FADE: A Dataset for Detecting Falling Objects Around Buildings in Video

Zhigang Tu, *Senior Member, IEEE*, Zhengbo Zhang, Zitao Gao, Chunluan Zhou, Junsong Yuan, *Fellow, IEEE*, and Bo Du, *Senior Member, IEEE*

***Abstract*—Objects falling from buildings, a frequently occurring event in daily life, can cause severe injuries to pedestrians due to the high impact force they exert. Surveillance cameras are often installed around buildings to detect falling objects, but such detection remains challenging due to the small size and fast motion of the objects. Moreover, the field of falling object detection around buildings (FODB) lacks a large-scale dataset for training learning-based detection methods and for standardized evaluation. To address these challenges, we propose a large and diverse video benchmark dataset named FADE. Specifically, FADE contains 2,611 videos from 25 scenes, featuring 8 falling object categories, 4 weather conditions, and 4 video resolutions. Additionally, we develop a novel detection method for FODB that effectively leverages motion information and generates small-sized yet high-quality detection proposals. The efficacy of our method is evaluated on the proposed FADE dataset by comparing it with state-of-the-art approaches in generic object detection, video object detection, and moving object detection. The dataset and code are publicly available at https://fadedataset.github.io/FADE.github.io/**

***Index Terms*—Falling object detection, a large diverse video dataset, baseline method.**

I. Introduction

OVER the past few decades, with the continuous expansion of urbanization, high-rise buildings have sprung up. Some residents of these buildings throw objects from above without caution, potentially leading to repeated injuries and incidents. According to a report [1] of the U.S. Bureau of Labor Statistics, there are more than 50,000 "struck by falling objects" injuries every year in USA. To highlight the potential danger posed by objects falling from tall buildings, we present a specific example: If a 200-gram apple falls from a 30-meter-high building, the impact duration is approximately 0.01 seconds. Based on the momentum theorem [2], this results in an equivalent impact force of roughly 49.5 kilograms. Such incidents can pose serious threats to public safety (see Figure 1 for visualization). To mitigate such incidents, several countries have enacted laws prohibiting the act of throwing objects from buildings, including the USA [3], Singapore [4], and China [5].

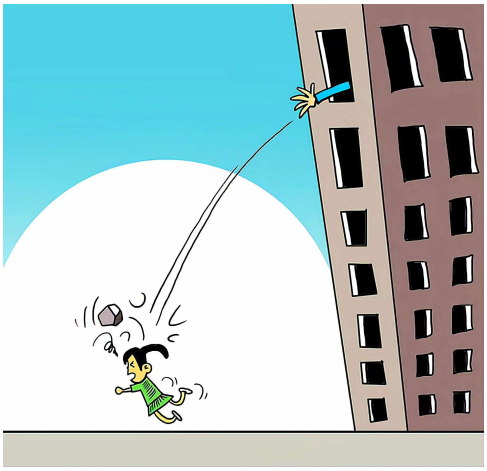

Fig. 1. Falling object incidents around buildings may pose serious risks to human life and public safety.

At the same time, intelligent video surveillance (IVS) has become a critical technology for ensuring public safety [6], fueled by recent advances in computer vision, including object detection [7], [8], [9], anomaly detection [10], [11], [12], [13], human identification [14], [15], [16], [17], tracking [18], [19], [20], [21], and video understanding [22], [23], [24], [25], [26]. These computer vision-based IVS methods provide benefits such as low cost, high accuracy, and reduced dependence on manual labor. Inspired by the successful applications of these techniques and supported by the availability of surveillance cameras around buildings, several research institutions and government agencies have begun exploring IVS algorithms for detecting falling object incidents around buildings (FODB). However, these IVS technologies often rely on large-scale datasets to train learning-based models, yet such a dataset is currently unavailable in the FODB domain. In fact, the FODB

Received 2 August 2024; revised 21 July 2025; accepted 30 August 2025. Date of publication 10 September 2025; date of current version 19 September 2025. This work was supported in part by the Natural Science Fund for Distinguished Young Scholars of Hubei Province under Grant 2022CFA075, in part by the National Natural Science Foundation of China (NSFC) under Grant 62106177, and in part by the Fundamental Research Funds for the Central Universities under Grant 2042023KF0180. The associate editor coordinating the review of this article and approving it for publication was Dr. Daniel Moreira. *(Corresponding authors: Zhengbo Zhang; Zitao Gao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by (Name of Review Board or Committee) (IF PROVIDED under Application No. xx, and performed in line with the (Name of Specific Declaration)).

Zhigang Tu, Zhengbo Zhang, and Zitao Gao are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zhengbozhang.1@gmail.com; gaozitao@whu.edu.cn).

Chunluan Zhou is with Ant Group Company Ltd., Beijing 100020, China.

Junsong Yuan is with the Computer Science and Engineering Department, The State University of New York at Buffalo, Buffalo, NY 14260 USA.

Bo Du is with the School of Computer Science, Wuhan University, Wuhan 430072, China.

Digital Object Identifier 10.1109/TIFS.2025.3607254

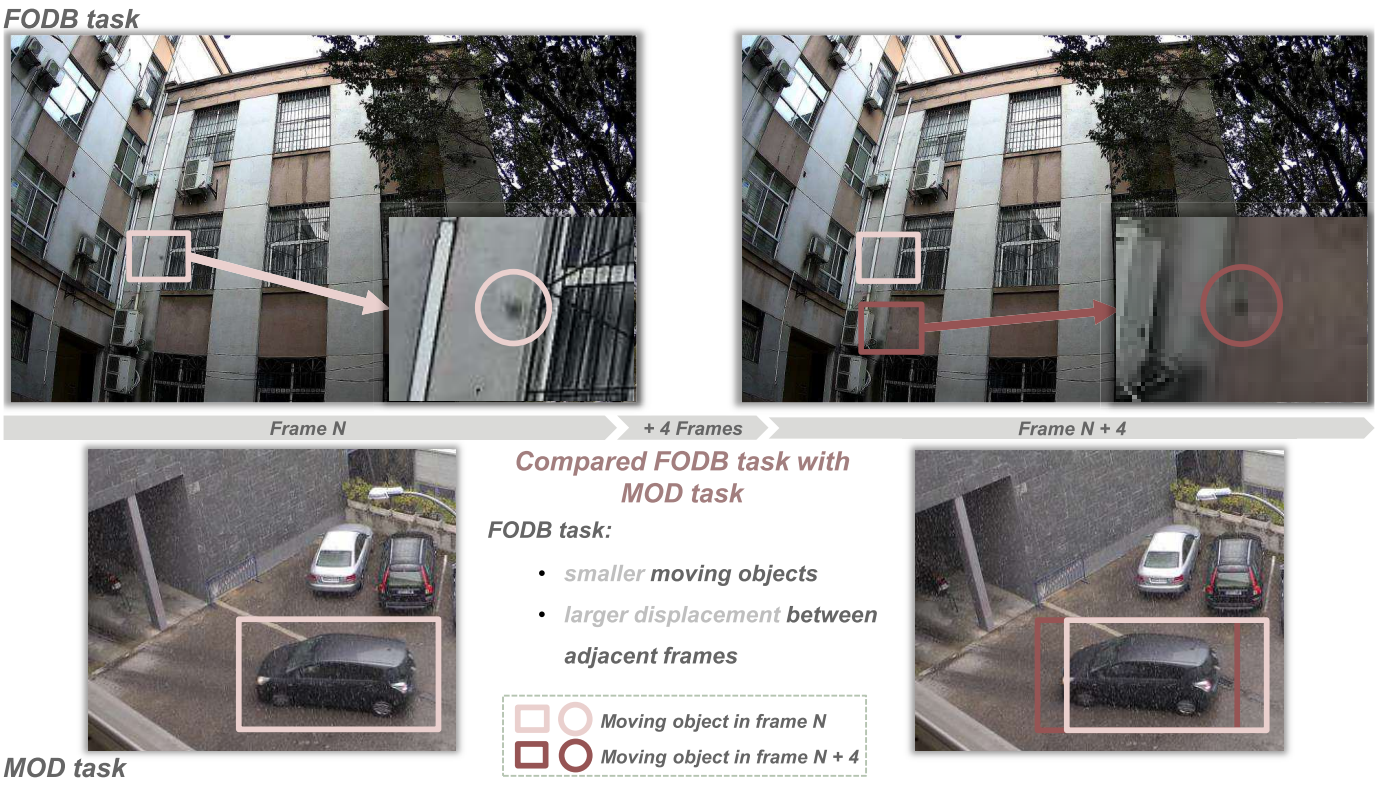

Fig. 2. Comparison the tasks of FODB and MOD. By observing four frames (two frames above are from our dataset, and two frames below are from the LASIESTA dataset [30]) from two video sequences, we can find that the moving object in FODB task is much smaller and has larger displacement between succesive frames. To see the falling object in the FODB task clearly, we enlarge the object and place it in the lower right corner of the video frame.

TABLE I

COMPARISON OF OUR FADE DATASET WITH THE PRIOR MOD DATASETS. "GT FRAMES" MEANS THE GROUND TRUTH FRAMES. THE DATASETS ARE SORTED BY THE PUBLISHING TIME IN ASCENDING ORDER. AS CAN BE SEEN, OUR FADE DATASET HAS THE LARGEST NUMBER OF VIDEOS, TOTAL FRAMES, GT FRAMES, AND SCENES

Dataset	Total videos ↑	Total frames	GT frames	Scenes	Frame-wise IOU (%)
SABS [27]	9	6,408	2,000	9	43
SegTrack V2 [31]	14	976	976	14	27
SBI [32]	14	5,024	14	14	31
GTFD [29]	25	1,067	1,067	7	37
CDnet 2014 [28]	31	90,000	90,000	11	33
LASIESTA [30]	48	18,425	18,425	4	35
DAVIS 2016 [33]	50	3,455	3,455	15	32
FADE (Ours)	**2,611**	**245,177**	**245,177**	**25**	**13**

task can be regarded as a special case of the moving object detection (MOD) task, and it may appear feasible to utilize existing MOD datasets, such as SABS [27], CDnet 2014 [28], GTFD [29], and LASIESTA [30], to train FODB algorithms. However, this approach is limited by the significant disparity between the two tasks. Specifically, moving (falling) objects in FODB are typically much smaller and move at higher speeds than the objects in standard MOD scenarios.

To better illustrate the differences between the FODB task and the MOD task, we present a visual comparison in Figure 2. As shown in Figure 2, moving objects in existing MOD datasets typically occupy large areas of the image, whereas falling objects in the FODB task are much smaller and cover only a small portion of the scene. Besides, falling objects exhibit rapid motion, resulting in (1) motion blur and (2) large displacements between adjacent video frames. These factors make falling objects difficult to detect and are rarely encountered in MOD datasets. Furthermore, none of the existing MOD datasets include categories corresponding to falling objects around buildings. Consequently, it is *necessary* to construct a large and diverse benchmark dataset to evaluate the performance of FODB algorithms. Such efforts are critical for advancing research and facilitating real-world deployment in FODB domain.

In this work, we present a video benchmark dataset for FODB, termed FADE, which comprises 2,611 videos captured across diverse scenes. Specifically, the dataset includes 245,177 annotated video frames spanning 25 scenes, 8 object categories, 4 video resolutions, 4 weather conditions, 3 camera angles, and 2 lighting conditions. As can be seen in Table I,

our dataset FADE has richer data compared to the existing MOD datasets. The size of the falling object in our FADE is small with a median area of about 20 pixels, which is also much smaller than the small object ($area < 32^2$ pixels) defined in the COCO dataset [34]. We also introduce a new baseline method, FADE-Net, which leverages motion cues to complement appearance features and incorporates a small-object mining region proposal network to generate high-quality proposals for small falling objects. We use the evaluation metrics in the MOD task to benchmark several methods of 3 different tasks (MOD, generic object detection, and video object detection) on our dataset. We also explore a new metric, time range overlap (TRO), to evaluate the performance of the detection methods on localizing the object falling incidents. The experimental results indicate that FODB is a challenging task and validate the effectiveness of our presented method.

The main contributions of our work are three-fold:

- We construct a new video dataset called FADE, which, in terms of application scenarios, is the first dataset for falling object detection around buildings (FODB). This dataset is large and diverse, which covers various scenes and complex conditions.
- We explore a new baseline method FADE-Net for the FODB task, which effectively utilizes motion cues and can generate high-quality proposals for small-sized falling objects detection.
- We evaluate the proposed FADE-Net, and other methods, *i.e.* MOD methods, generic object detection methods and video object detection methods, on our FADE dataset comprehensively, which can be served as a benchmark for future research on FODB.

II. Related Work

A. Moving Object Detection Dataset

Moving Object Detection (MOD) is a basic task in computer vision, and some MOD datasets have been released for training and testing the MOD algorithms. Early MOD datasets, such as Wallflower [35], SegTrack [36], and I2R [37], are limited in scale. For instance, Wallflower [35] is one of the earliest MOD datasets, containing only 7 video sequences, each with a single ground truth frame. SegTrack [36] consists of only 6 video sequences and 215 annotated frames. I2R [37] provides 10 video sequences, including scenes with dynamic backgrounds, challenging weather conditions, and gradual illumination variations. Later, some large scale and complicated MOD datasets are constructed. For instance, SABS [27] contains videos with 10 categories, and each video has 800 training frames. Some testing video frames in FBMS 59 [38] and SegTrack V2 [31] contain multiple moving objects. GTFD [29] provides a collection of 25 videos with both rigid and non-rigid moving objects. LASIESTA [30] consists of 45 videos and 18,425 video frames, recorded by the moving and static cameras. CDnet 2012 [39] and CDnet 2014 [28] serve as benchmarks for the IEEE Change Detection Workshop, with CDnet 2014 adding 22 videos and 5 new categories to CDnet 2012. Captured in the outdoor environment, BMC 2012 [40] includes a total of 29 real and synthetic videos in different weather conditions. DAVIS 2016 [33] supplies 50 high-quality and densely annotated videos. SBI [32] comprises 14 image sequences along with corresponding ground truth backgrounds and is the first dataset designed for evaluating background initialization MOD methods.

Unlike these MOD datasets, FADE is the first dataset for the falling object detection around buildings task which contains numerous videos and diverse scenes. The quantitative comparison between our FADE and the prior MOD datasets is shown in Table I.

B. Moving Object Detection

MOD has been extensively investigated due to its wide range of applications [41], [42], [43]. Many unsupervised MOD methods have been explored and can be broadly classified into two categories: background modeling and feature extraction. Specifically, for background modeling, parametric Gaussian mixture methods [9], [44], [45] have been proposed to represent the background in MOD. Barnich and Droogenbroeck [8] update the background by applying a novel random selection strategy. Baf et al. [46] apply Choquet integral [47] as an aggregation operator to aggregate color and texture features for MOD. Lin et al. [48] propose a dual-rate background modeling framework for foreground object detection, which leverages both short-term and long-term background models to enhance the accuracy of foreground inference. For feature extraction, Yang et al. [49] adopt an optical flow based method to detect moving objects. Zhou et al. [11] use motion information to enhance detection with motion fusion blocks that compressing video clips into a single image. Thanikasalam et al. [50] exploit a target-specific Siamese attention network that employs residual and channel attention modules to capture the global and channel-wise information of moving objects. Shang et al. [51], [52] consider MOD as a robust principal component analysis problem involving robust subspace learning and tracking. To address sudden and gradual background changes, Dong and DeSouza [53] explore a clustering feature space method to represent different background appearances.

The developments of deep learning [54], [55], [56] have greatly promoted the progress of supervised MOD methods. Early approaches can be broadly classified into six main categories: basic CNN [57], [58], [59], [60], [61], deep CNN [62], [63], 3D CNN [64], [65], ConvLSTM [66], FCN [67], and GAN [68], [69]. Specifically, CTFU-Net [70] hierarchically integrates the local feature extraction capabilities of CNNs with the global context modeling of Transformers to address dynamic scenes. TransBlast [71] employs an SVD-based subspace loss and Barlow Twins self-supervision to preserve foreground details while reducing the need for extensive annotations. GraphMOS-U [72] and GraphIMOS [73] are representative GNN-based approaches. GraphMOS-U [72] enables minimal-annotation underwater MOD by initializing Mask R-CNN with domain-specific feature fusion and Sobolev optimization. GraphIMOS [73] replaces transductive graphs with inductive block-diagonal GNNs to support real-time deployment on unseen videos.

Although numerous MOD methods have been proposed, few can be applied to the FODB task. In this work, we provide

(a) Fair (b) Cloudy (c) Overcast (d) Rainy

Fig. 3. To make our dataset design more accessible, we follow the commonly used Kaggle weather dataset [77] and categorize weather conditions into four broad types: fair, cloudy, overcast, and rainy.

benchmark results of state-of-the-art MOD methods on our proposed FODB benchmark dataset.

C. Falling Object Detection

There are a few works [74], [75], [76] concentrated on falling object detection. Specifically, [74] targets the prediction of potential falling objects (much like moving objects) in indoor environments by leveraging 3D point cloud data captured by distance sensors. [75] focuses on detecting objects which have already fallen (these are stationary) on railway tracks using ultrasonic sensors and signal coding sequence technology. [76] addresses falling hazards, that is, identifying stationary places a person might fall from or into, by analyzing the positional relationship between hazardous objects and workers at construction sites.

FODB is a critical task, as such incidents occur frequently, with approximately 50,000 cases reported annually in the United States [1]. These incidents pose fatal risks to pedestrians due to the high impact force of falling objects (see Section I). However, existing falling object detection methods [74], [75], [76] are not designed with small object detection and motion information utilization in mind, making them unsuitable for direct application to FODB scenarios, which involve detecting fast-moving small objects. Besides, the FODB field also lacks a public dataset, limiting both the training of learning-based methods and the evaluation of falling object detectors. To solve these issues, we construct the first large-scale benchmark dataset for FODB, and introduce a dedicated FODB method, FADE-Net.

III. The Proposed FADE Dataset

To advance research in the area of falling object detection around buildings (FODB), we construct and release the first benchmark dataset. In this section, we provide a detailed introduction to our FADE dataset, covering metadata, dataset construction, dataset splits and statistics, ethical considerations, licensing, maintenance plan, and evaluation metrics.

A. Metadata

To provide a comprehensive FODB dataset, we collect videos in diverse weather conditions, lighting cases, scenes, camera angles, and video resolutions. Example video frames are shown in our dataset (https://fadedataset.github.io/FADE.github.io/). Our metadata is defined as follows:

Object Category. To cover classes of the falling objects around buildings as many as possible, we collect 8 category of objects as follows: clothes, shoes, kitchen waste, books, spitballs, bottles, packaging bags, and packaging boxes.

Weather Condition. Different weather conditions lead to different light intensities, which affect the contrast between falling object and the background. To make our dataset design more accessible, we follow the commonly used Kaggle weather dataset [77] and classify weather conditions into four broad categories: fair, cloudy, overcast, and rainy, as illustrated in Figure 3.

Lighting Condition. Generally, when the light intensity is lower than 0.04 Lux, the image of the surveillance video will change from the RGB mode to the grayscale mode. Therefore, to ensure the generality of our dataset, we provide videos in both RGB mode and grayscale mode, corresponding to light intensities greater than and less than 0.04 Lux, respectively, as shown in Figure 4.

Scene. The scenes in our dataset are diverse. Specifically, the FADE dataset includes 18 distinct scenes that cover a wide range of environments where falling incidents may occur, such as classroom buildings, office buildings, dormitories, apartments, and buildings under construction, as illustrated in Figure 4. Notably, in our dataset, a scene refers to a specific viewpoint of a building. Across different scenes, the buildings appear distinct due to variations in the camera viewpoints.

Camera Angle. The videos in our dataset cover 3 different camera angles: 30°, 45°, and 60°. This design reflects real-world FODB surveillance setups, where cameras are typically installed 30 meters away from buildings, with different floors monitored using cameras positioned at varying angles.

Video Resolution. Usually, the surveillance video contains a variety of resolutions. To include multifarious data, we provide videos of 4 resolutions: 1280×720, 1920×1080, 2560×1440, and 2592×1520.

B. Dataset Construction

This section provides a detailed description of our dataset construction process, including data preparation, data

(a) Dormitories (RGB mode) (b) Classroom buildings (RGB mode) (c) Office buildings (RGB mode) (d) Classroom buildings (grayscale mode)

Fig. 4. Our dataset includes two modes (RGB and grayscale), captured across various scenes such as dormitories, classroom buildings, and office buildings.

collection, data format, and data annotation. Moreover, we provide dataset documentation, annotation guidelines, intended use cases, structured metadata, example videos, and evaluation code on our website https://fadedataset.github.io/FADE.github.io/index.html

1) Data Preparation: The diversity of object categories is important for training models for falling object detection around building. As stated above, the falling objects in our dataset span several common categories, including clothes, shoes, kitchen waste, books, spitballs, bottles, packaging bags, and packaging boxes. In addition, each category contains a diverse set of object instances. For example, the kitchen waste category includes items such as banana peels and uneaten apples.

2) Data Collection: We recruit seven volunteers and spend a year and a half to collect the video data. Specifically, we throw objects from the high-rise buildings of the corresponding university, and use a wide dynamic range camera, equipped with a $1/3''$ progressive scanning CMOS sensor and dot matrix LED infrared lamp to record the whole process of these events. The highest output video quality of the camera is 2592×1520 @ 30 FPS. We observe that some falling objects in our recorded video exhibit motion blur due to their high falling speed. Although recording videos with CMOS cameras equipped with global shutters and high FPS can mitigate this issue, the use of such specialized cameras presents two main drawbacks: (1) videos captured with these cameras often suffer from KTC noise [78], introducing visual artifacts; and (2) most existing surveillance systems deployed worldwide for monitoring falling objects use general-purpose, low-cost sensors. As a result, models trained on data from these specialized CMOS cameras may be less applicable to real-world FODB scenarios.

3) Data Format: The annotation format of our dataset is PASCAL VOC [79] style, which is one of the most popular dataset annotation formats. The detailed data format is provided at https://fadedataset.github.io/FADE.github.io/document.html

4) Data Annotation: The annotation process of our dataset lasts for half a year, including two rounds. In the first round, the novice annotator labels the video. In the second round, the expert annotator checks the annotation to improve the quality. Different from the pixel-by-pixel annotation method of the MOD datasets [27], [29], [32], we adopt the annotation manner used in the object detection task to generate the bounding box around each falling object.

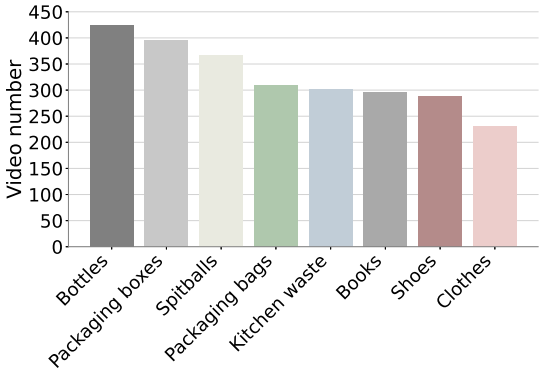

Fig. 5. Statistics of each object category's video number in our FADE dataset, sorted by descending order.

C. Dataset Division and Statistics

To provide a more intuitive understanding of our dataset, we present several quantitative statistics, including the dataset division, falling object sizes, the number of instances per object category, and the proportion of object area relative to the image.

1) Dataset Division: When splitting the dataset, we consider five attributes for each video: video resolution, scene, lighting condition, weather condition, and object category. It is worth noting that the videos in our training, validation, and testing sets collectively cover all labels from the five attributes mentioned above. To better test the generalization of FODB algorithms, the scenes in our training set, validation set, and testing set (except for the scenes captured in rainy days) are non-overlapped.

2) Dataset Statistics: We provide some statistical information about the constructed dataset. As shown in Figure 5, to make our FADE dataset with various falling object categories, we collect the videos covering falling objects of eight categories. The percentage of falling objects with different weights is shown in Figure 6.

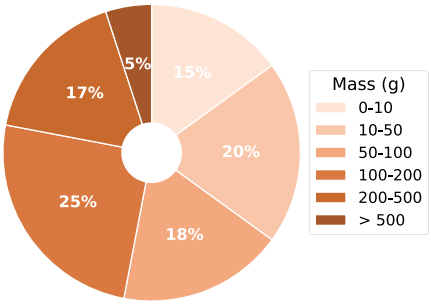

Fig. 6. Percentage of falling objects with different mass in our FADE dataset.

D. Dataset Ethics, License, and Maintenance Plan

To facilitate the use of our dataset, we provide details regarding its ethical considerations, licensing terms, and maintenance plan in this section.

1) Ethics: We have full ownership of these videos supplied in our FADE dataset, and all of the collected videos have been authorized for free use by the corresponding university. We inform the seven volunteers, who help us drop the objects from buildings, in advance that their personal information (e.g. faces and hands) may appear in the video, and we get their permission. We conduct two improvements to solve the ethical concerns in our dataset. 1) We have reviewed all the videos and found that 32 of them contain people. Among these, 21 videos include appearances of volunteers who assisted in the data collection process. For the 21 videos, the volunteers are informed of their appearance and have signed on an informed consent form to allow us to use the videos. The link of this signed informed consent form is: https://fadedataset.github.io/FADE.github.io/ethic.html. For the other 11 videos, we remove them from our FADE dataset. 2) We have obtained authorization from the corresponding university where the video data was captured. All videos in the updated FADE dataset are freely available for research on falling object detection. We upload the corresponding authorization files to our dataset website (https://fadedataset.github.io/FADE.github.io/ethic.html). Notably, the video in our dataset does not contain any personally identifiable information or offensive content, and we assume full responsibility in the event of any issues related to data licensing.

2) License: Our FADE dataset is published under the CC BY-NC-SA 4.0 license, which means everyone can use our dataset for the non-commercial research purpose. Our code is released under the Apache 2.0 license.

E. Evaluation Metrics

We use precision, recall, and F-measure to evaluate and compare the performance of methods across different tasks, including MOD, generic object detection, and video object detection. Since the object size in our dataset FADE is small, we treat a detection as a true positive if its IoU with the GT is larger than 0.3. The F-measure is defined as:

$$F\text{-}measure = \frac{(1+\beta^2)\cdot precision \cdot recall}{\beta^2 \cdot precision + recall}. \quad (1)$$

We set the positive real factor $\beta = 1$, when we calculate the F-measure.

Locating the falling incident temporally is crucial to find out the perpetrator who throws the falling object. We therefore design a useful metric named time range overlap (TRO) to evaluate the ability of the algorithm in this regard. The design of TRO is inspired by the DER (Diarization Error Rate) [83] in the speaker diarization task. The TRO is defined as:

$$TRO = \frac{TR_p \cap TR_{gt}}{TR_p \cup TR_{gt}}, \quad (2)$$

$$TR_p = [T_p^B, T_p^E], TR_{gt} = [T_{gt}^B, T_{gt}^E], \quad (3)$$

where TR_p indicates the predicted time range and TR_{gt} denotes the GT time range. T_p^B and T_{gt}^B are the predicted and the GT beginning time of a falling incident, respectively. T_p^E and T_{gt}^E separately indicates the predicted and the GT ending time of a falling incident.

IV. Proposed Method

Generic object detection methods rely heavily on appearance features to detect objects. However, in the FODB task, falling objects are typically small and often exhibit blurred appearances due to motion blur caused by high-speed movement, making them difficult to detect using appearance features alone. To address this limitation, we propose a method called FADE-Net, built upon Faster R-CNN [80] with FPN [81] (see Figure 7). Specifically, FADE-Net introduces a Moving Attention Module that incorporates motion cues to complement appearance features for improved detection of fast-moving objects. Additionally, a Small-Object Mining Region Proposal Network (SMRPN) is integrated in our FADE-Net to enhance the detection of small objects common in FODB scenarios. The details of MAM and SMRPN are described in the following sections.

A. Moving Attention Module (MAM)

Detecting falling objects around buildings based solely on appearance features is challenging, as these objects are often small and affected by motion blur due to their high speed. To enhance detection accuracy, we leverage the inherent motion characteristics of falling objects by designing a Moving Attention Module that incorporates motion information as a complementary cue to appearance features. Specifically, we first employ a Moving Object Prediction Module to extract a motion mask representing moving objects. This mask is then fed into the proposed Moving Attention Module, where it is fused with appearance features to produce more informative representations for falling object detection. The designs of the Moving Object Prediction Module and the Moving Attention Module are detailed in the following sections.

1) Moving Object Prediction Module: To obtain motion information as a complementary cue to appearance features for improved falling object detection, we design a Moving Object Prediction Module. Specifically, to balance robustness and real-time performance, we adopt the MOD method MOG2 [45] to generate moving object masks as motion cues. The masks are candidate moving object regions within the current frame.

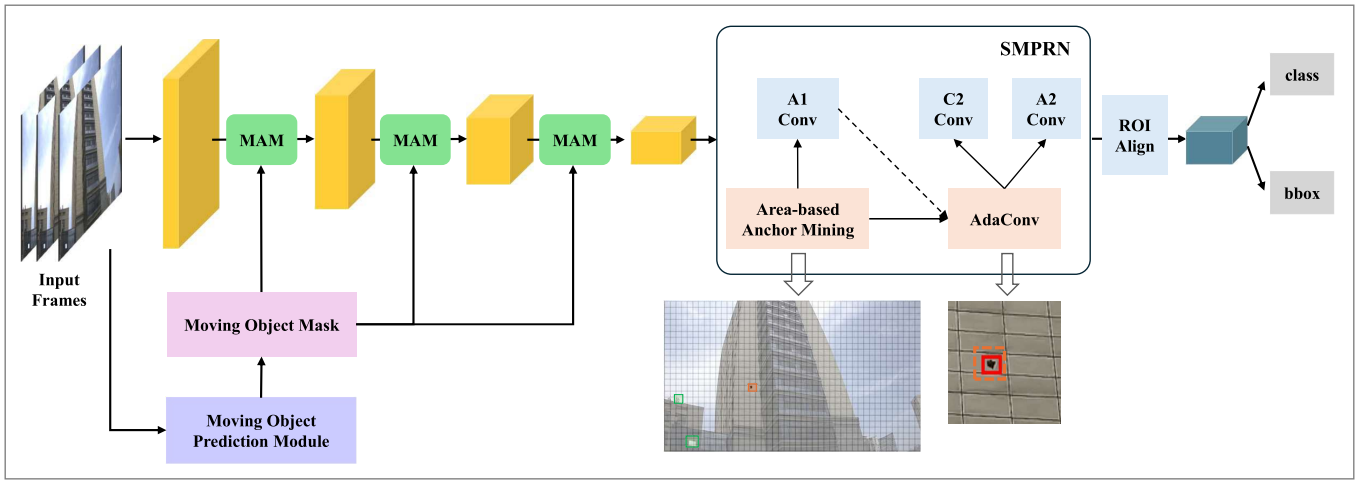

Fig. 7. Overview of our proposed method, which is built upon the Faster R-CNN [80] framework with Feature Pyramid Network [81].“MAM” denotes the Moving attention module. “C” and “A” denote classifier and anchor regressor, respectively. “Conv” and “AdaConv” indicate conventional convolution and the adaptive convolution [82] layers, respectively.

2) Moving Attention Module: The generated moving object mask, which serves as motion information, is fed into the Moving Attention Module and fused with appearance features to provide more robust representations for falling object detection. Specifically, in our Moving Attention Module, the appearance feature is represented by the feature map F from the previous layer, with dimensions $H \times W \times C$. To enhance the robustness of the appearance information, we apply both average pooling and max pooling to F before fusion.

We then concatenate the averaged and max-pooled feature maps with the moving object mask and feed them into a convolutional layer followed by a Sigmoid activation to fuse motion and appearance features. The fused output, after passing through the convolutional layer and Sigmoid activation, forms a more robust representation, referred to as the moving attention map, which serves as the output of our Moving Attention Module. The process is defined as:

$$M = \sigma(\text{Conv}(\text{Concat}(\text{AvgPool}(F), \text{MaxPool}(F), \text{Mask}))), \tag{4}$$

where M is the resulting moving attention map, and σ denotes the Sigmoid function. It is worth noting that our Moving Attention Module is applied at three locations, each positioned after the downsampling stages in our backbone. All Moving Attention Modules share the same moving object mask.

Overall, the proposed Moving Attention Module enables our detection model to effectively integrate appearance and motion information, resulting in more robust representations for detecting falling objects around buildings.

B. Small-Object Mining RPN (SMRPN)

After introducing the Moving Attention Module, we now present the Small-Object Mining RPN. While the Moving Attention Module enables the model to leverage both appearance and motion information for more robust detection of small objects, the small size of falling objects may still result in their features being lost during downsampling in the backbone. To address this challenge, we design the Small-Object Mining RPN (SMRPN), which integrates multi-level features and employs dynamic thresholds to ensure that small falling objects can be effectively captured by the model.

Specifically, SMRPN leverages features from all downsampled levels to perform the initial regression, enabling the model to effectively capture information across different object scales. We also adopt an adaptive convolution [82] to align anchor features with target features, allowing for more refined regression and ultimately generating high-quality proposals for small objects. In addition, to enhance the model's ability to detect typically small falling objects, we introduce an area-based anchor mining strategy with a dynamic threshold that adapts to object size in the first stage:

$$Threshold = \max\left(0.20, 0.15 + \alpha \cdot \log \frac{\sqrt{w \cdot h}}{5}\right), \tag{5}$$

where α is a scale factor, set to 0.2 by default in our task.

V. Experiments

In this section, we sequentially present the implementation details of our method, the main results, the ablation study, an analysis of the performance of optical flow-based methods, and visualization results.

A. Implementation Details

Our method is fine-tuned on a pre-trained Faster R-CNN [80] with a ResNet-50 [102] backbone. We use SGD as the optimizer with a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0005. The model is trained for 15 epochs with a batch size of 2 on our dataset. The Moving Attention Module consists of a convolution layer with a 7×7 kernel followed by a Sigmoid function. Its input includes a moving object mask predicted by MOG2 [45] along with its corresponding average pooling map and max pooling map. The kernel size of the convolution layer in our moving attention module is set to 7×7.

TABLE II

PERFORMANCE OF DIFFERENT METHODS ON THE TESTING SET OF OUR FADE DATASET. THE BEST AND SECOND BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY. DLA34 (DCNV2 [84]) [85] INDICATES THAT SOME OF THE CONVOLUTIONS IN THE DLA34 NETWORK ARE REPLACED BY DEFORMABLE CONVOLUTION V2 (DCNV2). FGFA (FLOWNET [86]) [87], FGFA (PWC-NET [88]) [87], AND FGFA (RAFT [89]) [87] RESPECTIVELY INDICATE FGFA BASED ON THE OPTICAL FLOW METHODS OF FLOWNET, PWC-NET, AND RAFT

Method	Type	F-measure	Precision	Recall	TRO	FPS
FADE-Net (Ours)	MOD-based Generic Object Detection	**72.08**	**73.52**	**70.69**	**51.77**	15.7
Faster R-CNN [80] + FPN [81]	Generic Object Detection	35.56	54.55	26.38	32.47	16.7
YOLOv5 [90]	Generic Object Detection	33.67	54.77	24.31	34.12	32.8
DLA34 (DCNv2 [84]) [85]	Generic Object Detection	22.57	20.04	25.83	24.69	22.7
DETR [91]	Generic Object Detection	29.98	49.47	21.51	26.80	7.9
swin-B [92]	Generic Object Detection	36.99	57.37	27.29	36.63	8.3
RT-DETR [93]	Generic Object Detection	<u>40.15</u>	<u>59.24</u>	30.36	38.62	28.7
MOG [94]	Moving Object Detection	17.96	12.85	29.80	20.91	<u>370.1</u>
MOG2 [45]	Moving Object Detection	24.55	19.03	34.57	48.03	331.8
GMG [95]	Moving Object Detection	2.71	1.55	10.89	14.71	97.7
Vibe [8]	Moving Object Detection	15.26	14.66	15.91	23.85	**488.5**
CNT [96]	Moving Object Detection	12.84	19.09	9.67	17.25	147.3
FMOD [97]	Moving Object Detection	2.48	6.73	1.52	16.09	289.7
KNN [98]	Moving Object Detection	0.83	0.42	30.48	34.20	188.5
GSOC [99]	Moving Object Detection	0.73	0.38	10.12	16.17	146.6
LSBP [100]	Moving Object Detection	0.20	0.10	6.73	12.35	158.0
MEGA [101]	Video Object Detection	5.20	2.71	<u>65.60</u>	<u>48.23</u>	8.3
FGFA [87] (FlowNet [86])	Video Object Detection	0.26	0.13	34.01	46.72	6.7
FGFA [87] (PWC-Net [88])	Video Object Detection	0.22	0.11	31.95	45.47	10.1
FGFA [87] (RAFT [89])	Video Object Detection	0.18	0.09	30.05	42.97	8.8

B. Main Results

To provide a comprehensive benchmark, we conduct extensive experiments on FADE, evaluating a range of state-of-the-art methods, including our proposed FADE-Net, five generic object detection methods, nine MOD methods, and two video object detection methods. We use the metrics defined in section III-E for performance evaluation. We specify the implementation details of all algorithms (requirements, hyperparameters, and training details) at https://github.com/Zhengbo-Zhang/FADE

As can be seen in Table II, our proposed FADE-Net achieves the highest performance across all four metrics. During the training process, the CNN generates feature maps that are significantly smaller than the original image, making it is challenging for generic object detection methods to capture features of small objects. However, the falling objects in our dataset are of small size. To address this challenge, our FADE-Net leverages multi-stage proposal refinement and an area-based anchor mining strategy, enabling more effective detection of small objects. Furthermore, motion blur caused by the fast motion of falling objects can reduce the recall of image-based detection models. To solve it, our FADE-Net integrates the proposed Moving Attention Module, which fuses motion information with appearance features to produce more robust representations for falling object detection.

Although MOG2 [45] is the best-performing MOD method, it still shows a substantial precision gap compared to our proposed FADE-Net. This is because FADE-Net effectively extracts and utilizes both appearance and motion features during inference, which significantly enhances falling object detection. In contrast, the MOD method can only update the model online during inference and lacks the ability to exploit various informative features. This leads MOG2 to detect certain moving objects that are not falling objects, such as clouds and shadows. The detection results of LSBP [100] contain numerous ghost regions, such as trailing areas behind fast-moving objects that lie outside their actual contours [103], resulting in the lowest precision and F-measure. MEGA [101] combines global semantic information with local localization cues and leverages more key frames in the video for detection, achieving the second-best recall and TRO performance. However, since MEGA is not designed for small object detection, its network struggles to capture the appearance features of small falling objects. As a result, MEGA yields low precision on the constructed FADE dataset.

The experimental results show that although our method is not the fastest in terms of inference speed, it achieves superior performance owing to the SMRPN and MAM modules, which are specifically designed for FODB. In terms of performance, our method significantly outperforms the second-best approach (F-measure 72.08 vs 40.15).

C. Ablation Study

In this part, we conduct ablation studies to evaluate the effectiveness of each module in our proposed method FADE-Net. As shown in Table III, both accuracy and recall are improved when adopting SMRPN. Notably, the recall is boosted by more than 20% (26.39% vs 46.51%). This indicates that multi-stage cascade refinement enhances the detection accuracy. Additionally, employing a dynamic area-based anchor mining strategy in the initial stage helps avoiding missed detection of small objects. The adoption of MAM resulted in a significant increase of 38.07% in the recall rate. This indicates that incorporating motion information in

TABLE III

PERFORMANCE OF DIFFERENT MODULES IN THE PROPOSED FADE-NET ON THE TESTING SET OF OUR FADE DATASET. WE USE A GTX 1080 TO EVALUATE THE FPS OF THE ALGORITHM. THE BEST PERFORMANCES ARE HIGHLIGHTED IN BOLD

Faster R-CNN+FPN	SMRPN	MAM	F-measure	Precision	Recall	TRO	FPS
√			35.62	54.78	26.39	32.50	**16.5**
√	√		53.42	62.75	46.51	39.19	16.2
√		√	61.72	59.21	64.46	47.25	15.9
√	√	√	**72.08**	**73.52**	**70.69**	**51.77**	15.7

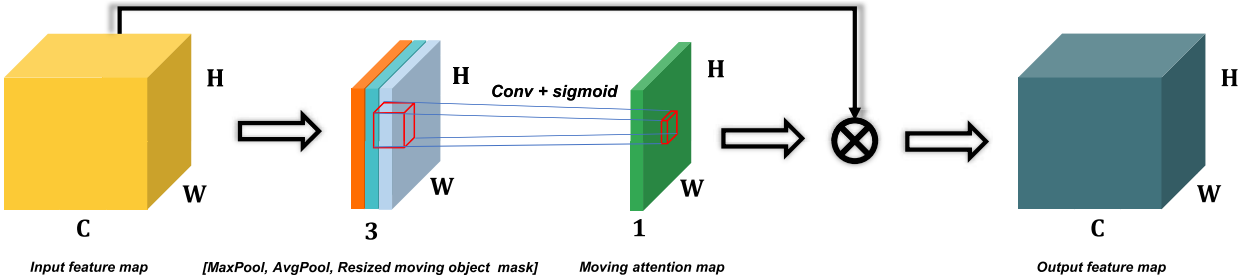

Fig. 8. Details of the Moving Attention Module. The Moving Attention Module is designed to fuse the appearance and motion information of objects. To improve the robustness of the appearance features, we apply average pooling and max pooling to the input appearance features before the fusion step.

our method enables effective capture of high-speed falling objects. The simultaneous usage of SMRPN and MAM further enhances the detection performance, suggesting that these two modules are complementary. Such improvement also shows that the ability to detect small objects and capture trajectories of moving objects contributes to the effectiveness of falling object detection around buildings. Besides, the experimental results show that introducing the SMRPN and MAM modules has minimal impact on the inference speed of our method. This is because the MAM module is built upon an efficient GPU-accelerated MOG algorithm, while the SMRPN functions primarily as a training-time strategy that employs dynamic thresholds to generate proposals tailored for small object detection.

D. Analysis of Optical Flow Based Methods in FODB

Optical flow is good at estimating motion information in video by capturing the pixel-level dense motion field. It is widely used in many video tasks, e.g. action recognition [106], [107], [108], [109] and MOD [101], and makes good performance. However, as shown in Table II, the optical flow based methods FGFA (FlowNet [86]) [87] (the original method using FlowNet to compute optical flow), FGFA (PWC-Net [88]) [87], and FGFA (RAFT [89]) [87], do not obtain the expected performance in falling object detection in our FADE (the F-measure of FGFA (RAFT) is the second worst).

We think that there are two main reasons for the poor performance of the optical flow based method in FODB task. The first reason is motion blur and occlusion. Appearance of a falling object constantly changes due to the motion blur caused by the fast movement as well as the occlusion caused by trees, neighboring buildings, et al. in the falling process. The second is large displacements. The falling object produces large displacements due to fast movement. With the challenges of motion blur, occlusion, and large displacements, the optical flow is hard to be estimated precisely.

E. Visualization Results

1) Optical Flow Images: As can be seen in Figure 9, we plot the optical flow images generated by 4 popular optical flow methods (FlowNet 2.0 [104], RAFT [89], PWC-Net [88], and TV-L1 [105]) to visualize the performance of the optical flow in our FADE dataset. The first three groups of optical flow images, estimated by the deep learning based optical flow algorithms [86], [88], [89], are very confusing, resulting in the background and the moving object being difficult to be distinguished. For the last group (*i.e.* last column) of optical flow images, which are captured by the traditional variational optical flow method [105], the movement of the static background is well estimated, but the motion of the moving objects is also poorly computed.

2) Visualization Results of Representative Methods: Visualization results of the representative moving object detection method (MOG2 [45]), image object detection method (Faster R-CNN [80]+FPN [81]), video object detection method (MEGA [101]), and our proposed FADE-Net are shown in Figure 10. Many detection results of MOG2 only capture part of the falling object, as MOG2 applies morphological opening in its post-processing, which tends to eliminate small detections. Thus, some small falling objects on the video frame cannot be detected. The accuracy of Faster R-CNN + FPN is higher than the other two algorithms. However, this method is also difficult to detect some small falling objects. In addition, the motion blur caused by the fast motion also increases the detection difficulty of the image-based detection model. The accuracy of MEGA is the worst. This method does not learn the appearance features and motion information of the falling object well during the training process, and the lack of non-maximum suppression processing further reduces its accuracy. Our proposed method FADE-Net achieves the best performance in all three video sequences, benefiting from seamlessly integrating generic object detection and motion information in videos. In addition, the designed Moving Attention Module

Fig. 9. Video frames in our FADE dataset and their corresponding optical flow field images estimated by different optical flow methods. (a) shows four consecutive video frames containing a falling object from a building in our dataset. (b), (c), (d), and (e) show the corresponding optical flow images computed by FlowNet 2.0 [104], RAFT [89], PWC-Net [88], and TV-L1 [105], respectively. To see the moving object in the falling object detection task clearly, we enlarge the object in the frame and place it in the lower left corner of the video frame.

enables the network to assign adaptive weights to different regions based on the motion information, effectively reducing false detections in the image.

F. Effect of Incorporating Long-Term Motion Information

Although our FADE-Net primarily focuses on utilizing short-term temporal motion information due to the limited computational capabilities of surveillance camera chips, in this section, we also explore the effectiveness of incorporating long-term motion information into our method.

Specifically, we revise our method to incorporate long-term temporal information to experiment with long-term temporal motion data. The revised method adopts a two-stream architecture: one stream captures short-term motion, while the other models long-term motion across the current frame and the preceding five frames. Here, we evaluate the revised model on the FADE dataset using a GTX 1080 GPU to simulate a realistic deployment scenario. The experimental results show that the performance of our original model is comparable to that of the revised model (see Table IV). We attribute this to the high velocity of falling objects, which causes large motion displacements over the long-term window, making it challenging for the model to extract reliable long-term motion cues.Moreover, we can observe from the experimental results that the revised model incurs a certain loss in inference efficiency compared to the original model (FPS 8.9 vs 15.7), which may limit its applicability in real-world falling object detection scenarios. Therefore, we focus on utilizing short-term temporal motion information.

TABLE IV
EFFECT OF INCORPORATING LONG-TERM MOTION INFORMATION IN THE PROPOSED FADE-NET. THE BETTER RESULT IS INDICATED IN **BOLD**

Method	F-measure	Precision	Recall	TRO	FPS
Ours w/o long-term	72.03	73.48	**70.70**	51.75	**15.7**
Ours w/ long-term	**72.08**	**73.52**	70.69	**51.77**	8.9

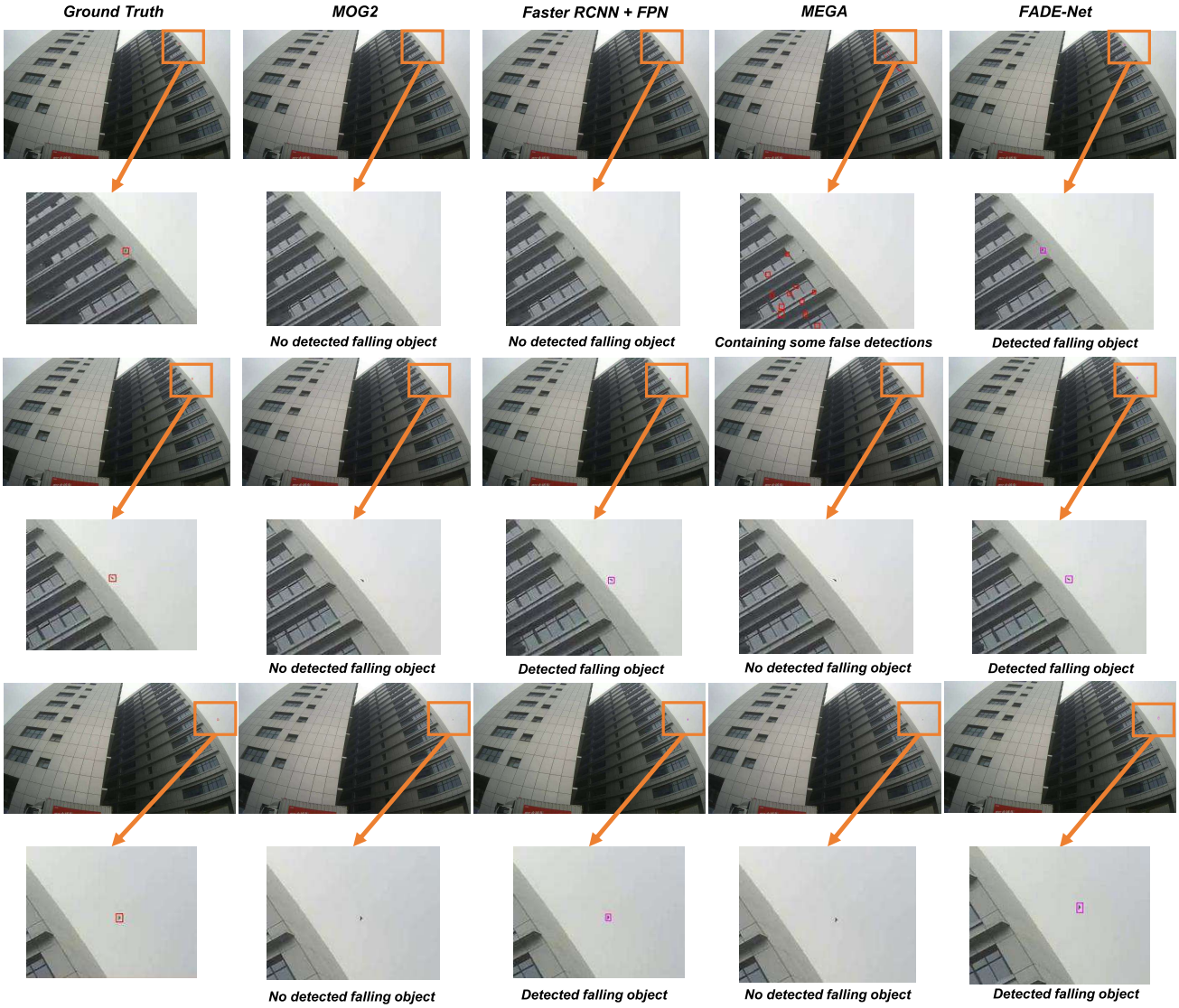

Fig. 10. We show ground-truth and the detection results of MOG2 [45], Faster R-CNN [80] + FPN [81], MEGA [101] and our proposed FADE-Net on three frames sampled from a video. MOG2, Faster R-CNN + FPN and MEGA are representative algorithms of moving object detection, image object detection, and video object detection respectively. In order to illustrate the difference between the detection results more clearly, we enlarge the regions around the falling objects in the frames.

G. Effect of Introducing Domain Adaptation Techniques Into Our Method

Since our falling object detection method is intended for real-world deployment and is expected to operate continuously, it may encounter unseen weather scenarios that were not present during training. To address this, we explore the integration of domain adaptation techniques [110], [111], [112], [113], commonly used in low-level vision tasks [110], [111], [112], [113], [114] such as image deraining [110], [111] and defogging [112], [113] to improve robustness, into our method and evaluate their effectiveness in this context.

Specifically, we design two variants of our method by integrating it with three different domain adaptation strategies, including an adversarial training-based approach using the Gradient Reversal Layer (GRL) [115], as well as a test-time domain adaptation methods, Tent [116]. To better evaluate these variants, we re-partitioned our proposed dataset so that the weather types in the training, validation, and test sets do not overlap. To be specific, the training set contains fair and cloudy conditions, the validation set contains overcast conditions, and the test set contains rainy conditions. Notably, this split is used only for this evaluation.

TABLE V

EFFECT OF DOMAIN ADAPTATION TECHNIQUES IN FADE-NET. WE USE A GTX 1080 TO EVALUATE THE FPS OF THE ALGORITHMS. THE BETTER RESULT IS INDICATED IN **BOLD**

Method	F-measure	Precision	Recall	TRO	FPS
Ours	72.08	73.52	70.69	51.77	**15.7**
Ours w/ GRL [115]	72.07	73.60	70.65	51.78	**15.7**
Ours w/ Tent [116]	**72.15**	**73.63**	**70.73**	**51.80**	10.7

The experimental results are presented in Table V above. From the results, we observe that the performance of Ours

and Ours w/ GRL is comparable. We believe this is because, although the domain adaptation technique (GRL) encourages the backbone to extract domain-invariant appearance features, in our task where the target objects are extremely small, effectively leveraging motion features is more critical than enhancing appearance features. This observation is further supported by the ablation study, which compares the effectiveness of the motion-focused Moving Attention Module and the appearance-focused Small-Object Mining RPN. In addition, as shown in Table V, our method benefits from the incorporation of the test-time domain adaptation method Tent. However, since Tent operates during inference, it introduces additional computational overhead and reduces the algorithm's FPS, making it less suitable for real-world deployment. Therefore, considering the overall trade-off between accuracy and inference speed, we adopt "Ours" as the final version.

VI. Conclusion

In this work, we propose a new large-scale video dataset termed FADE for falling object detection around buildings (FODB). It contains 2,611 videos and 245,177 video frames. The videos in FADE include various categories of objects and are captured under diverse scenes, weather conditions, lighting conditions, and video resolutions. Notably, different from the existing MOD datasets, our FADE is the first dataset specialized for FODB. Additionally, we introduce a new baseline method called FADE-Net, which seamlessly integrates motion information capturing and small-sized proposal mining into the detection network. Furthermore, to better evaluate the FODB algorithms, we design an evaluation metric called TRO, which measures the algorithm's ability to locate the beginning and ending times of falling incidents.

We provide a comprehensive benchmark that includes our FADE-Net baseline method, popular MOD methods, generic object detection methods, and video object detection methods. Extensive experimental results demonstrate that FODB is a challenging task in the presence of complex backgrounds and motion blur, and validating the effectiveness of our explored baseline method FADE-Net. The FADE dataset, with its diverse videos, will promote the progress of FODB and may also be useful for the investigation of MOD, generic object detection, and video object detection. In future, we will continue to refine and expand the dataset FADE, and explore more advanced FODB methods.

Acknowledgment

The numerical calculation was supported by the Super-Computing System in the Super-Computing Center of Wuhan University.

References

[1] U. B. Labor Statistics. (2019). *Commonly Used Statistics*. [Online]. Available: https://www.osha.gov/data/commonstats

[2] D. Morin, *Introduction to Classical Mechanics*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[3] (1990). *Code of Virginia*. [Online]. Available: https://law.lis.virginia.gov/vacode/title18.2/chapter4/section18.2-51.3/

[4] A. H. T. Council. (2023). *Laws*. [Online]. Available: https://www.ahtc.sg/by-laws/

[5] (2019). *Legal Document Issued By the Supreme People's Court*. [Online]. Available: https://www.chinadaily.com.cn/a/201911/14/WS5dcce956a310cf3e3557757b.html

[6] N. Babaguchi, A. Cavallaro, R. Chellappa, F. Dufaux, and L. Wang, "Guest editorial: Special issue on intelligent video surveillance for public security and personal privacy," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1559–1561, Oct. 2013.

[7] S. Aslani and H. Mahdavi-Nasab, "Optical flow based moving object detection and tracking for traffic surveillance," *Int. J. Electr., Comput., Energetic, Electron. Commun. Eng.*, vol. 7, no. 9, pp. 1252–1256, 2013.

[8] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

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

[10] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[11] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[12] G. Chen et al., "NeuroAED: Towards efficient abnormal event detection in visual surveillance with neuromorphic vision sensor," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 923–936, 2021.

[13] G. Yu, S. Wang, Z. Cai, X. Liu, E. Zhu, and J. Yin, "Video anomaly detection via visual cloze tests," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4955–4969, 2023.

[14] D. Peng, Z. Zhang, P. Hu, Q. Ke, D. K. Y. Yau, and J. Liu, "Harnessing text-to-image diffusion models for category-agnostic pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 342–360.

[15] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[16] K. L. Navaneet, R. K. Sarvadevabhatla, S. Shekhar, R. V. Babu, and A. Chakraborty, "Operator-in-the-loop deep sequential multi-camera feature fusion for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2375–2385, 2020.

[17] Y. Du, C. Lei, Z. Zhao, Y. Dong, and F. Su, "Video-based visible-infrared person re-identification with auxiliary samples," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1313–1325, 2024.

[18] Z. Zhang, L. Xu, D. Peng, H. Rahmani, and J. Liu, "Diff-tracker: Text-to-image diffusion models are unsupervised trackers," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, Apr. 2024, pp. 319–337.

[19] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.

[20] Z. Sun, J. Chen, L. Chao, W. Ruan, and M. Mukherjee, "A survey of multiple pedestrian tracking based on tracking-by-detection framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1819–1833, May 2021.

[21] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proença, "The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1696–1708, 2021.

[22] B. Degardin, V. Lopes, and H. Proença, "REGINA—Reasoning graph convolutional networks in human action recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5442–5451, 2021.

[23] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 1819–1831, 2022, doi: 10.1109/TMM.2022.3168137.

[24] Z. Tu et al., "Consistent 3D hand reconstruction in video via self-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9469–9485, Aug. 2023.

[25] J. Zhang, Z. Tu, J. Weng, J. Yuan, and B. Du, "A modular neural motion retargeting system decoupling skeleton and shape perception," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6889–6904, Oct. 2024.

[26] Z. Zhang et al., "Visual prompting for one-shot controllable video editing without inversion," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, May 2025, pp. 7784–7794.

[27] S. Brutzer, B. Höferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. CVPR*, Jun. 2011, pp. 1937–1944.

[28] Y. Wang, P. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDNet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit. workshops*, 2014, pp. 387–394.

[29] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 725–738, Apr. 2017.

[30] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Comput. Vis. Image Understand.*, vol. 152, pp. 103–117, Nov. 2016.

[31] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 2192–2199.

[32] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *Proc. ICIAP*. Cham, Switzerland: Springer, 2015, pp. 469–476.

[33] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Groß, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 724–732.

[34] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Mar. 2014, pp. 740–755.

[35] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf Comput. Vis.*, Apr. 1999, pp. 1–11.

[36] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, Nov. 2012.

[37] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[38] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.

[39] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2012, pp. 1–8.

[40] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre, "A benchmark dataset for outdoor foreground/background extraction," in *Proc. Asian Conf. Comput. Vis. Workshop*, Daejeon, South Korea, Nov. 2012, pp. 291–300.

[41] R. Cucchiara, C. Grana, G. Neri, M. Piccardi, and A. Prati, "The sakbot system for moving object detection and tracking," in *Video-Based Surveillance Systems*. Cham, Switzerland: Springer, 2002, pp. 145–157.

[42] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *Int. J. Soft Comput. Eng.*, vol. 2, no. 3, pp. 44–48, 2012.

[43] C. Poppe, S. De Bruyne, T. Paridaens, P. Lambert, and R. Van de Walle, "Moving object detection in the H.264/AVC compressed domain for video surveillance applications," *J. Vis. Commun. Image Represent.*, vol. 20, no. 6, pp. 428–437, 2009.

[44] S. Varadarajan, P. Miller, and H. Zhou, "Spatial mixture of Gaussians for dynamic background modelling," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 63–68.

[45] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. 17th Int. Conf. Pattern Recognit.*, Jun. 2004, pp. 28–31.

[46] F. El Baf, T. Bouwmans, and B. Vachon, "Fuzzy integral for moving object detection," in *Proc. IEEE Int. Conf. Fuzzy Syst. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1729–1736.

[47] H. Tahani and J. M. Keller, "Information fusion in computer vision using the fuzzy integral," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 3, pp. 733–741, Mar. 1990.

[48] M. K. S. Patil and P. G. A. Kulkarni, "Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 9, pp. 7–13, Sep. 2018.

[49] Y. Yang, A. Loquercio, D. Scaramuzza, and S. Soatto, "Unsupervised moving object detection via contextual information separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 879–888.

[50] K. Thanikasalam, C. Fookes, S. Sridharan, A. Ramanan, and A. Pinidiyaarachchi, "Target-specific Siamese attention network for real-time object tracking," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1276–1289, 2020.

[51] F. Shang, J. Cheng, Y. Liu, Z.-Q. Luo, and Z. Lin, "Bilinear factor matrix norm minimization for robust PCA: Algorithms and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2066–2080, Sep. 2018.

[52] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 32–55, Jul. 2018.

[53] Y. Dong and G. N. DeSouza, "Adaptive learning of multi-subspace for foreground detection under illumination changes," *Comput. Vis. Image Understand.*, vol. 115, no. 1, pp. 31–49, Jan. 2011.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 60, 2017, pp. 84–90.

[55] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2015, pp. 3431–3440.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[57] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2016, pp. 1–4.

[58] J. H. Giraldo, S. Javed, and T. Bouwmans, "Graph moving object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2485–2503, May 2022.

[59] T.-N. Le and A. Sugimoto, "Video salient object detection using spatiotemporal deep features," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5002–5015, Oct. 2018.

[60] P. W. Patil, K. M. Biradar, A. Dudhane, and S. Murala, "An end-to-end edge aggregation network for moving object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8149–8158.

[61] Z. Zhang, C. Zhou, and Z. Tu, "Distilling inter-class distance for semantic segmentation," 2022, *arXiv:2205.03650*.

[62] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.

[63] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Netw.*, vol. 117, pp. 8–66, Sep. 2019.

[64] M. Mandal, V. Dhar, A. Mishra, and S. K. Vipparthi, "3DFR: A swift 3D feature reductionist framework for scene independent change detection," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1882–1886, Dec. 2019.

[65] M. Mandal, V. Dhar, A. Mishra, S. K. Vipparthi, and M. Abdel-Mottaleb, "3DCD: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos," *IEEE Trans. Image Process.*, vol. 30, pp. 546–558, 2021.

[66] Y. Chen, J. Wang, B. Zhu, M. Tang, and H. Lu, "Pixelwise deep sequence learning for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2567–2579, Sep. 2019.

[67] D. Zeng and M. Zhu, "Background subtraction using multiscale fully convolutional network," *IEEE Access*, vol. 6, pp. 16010–16021, 2018.

[68] M. Sultana, A. Mahmood, T. Bouwmans, and S. K. Jung, "Complete moving object detection in the context of robust subspace learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–5.

[69] M. Sultana, A. Mahmood, S. Javed, and S. K. Jung, "Unsupervised deep context prediction for background estimation and foreground segmentation," *Mach. Vis. Appl.*, vol. 30, no. 3, pp. 375–395, Apr. 2019.

[70] T. Xia and Y. Yang, "CTFU-Net: CNN-transformer fusion U-shaped network for moving object detection," in *Proc. 3rd Int. Conf. Image Process. Media Comput. (ICIPMC)*, May 2024, pp. 44–50.

[71] I. Osman, M. Abdelpakey, and M. S. Shehata, "TransBlast: Self-supervised learning using augmented subspace with transformer for background/foreground separation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 215–224.

[72] W. Prummel, J. H. Giraldo, A. Zakharova, and T. Bouwmans, "Inductive graph neural networks for moving object segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Jun. 2023, pp. 2730–2734.

[73] M. Kapoor et al., "Graph-based moving object segmentation for underwater videos using semi-supervised learning," *Comput. Vis. Image Understand.*, vol. 252, Feb. 2025, Art. no. 104290.

[74] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu, "Detecting potential falling objects by inferring human action and natural disturbance," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 3417–3424.

[75] F. J. Álvarez, J. Urefia, M. Mazo, Á. Hernández, J. J. Garcia, and P. G. Donato, "Ultrasonic sensor system for detecting falling objects on railways," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2004, pp. 866–871.

[76] B. Yang, B. Zhang, Q. Zhang, Z. Wang, M. Dong, and T. Fang, "Automatic detection of falling hazard from surveillance videos based on computer vision and building information modeling," *Struct. Infrastruct. Eng.*, vol. 18, no. 7, pp. 1–15, Jul. 2022.

[77] N. Narayan. (2024). *Kaggle Weather Dataset*. [Online]. Available: https://www.kaggle.com/datasets/nikhil7280/weather-type-classification

[78] S. Lauxtermann, A. Lee, J. Stevens, and A. Joshi, "Comparison of global shutter pixels for CMOS image sensors," in *Proc. Int. Image Sensor Workshop*, Apr. 2007, pp. 1–8.

[79] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[80] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, vol. 28, 2015, pp. 91–99.

[81] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[82] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into high-quality region proposal network with adaptive convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–15.

[83] F. Yu et al., "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9156–9160.

[84] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.

[85] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

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

[87] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Mar. 2017, pp. 408–417.

[88] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, May 2018, pp. 8934–8943.

[89] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2020, pp. 402–419.

[90] G. Jocher. (2020). *YOLOv5*. [Online]. Available: https://github.com/ultralytics/yolov5

[91] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2020, pp. 213–229.

[92] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[93] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 35, Apr. 2024, pp. 16965–16974.

[94] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Cham, Switzerland: Springer, 2002, pp. 135–144.

[95] A. B. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *Proc. Amer. Control Conf. (ACC)*, Jun. 2012, pp. 4305–4312.

[96] S. Zeevi. (2016). *Backgroundsubtractorcnt*. [Online]. Available: https://github.com/sagi-z/BackgroundSubtractorCNT

[97] D. Rozumnyi, J. Matas, F. Šroubek, M. Pollefeys, and M. R. Oswald, "FMODetect: Robust detection of fast moving objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3521–3529.

[98] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, May 2006.

[99] G. S. Code. (2016). *Backgroundsubtractorgsoc*. [Online]. Available: https://docs.opencv.org/4.x/d4/dd5/classcv_1_1bgsegm_1_1BackgroundSubtractorGSOC.html

[100] L. Guo, D. Xu, and Z. Qiang, "Background subtraction using local SVD binary pattern," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1159–1167.

[101] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global–local aggregation for video object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10337–10346.

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2016, pp. 770–778.

[103] S. H. Shaikh, K. Saeed, and N. Chaki, *Moving Object Detection Using Background Subtraction*. Cham, Switzerland: Springer, 2014, pp. 15–23.

[104] E. Ilg, N. Mayer, T. Saikia, M. Keuper, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1–16.

[105] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L^1 optical flow," in *Proc. Joint Pattern Recognit. Symp.*, Cham, Switzerland: Springer, 2007, pp. 214–223.

[106] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 6299–6308.

[107] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal VLAD for video action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2799–2812, Jun. 2019.

[108] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.

[109] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[110] Z.-B. Chen and Y.-G. Wang, "DTT-Net: Dual-domain translation transformer for semi-supervised image deraining," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 1621–1625.

[111] Z. Pan, J. Wang, Z. Shen, S. Han, and J. Zhu, "Cross-domain collaborative learning for single image deraining," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118611.

[112] Q. Guo and M. Zhou, "Progressive domain translation defogging network for real-world fog images," *IEEE Trans. Broadcast.*, vol. 68, no. 4, pp. 876–885, Dec. 2022.

[113] X. Sun, Z. An, and Y. Liu, "D2SL: Decouple defogging and semantic learning for foggy domain-adaptive segmentation," 2024, *arXiv:2404.04807*.

[114] Z. Zhang, L. G. Foo, H. Rahmani, J. Liu, and D. W. Soh, "Performing defocus deblurring by modeling its formation process," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Apr. 2025, pp. 1–12.

[115] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 2, Jul. 2015, pp. 1180–1189.

[116] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," 2020, *arXiv:2006.10726*.