This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924188, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier xx.xxxx/ACCESS.20xx.DOI

A Framework for Knowing Who is Doing What in Aerial Surveillance Videos

FAN YANG[1,2], SAKRIANI SAKTI[1,2](Member, IEEE), YANG WU[1](Member, IEEE), SATOSHI NAKAMURA[1,2](Fellow, IEEE),

[1]Nara Institute of Science and Technology, Nara, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

Corresponding author: Yang Wu (e-mail: yangwu@rsc.naist.jp, wuyang0321@gmail.com).

ABSTRACT Ultra-high-resolution aerial videos are used to relieve the shortage of surveillance system in sparsely populated regions. For realistic application purpose, it is important to automatically analyze "who is doing what?" in such videos. Although atomic visual action (AVA) detection has been successfully used to recognize "who is doing what?" in movie data, it is challenging to adapt it to ultra-high-resolution aerial videos, where the target persons are relatively tiny and sparsely located. Besides, due to the lack of evaluation metrics, AVA detection has been evaluated by the single-label action, however, using multi-label actions in evaluation is more reasonable since several actions can be simultaneously performed by a person (*e.g.*, making a phone call and walking). To tackle these issues, we propose a novel framework for multi-label AVA detection in ultra-high-resolution aerial videos, and, introduce novel metrics for multi-label AVA detection evaluation. Experimental results demonstrate that our framework outperforms other methods for interpreting "who is doing what?" in our target task.

INDEX TERMS Aerial Surveillance Videos, Multi-label Atomic Visual Action Detection.

I. INTRODUCTION

Surveillance cameras are commonly installed in city regions to increase public safety. However, it is inapplicable to densely set up surveillance cameras in sparsely populated regions (*e.g.*, suburb), while the safety concern is needed therein. Considering the fact that some of the sparsely populated regions are not covered by tall trees or buildings, it is possible to periodically take surveillance videos by drones. Due to drones' mobility, a wide range of sparsely populated regions can be monitored at a low cost.

To facilitate the efficiency of surveillance analysis, it is desirable to automatically analyze "who is doing what?" in surveillance videos. In movie data, Atomic Visual Action (AVA) detection was proposed to detect the spatio-temporal location and action for each person [1], which means, "who is doing what?" can be determined at each frame in videos. Nonetheless, aerial surveillance videos have some special properties and existing AVA detection methods may not work properly on them. These special properties include: (1) to capture visual details from the sky, each frame of aerial surveillance videos is preferred to be an **ultra-high-resolution** image (*e.g.*, 2160×3840); (2) relative to the entire aerial image, each person appears to be a **tiny** object but could still contain a **large** amount of pixels, which are sufficient for obtaining his/her actions; (3) persons are **sparsely located**; (4) the drone could move **fast**, resulting in **significant relative position shift** of the targets even in adjacent frames.

To approach AVA detection in aerial surveillance videos, we specifically designed a new framework by constructing new modules to seamlessly integrate object detection, multi-object tracking, and action recognition (see FIGURE 1).

Object detection plays a fundamental role in AVA detection, which locates each person in a spatial domain by bounding boxes. An ultra-high-resolution aerial image, however, is too large to be the input of normal object detectors [2]–[5], while down-scaling it could impair detection performance. As an alternative approach, an ultra-high-resolution aerial image could be cropped into smaller patches before performing object detection. Some existing methods divide the entire aerial image into patches by a sliding window [6]–[8]. Although such methods have considerably improved object detection performance, they are inefficient when target objects are sparsely located. We propose a Clustering Region Proposal Network (C-RPN) to alleviate this issue. C-RPN works by only selecting patches that may include target

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924188, IEEE Access

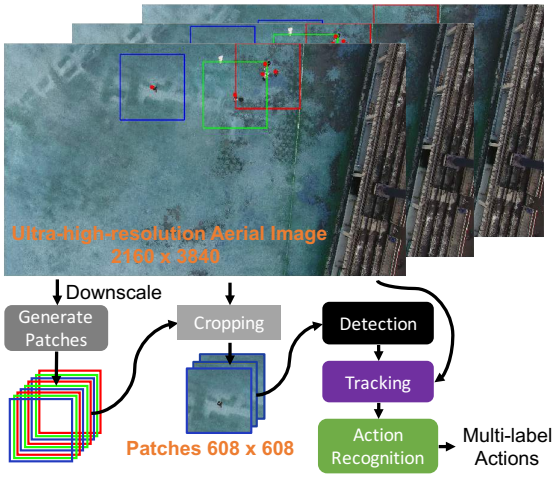


FIGURE 1: **Overview of proposed framework to know "who is doing what?" in aerial videos.**

objects. Subsequently, the number of selected patches could be fewer than using a sliding window when persons are sparsely located. In spite of that AVA detection estimates actions at each frame (*i.e.*, "is doing what"), spatio-temporal context is needed to obtain the person motion information. Generally, spatio-temporal tubes are used in AVA detection for such a purpose. Previous works [9], [10] obtain spatio-temporal tubes by extending bounding boxes from the central frame to nearby ones. In drone-recorded aerial videos, even if the absolute location of a person is static, its relative location may shift remarkably due to the drone movement. To eliminate the effect of drones' movement, we construct spatio-temporal tubes by a multi-object tracking method [11], and then align a spatio-temporal tube referred to its first frame. Since non-target objects might be included in the spatio-temporal tubes, action recognition performance could be affected. To tackle this issue, we assume the target person can be consistently observed in his/her spatio-temporal tube while others may not. Based on this assumption, we propose a novel Spatio-temporal Attention Module (STAM) to obtain attention for the target person in the spatio-temporal tube.

In addition, solely proposing the AVA detection framework is insufficient to fully analyze "who is doing what?" in aerial surveillance videos. In AVA detection, it is intuitive to consider that each person could take several actions simultaneously, which are corresponding to multi-label actions. For instance, a person could be making a phone call and walking at the same time. Due to the lack of evaluation metrics, AVA detection has been evaluated with only the single-label action for a while [1], [12]. Therefore, we provide novel metrics for multi-label AVA detection evaluation, which also contributes to the general AVA detection studies.

In summary, our **contributions** include: (1) proposing a novel framework for multi-label AVA detection on aerial surveillance videos, which outperforms other methods in our experiments; (2) providing novel metrics for multi-label AVA detection evaluation. To the best of our knowledge, existing metrics cannot be applied to multi-label AVA detection, and we are the first to introduce such metrics.

II. RELATED WORKS

In this section, we briefly discuss related works of object detection on aerial videos, AVA detection, and related datasets.

A. OBJECT DETECTION ON AERIAL VIDEOS

Detecting tiny objects is a nontrivial problem and many studies are trying to tackle it. Basically, there could be two cases in tiny object detection. One is the entire image has a low resolution and thereby the tiny objects only contain a few numbers of pixels. To improve the detection performance, amplification [13] and resolution enhancement [14] are applied. In another case, the object itself has plenty of pixels, but the object only constitutes a very small portion of the entire image so that it is relatively tiny. An ultra-high-resolution aerial image belongs to the second case and performing object detection on the original image size is desired.

Although the idea of transforming each frame of aerial videos into smaller patches for object detection has been around for some time [6]–[8], it is only recently that region proposals and clustering have been jointly applied to reduce the number of patches when objects are sparsely located [15]. Using the downsized aerial image, promising regions that may contain objects can be learned by density map regression. Based on the predefined patch size, these regions can be further clustered by their relative distances.

We assume that a good clustering strategy should satisfy two conditions: **first, reducing the number of patches; second, keeping the object appearance complete in patches**. However, to some extent, these two conditions work against each other. Solely satisfying the first condition may lead to an object being partially cropped, while assigning each object to a patch can effectively satisfy the second condition but may introduce redundant patches. In the previous study [15], grid-based clustering is used. Nevertheless, it is limited by predefined grid size and location, and thus objects may be incompletely cropped and further affect the bounding box detection. To resolve this issue, peak point Non-Max Suppression (NMS) and hierarchical clustering are used in our C-RPN, attempting to make every object complete in at least one patch.

B. ATOMIC VIDEO ACTION DETECTION

Atomic Video Action (AVA) detection concentrates on the study of the action unit in videos. Just like the word unit (*i.e.*, the unit of natural language processing) is explored to understand an article, the phoneme unit (*i.e.*, the unit of speech) is analyzed to understand human speaking, and the object detection/segmentation (*i.e.*, the unit of image) is studied to understand high-level visual tasks in images. AVA detection is important for understanding video scenes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924188, IEEE Access

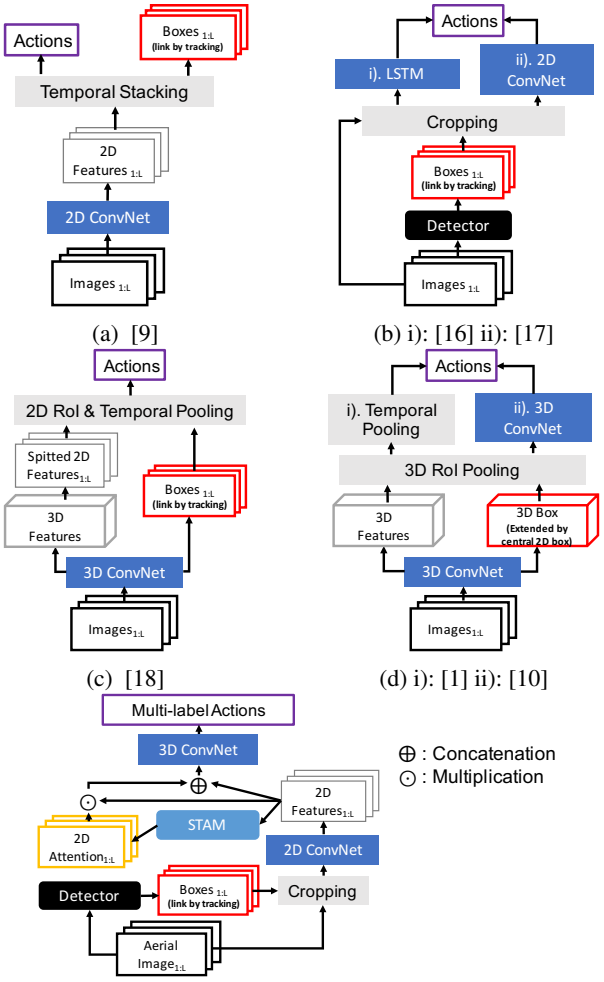


FIGURE 2: **A visualization of model structures for AVA detection when only RGB data is used.** L denotes the number of frames used in the model. i) and ii) represent different models that share the same structure at the beginning.

Several models that can be used for AVA detection are illustrated in FIGURE 2. FIGURE 2a, FIGURE 2c and FIGURE 2d learn actions and boxes by networks with end-to-end training, while FIGURE 2b and FIGURE 2e use independent detectors to generate boxes, and then connect boxes by trackers. This means generating spatio-temporal tubes and predicting actions are separate steps. Actions and boxes are jointly generated in FIGURE 2a, and boxes are linked to form tubes by an offline tracking. To better model the spatio-temporal information, FIGURE 2c and FIGURE 2d learn features by a 3D ConvNet. The main difference is that FIGURE 2c generates boxes and performs 2D Region of Interest (RoI) pooling for each frame, while FIGURE 2d extends the central frame boxes to adjacent frames. Additionally, FIGURE 2c and FIGURE 2d.i) apply temporal pooling to fuse features, while FIGURE 2d.ii) uses a 3D ConvNet to process features and obtain a better action recognition performance.

Owing to the divergence of the patch's local coordinate and the entire image's global coordinate, our inputs can only be aligned at the box level. Therefore, it is challenging to jointly detect bounding boxes and actions in our framework. Similar to FIGURE 2b, our framework (*i.e.*, FIGURE 2e) generates boxes by an independent detector and then connects boxes in the temporal domain by a multi-person tracking algorithm. Moreover, we propose a STAM to focus on the target object at each frame and use a 3D ConvNet for action recognition.

C. RELATED DATASETS

Other than AVA detection, the primary focus of aerial video study has been object detection and tracking [19]–[21]. In this paper, since we concentrate on **multi-label AVA detection in aerial videos**, we utilize Okutama-action dataset [12] for our experiments. The dataset comprises 43 minute-long drone-recorded aerial videos, with fully annotated bounding boxes in each frame and corresponding multi-label action classes. In all, there are 12 categories of human actions: Handshaking, Hugging, Reading, Drinking, Pushing/Pulling, Carrying, Calling, Running, Walking, Lying, Sitting and Standing. In the multi-label action annotation, one action class could associate with another one. For instance, "Reading" and "Sitting" could be assigned to the same person at the same time.

III. METHODOLOGY

Our proposed framework coherently generates patches, bounding boxes, spatio-temporal tubes, 2D CNN features, attention maps, and multi-label action classes (see FIGURE 3). Using a video frame of size 2160×3840, our C-RPN first generates patches of size 608×608. Based on selected patches, normal detectors (*e.g.*, YOLOv3-tiny [5]) can generate fine-grained bounding boxes for each person. After that, fine-grained bounding boxes are connected to form spatio-temporal tubes by a multi-person tracking algorithm (*e.g.*, Deep SORT [22]). Next, we sample L frames from spatio-temporal tubes and obtain their corresponding 2D CNN features. STAM then takes 2D CNN features to generate attention maps that focus on target persons. In the end, the concatenation of 2D CNN features and their multiplication with attention maps, are used to estimate multi-label action classes by a 3D ConvNet. For the overall processing, it is a special multi-label AVA detection that serves for aerial surveillance videos.

A. CLUSTERING REGION PROPOSAL NETWORK (C-RPN)

The Clustering Region Proposal Network (C-RPN) takes downsized aerial images (544×960) as its input. Since each person is relatively tiny compared with the aerial image, the coarse position of person could be modeled by a 2D Gaussian density map. The mean of 2D Gaussian is the centroid of a person and the covariance represents the uncertainty of this position, which is set to be roughly half of the bounding box size. Thus, coarse person locations can be learned by density map regression. Based on the predefined patch size, coarse

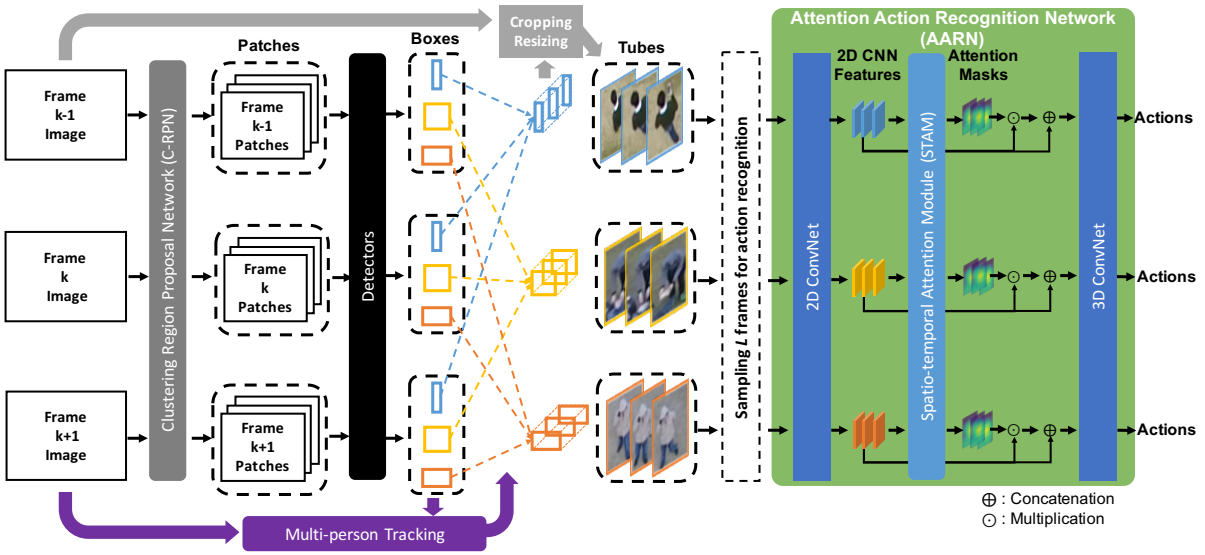


FIGURE 3: **Architecture of the proposed framework.** Given each ultra-high-resolution aerial image of size 2160×3840, C-RPN is utilized to select patches (608×608) that might contain persons. Based on selected patches, normal detectors are used to generate fine-grained bounding boxes for each person. After that, fine-grained bounding boxes are further connected to be spatio-temporal tubes by a multi-person tracking algorithm. Next, we sample L frames from spatio-temporal tubes and obtain their corresponding 2D CNN features. STAM then takes 2D CNN features to generate attention maps that focus on target persons. In the end, the concatenation of 2D CNN features and their multiplication with attention maps, are used to estimate multi-label action classes by a 3D ConvNet.

person locations can be further clustered by their relative distances and patches that may contain persons are generated (see FIGURE 4).

At frame k, let the network output of C-RPN be H_k^{pred} and its ground truth be H_k^{true}. When both H_k^{pred} and H_k^{true} have a row number of R and a column number of C, we can represent them by

$$\begin{aligned} H_k^{pred} &= \bigcup_{r=1}^{R}\bigcup_{c=1}^{C} h_{krc}^{pred}, \\ H_k^{true} &= \bigcup_{r=1}^{R}\bigcup_{c=1}^{C} h_{krc}^{true}, \end{aligned} \tag{1}$$

where r and c are the row index and column index of the heat map, respectively; h_{krc}^{pred} and h_{krc}^{true} denote the pixel at position $[r, c]$ of H_k^{pred} and H_k^{true}, respectively.

The h_{krc}^{true} is generated by

$$\begin{aligned} &h_{krc}^{true} = \\ &\sum_{i=1}^{N} exp\Big(- \frac{\big(r - p_{ki}(x) * s_1 * s_2\big)^2 + \big(c - p_{ki}(y) * s_1 * s_2\big)^2}{2\sigma_{ki}^2}\Big); \\ &h_{krc}^{true} = \begin{cases} 1, & if\ h_{krc}^{true} > 1; \\ h_{krc}^{true}, & else. \end{cases} \end{aligned} \tag{2}$$

where $[p_{ki}(x), p_{ki}(y)]$ are the center coordinates of the i^{th} ground-truth bounding box. Since the overlapping boxes may generate values larger than 1, we clip the maximum value of h_{krc}^{true} at 1. The downscale factor from original image to C-RPN input is denoted as s_1, and the down-sampling factor from C-RPN input to C-RPN output is denoted as s_2. In this work, we set $s_1 \approx 1/4$ (to be divisible by s_2) and $s_2 = 1/8$.

More specifically, σ_{ki}, $p_{ki}(x)$ and $p_{ki}(y)$ are generated by

$$\begin{aligned} \sigma_{ki} &= \frac{s_1 * s_2}{4}\Big((x_{ki}^{max} - x_{ki}^{min}) + (y_{ki}^{max} - y_{ki}^{min})\Big); \\ p_{ki}(x) &= \frac{s_1 * s_2}{2}(x_{ki}^{max} + x_{ki}^{min}); \\ p_{ki}(y) &= \frac{s_1 * s_2}{2}(y_{ki}^{max} + y_{ki}^{min}); \end{aligned} \tag{3}$$

where $[x_{ki}^{min}, y_{ki}^{min}, x_{ki}^{max}, y_{ki}^{max}]$ are corner positions of the i^{th} ground-truth bounding box at frame k. Here, σ_{ki} is roughly half size of the bounding box i at frame k.

We modify a penalty-reduced pixel-wise logistic regression with focal loss [23] and let it be our loss function $\mathcal{L}_{raw\_pos,k}$ as follows:

$$\begin{aligned} &\mathcal{L}_{raw\_pos,k} \\ &= -\sum_{r=1}^{R}\sum_{c=1}^{C} \begin{cases} \Big(1 - h_{krc}^{pred}\Big)^{\alpha} \log\Big(h_{krc}^{pred}\Big), \\ \qquad \text{if } h_{krc}^{true} = 1; \\ \big(1 - h_{krc}^{true}\big)^{\beta} \Big(h_{krc}^{pred}\Big)^{\alpha} \log\Big(1 - h_{krc}^{pred}\Big), \\ \qquad \text{otherwise}; \end{cases} \end{aligned} \tag{4}$$

where α and β are hyper parameters for focal loss and we follow work [23] to set α and β to be 2 and 4, respectively.

Ideally, each object center is a peak point on this density map, thus, we can apply peak point Non-Max Suppression (NMS) to obtain corresponding peak points. Nonetheless, there is no magic in the network of C-RPN, and it is still suffering the dilemma of detectors in setting a confidence threshold: better precision, or better recall. In C-RPN, although false-positive (FP) peak points may generate redun-

dant patches, such a redundancy has little effect on the final fine-grained object detection. Therefore, we set a low confidence threshold for peak point NMS to obtain peak points, regardless of it may end up with low precision and high recall.

Because peak points could be sparsely distributed, grouping neighboring peak points to guide patch generalization can reduce the number of patches. As we have discussed in the section of related works, grid-based clustering may not fit our requirements as it may incompletely crop person appearance in all patches. To make a trade-off between reducing the number of patches and preserving the objects appearance, we choose hierarchical clustering. In hierarchical clustering, by adjusting the threshold distance to generate suitable overlapping regions dynamically, we could make person appearance complete in at least one patch.

We do not need to specify how many persons are included in each patch, because another object detector (*e.g.*, YOLOv3) will take patches as inputs to generate bounding box for each person. Since overlapping patches could be generated, we not only have duplicated boxes in the same patch, but also have duplicated boxes on the overlapping regions between patches. In our approach, therefore, we only perform bounding box NMS once after transferring bounding boxes from the patch coordinate to the original aerial image coordinate.

B. ATTENTION ACTION RECOGNITION NETWORK (AARN)

In our approach, Deep SORT [22], a multiple-object tracking method, is employed to link bounding boxes into spatio-temporal tubes. Deep SORT takes an IoU (Intersection over Union) descriptor, an appearance descriptor, and a Kalman filter to perform bipartite bounding box assignments across frames. The appearance descriptor, which is used to overcome occlusion and long-time tracking issues, is a CNN network trained on a person re-identification dataset [24] by a Cosine Softmax Classifier [25].

After obtaining the spatio-temporal tube for each person, we obtain their actions at each frame by a novel Attention Action Recognition Network (AARN). Since AVA detection focuses on instantaneous actions other than long-term actions, we only take a short-term temporal context and sample L frames from each spatio-temporal tube for action recognition. Frames within 2 seconds (*i.e.*, 60 frames in 30 FPS videos) ahead of the target frame are excluded. For a person whose track ID is n, we denote the earliest and latest frames in the corresponding spatio-temporal tube as k_{min} and k_{max}, respectively. Setting k_{max} as the target frame, then L frames are sampled to form a set $\{x_0^n,\ x_1^n,\ ...,\ x_L^n\} \in X_{k_{max}}^n$ for action recognition. The details of our online sampling strategy are described in Algorithm 1.

Instead of directly processing RGB data $X_{k_{max}}^n$ by 3D ConvNet, we extract their corresponding 2D CNN features $\{f_1^n, f_2^n, ..., f_L^n\} \in F_{k_{max}}^n$ at the first step. Then, we proposed a Spatio-temporal Attention Module (STAM), which

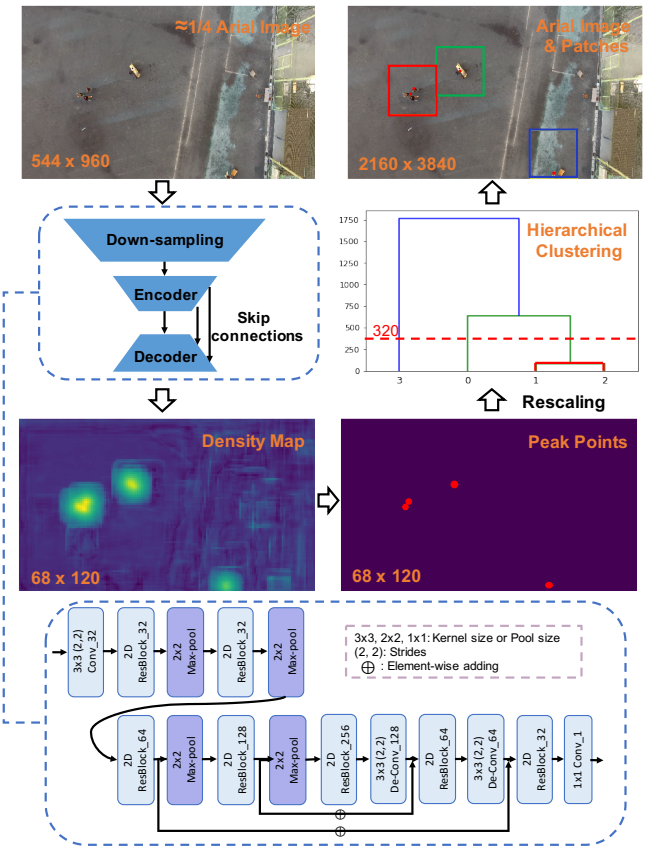


FIGURE 4: **The demonstration of generating patches by C-RPN.** The downscale factor $s_1 \approx 1/4$, and the down-sampling factor $s_2 = 1/8$.

Algorithm 1: On-line sampling from a spatio-temporal tube

Input : Spatio-temporal tube $T_{[k_{min}:k_{max}]}^n$

if $len(T_{[k_{min}:k_{max}]}^n) < L$ **then**

  $X_{k_{max}}^n \leftarrow \{T_{[k_{min}:k_{max}]}^n$ + Repeat Padding with $T_{k_{max}}^n\}$;

else

  $\delta = len(T_{[max(k_{min},k_{max}-60):k_{max}]}^n)//L$;

  $X_{k_{max}}^n \leftarrow \{$Randomly choose L frames from $T_{[max(k_{min},k_{max}-60):k_{max}]}^n$ with the interval $\delta\}$.

Output: $X_{k_{max}}^n$

is a 3D encoder-decoder with skip connections, to generate attentions maps $\{a_1^n, a_2^n, ..., a_L^n\} \in A_{k_{max}}^n$ by encoding and decoding the global spatio-temporal representation of $F_{k_{max}}^n$. After that, we perform element-wise multiplication between $F_{k_{max}}^n$ and $A_{k_{max}}^n$, and concatenate with $F_{k_{max}}^n$ to obtain a representation that can selectively focus on the target person across all frames. Finally, aforementioned 2D CNN features are stacked to be 3D CNN features, which are then fed to a 3D ConvNet to estimate multi-label action classes (see FIGURE 6).

Although it is common to utilize optical flow for action recognition, we do not use it in our framework. In drone-

FIGURE 5: Visualizations of optical flow maps generated by PWC-Net [26], using Okutama-action Dataset. Due to the tiny size of person and the drone camera movement, it is challenging to obtain person motion information from the optical flow.

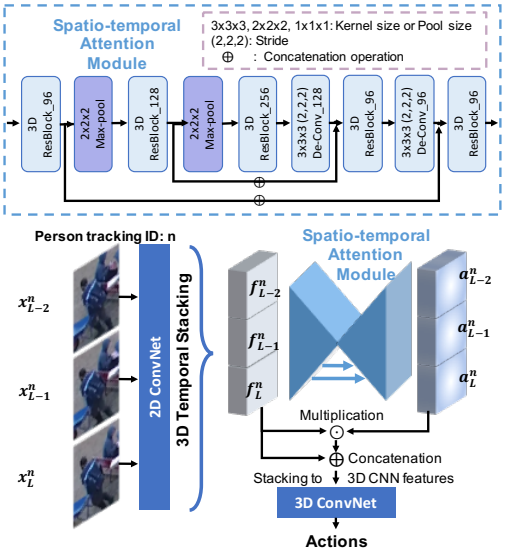


FIGURE 6: **An illustration of proposed Attention Action Recognition Network (AARN), with its Spatio-temporal Attention Module (STAM)**. Three frames are used in this illustration, where $\{x^n_{L-2}, x^n_{L-1}, x^n_L\}$ are RGB features sampled by Algorithm 1, and they are fed to 2D ConvNet to generate 2D CNN features $\{f^n_{L-2}, f^n_{L-1}, f^n_L\}$. STAM takes stacking 2D CNN features to obtain corresponding attention maps $\{a^n_{L-2}, a^n_{L-1}, a^n_L\}$. The multiplication results of 2D CNN features and attention maps are concatenated with 2D CNN features again, and then be used to estimate multi-label actions by a 3D ConvNet.

recorded aerial videos, even if the absolute location of an instance is static, its relative location may have a huge change across nearby frames, which is caused by the drone camera movement and tiny object size. In Okutama-action data, we use a state-of-the-art optical flow generator [26] to produce optical flows between nearby frames, and show them in FIGURE 5. We can see, it is hard to identify the movement of each person in the optical flow map.

IV. EVALUATION METRICS FOR MULTI-LABEL AVA DETECTION

The evaluation metrics for object detection and multi-label classification have been well studied separately [27], [28], but the problem remains on how to associate them together for multi-label AVA detection evaluation.

A simple approach could be evaluating the "person" object detection performance for all detected samples and then evaluating the multi-label action recognition performance for positively detected samples. For instance, assuming that a predicted sample is positive when IoU ≥ 0.5 for the predicted and ground-truth bounding boxes, we can apply $h.l.@0.5$, which corresponds to Hamming Loss associated with IoU ≥ 0.5, to measure its multi-label classification performance. Below, we show how the $h.l.@0.5$ is extended from the original Hamming Loss.

$$h.l.@0.5 = \frac{1}{N_{persons@0.5}} \frac{1}{N_{labels}} \sum_{i=1}^{N_{person@0.5}} \sum_{l=1}^{N_{labels}} Y^{i,l}_{true} \; XOR \; Y^{i,l}_{pred}, \quad (5)$$

where XOR is an exclusive-or operation and N_{labels} stands for the number of action categories. Y_{true} and Y_{pred} are boolean arrays that denote the ground truth and predicted labels, respectively. The number of positively detected samples are represented by $N_{persons@0.5}$. To help understand the above metrics, we illustrate how $h.l.@0.5$ is calculated by a toy example in FIGURE 7.

Due to the complexity of multi-label classification evaluation, usually more than one criterion is included to inspect the performance from different perspectives [28]. By applying the same modification as $h.l.@0.5$, we propose another three criteria as follows.

- $co.@0.5$: this is extended from Coverage, and it evaluates how far on average it is necessary to go through the ranked scores to cover all true labels for positive samples (IoU≥ 0.5).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924188, IEEE Access

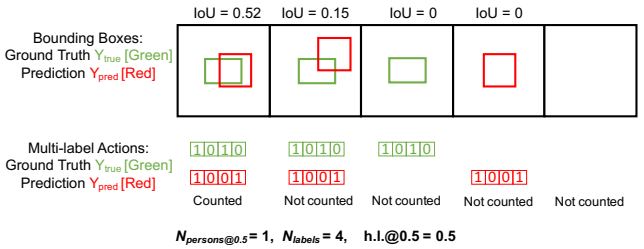


FIGURE 7: **An example of calculating h.l.@0.5.** Only the first case with IoU=0.52 is considered as a positively detected sample, and therefore the overall h.l.@0.5=0.5.

- $r.l.@0.5$: this is extended from Ranking Loss, and it evaluates the average proportion of label pairs that are incorrectly ordered for a positive sample (IoU≥ 0.5).
- $o.e.@0.5$: this is extended from One Error, and it evaluates the average number of top-ranked predicted labels that are not true labels for the positive sample (IoU≥ 0.5).

V. EXPERIMENTS

A. TRAINING AND TESTING SETUP

By following the previous work [12], Okutama-action dataset is split into a training set with 33 aerial videos and a testing set with 10 aerial videos.

For C-RPN, the Adam [29] optimizer with a learning rate 0.001 is applied for the first 50 epochs and then the learning rate is changed to be 0.0001 for another 150 epochs. The batch size is set up to be 8. Images and their corresponding density maps are jointly augmented by Albumentations [30].

We perform a peak point detection on a validation set (*i.e.*, 20% of the training set) and find that a density map can reach the confidence of $0.5 \sim 1.0$ and $0.0 \sim 0.1$ at the target and the non-target positions, respectively. To reach a high recall on the testing set, we set the peak point NMS confidence threshold as 0.3. We search the maximum person bounding boxes size in Okutama-action dataset to decide the distance threshold in peak point NMS. More specifically, the maximum person bounding box size is about 200 on the original size image. Considering the total downscale from the original size image (2160×3840) to the output density map (68×120) is about 32, the maximum person size on output density map is about 6. Since distance threshold should be an odd number, we take value 5 here. Using Python code, peak point NMS can easily be implemented by

```
from scipy.ndimage import maximum_filter
Peaks_map = (H_pred>0.3)*
          (H_pred==maximum_filter(H_pred,
           footprint = np.ones((5,5))))
```

Listing 1: Peak piont NMS

For other detectors used for comparison, as R-FCN-ResNet50 [3], Retinanet-ResNet50 [4], SSD-ResNet50 [2] and YOLOv3-tiny [5], we take their pre-trained weights on COCO dataset [31] and fine-tune them on our experimental datasets by their default training strategy.

To train AARN, we equally sample 64 ground-truth spatio-temporal tubes from each action class, and then sample $X^n_{k_{max}}$ from each spatio-temporal tube (see Algorithm 1). As only part of the training samples are included in one epoch training, it takes more iterations to get converged. We also apply the Adam optimizer for it, with learning rate 0.001, 0.0001, and 0.00001 for each 500 epochs. The batch size is set up to be 16. We perform the same data augmentation, *i.e.*, flipping, rotation, resizing, and cropping to all samples in $X^n_{k_{max}}$. During the inference process, Algorithm 1 is applied again to obtain inputs for the inference process.

Even though we are working on AVA detection with large-size aerial videos, our framework decomposes the whole problem into multiple simple tasks. Thus, all our experiments can be implemented on a single NVIDIA TITAN X GPU.

B. PERFORMANCE EVALUATION

Our proposed metrics evaluate the multi-label AVA detection performance by two steps. Firstly, we evaluate person detection performance, by using the mAP@0.5 metrics [31]. Secondly, we evaluate multi-label action recognition performance for positively detected samples (*i.e.*, a sample with mAP $\geqslant 0.5$). We jointly inspect the performance of two steps to obtain the overall multi-label AVA detection performance.

1) Detection Evaluation

For the person detection evaluation, our main purpose is to verify three assumptions: (1) compared with detectors that work on the downsized aerial image (608×608 with padding), although using our proposed C-RPN may take more running time, it should improve the person detection performance; (2) compared with partitioning the entire aerial image (2160×3840) into patches with a sliding window [8], our C-RPN should be faster when persons are sparsely located; (3) in contrast to grid-based clustering [15], using hierarchical clustering with a proper distance threshold can keep the complete appearance in at least one patch so that our method can achieve better person detection performance.

For detectors that take the entire aerial image as input, we standardize their input size to be 608×608 by padding, since it is difficult to train and test a detector with larger input size. When the sliding window passes the aerial image margin, we pad zeros to the inputs. To reach a fast speed, we choose YOLOv3-tiny [5] as the base detector in our framework. Although our default setting is hierarchical-clustering C-RPN, for a fair comparison with previous work [15], we form a grid-clustering C-RPN by solely replacing the clustering method.

The qualitative results of our patch generation and bounding box estimation are shown in FIGURE 8 and FIGURE 9, respectively. The quantitative results of the Okutama-action testing set are shown in TABLE 1. Taking the original-size aerial images (2160×3840), our C-RPN + YOLOv3-tiny achieves 85.2 mAP@0.5 in terms of "person" object detection, which remarkably outperforms detectors that utilize downsized aerial images. Besides, by using C-RPN,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924188, IEEE Access

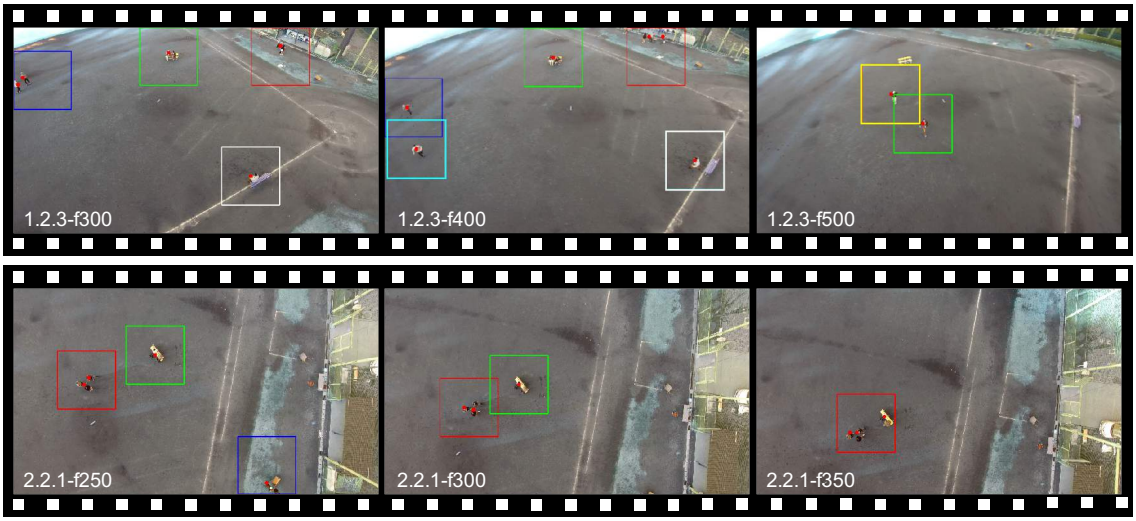


FIGURE 8: **Patch proposals in Okutama-action testing sets, which are generated by C-RPN.** Generated peak points are marked by red, and patches are enclosed by colorful rectangles. The first row shows three sequential frames (*i.e.*, 300, 400 and 500) in video 1.2.3. The second row shows three sequential frames (*i.e.*, 250, 300 and 350) in video 2.2.1. To efficiently cover target persons, clusters automatically merge and split, based on the relative distance within peak points.

Method	mAP@0.5↑	Speed for entire image (FPS)↑	Average patches↓
Using entire downscale image of size 608×608			
R-FCN-ResNet50 [3]	53.5	6	-
Retinanet-ResNet50 [4]	56.3	10	-
SSD-ResNet50 [2]	52.3	18	-
YOLOv3-tiny [5]	52.4	**120**	-
Using Patches of size 608×608 (without downsizing)			
Sliding (stride=388,404) [8]+YOLOv3-tiny	82.0	3	45.0
Sliding (stride=580,580) [8]+YOLOv3-tiny	79.4	5	28.0
C-RPN (Grid:$grid_{size}$=216×384) [15]+YOLOv3-tiny	77.5	25	3.9
C-RPN (Grid:$grid_{size}$=270×480) [15]+YOLOv3-tiny	78.3	28	3.7
C-RPN (Hierarchical:$d_{threshold} = 128$)+YOLOv3-tiny	85.0	26	3.8
C-RPN (Hierarchical:$d_{threshold} = 320$)+YOLOv3-tiny	**85.2**	30	3.1
C-RPN (Hierarchical:$d_{threshold} = 512$)+YOLOv3-tiny	82.9	38	**2.2**

TABLE 1: **"Person" object detection performance on Okutama-action dataset.** The symbol ↑(↓) indicates that the larger(smaller) the value, the better the performance.

the final object detection performance is even better than using a sliding window, since some ambiguous background might be excluded by C-RPN in advance. Last but not least, because we try to make the person appearance complete in at least one patch, the performance of hierarchical-clustering C-RPN outperforms grid-clustering C-RPN [15]. Moreover, we quantitatively calculate the average number of patches generated by each method in Okutama-action testing set. When hierarchical-clustering C-RPN reach the best detection performance, it only generates 3.1 patches averagely on Okutama-action testing set, which is more efficient than sliding window approach and similar to the grid-clustering C-RPN. Therefore, our approach can achieve a comparable speed of 30 FPS on the full resolution data.

2) Multi-label AVA Detection Evaluation

Better-AVA model [10] is one of the state-of-the-art models for AVA detection on movie data. It performs AVA detection for the central frame and need an odd number of frames as its inputs. We modify it to jointly estimate multi-label actions and bounding boxes. Its inputs are L frames of downscale aerial images (608×608 with padding), which are sampled near the target frame. Due to the limitation of our computational resource, we choose $L = 5$ for it.

To inspect whether our AARN can improve the action recognition performance by introducing spatio-temporal attention, we construct an ablation study by replacing AARN with I3D [32] and Lite ECO [33] in our framework. The results of applying our proposed metrics are shown in TABLE 2.

Compared with Better-AVA, our framework achieves better performance in both person detection and multi-label action recognition. Besides, our framework is faster than Better-AVA on our target task. Considering our framework decomposes the whole pipeline into several independent steps, less memory cost is needed in our framework.

Through introducing spatio-temporal attentions, our AARN performs better than I3D and Lite ECO, in terms of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924188, IEEE Access

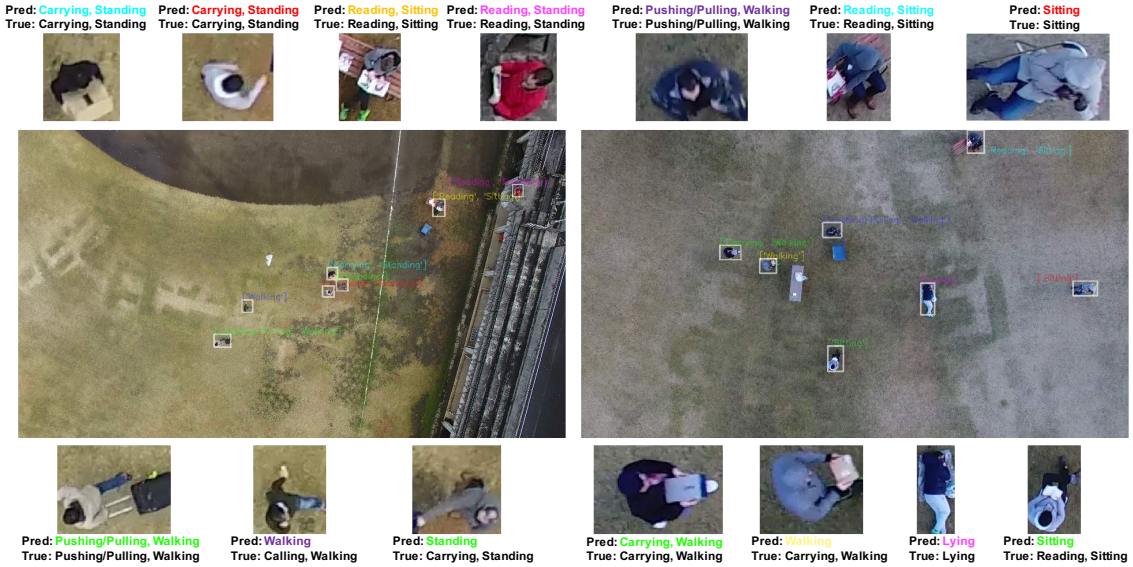


FIGURE 9: **Examples of multi-label AVA detection results in our framework.**

Method	h.l.@0.5↓	co.@0.5↓	r.l.@0.5↓	o.e.@0.5↓	mAP@0.5↑	Speed (FPS)↑
Off-line multi-label AVA detection						
Input: $L \times 608 \times 608$ (L frames of downsized aerial images with padding)						
Better-AVA [10] (L=5)	0.20	3.73	0.19	0.35	54.1	8
On-line multi-label AVA detection						
Input: $L \times 96 \times 96$ (L frames of cropped images from spatio-temporal tube)						
Replacing AARN by I3D [32] in our framework (L=8)	0.14	3.45	0.14	0.28	**85.2**	14
Replacing AARN by Lite ECO [33] in our framework (L=8)	0.15	3.46	0.13	0.27	**85.2**	**15**
Our framework (L=8)	**0.13**	**3.38**	**0.11**	**0.25**	**85.2**	14

TABLE 2: **Multi-label AVA detection results.** The symbol ↑(↓) indicates that the larger(smaller) the value, the better the performance. Only RGB data is used in this test. Note, we choose $L = 5$ for Better-AVA due to computation memory limitation and it has to be an odd number. While other models utilize $L = 8$ since instantaneous actions are defined in AVA detection. Except for Better-AVA, other action detection models use bounding boxes that are generated by C-RPN + YOLOv3-tiny, which achieves mAP@0.5=85.2.

action recognition in our target task. Examples of attention maps generated by STAM can be visualized in FIGURE10, which shows that STAM can learn to focus on the target person in an unsupervised manner.

FIGURE 10: **Visualization of attentions for the target person.** We assume that the target person consistently appears in his/her spatio-temporal tube while others may not. The attention mask is learned in an unsupervised manner.

VI. CONCLUSION

The aerial surveillance videos make it possible to increase public safety in sparsely populated regions. To automatically analyze "who is doing what?" in such videos, we specifically propose a novel multi-label AVA detection framework and corresponding evaluation metrics. Our framework gives the flexibility to replace its detector and tracker based on the need, which makes it possible to train and infer all modules on a single GPU. Thus, our framework can be more suitable than existing solutions for multi-label AVA detection in aerial videos. On a final note, our proposed evaluation metrics are not limited to aerial videos, and other AVA detection tasks can also leverage such metrics to perform a reasonable evaluation.

REFERENCES

[1] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar et al., "Ava: A video dataset of spatio-temporally localized atomic visual actions," CoRR, abs/1705.08421, vol. 4, 2017.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.

[3] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in Advances in neural information processing systems, 2016, pp. 379–387.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for

dense object detection," IEEE transactions on pattern analysis and machine intelligence, 2018.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[6] R. Porter, A. M. Fraser, and D. Hush, "Wide-area motion imagery," IEEE Signal Processing Magazine, vol. 27, no. 5, pp. 56–65, 2010.

[7] P. C. Hytla, K. S. Jackovitz, E. J. Balster, J. R. Vasquez, and M. L. Talbert, "Detection and tracking performance with compressed wide area motion imagery," in Aerospace and Electronics Conference (NAECON), 2012 IEEE National. IEEE, 2012, pp. 163–170.

[8] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," arXiv preprint arXiv:1805.09512, 2018.

[9] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," ICCV, Oct, vol. 2, 2017.

[10] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "A better baseline for ava," arXiv preprint arXiv:1807.10066, 2018.

[11] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," arXiv preprint arXiv:1603.00831, 2016.

[12] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in 1st Joint BMTT-PETS Workshop on Tracking and Surveillance, CVPR, 2017, pp. 1–8.

[13] P. Hu and D. Ramanan, "Finding tiny faces," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 951–959.

[14] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," 2018.

[15] R. LaLonde, D. Zhang, and M. Shah, "Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information," in Computer Vision and Pattern Recognition, 2018.

[16] J. He, Z. Deng, M. S. Ibrahim, and G. Mori, "Generic tubelet proposals for action localization," in IEEE Winter Conference on Applications of Computer Vision. IEEE, 2018, pp. 343–351.

[17] Z. Li, W. Wang, N. Li, and J. Wang, "Tube convnets: Better exploiting motion for action recognition," in IEEE International Conference on Image Processing. IEEE, 2016, pp. 3056–3060.

[18] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (t-cnn) for action detection in videos," in IEEE international conference on computer vision, 2017.

[19] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in European conference on computer vision. Springer, 2016, pp. 549–565.

[20] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in European conference on computer vision. Springer, 2016, pp. 445–461.

[21] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," arXiv preprint arXiv:1804.07437, 2018.

[22] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in IEEE International Conference on Image Processing. IEEE, 2017, pp. 3645–3649.

[23] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proceedings of the European Conference on Computer Vision, 2018, pp. 734–750.

[24] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in European Conference on Computer Vision. Springer, 2016, pp. 868–884.

[25] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in 2018 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2018, pp. 748–756.

[26] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934–8943.

[27] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, no. 2, pp. 303–338, 2010.

[28] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," ACM Computing Surveys, vol. 47, no. 3, p. 52, 2015.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[30] E. K. V. I. I. A. Buslaev, A. Parinov and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," ArXiv e-prints, 2018.

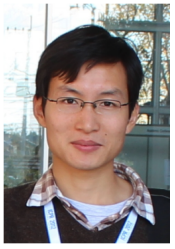
[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017, pp. 4724–4733.

[33] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in Proceedings of the European Conference on Computer Vision, 2018, pp. 695–712.

FAN YANG received a B.S. degree and a M.S degree from Nanjing University and Nara Institute of Science and Technology in 2012 and 2018, respectively. He is currently a Ph.D. candidate at Nara Institute of Science and Technology. His research focus on video processing.

SAKRIANI SAKTI received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATRNICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. In 2011-2017, she was an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. Currently, she is a research associate professor at NAIST, as well as a research scientist at RIKEN, the Center of for Advanced Intelligent Project AIP, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. She is also the officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition synthesis, spoken language translation, affective dialog system, and cognitive communication.



YANG WU received a BS degree and a Ph.D degree from Xi'an Jiaotong University in 2004 and 2010, respectively. From Sep. 2007 to Dec. 2008, he was a visiting student in the GRASP lab at University of Pennsylvania. From 2011 to 2014, he was a program specific researcher at the Academic Center for Computing and Media Studies, Kyoto University. Within this period, he was an invited academic visitor at the Big Data Institute of University College London from Jul. 2014 to Aug. 2014. He is currently an assistant professor of the NAIST International Collaborative Laboratory for Robotics Vision, Institute for Research Initiatives, Nara Institute of Science and Technology. His research is in the fields of computer vision, pattern recognition, and image/video search and retrieval, with particular interests in human-centric computer vison problems (detection, tracking, pose estimation, recognition, re-identification, action, activities, etc.). He is also interested in pursuing general data analysis models (with various kinds of supervision) for real applications.

SATOSHI NAKAMURA is a Professor of Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Project Leader of Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007- 2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampolli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.