Interactivity Proposals for Surveillance Videos

Shuo Chen, Pascal Mettes, Tao Hu and Cees G. M. Snoek
University of Amsterdam

ABSTRACT

This paper introduces spatio-temporal interactivity proposals for video surveillance. Rather than focusing solely on actions performed by subjects, we explicitly include the objects that the subjects interact with. To enable interactivity proposals, we introduce the notion of *interactivityness*, a score that reflects the likelihood that a subject and object have an interplay. For its estimation, we propose a network containing an interactivity block and geometric encoding between subjects and objects. The network computes local interactivity likelihoods from subject and object trajectories, which we use to link intervals of high scores into spatio-temporal proposals. Experiments on an interactivity dataset with new evaluation metrics show the general benefit of interactivity proposals as well as its favorable performance compared to traditional temporal and spatio-temporal action proposals.

KEYWORDS

interactivity detection; action detection; proposal generation

ACM Reference Format:
Shuo Chen, Pascal Mettes, Tao Hu and Cees G. M. Snoek. 2020. Interactivity Proposals for Surveillance Videos. In *2020 International Conference on Multimedia Retrieval (ICMR'20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3372278.3390680

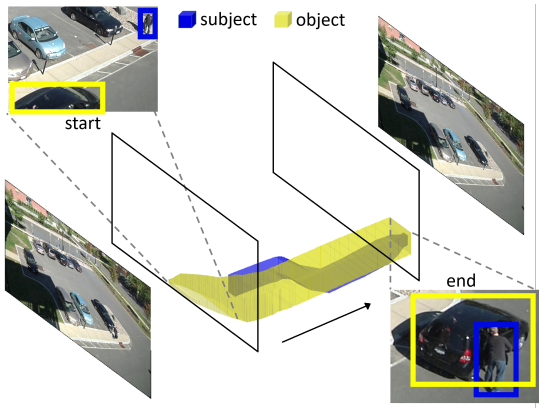


Figure 1: Interactivity proposals encapsulate a subject and object trajectory with the same start and end time. In this paper we define, generate and evaluate this new type of proposals for video surveillance.

1 INTRODUCTION

The goal of this paper is to generate spatio-temporal proposals that capture the interaction between subjects and objects in surveillance videos. Spatio-temporal proposals in videos are generally focused on actions [15, 17, 21, 41, 49], *i.e.*, centered around subjects only. The objects with which actions might interact are generally ignored or only used implicitly. In surveillance settings, interactions between subjects and objects are key, because they denote important events to analyze. Think about a person entering a car or loading gear into a trunk. Since surveillance videos may contain several events that happen simultaneously, localizing the temporal extent of an interactivity is insufficient; spatial localization is mandatory. We aim to explicitly capture subjects performing actions, and the objects with which they interact, in space and time. We focus on the proposal generation step, where a video is split into spatio-temporal segments, upon which detection algorithms can be applied.

To arrive at spatio-temporal interactivity proposals, we take inspiration from *objectness* [1] and *actionness* [8, 44]. These approaches estimate the likelihood of object presence in a spatial region or action presence in a spatio-temporal region. Based on the likelihood, object or action proposals can be generated. Subsequently, such proposals are scored by classifiers to obtain object or action detections. Here, we take this line of work further and introduce *interactivityness*. Rather than estimating the individual likelihoods of objects or subjects performing an action, we estimate when and where subjects and objects are jointly occurring and are also in interaction. Akin to objectness and actionness, we use interactivityness to obtain interactivity proposals, which we define as pairs of subject and object trajectories with the same start and end time, see Figure 1.

We make three contributions in this paper. First, we introduce the new task of spatio-temporal interactivity proposal generation in surveillance videos. Second, we introduce an interactivity network. This network estimates the interactivityness between a subject and object using an interactivity module that models the context around subjects and objects, as well as a geometric encoding that models the spatial relations of the pair. Third, we set up an interactivity proposal evaluation, including a dataset distilled from the ActEV surveillance benchmark [2] and interactivity evaluation metrics. Experiments on this evaluation show the effectiveness of our approach, outperforming existing approaches from the temporal and spatio-temporal action proposal literature. We will make the dataset, evaluation protocols, and code publicly available.

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390680

2 RELATED WORK

2.1 Action proposals

Temporal action proposals. Proposal methods for temporal action localization form an active research topic [4, 10, 12, 13, 25, 27, 32, 51, 51]. Escorcia *et al.* [10] utilize LSTMs on extracted CNN features to capture temporal information. Buch *et al.* [4] adopt the C3D network architecture as a feature extractor with a gated recurrent unit to capture long-term temporal information. Gao *et al.* [13] collect proposal candidates through a sliding window, which utilizes unit-level information for training. For each proposal, the average unit representation is adopted as proposal representation. Afterwards, temporal regression is performed on the unit-level to refine the start and end times of the proposals. Zhao *et al.* [51] generate actionness for each frame and group continuous frames with high actionness as proposals. All temporal action proposal methods use whole frames as input. In outdoor surveillance settings, many action and interactions can occur at the same time, hence using whole frames as input is not precise enough. Therefore, we target interactivity proposals in both space and time.

Spatio-temporal action proposals. Spatio-temporal action proposals target the spatio-temporal locations of subjects in videos [15, 17, 21, 30, 41, 49]. One common manner to obtain spatio-temporal action proposals is by clustering local voxels or dense trajectories in a hierarchical manner [21, 30, 41]. Yu *et al.* [49] generate generic action proposals in unconstrained videos by linking subject detections over time. He *et al.* [17] propose a tupelet proposal network for action detection, which adopts Faster RCNN [33] to collect boxes with high action scorse. They link the highest scoring boxes to obtain tubelet proposals. Gleason *et al.* [15] generate spatio-temporal cuboid proposals by clustering detected boxes in spatio-temporal regions, followed by jittering to collect more proposals for better recall. Where current spatio-temporal proposal methods focus on actions only, we target spatio-temporal proposals of both subjects and objects. More concretely, where a spatio-temporal action proposal is described by a single tube, a spatio-temporal interactivity proposal is described by two tubes with the same start and end time. The tubes represent a subject and an object that should be in interaction.

2.2 Visual human-object interaction

A wide range of works have investigated the relationship between humans (subjects) and objects [5, 11, 14, 48, 50] in images. Gkioxari *et al.* [14] learn to predict an action-specific density over object locations using detected subjects. Chao *et al.* [5] capture interaction information in images by measuring relative location information between boxes. Xu *et al.* [48] utilize semantic regularities for human-object interaction detection in images with knowledge graphs. Gao *et al.* [11] propose an instance-centric attention module that learns to dynamically highlight regions in an image conditioned on the appearance of each instance. Prest *et al.* [31] previously studied human-object interaction in actor-centric videos, such as *Drinking* and *Smoking*. In this setting, the person boxes generally cover the object boxes. In the surveillance domain, we aim for proposals of interactivities with unique boxes for persons and objects by focusing on the surveillance domain. Wang *et al.* [43] also investigate interactions in videos, but do so for agent-object animations, while we focus on interactivity detections by proposals.

2.3 Video Surveillance

Recognition in video surveillance is a long-standing challenge [6, 23, 24, 28, 40, 42, 46, 52]. Surveillance settings are often indoor with an explicit focus on subjects, as exemplified by the recent benchmark of Zhao *et al.* [52]. The works of Maguell *et al.* [36, 37] relates to our work as they focus on tracking loitering activities across multiple surveillance cameras. Our work focuses on capturing interactivity on single surveillance camera, without considering the explicit interactivity class. The works of Walker *et al.* [42] and Misra *et al.* [28] also relate to our work in that both tackle object localization in space and time. In this work, we focus on outdoor surveillance videos with the ActEv benchmark [2] and we focus on jointly capturing the spatio-temporal localization of subjects and objects in interaction. For spatio-temporal action detection, several datasets have been introduced, such as AvA [16], UCF-Sports [34], and J-HMDB51 [22]. Current datasets are commonly focused on human-centric actions in non-surveillance domains. Only the annotations of subjects is provided, while the spatio-temporal annotations of objects are absent. Hence, we will not consider these datasets for our experiments. Instead, we will set up an interactivity proposal evaluation, including a dataset distilled from the ActEV surveillance video benchmark [2] and interactivity evaluation metrics.

3 METHOD

In order to obtain interactivity proposals from an input video, our approach consists of three components: 1). obtaining interactivity candidates, 2). computing interactivityness, and 3). generating interactivity proposals. The overview of our method is sketched in Figure 2. We will describe each component in detail next.

3.1 Obtaining interactivity candidates

We first generate an over-complete set of interactivity candidates, where each candidate denotes a pair of subject and object trajectories that potentially interact. Due to the possibly overwhelming number of subjects and objects in a surveillance video, evaluating all possible subject and object pairs is infeasible. Physically, a subject can only interact with an object when they are close enough at some point in time. Hence, in most cases, the interactivity only happens when the subject and the object are in close contact with each other.

Suppose we have obtained N subject trajectories and M object trajectories in a video. Each trajectory has consecutive bounding boxes, *e.g.* , the subject trajectory $t_s = \{b_s^1, b_s^2, ..., b_s^n\}$ has n boxes and the object trajectory $t_o = \{b_o^1, b_o^2, ..., b_o^m\}$ has m boxes. A box $b \in \mathbb{R}^4$ is denoted by the leftmost, topmost, rightmost, and bottommost coordinates. For each frame f in the video, we calculate the Intersection over Union (IoU) between subject box $b_s^f \in t_s$ and object box $b_o^f \in t_o$. If they overlap with each other, *i.e.*, , their IoU score is larger than zero at any point in time, we deem the pair as a potential interactivity. In addition, we compute a union box that

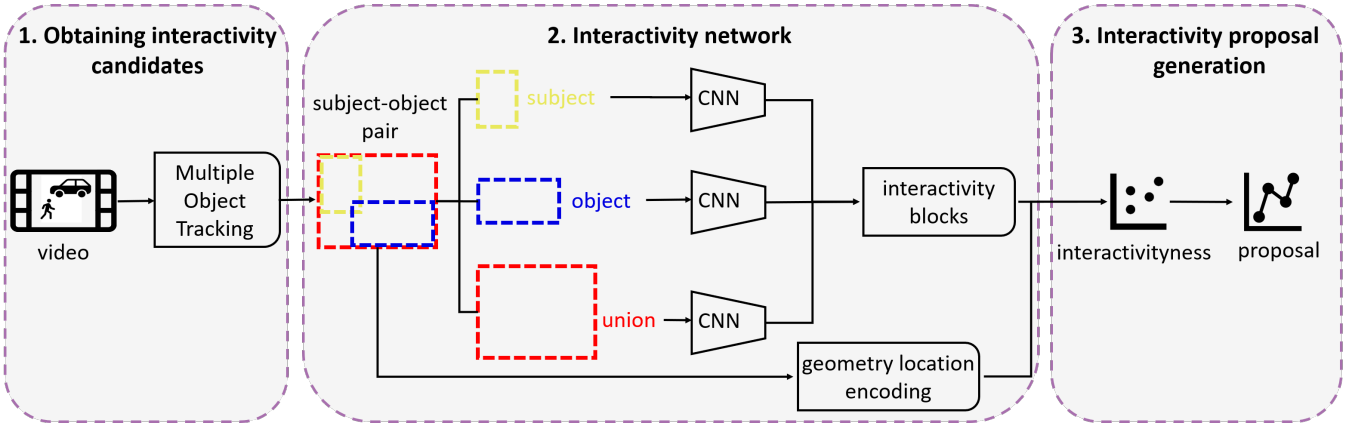


Figure 2: Method overview. During testing, we first obtain interactivity candidates by detecting and tracking subjects and objects in a surveillance video. For each frame of each subject-object pair, we input a subject-, object- and union-box to our interactivity network and obtain their interactivityness. Finally, we group continuous regions with high interactivityness to generate spatio-temporal interactivity proposals.

tightly unifies the subject and object boxes as follows:

$$b_u^f = \big(\min(b_s^f[0], b_o^f[0]),\ \min(b_s^f[1], b_o^f[1]),\ \max(b_s^f[2], b_o^f[2]),\ \max(b_s^f[3], b_o^f[3])\big). \quad (1)$$

We add the union boxes to the subject-object pairs and obtain k interactivity candidates, each consisting of a triplet of spatio-temporal trajectories, *e.g.* for temporal length k candidate c is denoted as $c = \{(b_u^1, b_s^1, b_o^1), (b_u^2, b_s^2, b_o^2), ..., (b_u^k, b_s^k, b_o^k)\}$.

This procedure is performed for test videos to obtain an initial pool of candidates. During training, we use ground truth trajectories of subjects and objects that are known to interact. The interactivity label itself is ignored, only the trajectories are used.

3.2 Interactivity network

Given a subject-object pair from our candidate pool, we need to detect whether this pair has any interactivity. If so, we also want to know when it starts and ends. Here we train a binary classifier to estimate the interactivity likelihoods, called interactivityness, for each triplet of boxes in each frame of the pair. The frame-level interactivityness scores will be used to generate our final spatio-temporal interactivity proposals. The main idea of our method is to capture interaction information to aid recognition. We achieve the goal in two ways: (1) We propose the interactivity block, an attention mechanism to compute interactions between the subject, object and union box features. The union box provides spatial contextual information, which is beneficial to recognize interactivity. (2) We encode the geometric relation between the subject and object. The relative positions of subjects and objects change over time and therefore provide useful information.

Interactivity block. In surveillance videos, the subjects and objects are usually small due to the high camera position. So the context information around subject and object is important to capture. At the same time, the network should focus on the subject and object during feature extraction. Therefore, the interactivity block should use union features to support subject and object features. Inspired by the non-local operation in action recognition [45], we design an interactivity block to capture small region features (namely subjects and objects) and context region feature (their union). We use two interactivity blocks: one to capture the attention between the subject features and the union features, and one for the attention between the object features and the union features. From the above we know a subject-object pair is composed of continuous triplet boxes $c = \{(b_u^1, b_s^1, b_o^1), (b_u^2, b_s^2, b_o^2), ..., (b_u^k, b_s^k, b_o^k)\}$. For each frame, the three boxes are first fed to a backbone convolutional neural network to extract features. For frame f, we obtain three box features: union box features F_u^f, subject box features F_s^f and object box features F_o^f. The three features then form the input to the interactivity block. Let $F_c^f = (F_s^f, F_o^f, F_u^f)$ denote the combined feature set, then the two individual blocks are given as:

$$\mathrm{IB}_s(F_c^f) = c_1(sm(c_2(F_s^f)^T \times c_3(F_u^f)) \times c_4(F_u^f)) + F_s^f, \quad (2)$$

$$\mathrm{IB}_o(F_c^f) = c_1(sm(c_2(F_s^f)^T \times c_3(F_u^f)) \times c_4(F_u^f)) + F_o^f. \quad (3)$$

Here c_1, c_2, c_3, c_4 are convolutional layers with kernel size 1×1 and sm denotes the softmax function. The output dimensions of $c1, c2, c3, c4$ are 512. We also incorporate Dropout [39], Rescaling, Layer Normalization [3] and matrix transposition operations. The two interactivity blocks' convolutional layers share weights during training. The two blocks are combined as follows:

$$\mathrm{IB}(p) = \mathrm{IB}_s(p) + \mathrm{IB}_o(p). \quad (4)$$

The details of the interactivity blocks are illustrated in Figure 3. Interactivity block operations do not change the dimensionality of input feature. The dimensionality of input features F_s, F_o, F_u are all $\mathbb{R}^{C\times H\times W}$, the output feature $\mathrm{IB}(p)$ remains the same.

With the interactivity block, we force the network to focus on both the subject and the object. At the same time, useful contextual information is provided. The output of the function $\mathrm{IB}(p)$ is fed to an average pooling layer with kernel size 2, resulting in $F_p^f \in \mathbb{R}^C$.

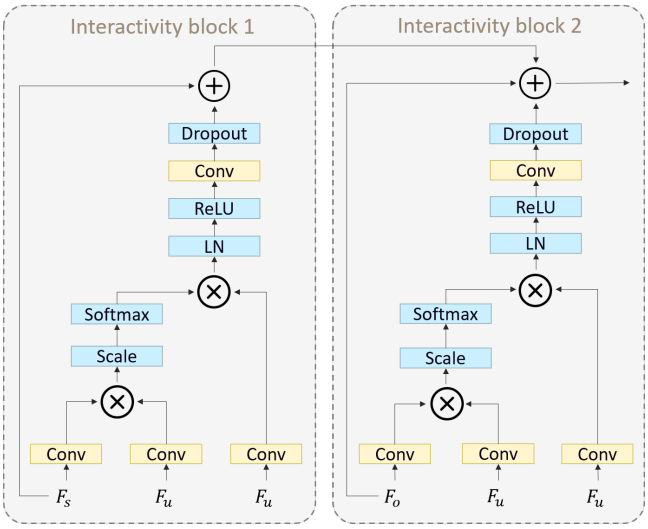


Figure 3: Interactivity block details. The two interactivity blocks share convolution layer weights with each other. The input are subject box feature f_s, object box feature f_o and union box feature f_u. Here $\oplus$ denotes element-wise sum and $\otimes$ denotes matrix product. LN is short for Layer Normalization.

Geometric location encoding. The aim of geometric location encoding is to capture the relative distance between the subject and object. Inspired by object detection in [20], we encode the relative geometric location in a subject-object pair using Eq. 5. For ease of notation, we now write each box using the topleft coordinate and width and height, *i.e.*, the subject box in f is denoted as (x_s, y_s, w_s, h_s) and the object box as (x_o, y_o, w_o, h_o), we compute the following geometry location features $F_g^f \in \mathbb{R}^8$:

$$F_g^f = \Big[\log\big(\frac{|x_s - x_o|}{w_s}\big), \log\big(\frac{|y_s - y_o|}{h_s}\big), \log\big(\frac{w_s}{w_o}\big), \log\big(\frac{h_s}{h_o}\big), \log\big(\frac{|x_o - x_s|}{w_o}\big), \log\big(\frac{|y_o - y_s|}{h_o}\big), \log\big(\frac{w_o}{w_s}\big), \log\big(\frac{h_o}{h_s}\big)\Big]. \tag{5}$$

We then concatenate F_p^f and F_g^f and score the feature:

$$s = \sigma\left(\left[F_p^f; F_g^f\right]\right), \tag{6}$$

where σ denotes the sigmoid classification and $[;]$ denotes the concatenate operation along channel dimension to get a representation of dimensionality $C + 8$.

Intearctivityness. The aim of the classification head is to output an interactivityness, a score that indicates the possibility of interaction happening in this triplet of boxes. During training, we first rely on a temporal sliding window along subject-object pairs to generate spatio-temporal interactivity proposal candidates. Then we calculate the temporal Intersection over Union (tIoU) between proposal candidates and ground truths. We collect two types of proposal samples: (1) positive proposals, *i.e.*, those overlap with the closest ground truth with at least 0.5 tIoU; (2) negative proposals, *i.e.*, those that do not overlap with any ground truth. Due to the sparsity of ground truth proposals, the number of negative proposals is much higher than the number of positive proposals. We adopt the weighted cross-entropy loss function to deal with this class imbalance:

$$\mathcal{L} = -\omega_y\big(y \log(s) + (1 - y) \log(1 - s)\big), \tag{7}$$

where s denotes the interactivityness output from Eq. 6, y the ground truth label, and ω_y the class-dependent weight used for balancing the positive and negative samples.

3.3 Interactivity proposal generation

For a subject-object pair, our network provides an interactivity score per frame. To generate spatio-temporal interactivity proposals, we rely on the 1D-watershed algorithm [35]. The main idea is to find continuous temporal segments with high interactivityness to generate proposals. The watershed algorithm was originally used as a segmentation method and later for temporal action proposal generation [51]. We first feed the boxes from the automatically computed candidate pairs to obtain frame-level interactivityness. Then, we regard the interactivityness score as a 1D terrain with heights and basins. This method floods water on this terrain with different *"levels"* (γ), resulting in a series of *"basins"* filled with water, named by $G(\gamma)$. Each obtained basin corresponds to a segment with high interactivityness. Starting from the initial basins, we merge consecutive basins until their length is above a temporal threshold τ. We uniformly sample τ and γ with step 0.05. By using multiple values for the two thresholds, multiple sets of regions are generated. We average the interactivityness for each region as the proposal score. We repeat this procedure for all selected pairs of subjects and objects. Finally, we apply non-maximum suppression on all generated proposals to remove redundant proposals. The final output is a set of spatio-temporal interactivity proposals for a video.

4 EXPERIMENTAL SETUP

4.1 KIEV dataset

To accommodate the new task of spatio-temporal interactivity proposals, we have distilled a subset from the NIST TRECVID ActEV (Activities in Extended Video) dataset, a collection of surveillance videos with spatio-temporal annotations for objects and subject [2]. ActEV is an extension of the VIRAT dataset [29]. Since not all actions in ActEV are interactions, we leverage a subset of ActEV that explicitly focuses on interactivities and call this the KIEV (Key Interactivities in Extended Video) dataset. KIEV includes high-resolution surveillance videos that are 1080p or 720p. In KIEV, the subject is a person and the object could be a person, vehicle or door. We select nine key interactivities from ActEV, namely *Closing, Closing Trunk, Entering, Existing, Loading, Opening, Opening Trunk, Unloading* and *Person Person Interaction*. Note that we do not use the interactivity labels in our approach, we are class-agnostic and are merely interested in recognizing their spatio-temporal locations. The training set has a duration of 2 hours and 17 minutes, divided over 51 long

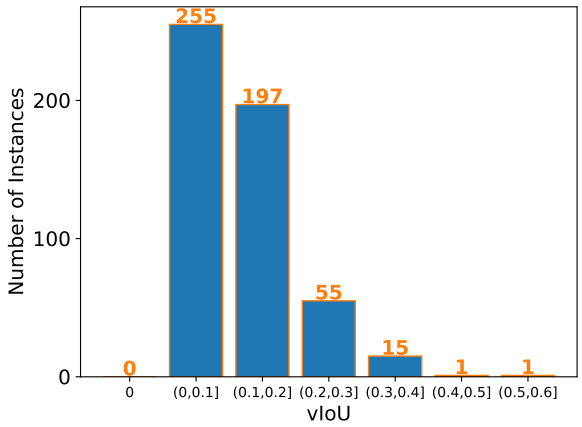


Figure 4: Histogram of vIoU between subject trajectory and object trajectory in interactivity proposal instances of KIEV. For all interactivity instances, the subjects overlap with the objects. Most overlap with vIoU from 0 to 0.2. This enforces our choice of generating interactivity candidates based on overlap.

videos. The average size of bounding boxes in the training set is 264×142, only 2.6% of the pixels in any given image. The validation set has a duration of 1 hour and 47 minutes, divided over 47 long videos.

4.2 Implementation details

Object detection and tracking. We use Faster R-CNN [33] with a ResNet-101 [19] backbone with dilated convolutions and feature pyramids [26] for person and vehicle detection. We use the model provided by [7]. The model is trained on the ActEV training set [2]. We apply this model on the unseen KIEV validation frames to obtain vehicle and person boxes. We rely on the Deep SORT tracking algorithm [47], to generate person and vehicle trajectories. During the tracking procedure, we use the boxes and Region of Interest [18] features from the detection model to link detected subjects and objects into trajectories.

Subject-object pairing. When pairing subjects and objects, we temporally extend each pair with three seconds in both directions. The temporal context is beneficial for recognizing interactivities. We also remove pairs whose duration is shorter than one second.

Interactivity network. We use the BN-Inception model provided by [51] as the feature extraction backbone. The model is pre-trained on ImageNet [9]. The interactivity network is inserted before the global average pooling layer. We use the features after the global_pool layer, whose dimensionality is 1024×7×7. After spatially pooling the feature from the interactivity network, we concatenate them with the geometric features and obtain a 1032-dimensional representation. The backbone, interactivity network, and interactivityness classifier are jointly optimized on the KIEV training set. All boxes are resized to 224×224 to meet the input dimension of BN-Inception. We train our model for 100 epochs using Adam with learning rate 1e-5, exponential decay rate 0.9, decay rate 0.999, and weight decay 5e-4. We follow [51] to set other parameters.

Proposal generation. A 1D Gaussian filter with kernel size 3 is applied to smooth the interactivityness sequence. We then apply non-maximum suppression with temporal overlap threshold 0.7 to filter out overlapping proposals.

Code. The dataset, evaluation protocols, and code are available at https://github.com/shanshuo/Interactivity_Proposals.

4.3 Evaluation metrics

We consider three evaluation metrics, which measure the temporal, spatial, and spatio-temporal proposal quality.

Average Temporal Recall. The first metric, Average Temporal Recall (ATR), measures the temporal alignment between proposals and ground truth interactivities. This metric is commonly used for temporal action proposals, *e.g.* [12, 13, 51]. A proposal is a true positive if its temporal intersection over union (tIoU) with a ground truth is greater than or equal to a given threshold. ATR is the mean of all recall values using tIoU between 0.5 to 0.9 (inclusive) with a step size of 0.05. AN is defined as the total number of proposals divided by the number of videos in the validation set. We report ATR_{25}, ATR_{50}, as well as the AUC (Area Under Curve) to see how well the proposal method works across all thresholds for number of proposals per video.

Average Spatial Recall. The second metric, Average Spatial Recall (ASR), is adapted from the AVA dataset [16]. We compare predicted boxes in each frame with ground truth boxes. If their overlaps are above a threshold of 0.5, we regard the predicted box as a true positive. We evaluate frame by frame to get the final recall.

Spatio-Temporal Recall. The third metric, Spatio-Temporal Recall, evaluates the spatio-temporal quality of an interactivity, inspired by [38]. To match a predicted interactivity proposal (t_s^p, t_o^p) to a ground truth interactivity (t_s^g, t_o^g), we require that the bounding-box trajectories overlap s.t. $\text{vIoU}(t_s^p, t_s^g) \geq 0.5$ and $\text{vIoU}(t_o^p, t_o^g) \geq 0.5$ and the proposal is not closer to another unmatched ground truth interactivity. The term vIoU refers to the voluminal Intersection over Union and is calculated as vIoU = (tube of overlap) / (tube of union). We report the spatio-temporal recall for the top 25 proposals (STR_{25}) and top 50 proposals (STR_{50}).

5 RESULTS

We consider three experiments: (i) we ablate the effectiveness of our interactivity networks, (ii) we assess the effect of automatic trackers over ground truth spatial locations, and (iii) we compare to other proposal methods.

5.1 Ablating the interactivity network

In the first experiment, we evaluate the two core components of our interactivity network: the interactivity block and the geometric

interactivity block	geometric encoding	Average Temporal Recall ATR_{25}	ATR_{50}	AUC
		6.9	14.2	6.9
✓		10.9	15.7	10.1
	✓	10.6	15.5	9.6
✓	✓	**12.4**	**19.0**	**11.3**

Table 1: Ablating the interactivity network based on temporal average recall (%). Both the interactivity block and the geometric encoding aid the proposal quality. Their combination works best. The results prove the efficiency of our method.

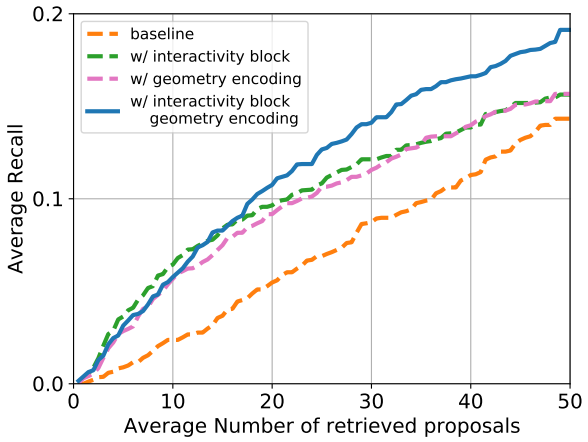


Figure 5: Ablating the interactivity network by increasing retrieved proposals. When using both the interactivity block and the geometry encoding we obtain best average recall.

encoding. The baseline method does not contain these two components. For the baseline we sum the subject feature, object feature and union feature obtained from CNN backbone together. Then we input the summed feature into classifier. We use the Average Temporal Recall as the evaluation metric. We rely on ground truth person and vehicle tubes as the subject and object trajectories to eliminate the influence of the tracker.

Interactivity block. Table 1 shows the effect of the interactivity block on the quality of the temporal interactivity proposal. We report the ATR_{25}, ATR_{50}, and AUC. The interactivity block improves ATR_{25} by 4 percent points, ATR_{50} by 1.5 and AUC by 3.2. This result indicates the interactivity block is an important element of the approach; capturing context around subjects and objects matters.

Geometry encoding. In Table 1, we also show the effect of the geometric encoding, as well as its combination with the interactivity block. After adding geometry encoding the AR_{25} is improved by 3.7, AR_{50} by 1.3, and AUC by 2.7. Combining the interactivity

Tracker	Temporal ATR_{25}	ATR_{50}	AUC	Spatio-Temporal STR_{25}	STR_{50}
ground truth	12.4	19.0	11.3	20.0	23.3
automatic	11.6	17.6	10.8	6.3	7.8

Table 2: Effect of automatic tracks on temporal and spatio-temporal proposal quality. For temporal recall, switching from ground truth to automatic trajectories has minimal effect on performance. For spatio-temporal recall, the scores naturally have a larger drop. Automatic tracks are robust enough for temporal proposal quality, but not for spatio-temporal quality.

block with the geometric encoding is most beneficial and results in improvements on all three metrics. Evidently, encoding the geometric relations between subjects and objects aids the quality of interactivity proposals.

Figure 5 shows the Temporal Average Recall as a function of the average number of retrieved proposals per video. The interactivity block and the geometric encoding improve the proposal quality scores. For their combination, the largest improvements are obtained when more proposals are generated. We conclude that the interactivity block and geometric encoding are important components of our method and we will report further experiments with their combination.

5.2 Effect of automatic tracks

Next, we evaluate the effect of using automatic tracks for subjects and objects on the interactivity proposal quality. We report both the temporal proposal quality (ATR) and spatio-temporal quality (STR) and show results in Table 2.

When evaluating the temporal dimension only, we find that automatic tracks are competitive with ground truth subject and object tubes. Indicating our method is temporally robust to noise in the spatial locations of subjects and objects. Table 2 also shows the spatio-temporal proposal quality is directly impacted by the switch from ground truth to automatic tracks. This is not surprising, since the spatio-temporal evaluation metric is very strict in its spatial evaluation; both the subject and object boxes need sufficient overlap. In Figure 7, we show a number of example proposals when using automatic trackers for the subject and object trajectories. The qualitative results indicate the difficult nature of the problem of finding spatio-temporal interactivities. Due to occlusions and tiny object sizes, there are some missed detection of interactivity in this dataset, as visualized in Figure 7c. Improved detection will positively affect interactivity proposal generation.

5.3 Comparison to prior work

In the third experiment, we compare our approach to several baselines from both the temporal and spatio-temporal action proposal literature, to show that proposing spatio-temporal interactivity locations can not be achieved by existing action proposal methods.

Baselines. We compare to two temporal proposal baselines and one spatio-temporal baseline. The first temporal proposal baseline

Method	ATR_{25}	ATR_{50}	AUC
Zhao *et al.* [51]	0.0	0.0	0.0
Gleason *et al.* [15]	1.4	1.6	1.2
Gao *et al.* [13]	8.1	12.4	7.4
This paper	**11.6**	**17.6**	**10.8**

Table 3: Temporal comparison of our interactivtyness proposals versus regular action proposals. Our method outperforms alternatives.

is TAG from Zhao *et al.* [51], which proposes temporal regions based on actionness grouping. The second temporal proposal baseline is TURN-TAP from Gao *et al.* [13], which is based on sliding windows. The spatio-temporal baseline is by Gleason *et al.* [15], who introduce a spatio-temporal proposal cuboid approach for actions. For a fair comparison, the input object boxes are the same as our approach.

Temporal comparison. Since temporal action proposal methods only provide the start and end times, we first compare our proposals to all baselines using the temporal quality metrics. The results are shown in Table 3 and Figure 6. Our approach performs better than all baselines. In comparison to the best scoring baseline of Gao *et al.* [13], our method improves the ATR_{25} by 3.5, the AR_{50} by 5.2, and the AUC by 3.4. The approaches of Zhao *et al.* [51] and Geo *et al.* [13] fail to generate efficient proposals in this setting because they take the whole frame as input. Since interactivities are only a small part of the video spatially, their representations hardly capture the precise interactions, as expected. These temporal action localization methods fail to solve the interactivity proposal problem. They are capable of localizing temporal boundaries but ignore spatial boundaries. Our approach operates locally in space, which allows for a better estimation of interactivities in time. The approach of Gleason *et al.* [15] does operate locally in space, but does not explicitly capture contextual and geometric relations between subjects and objects, which results in lower recall scores.

Spatio-temporal comparison. In Table 4, we also compare our approach to Gleason *et al.* [15] with respect to the spatio-temporal proposal quality. The results show that spatially, the baseline obtains an ASR of 8.4, while we reach a score of 61.5, a considerable gain. Furthermore, the spatio-temporal recall at both 25 and 50 proposals per video is 0 for the baseline, compared to 4.8 and 6.3 for our approach. The reason for this gap in performance is because the baseline generates cuboid-style proposals, leading to coarse spatial localization of subjects and objects. The cuboid-style proposals have low IoUs compared to trajectory-style ground truths. In our evaluation, we care about a precise dynamic alignment in space and time for subjects and objects. Our approach yields more accurate spatio-temporal interactivity proposals, be it the overall spatio-temporal recall is modest. Compared to Gleason *et al.* [15] we conclude that our approach is better equipped to find interactivities more precisely in space and time.

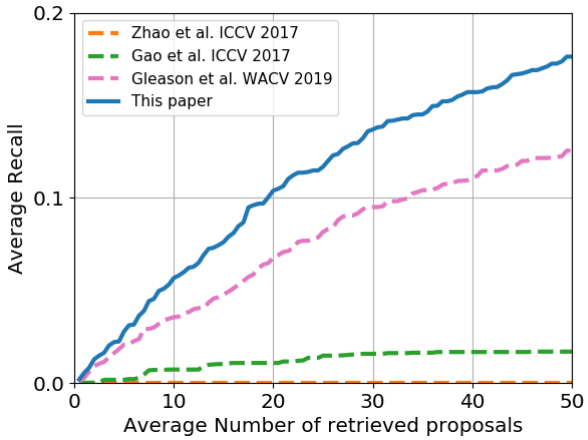


Figure 6: Temporal comparison of interactivtyness proposals versus regular action proposals under varying number of retrieved proposals. Modeling interactivityness rather than activity is beneficial.

Method	ASR	STR_{25}	STR_{50}
Gleason *et al.* [15]	8.4	0.0	0.0
This paper	**61.5**	**4.8**	**6.3**

Table 4: Spatio-temporal comparison of our interactivityness proposals versus a regular action proposal in terms of Recall (%). Explicitly modeling interactivity results in better spatio-temporal localization.

6 CONCLUSION

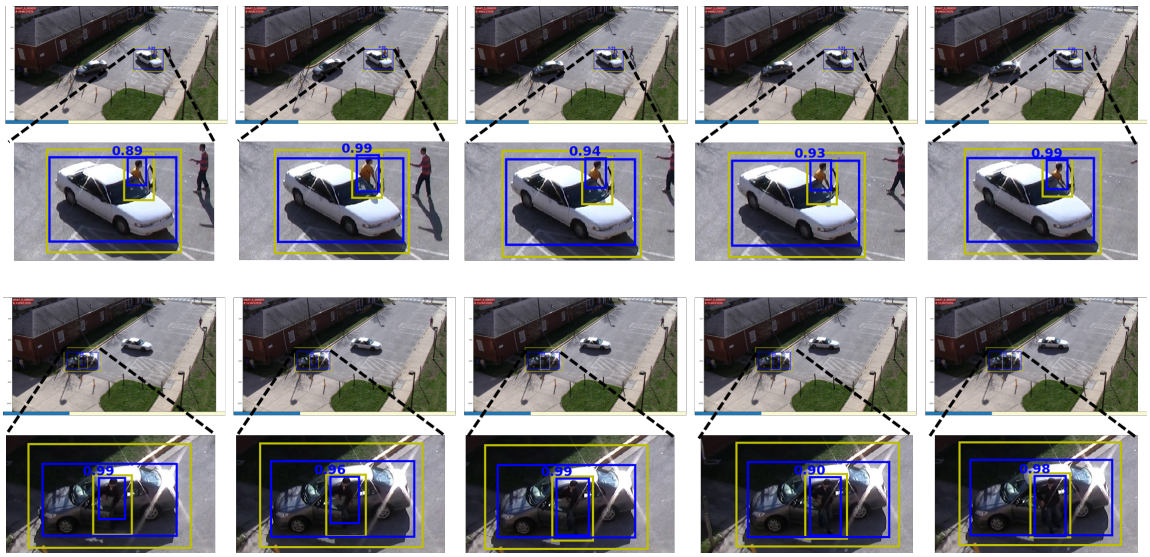
This paper introduces interactivity proposals for video surveillance. Rather than focusing on the actions of the subject only, our proposals capture the interplay between subjects and objects in space and time. To that end, we propose a network to compute interactivityness between subjects and objects from which we generate class-agnostic proposals. We evaluate the proposals on an interactivity dataset with new overlap metrics, where experiments show the improvement of our approach over traditional temporal and spatio-temporal action proposal methods. Overall, the results are far from perfect, indicating the challenging nature of the problem. To encourage further progress on recognizing interactivity proposals we make the dataset split, evaluation metrics, and code publicly available.

REFERENCES

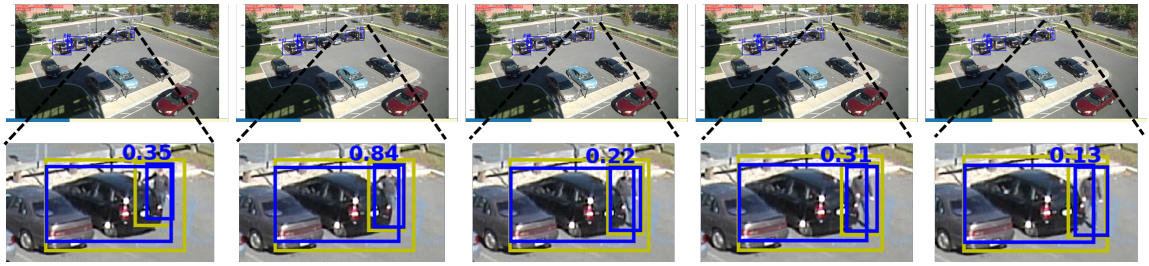
[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. 2012. Measuring the objectness of image windows. *PAMI* (2012).

[2] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, and Saverio Blasi. 2018. TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. In *TRECVID*.

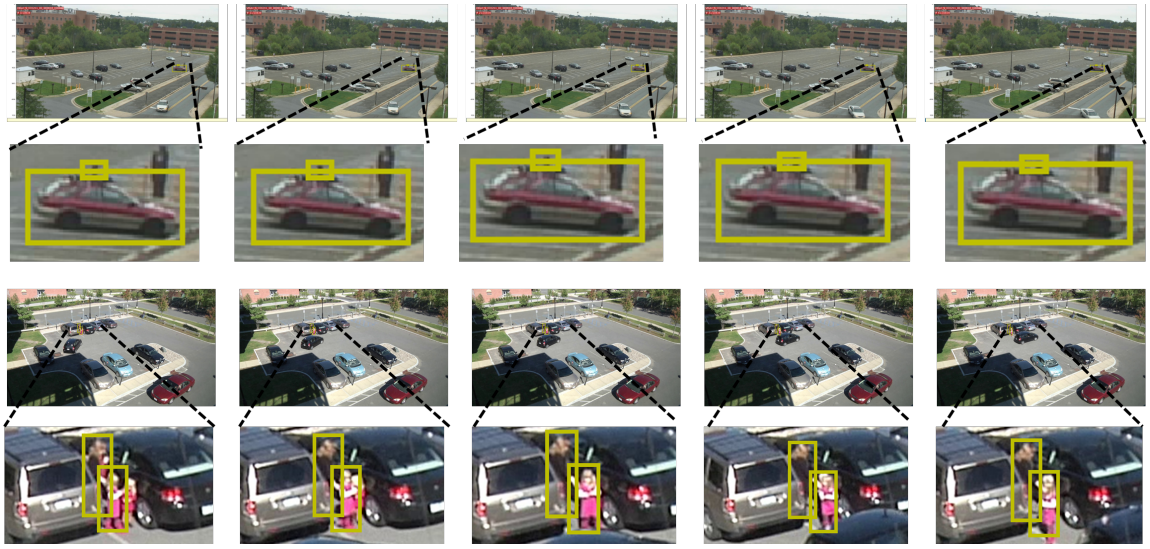
[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).



(a) Successful cases.



(b) Imperfect case.



(c) Failed cases.

Figure 7: Qualitative results. (a). The top two examples show successful cases, where the proposal highly overlaps in space and time with the ground truth. From top to bottom the interactivities are Entering, Exiting, Closing, Entering and Person Person Interaction. Note that we do not output labels. Here the labels are only for clarifying. The bottom two examples show failure cases, (b). occlusion and (c). small object sizes either result in a low interactivityness or even missed subject and object trajectories. These failure cases highlight the difficult nature of finding interactivities in outdoor settings.

[4] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. 2017. SST: Single-stream temporal action proposals. In *CVPR*.

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *WACV*.

[6] Jia Chen, Jiang Liu, Junwei Liang, Ting-Yao Hu, Wei Ke, Wayner Barrios, Dong Huang, and Alexander G Hauptmann. 2019. Minding the Gaps in a Video Action Analysis Pipeline. In *WACV workshop*.

[7] Jia Chen, Jiang Liu, Junwei Liang, Ting-Yao Hu, Wei Ke, Wayner Barrios, Dong Huang, and Alexander G Hauptmann. 2019. Minding the Gaps in a Video Action Analysis Pipeline. In *WACV Workshop*.

[8] Wei Chen, Caiming Xiong, Ran Xu, and Jason J Corso. 2014. Actionness ranking with lattice conditional ordinal random fields. In *CVPR*.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[10] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *ECCV*.

[11] Chen Gao, Yuliang Zou, and Jia-Bin Huang. 2018. ICAN: Instance-centric attention network for human-object interaction detection. In *BMVC*.

[12] Jiyang Gao, Kan Chen, and Ram Nevatia. 2018. CTAP: Complementary temporal action proposal generation. In *ECCV*.

[13] Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ramakant Nevatia. 2017. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. (2017).

[14] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *CVPR*.

[15] Joshua Gleason, Rajeev Ranjan, Steven Schwarcz, Carlos Castillo, Jun-Cheng Chen, and Rama Chellappa. 2019. A Proposal-Based Solution to Spatio-Temporal Action Detection in Untrimmed Videos. In *WACV*.

[16] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*.

[17] Jiawei He, Zhiwei Deng, Mostafa S Ibrahim, and Greg Mori. 2018. Generic tubelet proposals for action localization. In *WACV*.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[20] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *CVPR*.

[21] Mihir Jain, Jan Van Gemert, Hervé Jégou, Patrick Bouthemy, and Cees G M Snoek. 2017. Tubelets: Unsupervised action proposals from spatiotemporal super-voxels. *IJCV* (2017).

[22] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. 2013. Towards understanding action recognition. In *ICCV*.

[23] Tae Soo Kim, Mike Peven, Weichao Qiu, Alan Yuille, and Gregory D Hager. 2019. Synthesizing Attributes with Unreal Engine for Fine-grained Activity Analysis. In *WACV Workshop*.

[24] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. 2019. Peeking into the future: Predicting future person activities and locations in videos. In *CVPR*.

[25] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *ICCV*.

[27] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. 2019. Multi-granularity Generator for Temporal Action Proposal. In *CVPR*.

[28] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. 2015. Watch and learn: Semi-supervised learning for object detectors from video. In *CVPR*.

[29] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*.

[30] Dan Oneata, Jérôme Revaud, Jakob Verbeek, and Cordelia Schmid. 2014. Spatio-temporal object detection proposals. In *ECCV*.

[31] Alessandro Prest, Vittorio Ferrari, and Cordelia Schmid. 2012. Explicit modeling of human-object interactions in realistic videos. *PAMI* 35, 4 (2012), 835–848.

[32] Haonan Qiu, Yingbin Zheng, Hao Ye, Yao Lu, Feng Wang, and Liang He. 2018. Precise temporal action localization by evolving temporal proposals. In *ICMR*.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.

[34] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. 2008. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition.. In *CVPR*.

[35] Jos BTM Roerdink and Arnold Meijster. 2000. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta Informaticae* (2000).

[36] Maguell LTL Sandifort, Jianquan Liu, Shoji Nishimura, and Wolfgang Hürst. 2018. An entropy model for loiterer retrieval across multiple surveillance cameras. In *ICMR*.

[37] Maguell LTL Sandifort, Jianquan Liu, Shoji Nishimura, and Wolfgang Hürst. 2018. VisLoiter+: An entropy model-based loiterer retrieval system with user-friendly interfaces. In *ICMR*.

[38] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. In *ICMR*.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* (2014).

[40] Eran Swears, Anthony Hoogs, Qiang Ji, and Kim Boyer. 2014. Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *CVPR*.

[41] Jan C Van Gemert, Mihir Jain, Ella Gati, and Cees G M Snoek. 2015. APT: Action localization proposals from dense trajectories.. In *BMVC*.

[42] Jacob Walker, Abhinav Gupta, and Martial Hebert. 2014. Patch to the future: Unsupervised visual prediction. In *CVPR*.

[43] He Wang, Sören Pirk, Ersin Yumer, Vladimir G Kim, Ozan Sener, Srinath Sridhar, and Leonidas J Guibas. 2019. Learning a Generative Model for Multi-Step Human-Object Interactions from Videos. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 367–378.

[44] Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool. 2016. Actionness estimation using hybrid fully convolutional networks. In *CVPR*.

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *CVPR*.

[46] Xiaoyang Wang and Qiang Ji. 2014. A hierarchical context model for event recognition in surveillance video. In *CVPR*.

[47] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *ICIP*.

[48] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. 2019. Learning to Detect Human-Object Interactions With Knowledge. In *CVPR*.

[49] Gang Yu and Junsong Yuan. 2015. Fast action proposals for human action detection and search. In *CVPR*.

[50] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. 2019. On Exploring Undetermined Relationships for Visual Relationship Detection. In *CVPR*.

[51] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *ICCV*.

[52] Zhicheng Zhao, Xuanchong Li, Xingzhong Du, Qi Chen, Yanyun Zhao, Fei Su, Xiaojun Chang, and Alexander G Hauptmann. 2018. A unified framework with a benchmark dataset for surveillance event detection. *Neurocomputing* (2018).