparameter $\zeta$ is robust within the appropriate range and the proposed STAloss can be nearly regarded as hyperparameter-free.

## 6. Conclusion

In this work we make attempts on a rarely explored task named insubstantial object detection which is completely different fropm revious detection tasks. Insubstantial objects have indistinct boundary and amorphous shape, and may appear to be invisible in the background due to the lack of color information. In consideration of this, a feasible way is exploiting spatio-temporal features to compensate the feature absence of single static frames. But the feature extraction of previous video-level detection methods mainly rely on 2D-CNN and this paradigm may be unsatisfactory on IOD. To drive progress on this challenging task, we collect the IOD-Video dataset which consists of 600 videos (141,017 frames) and construct the spatio-temporal aggregation framework from two aspects: First we measure the detection capacity of different action recognition backbone and reveal temporal shift models perform best; Second the STAloss is designed to pull the prediction boxes of each frame along the temporal dimension together. As shown in Fig. 6, compare with CenterNet, our baseline method (TEA backbone + STAloss) is robust to light and shade variations and severe deformation with spatio-temporal aggregation introduced. Nevertheless, there are still large room for our baseline method to improve. Additionally, IOD may benefit the study of infrared dim small objection detection [2], ground-glass nodule diagnosis in medical [54], Synthetic Aperture Radar (SAR) detection [83] and some partially-occluded targets in RGB images [86], which also have the similar characteristic of indistinct boundary. We consider to extend the IOD-Video dataset to the multispectral bands in the future, which will make it possible to distinguish the specific substance of detected objects. The proposed IOD-Video dataset and our baseline approach are expected to drive progress on the study of challenging IOD task.

**Acknowledgment.** This research was supported by National Natural Science Foundation of China under Grant (62025108), we thank *ZHIPUTECH* for the multispectral camera supports and data collection.

## References


[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 73–80. IEEE, 2010. 1

[2] Tae-Wuk Bae. Spatial and temporal bilateral filter for infrared small target enhancement. *Infrared Physics & Technology*, 63:42–53, 2014. 8

[3] Olivier Barnich and Marc Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6):1709–1724, 2010. 2

[4] 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. 2

[5] Junchi Bin, Choudhury A Rahman, Shane Rogers, and Zheng Liu. Tensor-based approach for liquefied natural gas leakage detection from surveillance thermal cameras: A feasibility study in rural areas. *IEEE Transactions on Industrial Informatics*, 17(12):8122–8130, 2021. 3

[6] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004. 2, 7

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 5, 7

[8] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 2

[9] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7814–7823, 2018. 2

[10] 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. 3

[11] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV-IEEE International Conference on Computer Vision*, pages 3218–3226. IEEE, 2015. 2

[12] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 2

[13] 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. 3

[14] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6678–6687, 2019. 3

[15] 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. 3

[16] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. Spatio-temporal channel correlation networks for action classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–299, 2018. 2

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[18] Quanfu Fan, Chun-Fu Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. *arXiv preprint arXiv:1912.00869*, 2019. 2, 5, 7

[19] 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. 2

[20] Qichuan Geng, Hong Zhang, Na Jiang, Xiaojuan Qi, Liangjun Zhang, and Zhong Zhou. Object-aware feature aggregation for video object detection. *arXiv preprint arXiv:2010.12573*, 2020. 3

[21] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015. 1

[23] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 5, 7

[24] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 3

[25] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. Progressive sparse

local attention for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3909–3918, 2019. 3

[26] Nathan Hagen, Robert T Kester, Christopher G Morlier, Jeffrey A Panek, Paul Drayton, Dave Fashimpaur, Paul Stone, and Elizabeth Adams. Video-rate spectral imaging of gas leaks in the longwave infrared. In *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XIV*, volume 8710, page 871005. International Society for Optics and Photonics, 2013. 1

[27] 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. 3

[28] 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

[29] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 2

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[31] 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. *arXiv preprint arXiv:2105.10920*, 2021. 3

[32] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017. 3

[33] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007. 1

[34] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 3

[35] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 7

[36] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2000–2009, 2019. 2

[37] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020. 3

[38] 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*, 2021. 2

[39] Ruibing Jin, Guosheng Lin, Changyun Wen, Jianliang Wang, and Fayao Liu. Feature flow: In-network feature flow estimation for video object detection. *Pattern Recognition*, 122:108323, 2022. 3

[40] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 3, 4

[41] 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

[42] Junho Koh, Jaekyum Kim, Younji Shin, Byeongwon Lee, Seungji Yang, and Jun Won Choi. Joint representation of temporal image sequences and object motion for video object detection. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13370–13376. IEEE, 2021. 3

[43] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *European Conference on Computer Vision*, pages 345–362. Springer, 2020. 5, 7

[44] Changhai Li, Huawei Chen, Jingqing Lu, Yang Huang, and Yingying Liu. Time and frequency network for human action detection in videos. *arXiv preprint arXiv:2103.04680*, 2021. 3

[45] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–318, 2018. 3

[46] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020. 2, 5, 6, 7

[47] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision*, pages 68–84. Springer, 2020. 3, 4, 5, 6, 7

[48] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2, 5, 7

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[50] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 6

[51] Yu Liu, Fan Yang, and Dominique Ginhac. Acdnet: An action detection network for real-time edge computing based on flow-guided feature approximation and memory aggregation. *Pattern Recognition Letters*, 145:118–126, 2021. 3

[52] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11669–11676, 2020. 2

[53] Hao Luo, Lichao Huang, Han Shen, Yuan Li, Chang Huang, and Xinggang Wang. Object detection in video with spatial-temporal context aggregation. *arXiv preprint arXiv:1907.04988*, 2019. 3

[54] Jie Mei, Ming-Ming Cheng, Gang Xu, Lan-Ruo Wan, and Huan Zhang. Sanet: A slice-aware network for pulmonary nodule detection. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 8

[55] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021. 3

[56] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *European conference on computer vision*, pages 744–759. Springer, 2016. 3, 4

[57] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 6

[59] 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. 4

[60] Manuel Sarmiento Calderó, David Varas, and Elisenda Bou-Balust. Spatio-temporal context for action detection. *arXiv e-prints*, pages arXiv–2106, 2021. 3

[61] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11966–11973, 2020. 2, 5, 7

[62] 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. 3

[63] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7

[64] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999. 2

[65] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020. 2

[66] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 3

[67] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[68] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[69] Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool. Actionness estimation using hybrid fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2016. 3

[70] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021. 2, 5, 7

[71] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2, 5

[72] 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. 3

[73] Xinggang Wang, Zhaojin Huang, Bencheng Liao, Lichao Huang, Yongchao Gong, and Chang Huang. Real-time and accurate object detection in compressed video by long short-term feature aggregation. *Computer Vision and Image Understanding*, 206:103188, 2021. 3

[74] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 3

[75] 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. 3

[76] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *European Conference on Computer Vision*, pages 440–456. Springer, 2020. 3, 6

[77] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the Eu-*

*ropean Conference on Computer Vision (ECCV)*, pages 485–501, 2018. 2

[78] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 5, 7

[79] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019. 3

[80] Chun-Han Yao, Chen Fang, Xiaohui Shen, Yangyue Wan, and Ming-Hsuan Yang. Video object detection via object-level temporal aggregation. In *European conference on computer vision*, pages 160–177. Springer, 2020. 3

[81] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 6

[82] Can Zhang, Yuexian Zou, Guang Chen, and Lei Gan. Pan: Towards fast action recognition via learning persistence of appearance. *arXiv preprint arXiv:2008.03462*, 2020. 2

[83] Tianwen Zhang, Xiaoling Zhang, Jun Shi, and Shunjun Wei. Hyperli-net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:123–153, 2020. 8

[84] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6566–6575, 2018. 5

[85] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020. 6

[86] Zhaohui Zheng, Rongguang Ye, Ping Wang, Jun Wang, Dongwei Ren, and Wangmeng Zuo. Localization distillation for object detection. *arXiv preprint arXiv:2102.12252*, 2021. 8

[87] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 4, 6, 7

[88] 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. 3

[89] 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

[90] 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

[91] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020. 2, 5