Exploiting Temporal Context for Tiny Object Detection

Christof W. Corsel[1], Michel van Lier[2], Leo Kampmeijer[2], Nicolas Boehrer[2], Erwin M. Bakker[1]

[1]LIACS, Leiden University, The Netherlands
[2]TNO Intelligent Imaging, The Netherlands
c.w.corsel@umail.leidenuniv.nl, {michel.vanlier, leo.kampmeijer, nicolas.boehrer}@tno.nl, e.m.bakker@liacs.leidenuniv.nl

Abstract

In surveillance applications, the detection of tiny, low-resolution objects remains a challenging task. Most deep learning object detection methods rely on appearance features extracted from still images and struggle to accurately detect tiny objects. In this paper, we address the problem of tiny object detection for real-time surveillance applications, by exploiting the temporal context available in video sequences recorded from static cameras. We present a spatio-temporal deep learning model based on YOLOv5 that exploits temporal context by processing sequences of frames at once. The model drastically improves the identification of tiny moving objects in the aerial surveillance and person detection domains, without degrading the detection of stationary objects. Additionally, a two-stream architecture that uses frame-difference as explicit motion information was proposed, further improving the detection of moving objects down to 4×4 *pixels in size. Our approaches outperform previous work on the public WPAFB WAMI dataset, as well as surpassing previous work on an embedded NVIDIA Jetson Nano deployment in both accuracy and inference speed. We conclude that the addition of temporal context to deep learning object detectors is an effective approach to drastically improve the detection of tiny moving objects in static videos.*

1. Introduction

Deep learning object detection models have shown impressive results in applications such as autonomous driving [39], search and rescue [14], pedestrian detection [22], and surveillance [18]. However, the detection of tiny, low-resolution objects consisting of only a few pixels remains a challenge [26].

In real-time surveillance systems such as wind farm monitoring [50] and aerial surveillance [32, 49], limited camera resolution and large distances between camera and targets require object detection methods to identify objects with a tiny pixel footprint. The costs of acquiring and processing high-resolution data required for these applications is computationally and monetarily expensive, driving researchers to develop methods to detect objects at lower image resolutions [38].

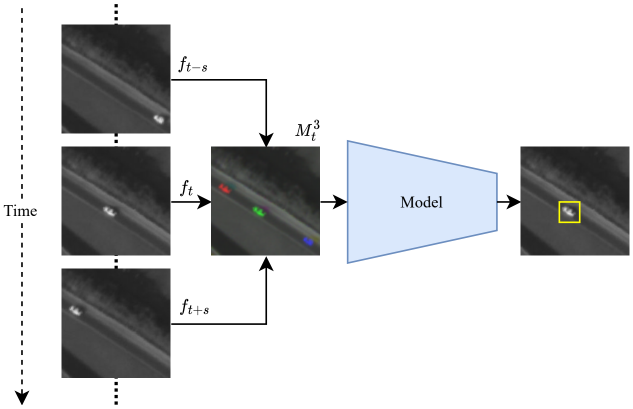

Figure 1: Overview of our proposed spatio-temporal object detection approach. Three video frames are combined into a 3-channel image. A deep learning object detector detects objects by exploiting the temporal context.

The small, indistinctive appearance features of target objects poses to be a challenge in the Tiny Object Detection (TOD) field. Their low signal-to-noise ratio makes the objects difficult to distinguish from background [51]. Other challenges include densely clustered objects and large backgrounds. Furthermore, tiny objects are underrepresented in large-scale object detection benchmarks like MS COCO [24]. Recently, research into improving the detection of tiny objects has increased. However, the detection accuracy of state-of-the-art general object detectors like Cascade R-CNN [6] or Deformable DETR [55] is lacking when applied to TOD datasets [13].

Methods specifically designed for TOD focus on improving appearance feature collection [23, 38, 54]. They often ignore temporal context, even though this information is available for many surveillance applications that record and process video streams. Temporal information is frequently used in the Moving Object Detection (MOD) field [40]. However, these methods are sensitive to noise and struggle to detect stationary objects. Deep learning models designed for video object detection often employ feature aggregation between frames [12, 30, 56]. These methods were found to be ineffective for TOD, where object feature correlation between frames is difficult [4].

In this paper, we aim to combine the advantages of both deep learning and MOD approaches by allowing a deep learning object detection model to exploit temporal information (Figure 1). Based on the real-time YOLOv5 [19] detector, we present single-stream and two-stream spatio-temporal detectors, which do not require computationally expensive additions to the model architecture like optical flow modules [56], Long-Short-Term-Memory (LSTM) layers [11] or tracker modules [4]. Contrary to MOD methods, our method also allows for the detection of stationary objects. In this work, the following contributions are made:

- A spatio-temporal detector based on YOLOv5 is introduced. The model exploits temporal context to improve the detection of tiny moving objects without deteriorating performance on stationary objects.
- A two-stream architecture which uses frame differencing is proposed to extract explicit motion information and further enhance the detection of moving objects.
- Experiments are conducted on the public WPAFB WAMI dataset, as well as on our own-recorded person detection dataset. The proposed methods outperform previous work, showing impressive performance on tiny objects smaller than 4×4 pixels.
- Various model architecture sizes are evaluated and shown to outperform previous work in terms of accuracy and inference time on the embedded NVIDIA Jetson Nano platform.

2. Related Works

Typically, object detection systems use single frames as input, relying on spatial and appearance information. These general deep learning systems can be subdivided into two-stage and single-stage networks. Two-stage networks like Cascade R-CNN [6] and Hybrid Task Cascade [10] use a Region Proposal Network (RPN) to preselect areas of interest in the image, on which the second stage performs further object classification and localization. Single-stage methods such as the anchorbox-based YOLO family [33, 34, 35, 3, 19] or anchorbox-free methods like TOOD [15], FCOS [42] and YOLOX [16], merge these two stages to increase inference speed. Recently, vision transformer-based models like DETR [8] have shown impressive results on benchmarks like MS COCO [24]. However, their performance on small objects was found to suffer due to the low object feature resolution. Strides have been made to solve this issue in works like Deformable-DETR [55] and DINO [53]. However, their applicability to the TOD domain is yet to be proven.

In order for single-frame methods to better deal with tiny objects, various approaches have been proposed. Multiscale networks, derived from Feature Pyramid Network (FPN) [23], merge features from various stages and downscaling levels in the network to improve the multiscale capabilities of the models [9]. It allows models to utilize features from early layers, which contain the detailed spatial information that is crucial for detecting tiny objects. Multiscale architectures like PANet [25] or BiFPN [41] have been prevalent in state-of-the-art object detection models.

Another approach is the use of Super Resolution (SR) networks to upscale the low-resolution objects. Some methods use SR models as a pre-processing step in the object detection pipeline, by increasing image resolution and recovering features of tiny objects [38, 45, 48]. Other methods use SR models to upscale and classify the detector output to filter false positives [2, 54]. SR methods require computationally expensive upscaling modules or are specific to certain object types, limiting their general applicability.

Non-deep-learning MOD algorithms have been widely used to detect tiny objects in Wide Area Motion Imagery (WAMI) data. These methods are often based on frame differencing [20] or background subtraction [29]. Frame differencing techniques calculate pixel-wise intensity differences between frames to highlight moving objects. Background subtraction compares the current frame against a created background model to detect differences. Using the median image of a sequence of 10 consecutive images was found to be an accurate way to remove background noise [36]. However, computationally expensive techniques like graph matching [47] are used to detect and track moving objects. Other downsides of these methods are the requirement of accurate frame registration and the sensitivity to noise and parallax effects. Furthermore, these methods cannot detect stationary objects [40].

To exploit temporal context found in video data, Video Object Detection (VOD) methods often use feature aggregation to combine object features from multiple frames. FFAVOD [30] applies feature aggregation by merging features from sequence frames surrounding the target frame using 1×1 Convolutional Neural Network (CNN) layers. FGFA [56] uses optical flow vectors to warp sampled frames to overlap with target frames. In MEGA [12], global video context is used by sampling frames from the complete video sequence. These methods were found to increase performance on the ImageNet-VID [37] dataset, with models being able to better deal with challenges like motion blur, rare object poses and occlusion. However, previous work found that these approaches do not translate well to TOD datasets, where correlating object features from multiple frames is difficult due to their small size [4].

Thus, spatio-temporal models designed for TOD applications generally employ different techniques. ClusterNet [21] generates an object heatmap by stacking five grayscale frames and processing them with a two-stage CNN model through a coarse-to-fine approach. The first stage aims to find general regions in the input image where moving objects are to be expected. These regions are further processed by a second CNN to locate individual objects. The method outperformed previous MOD methods on the WPAFB 2009 WAMI dataset [1], showing the potential of spatio-temporal methods for TOD applications. TrackNet [17] uses a similar heatmap-based approach for tracking ball positions in sports applications. It uses a stack of three RGB images which are processed by a 2D CNN network to detect fast moving, small objects. A deconvolutional network is used to generate the object location heatmaps. T-RexNet [7] is a spatio-temporal TOD model for embedded applications. It extracts motion information from three frames by explicit frame differencing, highlighting the difference in frames for the model. After this, it uses two CNN streams to process appearance and motion data.

Inspired by spatio-temporal TOD methods, we introduce temporal context to a YOLOv5 object detector by using a multi-frame input. Our approach enables spatio-temporal object detection using standard object detection model architectures, and does not require heatmap-based outputs or computationally expensive feature aggregation modules.

3. Methods

The YOLOv5 [19] object detection model was chosen as the basis for our approach for its high accuracy and real-time inference speed. The model is easily scalable to various model sizes which trade off detection accuracy for inference speed, allowing it to be applied to a wide range of hardware.

3.1. YOLOv5 Overview

YOLOv5 is part of the popular YOLO family of single-stage object detectors [33, 34, 35, 3, 19], providing accurate, real-time object detection. The model uses the CSP-Darknet53 backbone [44, 3] and a PANet [25] architecture to provide multiscale detections. Both backbone and head architectures consist of Cross Stage Partial [44] C3 modules. By adjusting the number of channels (width) in the C3 modules and the number of such modules in the network (depth), the architecture can be scaled to balance inference speed and detection accuracy. YOLOv5 provides architecture scales similar to those presented by Scaled-YOLOv4 [43]. The models are defined as *Nano* (YOLOv5n), *Small* (YOLOv5s), *Medium* (YOLOv5m), *Large* (YOLOv5l), and *Extra Large* (YOLOv5x). For our approach, we employed the YOLOv5x architecture for its high detection accuracy.

3.2. T-YOLOv5: Exploit Temporal Context

To introduce temporal context to the model, a multi-frame model input was used by sampling three consecutive frames from the video sequence. For each current frame f_t, two additional support frames f_{t-s}, f_{t+s} are sampled with temporal frame shift s. Edge cases caused by video boundaries are dealt with by duplicating the current frame f_t. The three frames are preprocessed into a single, 3-channel image M_t^3 by extracting a single gray-scale channel from each frame and stacking them. The three channels of the final input image are defined as:

$$M_t^3 = (f_t, f_{t-s}, f_{t+s}) \tag{1}$$

By embedding the temporal frames into the channels of the input image, the standard 2D CNN model architecture for three-channel RGB images can be utilized. The model is trained to output bounding box detections positioned with regard to the middle frame f_t, and does not require the additional support-frames to be labelled. Our approach trades colour information for temporal context, which we argue is a worthwhile trade-off for tiny object detection.

Data augmentation techniques [19] are applied during training to improve dataset variation. We propose Temporal Data Augmentation techniques which adapt augmentations for still images to the spatio-temporal detection domain. With Temporal Data Augmentation, all augmentations are equally applied to all sampled input frames, ensuring the absolute difference between the sampled frames remains the same. For augmentation methods that require other dataset samples, MixUp [52] and Mosaic [3], random sequences are sampled with the same temporal shift and combined with the original sequence. For Temporal Mosaic Augmentation, three additional random sequences are sampled and channel-wise mosaics are created. The mosaics undergo further data augmentations, where the same augmentations are applied to each mosaic. Finally, the resulting images are merged to integrate temporal information into the model input.

Temporal YOLOv5 (T-YOLOv5) can learn temporal features as the model processes a sequence of frames at once. The intensity difference between the frames will highlight moving objects, as their position changes frame-to-frame. The distinction between background and moving objects is largest in scenes without global camera motion, where the background overlaps accurately. This makes our approach useful for surveillance applications with static cameras. For applications with moving cameras, such as drone surveillance, global motion needs to be removed by using frame registration [36], similar to techniques applied for MOD approaches [21, 40]. However, in contrast to MOD methods, our approach can also detect stationary objects due to the inclusion of appearance features. Compared to previous spatio-temporal models, it does not require explicit motion features generated by optical flow or frame-difference and can operate on the raw video frames. Although we focus on single-class object detection in this work, our approach can be used for multi-class detection applications.

3.3. T2-YOLOv5: Two-Stream Approach

Using a frame sequence input allows our model to use motion information to detect objects. To further enhance this capability, we add a second stream to our model that is tasked with generating motion-only features, following the method presented by T-RexNet [7]. The additional stream allows part of the model to specialize to extract features from motion-only images. The images are generated by calculating the absolute difference of the input frames. The extracted motion-only features are later combined with appearance features from the main stream, which processes the original input frames. The input image M_t^2 of the second stream is the 2-channel, absolute frame difference between the key-frame and support-frames:

$$M_t^2 = (|f_{t-s} - f_t|, |f_{t+s} - f_t|) \tag{2}$$

As shown in Figure 2, the YOLOv5s model backbone [19] was used for the second stream to prevent a large computational overhead. Since YOLOv5 predicts bounding boxes on three different feature map scales, the features from the motion stream are combined with those of the main stream using a concatenation block at each scale. This allows motion-only features to be used by all detection output heads.

4. Experiments

4.1. Datasets

To validate our approach for aerial surveillance, we utilize the public WPAFB 2009 [1] dataset. For person detection, public surveillance datasets like UA-DETRAC [46] were found to lack a sufficient amount of tiny objects, as these objects are ignored during labelling. Therefore, we construct a custom tiny object detection dataset from the public VIRAT-Ground surveillance dataset [27], and augment it with our own acquired dataset called TwitCam.

WPAFB 2009 [1]: A publicly available aerial WAMI dataset with labelled vehicle tracks. The dataset consists of a gray-scale video sequence recorded by a matrix of six overlapping $1.25Hz$ camera sensors. The images are available at various resolution scales ($R0-R5$), where the highest quality images ($R0$) are $26K \times 21K$ pixels. Subsequent resolution scales decrease resolution by a factor of 2. In contrast to previous work, multiple resolution scales ($R0-R3$) are used, where object sizes range from 30×30 to 4×4 pixels. We argue that this provides better insight into the model's performance on tiny objects. Scales $R4$ and $R5$ were omitted from testing, as many target objects were no longer visible due to the low image resolutions.

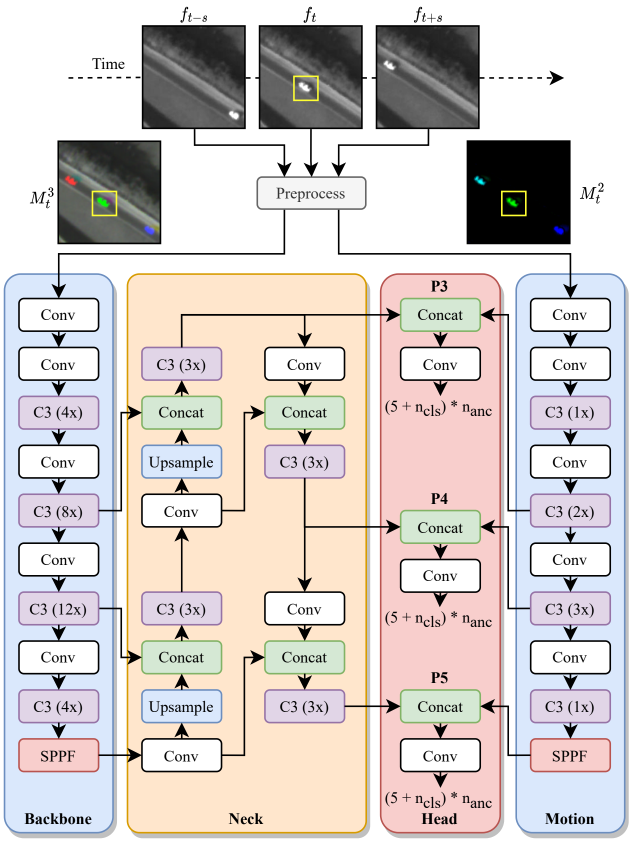

Figure 2: T2-YOLOv5 architecture. Explicit motion information extracted by frame-differencing is processed by a second motion stream. The extracted features are combined at three detection scales $P3$ to $P5$.

Due to the significant size of the dataset frames, Areas Of Interests (AOIs) are selected according to the procedure described in previous work [21, 7]. We register the frames to remove global motion and extract AOIs 1, 2, and 3 for our experiments. An overview of the used AOIs can be seen in Figure 3. The size of the AOIs range from 2300×2300 pixels on scale $R0$, to 287×287 pixels on scale $R3$. To enable detections using bounding boxes, the single-point ground truth labels were converted to fixed size bounding boxes. The bounding box sizes were set to 30×30, 15×15, 8×8, and 4×4 pixels for the scales $R0$ to $R3$ respectively.

Previous work often removes static objects from the

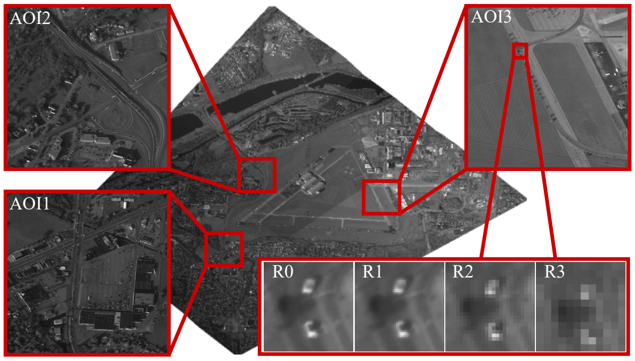

Figure 3: WPAFB [1] selected AOIs and object examples.

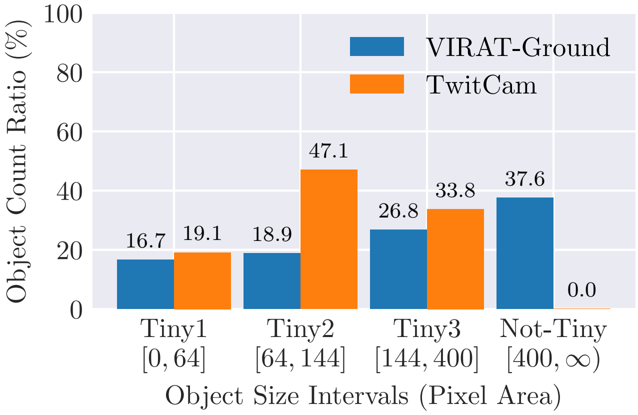

Figure 4: Object size ratio statistics of downscaled datasets.

dataset to focus on moving object detection [40, 7]. For comparisons, static objects were filtered following previous work [7]. Objects that move less than $15px$ between two consecutive frames of the $R0$ set were removed from the dataset. For the lower resolution scales, the same ground truth set was used. As our method allows for the detection of stationary object by using the appearance information of objects located in the frame, additional experiments that include stationary objects were performed. We found that static objects are labelled inconsistently, with many objects remaining unlabelled. Thus, missing labels were manually added to create the *persistent WPAFB* dataset.

VIRAT-Ground [27]: This surveillance dataset consists of 329, 30 fps video clips from 11 unique camera locations in city environments. A selection of the clips was used based on their camera location and label quality. As the supplied detection labels contained errors and missed objects, we manually adjusted the dataset labels. The resulting train set consists of 10 sequences and a total of 18K frames. The dataset classes used for training were *person* and *vehicle*. To increase the number of small objects in the dataset, we have decreased the image resolution by a factor of 4, using bicubic downsampling.

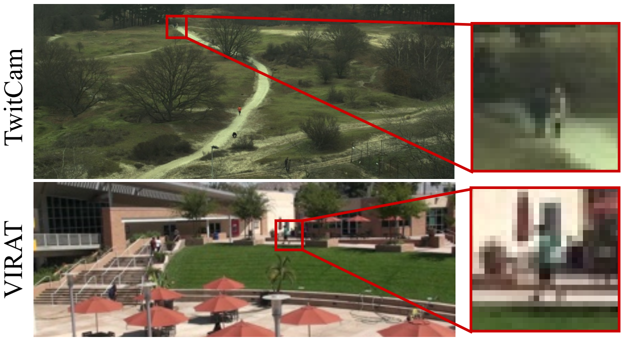

Figure 5: TwitCam and VIRAT dataset samples.

TwitCam: This person detection dataset contains three, 2 fps sequences of a natural environment. The set consists of 3000 frames and 19K labelled persons. We adapt this data for TOD by applying bicubic downsampling with a factor of 4, leading to a frame size of 1280×410 pixels. This downsampling causes most objects to be smaller than 20×20 pixels. Figure 4 and 5 show the object size distributions and dataset samples of the downscaled TwitCam and VIRAT datasets.

4.2. Metrics

$\mathbf{F_1}$ score: For the aerial surveillance dataset experiments, we provide the F_1 scores following previous work [7, 21]. A detection is classified as True Positive (TP) if the distance between its bounding box middle coordinate and that of a ground truth label is below a pixel distance threshold. This threshold value is set to $20px$, $10px$, $5px$, $3px$ for the WPAFB $R0$ to $R3$ resolution scales.

Average Recall (AR): For the comparisons on the *persistent WPAFB* dataset, we employ AR as defined by the MS COCO [24] evaluation standard. A detection is classified TP, if its Intersection Over Union (IOU) exceeds the threshold of 50%. A ground truth object is classified as $static$, if it moves less than 10 pixels between two consecutive frames on the *R0* scale. Otherwise, it is classified as $moving$. The same classification is used for evaluation on the lower resolution scales.

Average Precision (AP): We follow the MS COCO [24] protocol for calculating AP scores. As the person detection datasets contain multiscale objects, AP is reported on various object size intervals as defined in previous work [51]. The intervals are defined as $tiny1 : [0, 64]$, $tiny2 : [64, 144]$, $tiny3 : [144, 400]$. The IOU threshold was set to 50%.

4.3. Setup

Model Training: Provided YOLOv5 [19] models trained on MS COCO [24] were used to fine-tune our models for the target datasets and architectures. Our T-YOLOv5

Table 1: Performance comparison on AOI 2 of the WPAFB dataset on resolution scales $R0$ to $R3$. Results indicated by (*) were retrieved from their respective paper. Best results in each column are shown in bold.

Methods	**R0**		**R1**		**R2**		**R3**		**Params**	**Inference Time**
	F_1	AP_{50}^{all}	F_1	AP_{50}^{all}	F_1	AP_{50}^{all}	F_1	AP_{50}^{all}	(M)	(ms)
ClusterNet * [21]	95.1	-	-	-	-	-	-	-	-	-
T-RexNet * [7]	91.0	-	-	-	-	-	-	-	2.4	70.3
YOLOv5x [19]	89.6	91.4	85.7	86.1	65.9	56.5	49.0	23.9	86.7	347.7
Median BG + N * [40]	89.0	-	-	-	-	-	-	-	-	-
MOD + T	81.3	-	79.7	-	67.6	-	38.2	-	-	-
T2-YOLOv5	**95.5**	**95.7**	**94.0**	**95.0**	**92.1**	**89.7**	**85.0**	**70.0**	90.3	384.6
T-YOLOv5	95.4	95.6	93.7	94.8	91.4	89.1	83.6	64.4	86.7	347.7
T2-YOLOv5l	95.5	95.7	93.7	94.8	91.1	89.6	83.8	65.5	50.2	239.1
T2-YOLOv5m	95.3	95.4	93.6	94.5	90.9	87.9	81.4	57.0	25.0	147.1
T2-YOLOv5s	94.8	95.4	92.7	93.4	89.4	85.9	69.3	38.9	11.2	88.5
T2-YOLOv5n	94.1	94.8	92.2	92.9	87.1	79.8	60.2	24.5	2.8	52.9
T-YOLOv5l	95.2	95.6	93.4	94.2	90.7	88.7	81.7	58.5	46.5	202.2
T-YOLOv5m	95.2	95.6	93.2	93.8	90.7	87.1	77.2	49.1	21.2	112.2
T-YOLOv5s	94.7	95.1	92.7	93.4	88.9	83.3	62.2	28.4	7.2	53.5
T-YOLOv5n	93.6	94.5	91.9	92.4	85.5	75.5	51.7	14.0	**1.9**	**32.0**

models were fine-tuned for 300 epochs using the SGD optimizer [5] with an initial learning rate of 3.34×10^{-3} and weight decay of 2.5×10^{-4}. For T2-YOLOv5, the fine-tuned T-YOLOv5 weights were transferred to the main stream. The provided YOLOv5s model trained on MS COCO was used to initialize the motion-only stream. The model was further fine-tuned for 300 epochs using the same parameters as used for the T-YOLOv5 models.

We additionally train smaller versions of our models following the architecture sizes as defined by YOLOv5. For the T2-YOLOv5, the motion-only stream followed the *Small* scale for the *Large*, *Medium*, and *Small* models. For the *Nano* model, the motion-only stream was adjusted to match the YOLOv5n architecture size.

Aerial Surveillance Experiment: Following literature [7], we train our model on *AOI1* and *AOI3* and perform the evaluation on *AOI2* of the WPAFB dataset [1]. Additionally, we provide comparisons on four resolution scales. For the scales $R0$ and $R1$, image tiling [28] was applied during training due to the large frame sizes. The frames are split into 640×640 pixel tiles with 5% overlap. For the lower resolution scales $R2$ and $R3$, the input image size was set to 575×575 and 287×287, respectively. During evaluation, the full image resolution was used without tiling. For the temporal models, temporal frame shift s was set to $0.8s$, matching the dataset frame rate.

As previous work do not present results on the lower WPAFB resolution scales, we present results of a custom non-deep-learning MOD method applied to these scales. This method (MOD+T) utilizes background subtraction [29] and a tracking algorithm [31] which associates objects between frames.

Person detection Experiment: For the person detection experiments, all models are trained on the VIRAT-Ground dataset and evaluated on the TwitCam dataset. 320×320 tiling with an overlap of 5% was used during training. During evaluation, an input size of 1280×1280 pixels was used. The temporal shift parameter s was set to $0.5s$ to match the frame rate of the TwitCam dataset.

Embedded Deployment: Our models were deployed on the NVIDIA Jetson Nano development board to investigate their inference speed on embedded hardware. Models were optimized for this hardware by converting them to the TensorRT (TRT) framework, following previous work [7]. Models using the half-precision (FP16) representation were benchmarked with an input size of 512×512 pixels. Additionally, the board's performance configuration was set to the high-performance *Max-N* mode.

4.4. Results

Aerial Surveillance Results: Table 1 provides comparisons on the WPAFB 2009 dataset [1] between our temporal approaches, the still image YOLOv5 [19], and previous work [21, 7, 40]. Results show that our approaches outperform pervious work on the *R0* scale. On this scale, our T-YOLOv5 model improves over the still-image YOLOv5 model by 6.5% in F_1 score and 4.6% in AP score. However, as resolution is reduced, this increases to 70.6% and

Table 2: AR score comparison on the *persistent WPAFB* dataset. No static objects are removed from the ground truth set.

Methods	**R0**		**R1**		**R2**		**R3**	
	AR_{50}^{moving}	AR_{50}^{static}	AR_{50}^{moving}	AR_{50}^{static}	AR_{50}^{moving}	AR_{50}^{static}	AR_{50}^{moving}	AR_{50}^{static}
T2-YOLOv5	**96.2**	70.8	**95.1**	63.9	**89.0**	42.3	**54.4**	2.4
T-YOLOv5	96.1	68.2	94.7	64.8	88.5	42.0	49.5	**2.9**
YOLOv5x [19]	91.5	**71.0**	82.3	**65.6**	72.0	**42.6**	27.1	**2.9**

169% on the $R3$ scale. This demonstrates the importance of using temporal context when dealing with tiny moving objects. T2-YOLOv5 improves performance over T-YOLOv5 for all scales, with the largest improvement seen on the *R3* scale. The addition of the second motion-only stream increases F_1 score by 1.7% and AP by 8.7%. It shows that the addition of explicit motion information is especially useful for detecting tiny objects with little appearance features.

In addition, our approach is scalable to smaller model architectures. The smallest T-YOLOv5n model's F_1 score decreased by only 2% compared to the best performing model on the $R0$ scale. However, we find that as resolution shrinks, the detection performance of smaller models quickly deteriorates. This shows that larger model architectures are necessary to detect tiny objects accurately.

Table 2 provides a comparison of our approaches against YOLOv5 [19] on the *persistent WPAFB* dataset. Our approaches outperform YOLOv5 on moving objects, even when many stationary objects are included. As shown previously, T2-YOLOv5 is especially effective for detecting tiny objects, exceeding YOLOv5 by 101% on the *R3* scale. Furthermore, the AR results of our approaches for static objects are comparable to those of the YOLOv5 models. This shows our approach can successfully use spatio-temporal information to enhance the detection of moving objects without deteriorating the detections of static objects.

Embedded Deployment Results: Table 1 additionally reports the inference speeds of our models on the NVIDIA Jetson Nano platform after TensorRT conversion. The T-YOLOv5n model outperforms competing small-scale model T-RexNet [7] by 2.9% in F_1 score on the *R0* scale whilst more than halving the inference time. It shows our approach can be effectively applied to embedded applications.

Person Detection Results: An AP performance comparison between our approach and the YOLOv5 baseline for our person detection TwitCam dataset is presented in Table 3. Our T-YOLOv5 model outperforms YOLOv5 by 50% in overall AP score. The largest increase in performance is attributed to the smallest objects smaller than 8×8 pixels in size, with T-YOLOv5 increasing AP^{tiny1} by 90%. Similar to the results for aerial surveillance, our T2-YOLOv5 model further improves AP^{all} by 2.9% on this dataset compared to T-YOLOv5. Likewise, the largest improvement of 7.2% is seen on the smallest $tiny1$ object size category. Additionally, qualitative results visualized in Figure 6 show that our approach aids the detection of distant and partially occluded objects.

Table 3: AP results on the TwitCam dataset.

Methods	AP_{50}^{all}	AP_{50}^{tiny1}	AP_{50}^{tiny2}	AP_{50}^{tiny3}
T2-YOLOv5	**79.1**	**45.9**	**87.3**	**90.0**
T-YOLOv5	76.9	42.8	85.3	89.6
YOLOv5x [19]	51.2	22.5	57.2	67.7

Table 4: Performance comparison of model input representations on the TwitCam dataset.

Methods	AP_{50}^{all}
T2-YOLOv5	**79.1**
T-YOLOv5	76.9
Motion-Augmented [7]	73.6
Temporal Colour	70.1
T-YOLOv5 + Motion	67.0
Colour [19]	51.2
Gray-Scale	49.6

4.5. Ablation Study

On the input representations: Besides three gray-scale frames, other input representations were considered. Table 4 provides the experimental results for the following representations of the TwitCam dataset: Gray-Scale: Three channel gray-scale representation of the current frame. Motion-augmented [7]: Three channel input consisting of a single gray-scale channel and two motion-only channels generated by frame difference: $M_t^3 = (f_t, |f_{t-s} - f_t|, |f_{t+s} - f_t|)$. Temporal Colour: Instead of using three gray-scale images, three RGB images are stacked into a 9-channel input so colour information is not lost. T-YOLOv5 + Motion: Merges the T2-YOLOv5 input representation into a single stream: $M_t^5 = (f_t, f_{t-s}, f_{t+s}, |f_{t-s} - f_t|, |f_{t+s} - f_t|)$. Experiments show the addition of colour or motion data do not improve the detection performance on these datasets.

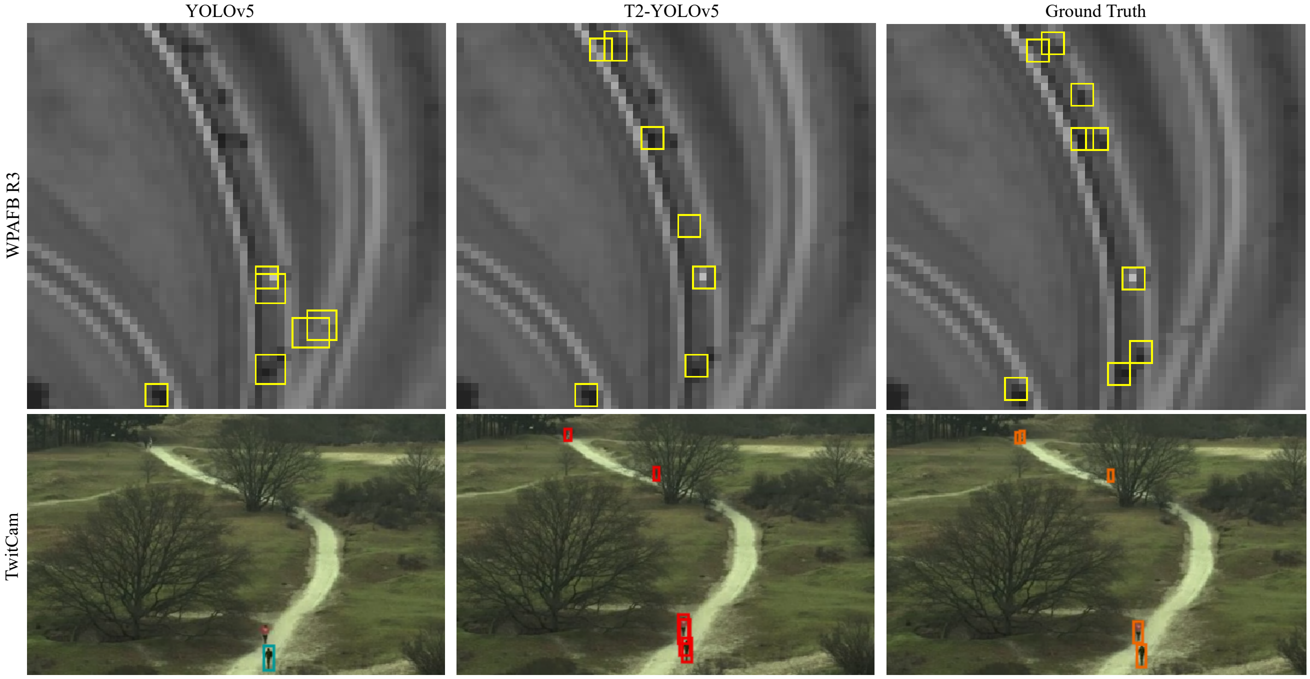

Figure 6: Qualitative detection results on crops of WPAFB $R3$ and TwitCam datasets. T2-YOLOv5 detects more objects and eliminates false positives. Furthermore, it enables partially occluded objects to be detected.

Table 5: AP_{50}^{all} performance comparison on number of sampled support frames and sample strategy.

	# frames	0	2	4	6
TwitCam	balanced	49.6	**76.9**	70.7	71.5
	online	49.6	76.7	67.8	61.8
WPAFB R3	balanced	23.9	**64.4**	55.9	62.5
	online	23.9	61.8	49.7	56.6

On the support frame sampling: In this experiment, we investigate the impact of the frame sampling strategy and the number of sampled support frames on the detection performance. The balanced sampling strategy keeps the target frame in the middle of the sequence. The online strategy [30] places the target frame at the end of the sequence, so predictions are made on the most recent frame. This removes the detection delay of the balanced method, where the detector must wait for future frames before making predictions. The number of sampled temporal frames are varied from 0 to 6 with a step size of 2. Table 5 presents the numerical results on the $R3$ scale of the WPAFB dataset [1] and TwitCam. We find that the balanced sampling strategy is superior to the online method. Furthermore, sampling more than 2 support frames does not improve detection performance.

5. Conclusion

This work shows that including temporal context is an effective technique to improve the tiny object detection performance of deep learning object detectors. We present a spatio-temporal network based on YOLOv5 for aerial surveillance and person detection applications. Our approach enables the detector to exploit temporal context by using three temporal gray-scale channels as model input. Additionally, we propose a two-stream network that utilizes motion-only information extracted by frame-differencing to enhance the detection of tiny moving objects. Our approaches were shown to outperform previous work on the WPAFB 2009 dataset. Furthermore, the detection of tiny moving objects improved over the still-image YOLOv5 baseline, without deteriorating the performance on stationary objects. Our approach is scalable to various network architecture sizes, exceeding competing detectors suitable for embedded applications in both accuracy and inference speed. In addition, the still-image baseline was outperformed on our recorded person detection dataset, showing the general applicability of our approach.

Acknowledgement

Some of the research presented in this paper was part of the Caviar/Mantis project, funded by the PENTA consortium.

References

[1] AFRL. Wright-patterson air force base (wpafb) dataset., 2009.

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 210–226, Cham, 2018. Springer International Publishing.

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[4] Brais Bosquet, Manuel Mucientes, and Víctor M. Brea. Stdnet-st: Spatio-temporal convnet for small object detection. *Pattern Recognition*, 116:107929, 8 2021.

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.

[6] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[7] Alessio Canepa, Edoardo Ragusa, Rodolfo Zunino, and Paolo Gastaldo. T-rexnet—a hardware-aware neural network for real-time detection of small moving objects. *Sensors*, 21:1252, 2 2021. Use frame differencing techniques and small mobilenet SSD model to allow for small object detection.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[9] Guang Chen, Haitao Wang, Kai Chen, Zhijun Li, Zida Song, Yinlong Liu, Wenkai Chen, and Alois Knoll. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(2):936–953, 2022.

[10] K. Chen, W. Ouyang, C. Loy, D. Lin, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, and J. Shi. Hybrid task cascade for instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4969–4978, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[11] Xingyu Chen, Junzhi Yu, and Zhengxing Wu. Temporally identity-aware ssd with attentional lstm. *IEEE Transactions on Cybernetics*, 50:2674–2686, 6 2020.

[12] Y. Chen, Y. Cao, H. Hu, and L. Wang. Memory enhanced global-local aggregation for video object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10334–10343, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[13] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *arXiv preprint arXiv:2207.14096*, 2022.

[14] Zsolt Domozi, Daniel Stojcsics, Abdallah Benhamida, Miklos Kozlovszky, and Andras Molnar. Real time object detection for aerial search and rescue missions for missing persons. In *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*, pages 000519–000524, 2020.

[15] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[16] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[17] Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsì-Uí İk, and Wen-Chih Peng. Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019.

[18] Sudan Jha, Changho Seo, Eunmok Yang, and Gyanendra Prasad Joshi. Real time object detection and trackingsystem for video surveillance system. *Multimedia Tools and Applications*, 80(3):3981–3996, 2021.

[19] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Feb. 2022.

[20] Mark Keck, Luis Galup, and Chris Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 441–448, 2013.

[21] Rodney LaLonde, Dong Zhang, and Mubarak Shah. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, pages 1–1, 2017.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] 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.

[25] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[27] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, 2011.

[28] F. Ozge Unel, Burak O. Ozkalayci, and Cevahir Cigla. The power of tiling for small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[29] Kannappan Palaniappan, Mahdieh Poostchi, Hadi Aliakbarpour, Raphael Viguier, Joshua Fraser, Filiz Bunyak, Arslan Basharat, Steve Suddarth, Erik Blasch, Raghuveer M. Rao, and Guna Seetharaman. Moving object detection for vehicle tracking in wide area motion imagery using 4d filtering. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2830–2835, 2016.

[30] Hughes Perreault, Guillaume-Alexandre Bilodeau, Nicolas Saunier, and Maguelonne Héritier. Ffavod: Feature fusion architecture for video object detection. *Pattern Recognition Letters*, 151:294–301, 11 2021.

[31] N. Prabhakar, V. Vaithiyanathan, Akshaya Prakash Sharma, Anurag Singh, and Pulkit Singhal. Object tracking using frame differencing and template matching. *Research Journal of Applied Sciences, Engineering and Technology*, 4(24):5497–5501, Dec 2012.

[32] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[35] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[36] Vladimir Reilly, Haroon Idrees, and Mubarak Shah. Detection and tracking of large number of targets in wide area surveillance. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 186–199, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[38] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[39] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand, and Ahmad El-Sallab. Modnet: Motion and appearance based moving object detection network for autonomous driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2859–2864, 2018.

[40] Lars Wilko Sommer, Michael Teutsch, Tobias Schuchert, and Jürgen Beyerer. A survey on moving object detection for wide area motion imagery. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[41] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635, 2019.

[43] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13029–13038, 2021.

[44] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[45] Zhuang-Zhuang Wang, Kai Xie, Xin-Yu Zhang, Hua-Quan Chen, Chang Wen, and Jian-Biao He. Small-object detection based on yolo and dense block via image super-resolution. *IEEE Access*, 9:56416–56429, 2021.

[46] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 2020.

[47] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng Han. Vehicle detection and tracking in wide field-of-view aerial video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 679–684. IEEE, 2010.

[48] Mowen Xue, Theo Greenslade, Majid Mirmehdi, and Tilo Burghardt. Small or far away? exploiting deep super-resolution and altitude data for aerial animal surveillance. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 509–519, 2022.

[49] Tao Yang, Xiwen Wang, Bowei Yao, Jing Li, Yanning Zhang, Zhannan He, and Wencheng Duan. Small moving vehicle detection in a satellite video of an urban area. *Sensors*, 16(9), 2016.

[50] Ryota Yoshihashi, Rei Kawakami, Shaodi You, Tu Tuan Trinh, Makoto Iida, and Takeshi Naemura. Finding a needle in a haystack: Tiny flying object detection in 4k videos using a joint detection-and-tracking approach, 2021.

[51] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[52] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 10 2017.

[53] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[54] Yongqiang Zhang, Mingli Ding, Yancheng Bai, and Bernard Ghanem. Detecting small faces in the wild based on generative adversarial network and contextual information. *Pattern Recognition*, 94:74–86, 2019.

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[56] 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 (ICCV)*, Oct 2017.