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:

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.
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 actions 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 tubelet proposal network for action detection, which adopts Faster R-CNN [33] to collect boxes with high action score. 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] relate 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 JHMDB51 [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 $t_s = \{b^1_s, b^2_s, ..., b^n_s\}$ has $n$ boxes and the object trajectory $t_o = \{b^1_o, b^2_o, ..., b^m_o\}$ 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 $b^f_s \in t_s$ and object box $b^f_o \in t_o$. 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
We add the union boxes to the subject-object pairs and obtain =

\[ t^F = \text{IB}(p) = \text{IB}_s(p) + \text{IB}_o(p). \] (4)

The details of the interaction blocks are illustrated in Figure 3. Interaction 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 \(\text{IB}(p)\) remains the same.

With the interaction 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 \(\text{IB}(p)\) is fed to an average pooling layer with kernel size 2, resulting in \(F'_p \in \mathbb{R}^C\).
We then concatenate \( F \) where we calculate the temporal Intersection over Union (tIoU) between

\[
\sigma \text{ denotes the sigmoid classification and } [ , ] \text{ denotes the concatenation operation along channel dimension to get a representation of dimensionality } C + 8.
\]

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:

\[
\mathcal{L} = -\omega_y (y \log(s) + (1 - y) \log(1 - s)),
\]

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

### 3.3 Interactivityness proposal generation

For a subject-object pair, our network provides an interactivityness 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” (\( y \)), 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 \( \tau \). We uniformly sample \( \tau \) and \( y \) 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
We also remove pairs whose duration is shorter than one second.

We apply this model on the unseen KIEV validation frames to obtain a 1032-dimensional representation. The backbone, interactivity network, and interactivityness classifier are jointly optimized on the KIEV training set.

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.

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.


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 ATR25, ATR50, 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 (STR), evaluates the spatio-temporal quality of an interactivity, inspired by [38]. To match a predicted interactivity proposal \((t^p_o, t^p_s)\) to a ground truth interactivity \((t^g_o, t^g_s)\), we require that the bounding-box trajectories overlap s.t. \(\text{vIoU}(t^p_o, t^p_s) \geq 0.5\) and \(\text{vIoU}(t^p_o, t^p_s) \geq 0.5\) and the proposal is not closer to another unmatched ground truth interactivity. The term \(\text{vIoU}\) refers to the voluminal Intersection over Union and is calculated as \(\text{vIoU} = (\text{tube of overlap}) / (\text{tube of union})\). We report the spatio-temporal recall for the top 25 proposals (STR25) and top 50 proposals (STR50).

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...
Average Temporal Recall

<table>
<thead>
<tr>
<th>Interactivity Block</th>
<th>Geometric Encoding</th>
<th>ATR_{25}</th>
<th>ATR_{50}</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>6.9</td>
<td>14.2</td>
<td>6.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>10.3</td>
<td>15.7</td>
<td>10.1</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>12.4</td>
<td>19.0</td>
<td>11.3</td>
</tr>
</tbody>
</table>

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 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.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Temporal ATR_{25}</th>
<th>Temporal ATR_{50}</th>
<th>Temporal AUC</th>
<th>Spatio-Temporal STR_{25}</th>
<th>Spatio-Temporal STR_{50}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ground truth</td>
<td>12.4</td>
<td>19.0</td>
<td>11.3</td>
<td>20.0</td>
<td>23.3</td>
</tr>
<tr>
<td>automatic</td>
<td>11.6</td>
<td>17.6</td>
<td>10.8</td>
<td>6.3</td>
<td>7.8</td>
</tr>
</tbody>
</table>

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.

### 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
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


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.


[17] Jiawei He, Zhiwei Deng, Mostafa S Ibrahim, and Greg Mori. 2018. Generic tubelet generator for temporal action proposal. In CVPR.


[19] 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.


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

[29] Nicole Wojke, Alex Dewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In IJCV.

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


