Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Human-Scene Network: A Novel Baseline with Self-rectifying Loss for Weakly Supervised Video Anomaly Detection

Snehashis Majhi[a,**], Rui Dai[a], Quan Kong[b], Lorenzo Garattoni[c], Gianpiero Francesca[c], François Brémond[a]

[a] *INRIA, 2004 Rte des Lucioles, Valbonne, France*
[b] *Woven Planet Holdings, 3-2-1 Nihonbashimuromachi, Chuo-ku, Tokyo, Japan*
[c] *Toyota Motor Europe, 60 Av. du Bourget, Brussels, Belgium*

ABSTRACT

Video anomaly detection in surveillance systems with only video-level labels (*i.e. weakly supervised*) is challenging. This is due to (i) the complex integration of a large variety of scenarios including human and scene-based anomalies characterized by subtle or sharp spatio-temporal cues in real-world videos and (ii) non-optimal optimization between normal and anomaly instances under weak supervision. In this paper, we propose a Human-Scene Network to learn discriminative representations by capturing both subtle and strong cues in a dissociative manner. In addition, a self-rectifying loss is proposed that dynamically computes the pseudo-temporal annotations from video-level labels for optimizing the Human-Scene Network effectively. The proposed Human-Scene Network optimized with self-rectifying loss is validated on three publicly available datasets *i.e.* UCF-Crime, ShanghaiTech, and IITB-Corridor, outperforming recently reported state-of-the-art approaches on five out of the six scenarios considered.

1. Introduction

Anomaly detection in real-world videos is a crucial computer vision task thanks to its potential applications in smart cities empowering timely anomaly prevention and investigation. This problem remains unsolved due to the scarcity of spatio-temporally annotated data and the sparsity in the occurrence of anomaly events. In consequence, earlier popular methods (Cong et al., 2013; Adam et al., 2008; Kim and Grauman, 2009; Yu et al., 2021; Wang, 2019; Ramachandra and Jones, 2020; Kim and Grauman, 2009; Zhao et al., 2011; Liu et al., 2018; Lu et al., 2013a; Sun et al., 2020; Zaheer et al., 2022; Hasan et al., 2016) learn a uni-class (*i.e. only normal class that is easy to acquire*) encoder-decoder network for learning the global temporal regularity and treat abnormality as an out-of-distribution detection (OOD) w.r.t. the learned normal distribution. As a matter of fact, these methods fail to learn generalized representations for all possible normal scenarios and hence cause false alarms for unseen normal ones. In light of superior generalization capabilities and detection performance, weakly-supervised video anomaly detection (WSVAD) methods (Sultani et al., 2018; Zhong et al., 2019; Wu et al., 2020) have recently gained popularity. These methods learn from both normal and anomaly distributions to optimize the separability among the classes with only video-level labels.

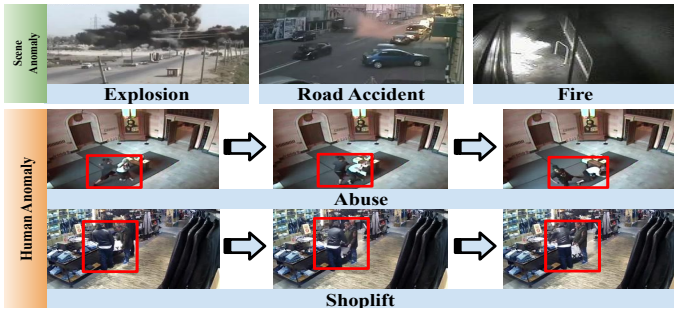

Fig. 1. Visualization of scene and human centric anomalies in CCTV videos. The scene-centric anomalies (*row-1*) contain sharp changes of spatio-temporal cues. But human-centric anomalies may carry strong local motion (*row-2: abuse*) or characterized by subtle features (*row-3: shoplift*).

Despite the prosperity in mainstream WSVAD approaches (Sultani et al., 2018; Zhu and Newsam, 2019; Zhang et al., 2019; Lin et al., 2019; Zhong et al., 2019; Wu et al., 2020;

**Corresponding author:
e-mail: snehashis.majhi@inria.fr (Snehashis Majhi)

Zaheer et al., 2020), their performance is still limited due two major challenges: (a) complex real-world abnormalities: difficulties in obtaining discriminative spatio-temporal representations with only video-level labels, when there exists an integration of local human-centric anomalies (*Abuse, Shoplifting, Stealing, Robbery, Vandalism*) with the global scene-centric anomalies (*Explosion, Road Accidents, Fire, Burglary*), as illustrated in Fig. 1; (b) non-optimal separation between normal and anomaly classes: obscure consideration of normal and anomaly instances (or temporal segments) in a long untrimmed video labeled as "anomaly" for optimizing the separation among the classes leads to non-optimality.

To address the above challenges, attempts have been made by (Sultani et al., 2018; Wu et al., 2020). They first extract features using a 3D ConvNet and then learn an MLP ranker by multiple instance learning (MIL) based optimization. Since many previous methods consider only global feature representation (i.e. features extracted from the whole frame) for optimizing the MLP ranker, they still lag in detecting human-centric anomalies. This is mainly due to the different features characterizing the anomalies. Scene-centric anomalies are characterized by strong appearance and global motion cues, whereas human-centric anomalies have rather subtle and local motion patterns. As a result, global features fail to capture the human-centric subtle cues although they succeed in characterizing scene-centric cues. Furthermore, for optimizing the MLP ranker, earlier WSVAD approaches adapt a classical MIL loss proposed by (Sultani et al., 2018) which selects two instances based on the presence of abnormality (*i.e. one each from normal and anomaly videos*) to take part in the optimization process. We believe such optimization can perform well when anomalies are short in duration like explosions and road accidents, but it may fail drastically when anomalies last longer like shoplifting.

In contrast, we propose a novel Human-Scene Network (HSN) comprising two decoupled sub-networks (subNet) *i.e scene and human subNets*, which are optimized independently with a soft-selection module as the key component for addressing the stringent requirement of WSVAD. The decoupled design of the network helps in learning discriminative representations for scene and human-centric anomalies in a mutually exclusive manner, thus enabling each subNet to detect either coarse scene-centric or fine-grained human-centric anomalies effectively. Now, instead of treating both subNets as independent decision modules to detect anomalies, we propose a class-agnostic soft-selection coupler to choose between the scene and human-centric subNets for a given video. The key difference between the previous WSVAD methods and our approach is outlined in Fig. 2. Since the overall performance of the proposed network can be affected due to the limitations of classical MIL loss, we propose a self-rectifying loss function for enhanced optimization with video-level labels. Unlike the earlier loss, the proposed loss not only ensures video-level context maximization between normal and anomaly classes, but also performs an instance-level maximization by dynamically choosing the optimal number of instances empowered by a self error-minimization procedure. Our main contributions are :

- A novel Human-Scene Network (HSN) is proposed as a baseline framework for effectively detecting scene and human centric anomalies under weak-supervision.
- A new self-rectifying loss function is proposed for ensuring superior separation between normal and anomaly instances by overcoming the drawbacks of the previous MIL loss.
- An exhaustive experimental analysis is performed to corroborate the robustness of HSN along with the self-rectifying loss on three competitive datasets UCF-Crime (Sultani et al., 2018), ShanghaiTech (Lu et al., 2013b) and IITB-Corridor (Rodrigues et al., 2020) datasets, outperforming previous approaches on five out of the six scenarios considered.

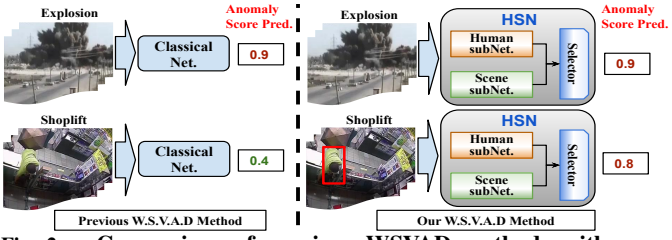

Fig. 2. Comparison of previous WSVAD methods with ours. Previous methods consider only scene-centric features (*i.e. from the whole frame*) and they fail to detect complex human-centric anomalies. In contrast, our HSN can learn from both cases and obtain a better performance.

2. Related Work

Weakly-supervised anomaly detection. It has been extensively studied in the past few years (Sultani et al., 2018; Zhong et al., 2019; Wu et al., 2020; Zhang et al., 2019; Wan et al., 2020; Zhu and Newsam, 2019; Lin et al., 2019; Zaheer et al., 2020). Majority of previous works follow the multiple instance learning (MIL) approach introduced by (Sultani et al., 2018) to overcome the drawbacks of traditional unsupervised one-class learning anomaly detection methods (Adam et al., 2008; Kim and Grauman, 2009; Cong et al., 2013; Ramachandra and Jones, 2020; Kim and Grauman, 2009; Zhao et al., 2011; Liu et al., 2018; Lu et al., 2013a; Sun et al., 2020). In weakly-supervised anomaly detection task only video-level labels are provided for learning. (Sultani et al., 2018) only extract off-the-shelf global scene features from a pre-trained 3D ConvNet backbone (Tran et al., 2015; Carreira and Zisserman, 2017) and aim at training a classification network through a classical ranking loss function. Although superior separation between normal and anomaly instances is ensured by the ranking loss than that of unsupervised anomaly detection methods by choosing only the maximum scoring segment of both normal and anomaly videos for optimization but (Sultani et al., 2018) were able to produce limited detection performance. This is because they only focus on scene level global features ignoring the object or human level features and temporal context modeling.

Human and Scene Centric Feature Combination. In order to detect both human and scene-centric anomalies characterized by divergent cues, effectively combining the local object-centric subtle features and global scene-centric sharp features are necessary. However, the development and validation of such methods are limited in real-world large-scale WSVAD datasets.

This is majorly due to the difficulties involved in combining divergent cues by highlighting the salient features. Whereas, several object-centric anomaly detection methods (Ionescu et al., 2019; Hinami et al., 2017; Roy et al., 2021; Wang et al., 2020; Liu et al., 2023) exist in unsupervised video anomaly detection (UVAD) methods. In UVAD, the object-centric methods are quite straightforward as the datasets considered are small-scale, the abnormal events are acted, and the event of interest occurs in the object-localized spatial regions. As a result, considering only object trajectories as input to the model can be sufficient to obtain desired performances. In contrast, the anomaly localized regions in real-world scenarios can be human or scene-centric. To address this, (Purwanto et al., 2021) combines the global scene-level feature and local patch-based object features via a self-attention mechanism. However, processing both local and global features jointly often tends to overlook the subtle object/human cues. Thus, a decoupled learning framework is required to encourage the subtle and sharp abnormal cues.

Temporal context modeling for long and short length anomalies. To detect long and short length anomalies, temporal context modeling is another crucial aspect. For this, authors in (Zhang et al., 2019) utilize TCN (Lea et al., 2016) in MIL based approach to learn temporal dependency encoding for anomaly instances at the feature level. Another approach (Zhu and Newsam, 2019) combines global optical flow features obtained from PWCNet (Sun et al., 2018) with the RGB feature map for discriminating the anomaly instances that exhibit strong motion. Since these methods only enable to capture the short term sharp temporal variations, it can only aid to detect short length anomalies. In contrast, authors in (Tian et al., 2021) proposed a multi-scale temporal convolution network (MTN) for global temporal dependency modeling between normal and anomaly segments. Recently, (Zhou et al., 2023) and (Chen et al., 2023b) adopt transformer-based global-local and focus-glance blocks respectively to capture long and short-term temporal dependencies in normal and anomalous videos. However, as these methods (Chen et al., 2023b; Zhou et al., 2023; Tian et al., 2021) follow a magnitude-based optimization, they only encourage the sharp abnormal cues of short anomalies to take part in temporal modeling. Further, feature magnitudes-based optimization is influenced by only strong spatio-temporal variation across temporal segments leading to ineffective separability for subtle and local anomalies. Two key drawbacks observed in the above approaches are: (i) consideration of less separable global features to capture real-world anomalies, (ii) optimization with classical ranking loss which only considers selected instances for optimization and does not ensure optimal separability.

To combat the limitation of classical ranking loss in obtaining discriminative features, (Zhong et al., 2019) and (Feng et al., 2021) aim at training the 3D ConvNet backbone by generating pseudo-temporal annotations through a cleaning noisy labels approach. (Zhong et al., 2019) capture temporal consistency in a GCN and (Feng et al., 2021) use a deep MIL ranker to generate pseudo temporal annotations. Although they obtain higher performance than MIL approaches, they are still limited by learning the global features in 3D ConvNet and in the pseudo annotation generator. A drawback of these approaches is that they operate in a two-step manner, where the first step is to generate pseudo labels followed by 3D ConvNet optimization. Again, as noisy pseudo labels can be the result of non-discriminative global features, it can mislead the optimization of 3D ConvNet backbone. Thus, we propose a novel method: Human Scene Network (HSN), which considers both local and global spatio-temporal cues for obtaining discriminative representation in real-world scenarios. In addition, we also propose a self-rectifying loss to optimize HSN, which generates pseudo-labels to select the optimal number of instances in MIL optimization paradigm for maximum separability.

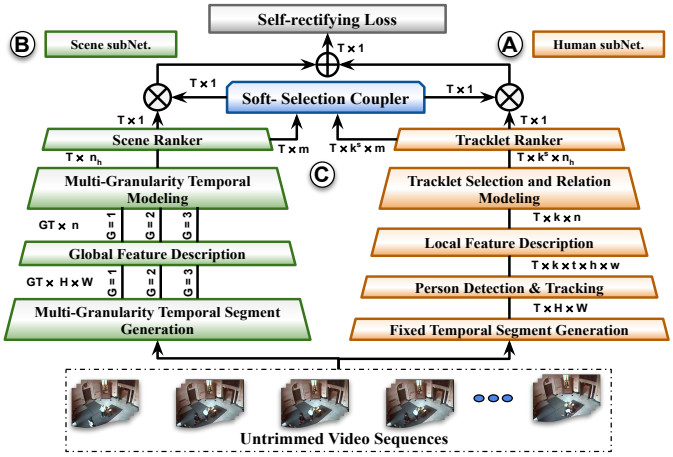

Fig. 3. Human-Scene Network (HSN): It comprises of three key building blocks *i.e.* (A) Human subNet, (B) Scene subNet, and (C) Soft-Selection Coupler. HSN inputs a long video sequence that is processed by (A) and (B) to learn either human or scene centric representations synchronously. Afterwards, (C) is optimized to compute a coupled selection factor based on the representation learned by (A) and (B) to focus on either of them.

3. Human-Scene Network (HSN)

The overview of the proposed HSN is delineated in Fig. 3. Its three key components, *human subNet , scene subNet, and soft-selection coupler* are designed to precisely detect real-world anomalies when an untrimmed video is given as input. A detailed description of each component in HSN along with the optimization strategy is given in the following subsections.

3.1. Scene-subNet

The objective of the Scene-subNet (SsN) is to learn discriminative global representations characterizing scene-centric anomalies. The SsN ensures to capture strong appearance and global motion cues with its four salient building blocks as elaborated below.

3.1.1. Multi-Granularity Temporal Segment Generation:

Primarily, this section considers a new temporal segment generation strategy ideal for WSVAD. A key drawback in earlier work (Sultani et al., 2018) is that they generate fixed scale (namely granular) temporal segments (*say* T) for each video V. By dividing V of length T_l into T, where $T \ll T_l$, it misses out the fine temporal cues. For this, the short duration anomalies get suppressed by the neighbouring normal ones and hence the approach lags in detection performance. To preserve fine temporal cues, our strategy consists in dividing each video V at

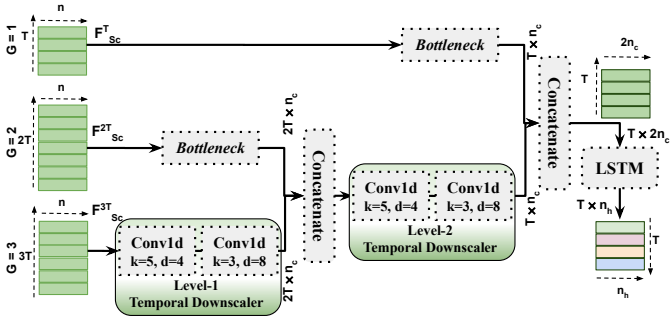

Fig. 4. Multi-Granularity Temporal Modeling (MGTM): It creates a temporal feature pyramid with *Temporal Downscaler* and *Bottleneck* modules followed by a *LSTM* cell to learn discriminative fine-grained and contextual temporal representations.

multiple granularities $G = 1, 2, 3$ for generating temporal segments of size T, $2T$ and $3T$ respectively. This temporal segment generation is performed by varying the sampling rate of V.

3.1.2. **Global Feature Description:**

For global feature description, *off-the-shelf* spatio-temporal features are extracted from a pre-trained 3D ConvNet for each temporal segment (GT_i) generated from V. A GT_i is a set of consecutive frames that have both background and foreground scene. For a given GT_i, the 3D ConvNet extracts a feature map of dimension $c \times n$, where c is the number of 64-frames clips inside GT_i and n is the channel size. Since multiple 64-frame clips can be present inside a GT_i, a `max-pooling` is done over c to get a uni feature map per GT_i. So, for a given V containing GT segments, a feature map F^{GT}_{Sc} of dimension $GT \times n$ is obtained from this block, where $G = 1, 2, 3$.

3.1.3. **Multi-Granularity Temporal Modeling :**

As short and long anomalies are characterized by sharp and progressive change in spatio-temporal cues respectively, it is crucial to encode the contextual and fine-grained temporal dependencies respectively. Hence, a higher-granular feature map $F^{3T}_{Sc} \in \mathbb{R}^{3T \times n}$ is desirable for short anomalies as it can preserve fine temporal details to capture sharp changes. Similarly, a lower-granular feature map $F^{T} \in \mathbb{R}^{T \times n}$ can succeed in providing the context for long anomalies. So, to effectively learn the representations for both long and short anomalies from the pre-computed feature maps F^{T}_{Sc}, F^{2T}_{Sc}, F^{3T}_{Sc}, a multi-granularity temporal modeling (MGTM) block is proposed in SsN as shown in Fig. 4.

MGTM aims at building a temporal feature pyramid followed by dependency modeling thanks to its three modules: (i) temporal downscaler (ii) bottleneck layer and (iii) LSTM cell to encode the contextual and fine-grained temporal dependencies. MGTM obtains the temporal feature pyramid by aggregating the feature maps from multiple temporal granularities *i.e.* $G = 1, 2, 3$. The aggregation is done by first reducing the temporal dimension of higher granular feature maps (say F^{3T}_{Sc}) and then combining with the successive lower granular feature maps (say F^{2T}_{Sc}). The temporal downscaler block ensures to reduce the temporal dimension by a factor of T with its two sequential 1D convolution layers with kernel size $k \in \{5, 3\}$ and dilation rate $d \in \{4, 8\}$ respectively. The temporal downscaling is performed by encapsulating the features from neighboring temporal segments with the increased receptive field of 1D dilated convolution layers. The MGTM has two-levels of temporal downscaler *i.e.* level-1 between $G = 3$ and 2, level-2 between $G = 2$ and 1. Further, to combine the temporally down scaled feature with lower granular features F^{GT}_{Sc} (where $G = 1, 2$), first F^{GT}_{Sc} is applied to a bottleneck module having one layer of 1D convolution with kernel size $k = 1$ and then a `concatenate` operation is performed between the feature maps. At the end of the final `concatenate` operation that combines the features from $G = 2$ and 1, it results in a feature map of dimension $T \times 2n_c$ (where T= temporal dimension similar to $G = 1$, n_c = number of convolution filters in final bottleneck and temporal downscaler module). The resultant feature map is subsequently input to a `many-to-many` LSTM cell with n_h hidden neurons for global temporal dependency encoding (F^{*}_{Sc}). Since the obtained temporal encodings are enriched by fine-grained and contextual temporal dependency modeling at multiple granularities, it ensures a better discriminative representation among the temporal segments T_i.

3.1.4. **Scene Ranker :**

The scene ranker is a multi-layer perceptron (MLP) with three `fully-connected` (FC) layers which inputs the $F^{*}_{Sc} \in \mathbb{R}^{T \times n_h}$ to assign anomaly ranks (*or* scores) to each temporal segment. For this, the final layer of MLP has a single neuron with *sigmoid* activation to rank each temporal segment independently. Finally the scene ranker outputs a detection score map D_{Sc} of dimension $T \times 1$ to be used in anomaly detection.

3.2. Human-subNet

The objective of the Human-subNet (HsN) is to learn discriminative local representation characterising human-centric anomalies. In real-world situations human-centric anomalies can either contain subtle or sharp appearance and motion cues. The HsN ensures to learn a local discriminative representation in all possible scenarios with its four major building blocks as described below.

3.2.1. **Fixed Temporal Segment Generation:**

Unlike SsN, here the video V is divided into a fixed number of temporal segments, (*say* T). This is because in the following blocks of HsN, humans are detected and tracked in each temporal segment. So, humans are tracked as long as possible making the multi-temporal granularities non-suitable. For instance, with a high granularity value (such as $G = 3$), the length of a segment T_i is reduced by a multiplication factor of G, hence the tracking may be lost in such reduced temporal duration. Thus, for a given V, HsN outputs T temporal segments of spatial resolution $H \times W$ which are provided to the subsequent block.

3.2.2. **Human Detection and Tracking:**

In order to learn a local representation in HsN a pre-requisite is to obtain human bounding boxes (BBox) and their corresponding trajectories (*namely tracks*) with a sufficient quality. For this, a pre-trained human detector and tracker network (ByteTrack (Zhang et al., 2022b)) is used to extract humans BBox and the tracks in each segment T_i. So, for T segments generated from video V with k humans present in the scene, this block outputs a tracklet map of dimension $T \times k \times t \times h \times w$, where t is the maximum track duration of each tracklet and $h \times w$ is the BBox dimension.

3.2.3. **Local Feature Description:**

A pre-trained 3D ConvNet is used to sequentially extract spatio-temporal features for each tracklet k_j present in each temporal segment T_i. Since the 3DConvNet used in HsN is bounded by the input size of 64 frames, each with resolution 224×224 , the tracklet map is resized and padded to meet the input requirements. For a given tracklet k_j present in T_i, the track duration t can be larger than 64-frame sequences. So, a `max-pooling` operation is performed to obtain a n-dimensional unitary feature vector per tracklet k_j in each T_i. The obtained local feature map $F_{Tr} \in \mathbb{R}^{T\times k\times n}$ is fed to the subsequent block to encode the dependency among tracklets.

3.2.4. **Tracklet Selection and Relation Modeling:**

A key intricacy of encoding the individual tracklets behaviour and relations among them lies in the number of people present in the video. With a large number of people, the relation modeling seems difficult and increases the model complexity as well. For this, a tracklet selection (TS) strategy targeted only for the anomaly detection task is proposed here to filter out the salient tracklets, followed by a relation modeling method to obtain local discriminative representation in an effective and efficient manner.

The TS filters out k^s salient people out of k in F_{Tr}, where k^s people has significantly distinctive behaviour (*assuming sharp change in the spatio-temporal feature space as distinctive*) than the remaining others(*i.e.* $k - k^s$). For a given spatio-temporal feature vector X corresponding to tracklet k_j, the distinctive behaviour is defined by computing the feature magnitude (FM) of X that captures the variation of appearance and motion cues in the spatio-temporal feature space. Formally, $FM(X) = \sum\|X\|_2$. So, the proposed selection strategy starts by computing FM for all k followed by arranging them in descending order and then keeps top k^s to be used in relation modeling, as computed by $TS(F_{Tr}) = \max(\sum_{j=1}^{j=k}\|F_{Trj}\|_2)$. With the top k^s tracklets where $FM(k^1) \leq FM(k^2) \leq \ldots \leq FM(k^s)$, a LSTM cell with n_h hidden neurons is used to output a fixed order dependency encoded feature map F^*_{Tr} as a relation modeling among the selected tracklets.

3.2.5. **Tracklet Ranker:**

The tracklet ranker is identical to the scene ranker which inputs $F^*_{Tr} \in \mathbb{R}^{T\times k^s\times n}$ to assign anomaly scores to each tracklet present in temporal segments. The difference lies in obtaining the temporal detection score map D_{Tr} of dimension $T\times 1$ which is computed by applying a `max-pooling` operator over k^s.

3.3. Soft-Selection Coupler

In order to effectively detect both human and scene centric anomalies by selecting either HsN or SsN, a Soft selection coupler (SSC) is proposed in HSN. The SSC shown in Fig. 5 computes two selection factors *i.e.* S_{HsN} and S_{SsN} by inputting the intermediate representations of tracklet ranker $F_T \in \mathbb{R}^{T\times k^s\times m}$ and Scene ranker $F_S \in \mathbb{R}^{T\times m}$, where T is the number of temporal segments, k^s is the number of tracklets, and m is the channel size. Essentially, SSC has three blocks: (i) segment-level, (ii) video-level, and (iii) final selection block to output S_{HsN} and S_{SsN}. The *segment-level selection* block computes two attention weights *i.e.* $A^S_{HsN} \in \mathbb{R}^{T\times 1}$ and $A^S_{SsN} \in \mathbb{R}^{T\times 1}$ for each temporal segment T_i signifying the weighted association of each T_i to HsN and SsN. In this block, a `max-pooling` operator is first applied over k^s dimension of F_T to identically match with F_S dimension. The output of max-pooling operation is denoted by F^M_T. Then, F^M_T and F_S are projected to two parallel `FC` layers followed by a *ReLu* activation for latent space representation. This dissociative latent representations are coupled by a `concatenate` operator. The coupled representation is then applied to two parallel `sigmoid` activated `FC` layers, each having one unit to compute A^S_{HsN} and A^S_{SsN} in a mutually exclusive manner. The segment-level selection block outputs A^S_{HsN} and A^S_{SsN} by learning the fine-grained features of a video.

Dissimilar to this, a *video-level selection* block encodes the association of a video by computing two attention weights *i.e.* $A^V_{HsN} \in \mathbb{R}^{1\times 1}$ and $A^V_{SsN} \in \mathbb{R}^{1\times 1}$ from the contextual temporal representation. For this, temporal contextual initialization is done by applying a `average-pooling` operation over T to the F^M_T and F_S feature maps. Then the contextual representations are entangled by a `concatenate` operation. This feature map is then projected to two parallel `sigmoid` activated `FC` layers, each having one units to compute A^V_{HsN} and A^V_{SsN} in a mutually exclusive manner. The output from segment-level and video-level selection blocks are masked in the *final selection* block to obtain $S_{HsN} \in \mathbb{R}^{T\times 1}$ and $S_{SsN} \in \mathbb{R}^{T\times 1}$. For masking, the A^V_{HsN} and A^V_{SsN} are first *inflated* across T to match the dimension of A^S_{HsN} and A^S_{SsN} and then the corresponding weights are combined by a *Hadamard product*.

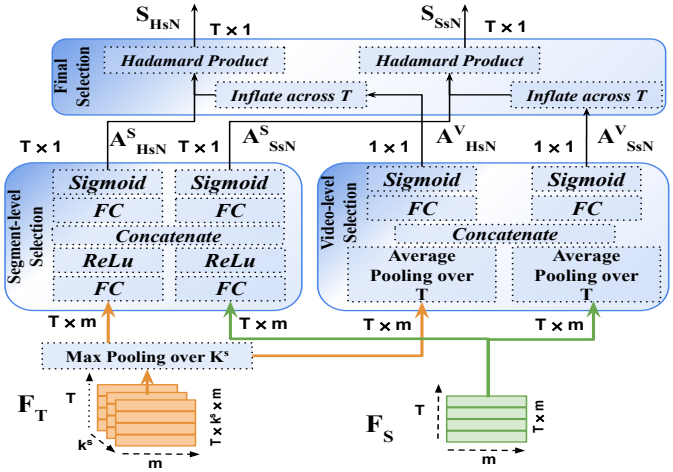

Fig. 5. Soft-Selection Coupler (SSC): It learns two selection factors S_{HsN} and S_{SsN} to couple the local and global representations learned by HsN and SsN respectively.

Coupled Detection Score: The final detection score (D) of HSN is computed by coupling the selection factors S_{HsN} and S_{SsN} with the human and scene subNet detection scores *i.e* D_{Tr} and D_{Sc} respectively. The coupling is done by multiplying the selection factors with the corresponding detection scores and then the results are added to obtain D. Formally, $D = (S_{HsN} \times D_{Tr}) + (S_{SsN} \times D_{Sc})$.

3.4. Optimization of Human-Scene Network

The proposed human-scene network (HSN) is end-to-end trainable (excluding the local and global feature extractors) with a novel self-rectifying loss. Although HSN follows multiple instance learning (MIL) optimization paradigm similar to

earlier WSVAD works (Sultani et al., 2018), the error computation procedure differs significantly. In this, due to the unavailability of precise temporal annotation for each video, the error is computed by considering two bags of instances, namely D_a and D_n. D_a and D_n are collection of detection scores (D) corresponding to the temporal segments extracted from anomaly and normal video sequences respectively.

Self-rectifying Loss: It takes inspiration from deep metric learning methods (Chopra et al., 2005; Schroff et al., 2015) to ensure a separable margin between normal and anomaly instances. Unlike (Chopra et al., 2005; Schroff et al., 2015) it considers detection scores D_a and D_n instead of feature distance measure to compute the loss. Futher, it is designed in such a way that it not only ensures context level but also performs instance level score maximization by generating a pseudo temporal annotation for each segment. The aim of pseudo label generation is to choose the correct number of anomaly and normal instances to take part in the optimization. The pseudo temporal annotations are computed during each iteration of optimization (*i.e. one step*) to avoid two-step optimization paradigm (Zhong et al., 2019; Feng et al., 2021) (*where, first step generates pseudo annotations, second step performs separability maximization*). Since our loss function is based on one-step optimization and since the pseudo annotations computed at the initial iteration can be noisy, a *self-rectification* mechanism is also proposed to refine the labels during the optimization. The proposed self-rectifying loss is presented in equation (1), which is a weighted sum of two components L_C and L_I to perform context and instance level maximization respectively.

L_C first computes the `sum` of scores from all T instances present in D_a and D_n to obtain video context and then maximizes the separation among them to ensure context separability. With only L_C, it is not sufficient to ensure optimal separability between normal and anomaly classes, since video containing short anomalies will be suppressed by neighboring normal scores and hence can not be effectively optimized.

$$L_{SR}(D_a, D_n) = \underbrace{\lambda_1 \max\left(0, 1 - \sum_{i=1}^{T}(D_a^i) + \sum_{i=1}^{T}(D_n^i)\right)}_{L_C} + \underbrace{\lambda_2 \|Err(Correct) - Err(Noisy)\|}_{L_I} \tag{1}$$

$$Err(X) = \begin{cases} \underbrace{\dfrac{1}{T}\sum_{i=1}^{T}(D_n^i - P_{y_n}^i)^2, \quad \forall i,\ P_{y_n}^i = 0}_{MSE(D_n)} & \text{if } X = Correct \\ \underbrace{\dfrac{1}{T}\sum_{i=1}^{T}(D_a^i - P_{y_a}^i)^2, \quad \forall i,\ \text{if } D_a^i \le D_{ref} \text{ then } P_{y_a}^i = 0, \quad \forall i,\ \text{if } D_a^i > D_{ref} \text{ then } P_{y_a}^i = 1}_{MSE(D_a)} & \text{if } X = Noisy \end{cases} \tag{2}$$

For this, L_I performs the instance-level optimization in D_a and D_n by generating pseudo-temporal labels which are empowered by a self-rectification mechanism. Figure 7 shows

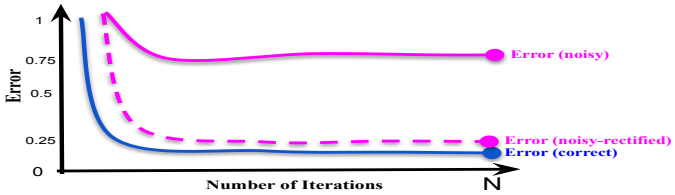

Fig. 6. Visualization of typical convergence plot for correct (*Blue*), noisy (*Pink*) and an ideal noisy-rectified (*Pink-Dotted*) pseudo labels based optimization.

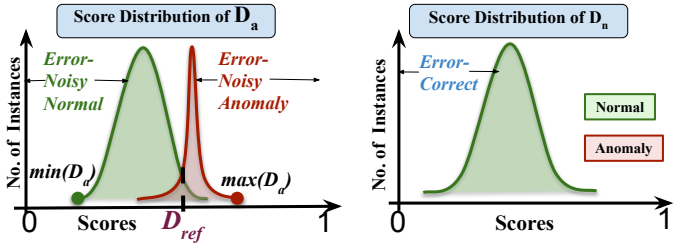

Fig. 7. A typical plot to show the score distributions of D_a and D_n.

a typical instance-level score distribution of D_a and D_n in a weakly-supervised setting. Since D_n contains no anomaly instances, the pseudo temporal label for segment i (Py_n^i) for D_n is `correct` (*i.e.* $\forall i, Py_n^i = 0$. Note that, we refer label = 0 for normal and 1 for anomaly instances. On the other hand, D_a contains a mixed distribution (*i.e.* normal and anomaly) with no prior knowledge, so the pseudo temporal labels for D_a (*i.e.* Py_a^i) are `noisy`. In D_a, the pseudo temporal labels Py_a^i are computed by comparing their prediction scores (D_a^i) to a dynamic reference point (D_{ref}). The D_{ref} for D_a is calculated by taking the average between the maximum and minimum score of D_a, $D_{ref} = (\max(D_a) + \min(D_a))/2$, which essentially provides a dynamic threshold point for each video characterizing different categories of anomalies. The quality of Py_a computation depends on the choice of an appropriate D_{ref}, which may be a rough approximation at the initial iteration. Hence, for optimal instance-level separation, it is necessary to rectify the noisy pseudo temporal labels (*i.e.* Py_a) with the help of the correct ones (*i.e.* Py_n). Figure 6 shows a typical convergence plot for `correct` and `noisy` pseudo label based optimization. Further, it shows an ideal convergence plot after the rectification is made to the noisy (*i.e.* `noisy-rectified`) pseudo labels. So, to rectify the noisy distribution, L_I minimizes the difference between the errors (Err) obtained from the correct pseudo labels and the noisy pseudo labels. It can be observed that $Err(Noisy)$ has a direct dependency with the D_{ref} point. So, by minimizing $\|Err(Correct) - Err(Noisy)\|$, the D_{ref} gets adjusted and reaches an optimal point where Py_a has a minimum noise. This self-rectification procedure enables the noisy pseudo labels to get rectified with the guidance of correct labels. Note that, for computing Err, we adopt `mean-squared-error (MSE)` between the predicted scores and their corresponding pseudo labels as portrayed in equation (2).

4. Experiments

4.1. Datasets

The experiments are conducted on three anomaly detection datasets, namely, UCF-Crime, ShanghaiTech, and IITB-Corridor as described below.

UCF-Crime (Sultani et al., 2018) : It is a diverse and large-scale dataset containing 1900 real-world surveillance videos from 13 types of anomaly activities. In this dataset anomaly activities may occur for long or short durations, which makes the detection problem challenging. It has 1610 videos for training, out of which 810 and 800 videos belong to anomaly and normal classes respectively. Similarly, for testing there are 290 videos containing 140 anomaly and 150 normal videos.

ShanghaiTech (Lu et al., 2013b): It is a medium scale dataset recorded in a university campus. Originally this dataset was designed for unsupervised anomaly detection, but we train-test protocol designed by (Zhong et al., 2019) for weakly-supervised settings. This dataset contains 437 videos in total out of which 175 normal and 65 anomaly videos are considered for training and 155 normal and 44 anomaly videos for testing.

IITB-Corridor (Rodrigues et al., 2020): It is also a medium scale dataset recorded in a corridor of IIT Bombay campus. It contains a total of 358 videos in standard protocol (Rodrigues et al., 2020) where 208 videos in the training set are normal videos only and the test set contains 10 normal videos 140 anomaly videos. This standard setting considering only normal videos in train set is not suitable for WSVAD. Thus, we reorganize the dataset to meet the requirement of WSVAD. The new training split contains both normal and anomaly classes by randomly moving 71 anomaly videos from standard test set to new anomaly class of train set followed by 147 normal videos from standard train set to new normal class of train set. Similarly, the new test split is a collection of 69 remaining anomaly videos of standard test set and 71 normal videos from standard train, test set. In summary, the new training and testing split contains 218 and 140 videos respectively.

4.2. Evaluation Metric

Following (Sultani et al., 2018; Tian et al., 2021), frame-level Receiver Operating Characteristics (ROC) and its corresponding Area Under the Curve (AUC) are used to evaluate the anomaly detection performance. For UCF-Crime dataset, category wise detection performance is also computed on both official test split (Sultani et al., 2018) and whole dataset (*i.e. using 5-fold cross validation*) to evaluate the robustness of detection performance in various critical situations. For the 5-fold cross-validation, we report the mean-AUC (mAUC) of all 5-folds to evaluate the method. Since for the 5-fold evaluation, we need the temporal annotation of the complete UCF-Crime dataset, we obtained it from (Wan et al., 2021).

4.3. Implementation Details

Temporal segment (T) division: Following earlier works (Sultani et al., 2018; Tian et al., 2021) for a fair comparison, we follow their setting and fixed the temporal segment (T) division $T = 32$ for training and testing videos. Although the scene subNet divides the video into multiple temporal granularities *i.e.* $G = 1, 2, 3$ to obtain T, $2T$, $3T$ temporal segments, the MGTM module decomposes the temporal dimension to T in the subsequent stages to be utilized further. **Global and local feature extraction**: In both global and local feature extraction, we use two variants of I3D (Carreira and Zisserman, 2017) backbone *i.e.* with InceptionV1 (I3D-Inc) (Szegedy et al., 2015) and with ResNet50 (I3D-Res) (He et al., 2016) which are pre-trained on the Kinetics-400 (Kay et al., 2017) dataset. For I3D-Inc and I3D-Res, we extract features from the *global_pool* and *mix_5c* layers yielding feature vectors of dimension 1024D and 2048D respectively. **MGTM block**: In MGTM, the *temporal downscaler block* has two sequential 1D convolution layers with kernel size $k \in \{5, 3\}$ and dilation rate $d \in \{4, 8\}$ respectively. The *bottleneck block* has one layer of 1D convolution with with kernel size $k = 1$ and and dilation rate $d = 1$. The number of hidden neurons($= n_h$) of LSTM cell used in MGTM is set to 256. **TSRM block**: For number of tracklets selection, we set the value of k^s to 10, 7, 7 in UCF-Crime, ShanghaiTech and IITB-Corridor datasets respectively. The number of hidden neurons($= n_h$) of LSTM cell used in TSRM block is set to 1024. **Scene and tracklet ranker block**: The number of neurons in three FC-layers of both scene and tracklet ranker are set to 96,32 and 1 respectively. **SSC block**: In soft-selection coupler (SSC) the value of m is set to 96 (*i.e.* the number of neurons in intermediate layer of scene and tracklet ranker) and also the number of neurons in initial FC layers of segment-level selection is set to 96. For the final FC layers of both segment-level and video level selection the number of neuron is set to 1. **HSN training**: The HSN is trained using Adam optimizer at a learning rate 0.0001 and with the loss weighting factors $\lambda_1 = \lambda_2 = 0.5$. We also randomly select 30 anomaly and 30 normal videos as a mini-batch and compute the gradient using reverse mode automatic differentiation on computation graph using Tensorflow. Then the loss is computed and back-propagated for the whole batch. For faster convergence of the HSN, we first train the human and scene subNet independently and with the pre-trained weights then we re-train whole network along with the soft-selection coupler. For each training, we use the self-rectifying loss for error computation. In UCF-Crime, ShanghaiTech and IITB-Corridor datasets we train up to 1500, 800, 1050 epochs respectively.

4.4. Ablation Study

A detailed and sequential ablation study is carried out in this section to quantify the two novel contributions *i.e.* human-scene network (HSN) and self-rectifying loss. For all ablation studies, UCF-Crime (Sultani et al., 2018) dataset is chosen as it has a good number of human and scene based anomalies.

Effectiveness of HSN : As HSN comprises of multiple building blocks, each block is evaluated in terms of anomaly detection performance (AUC) in the official test-split as shown in Table 1. At first, human and scene subNets are independently considered for experimentation. The temporal segment generation blocks, feature descriptors (*i.e.* I3D-ResNet50), and ranker block are inherently added to both subNets for determining detection performances. In human subNet, the human baseline experiment is performed with only the tracklets obtained from the people detection and tracking (PDT) method. Subsequently, by adding tracklet selection and relation modeling (TSRM) block to human subNet, it boosts the performance (+4.68%) significantly compared to the human baseline. Similarly, in scene subNet, the global scene (GS) is only taken at the beginning to define the scene baseline. On top of that, adding a multi-granularity

Human subNet.		Scene subNet.		SSC		AUC(%)
PDT	TSRM	GS	MGTM	SLS	VLS	
✓	✗	✗	✗	✗	✗	76.53
✓	✓	✗	✗	✗	✗	81.21
✗	✗	✓	✗	✗	✗	78.62
✗	✗	✓	✓	✗	✗	83.78
✓	✓	✓	✓	✓	✗	84.52
✓	✓	✓	✓	✓	✓	85.45

Table 1. Ablation on each component of HSN on UCF-Crime dataset in terms of AUC(%). Here, SLS:Segment-level selection, VLS:Video-level selection block of SSC.

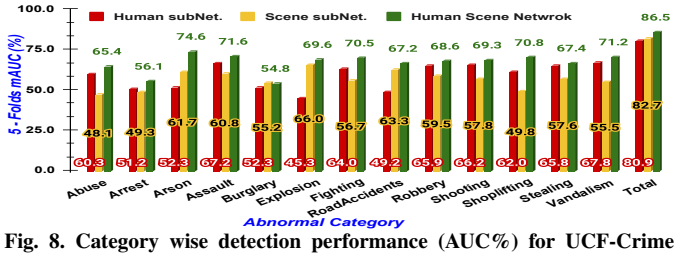

Fig. 8. Category wise detection performance (AUC%) for UCF-Crime dataset quantifying effectiveness of HSN.

temporal modeling (MGTM) block improves the detection performance (+5.16%) by a large margin. The performance boost for both human and scene subNets outlines the potentiality and significance of TSRM and MGTM blocks respectively.

Further, in order to verify the complementary nature of the representations learned by both subNets, we visualize the category wise detection performance (mAUC) by covering the entire UCF-Crime dataset with 5-folds evaluation, as reported in Fig. 8. For human-centric anomaly categories such as *abuse, arrest, assault, fighting, robbery, stealing, shooting, shopliftng, and vandalism* the performance of the human subNet is superior to the scene subNet and vice versa for the scene-centric anomaly category like *arson, explosion, road accident*. This portrays the necessity of both human and scene subNets to detect anomalies pertaining to respective categories. Then soft selection coupler (SSC) is introduced in HSN to take benefits from both subNets and to improve the overall detection performance. Table 1 shows that with only the segment-level selection (*i.e.* A^S_{HsN} and A^S_{SsN}) of SSC, the detection performance is improved by 3.31% and 0.74% compared to individual subNets in official test split. Finally by combining both segment-level and video-level selections in SSC (*i.e.* S_{HsN} and S_{SsN}), it further boosts the detection performance by 4.24% and 1.67% compared to that of only human and scene subNet respectively. From Fig. 8, it can be inferred that thanks to the decoupled design of HSN, distinctive features are learned for human and scene based anomalies and the SSC effectively combines the complementary representations to boost the overall performance.

Effectiveness of L_{SR}: In order to gain enhanced performance, optimization with L_{SR} plays a key role in HSN. To corroborate the effectiveness and adaptability of L_{SR}, an experimental ablation study is performed as shown in Table 2. We choose two recently reported MIL based methods *i.e.* (Sultani et al., 2018) and (Majhi et al., 2021b) for comparison. These methods were previously optimized with classical ranking loss (L). Optimizing the (Sultani et al., 2018) and (Majhi et al., 2021b) methods with L_{SR} gives 0.97% and 1.39% performance gain

Method	Optimization		AUC(%)
	Sultani et al. (2018)	L_{SR} [eq 1]	
Sultani et al. (2018)	✓	✗	77.42
	✗	✓	78.39
Majhi et al. (2021b)	✓	✗	81.88
	✗	✓	83.27
HSN	✓	✗	82.76
	✗	✓	85.45

Table 2. Ablation to show the AUC gain by L_{SR} loss in various methods evaluated on UCF-Crime dataset.

k^s	N-3	N-2	N-1	N	N+1	N+2	N+3
AUC (%)	82.56	83.61	84.72	**85.45**	85.21	85.18	85.16

Table 3. Ablation to show the impact of k^s tracklet selected in overall performance of UCF-Crime dataset.

respectively. Similarly, HSN optimized with L_{SR} boosts the detection performance by 2.69% compared to L. Since previous methods operate on coarse spatio-temporal features which have lower separabilities compared to HSN representation, the performance boost by L_{SR} is marginal in those methods. But in HSN, a superior discriminative representation is captured due to the decoupling of scene and human cues and hence it generates less noisy pseudo labels leading to greater separable features.

Impact of Tracking and Tracklet Selection: As our method captures human centric cues for anomaly detection, it relies on good quality people detection and tracking (PDT). Since any PDT can be configured to our method in a plug-and-play manner, we experiment with several PDTs and found that *ByteTrack* is effective and fast. Further, our method is resilient to human ID switches thanks to the TSRM module which selects the salient abnormal tracklets irrespective of their IDs. In TSRM, the number of tracklet selection (k^s) is a crucial hyperparameter that effects the overall performance. Hence, we first initialize k^s to $\lceil N \rceil$, where N = average number of tracks in the whole UCF-C dataset. Then, we linearly increase and decrease k^s to study its effect on the overall performance and found that it performs best when $k^s = \lceil N \rceil$ ($N = 10$ for UCF-C) as shown in Table 3.

Experiments on λ_1 and λ_2 : The impact of loss weighting factors λ_1 and λ_2 is shown in Table 4 and the best performance is achieved for equal probability weighting, where $\lambda_2 = 1 - \lambda_1 = 0.5$.

λ_1	0.3	0.4	0.5	0.6	0.7
AUC (%)	80.64	83.92	**85.45**	82.51	81.92

Table 4. Ablation to show the impact of loss weighting factors λ_1 and λ_2 on the overall performance of UCF-Crime dataset.

Datasets	UCF-Crime	ShanghaiTech	IITB-Corridor
D_{ref}	0.43	0.68	0.65

Table 5. Ablation to show the variability of D_{ref} across different datasets.

Variability of D_{ref} across datasets: Since the D_{ref} point is learnable automatically, it varies between 0.43 to 0.68 across three datasets considered (shown in Table 5). The learned D_{ref} point has low variability for similar data distributions like ShanghaiTech, IITB-Corridor and has moderate variability for diverse datasets like UCF-Crime. Since the UCF-Crime has both complex and diverse anomaly distribution with subtle and sharp motion cues, the D_{ref} is lower to encourage subtle abnormal temporal segments (that have lower detection scores D) that can take part in the optimization. However, the Shang-

Methods	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	RoadAcc.	Robbery	Shooting	Shoplifting	Stealing	Vandalism	Total
Official Test-Evaluation (Sultani et al., 2018) (AUC%).														
Tian et al. (2021)	55.93	59.16	65.32	70.71	70.11	45.28	70.02	55.94	69.84	73.63	70.8	75.17	64.31	84.30
Ours	59.21	60.02	67.86	91.95	62.19	44.57	68.36	47.01	72.37	71.63	75.78	71.18	43.26	85.45
5-Fold Evaluation.(mean-AUC%)														
Tian et al. (2021)	56.3	54.6	66.8	69.4	51.2	69.8	62.4	69.9	59.2	63.2	59.1	61.2	65.9	82.7
Ours	65.4	56.1	74.6	71.6	54.8	69.6	70.5	67.2	68.6	69.3	70.8	67.4	71.2	86.5

Table 6. Category wise performance comparison with a Tian *et al.* to show the AUC gain by our method in various categories of UCF-Crime dataset. We made the comparison in both official test split (Sultani et al., 2018) and 5-fold evaluation to justify the robustness of our method.

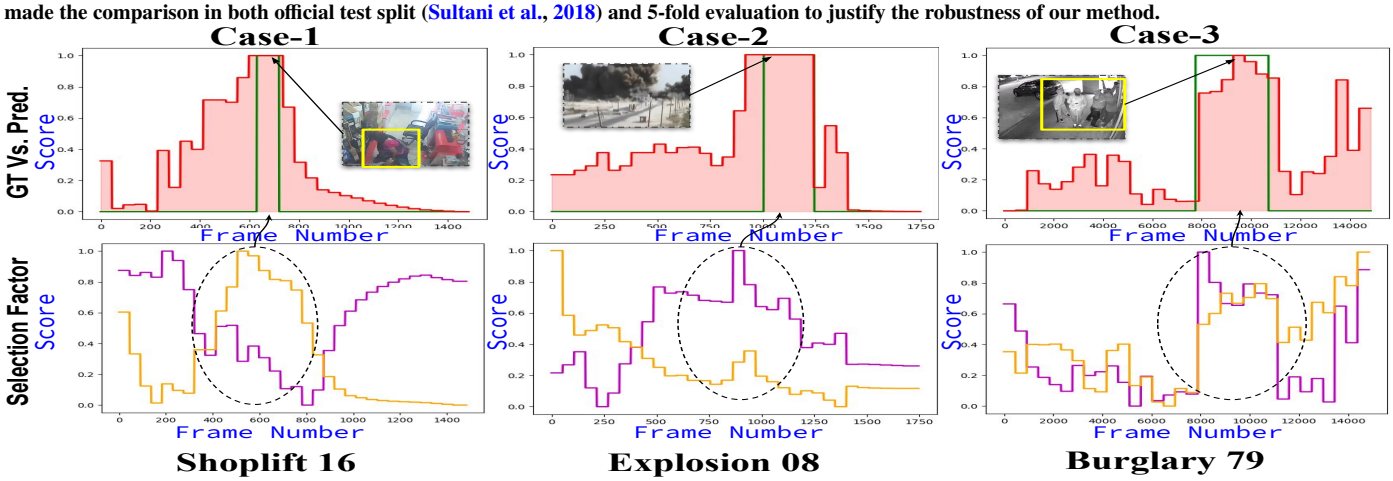

Fig. 9. Visualization of Ground truth (green shed) vs. prediction scores (red shed) for various cases in Row-1. Row-2 is the selection factors S_{HsN} (yellow plot) and S_{SsN} (violet plot) to focus on either HsN or SsN for effective coupling.

Methods	Feature	UCF-C	ST	IITB-C
		AUC(%)		
Sultani et al. (2018)	C3D	75.41	-	-
	I3D-Inc	77.42	80.02	74.59
Zhang et al. (2019)	C3D	78.66	-	-
Zhu and Newsam (2019)	C3D	79.00	-	-
Zhong et al. (2019)	C3D	81.08	76.44	-
	TSN-Inc	82.12	84.44	-
Feng et al. (2021)	I3D-Inc	82.30	-	-
Majhi et al. (2021a)	I3D-Inc	82.12	-	-
Wu et al. (2020)	I3D-Inc	82.44	85.38	-
Majhi et al. (2021b)	I3D-Inc	82.67	88.86	80.31
Tian et al. (2021)	I3D-Res	84.30	**97.21**	81.12*
Purwanto et al. (2021)	TRN	85.00	96.85	-
Lv et al. (2021)	TSN	85.38	-	-
Zhang et al. (2022a)	I3D-Res	83.17	97.62	-
Sapkota and Yu (2022)	I3D-Res	83.39	96.00	-
Li et al. (2022)	Video-SWIN	85.30	-	-
Chen et al. (2023a)	I3D-Res+Caption	84.90	-	-
Fan et al. (2024)	I3D-Res	84.05	95.00	-
Our	**I3D-Inc**	**84.33**	**93.72**	**84.12**
	I3D-Res	**85.45**	96.22	**86.98**

Table 7. State-of-the-art performance comparisons in terms of frame-level AUC on UCF-Crime (UCF-C), ShanghaiTech (ST), and IITB-Corridor (IITB-C) dataset. Note: * marked AUC is our implementation.

haiTech and IITB-Corridor datasets are small-scale with fewer diversities in anomalies, and thus, the D_{ref} is relatively higher to suppress the noisy temporal segments.

4.5. Qualitative Analysis

In Fig. 9, we show the prediction scores (Row-1) obtained from proposed method along with the selection factors (Row-2) corresponding to human and scene subNets (*i.e.* HsN and SsN) in three cases **(case-1: human centric, case-2: scene centric, case-3: both human and scene)** of anomaly scenarios. The samples "Shoplifting16", "Explosion08", and "Burglary79" give an overview of the performance for all three cases. In "Shoplifting16" where the anomaly is done by a woman, the effective detection score is majorly influenced by HsN since higher selection factor to HsN is assigned to the localized area. Similarly, for "Explosion08" where a bomb blast is recorded, we see that the detection performance is triggered due to a higher factor assigned to SsN. From both anomaly cases, the proposed method gives superior detection performance by effectively selecting either subNet. Moreover, we choose a third case where the anomaly is characterized by both human and scene localized areas. "Burglary79" is such a scenario where a group of criminals are illegally entering to a building by breaking the entrance.

4.6. State-of-the-art Comparison

Overall Performance Comparison : In Table 7, we compare the overall performance of the proposed method with recently reported state-of-the-art methods for UCF-Crime, ShanghaiTech, and IITB-Corridor datasets. For a fair comparison, as several feature extractor backbones are used to report AUC in previous methods, we report the results using two widely used backbones *i.e.* Inception-v1 I3D (I3D-Inc) and ResNet50 I3D (I3D-Res) for the three datasets. In UCF-Crime, our method outperforms the state-of-the-art I3D-Inc and I3D-Res based methods of (Majhi et al., 2021b) and (Tian et al., 2021) by +1.66% and +1.12% margin respectively. Further, as UCF-Crime contains complex human and scene-based anomalies, we consider existing **local-global object-centric** feature combination method (Purwanto et al., 2021) for comparison and found that our method improves the performance with I3D-Res feature backbone. Similarly, to precisely detect long-short

Methods	FLOPs(G)	Speed(FPS)
PDT (Zhang et al., 2022b)	281.9	30
Majhi et al. (2021b)(I3D-Inc)	108.1	267
Tian et al. (2021)(I3D-Res)	153.2	211
HSN(I3D-Inc) w-o PDT	108.7	227
HSN(I3D-Res) w-o PDT	153.7	159

Table 8. Complexity comparison of Human-Scene Network. Here, G: Giga, FPS: Frames-per-second, w-o: with-out.

length anomalies, we also consider **temporal modeling** based competitive methods (Zhang et al., 2022a; Tian et al., 2021; Li et al., 2022; Chen et al., 2023a; Fan et al., 2024) for comparison and found that our HSN is better than previous methods in complex UCF-Crime dataset. For ShanghaiTech dataset, although our method outperforms the recent I3D-Inc based method of (Majhi et al., 2021b) by +4.86%, it fails to achieve better performance than (Tian et al., 2021) with I3D-Res backbone. This is due to ShanghaiTech, that has only 65 anomaly videos collected from a small and focused data distribution for training. It contains only simple human anomalies (e.g. run, fall down, ride cycle etc.) characterized by strong motion, so only few events are not correctly detected. Due to the small number of anomaly samples, our HSN could not be sufficiently optimized to outperform methods dedicated to strong motion events. Further, to confirm the robustness of our method in an additional anomaly distribution, we have also validated our method on the IITB-Corridor dataset. In this, it surpasses the recent I3D-Inc method (Majhi et al., 2021b) and I3D-Res method (Tian et al., 2021) by +3.81% and +5.86% margin respectively.

Category Wise Performance Comparison : Further, we also compare the abnormal category wise performance with (Tian et al., 2021) in UCF-Crime dataset. First, we made the comparison in the official test split (Sultani et al., 2018) and found that our method is superior in detecting human-centric abnormal categories like *abuse, arrest, arson, assault, robbery, shoplifting* compared to (Tian et al., 2021) as reported in Table 6. However, the performance gain is not significant. This is due to the fewer number of samples in the official test set. For this, we perform the 5-fold cross-validation and report the mean-AUC (mAUC) of 5-folds for categories-wise performance comparison. Since it covers the entire UCF-Crime dataset, we believe it provides a more robust and justified performance w.r.t. the official test set (Sultani et al., 2018). From the 5-fold cross-validation, we found that our method outperforms (Tian et al., 2021) in all fine-grained and subtle human-centric anomalies (*abuse, arrest, arson. assault, fighting, robbery, shooting, shoplifting, stealing*) by a significant margin as reported in Table 6. As a result, our method surpasses (Tian et al., 2021) by 3.8% margin in total performance in K-fold evaluation. But, our method lies behind (Tian et al., 2021) on simple scene based anomalies (*explosion, road accidents*) where there exists a sharp change in the scene.

4.7. Network Complexity Analysis

In this section, a complexity analysis of HSN is performed to meet real-world applicability. Since HSN relies on people detection and tracking (PDT) method, the complexity of the PDT method (*ByteTrack* (Zhang et al., 2022b)) is evaluated first, and then the complexity of anomaly detection methods are reported and compared w.r.t FLOPs, and speed as shown in Table 8. It can be seen that our HSN is computationally competitive in terms of FLOPs and speed w.r.t. recently reported articles. For a fair comparison, evaluation is done on a single 2080Ti GPU. Considering recent CCTV cameras that operate in 30 FPS, HSN can detect anomalies effectively in near real-time.

5. Conclusion

In this work, we presented a novel human scene network (HSN) optimized by a self-rectifying loss function as a baseline to detect real-world anomalies under weak-supervision. The HSN ensures superior detection performance in complex real world scenarios for two reasons: First, the decoupled design of HSN comprising Human and Scene subNets followed by a soft-selection coupler can effectively learn local and global discriminative representations for human and scene-centric anomalies, respectively. Second, optimizing HSN with the proposed self-rectifying loss ensures a greater separability between the classes. From experimentation, it can be noted that the proposed method gains competitive performance in five out of six scenarios considered compared to the recently reported methods for three popular datasets.

Limitation-and-Contribution Trade-off : As HSN relies on people detection and tracking method for learning human centric representation, it induces slightly more complexity compared to the previous methods. However, it opens up new directions to analyse the complex abnormal scenarios in a more fine-grained way to address the real-world challenges which is majorly missing in earlier weakly-supervised methods. From the performance and complexity analysis prospective, our method establish a good trade off to meet the real-world applicability.

Acknowledgments

This work was supported by Toyota Motor Europe (TME) and the French government, through the 3IA Cote d'Azur Investments with the reference number ANR-19-P3IA-0002.

References

Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D., 2008. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE transactions on pattern analysis and machine intelligence 30, 555–560.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, W., Ma, K.T., Yew, Z.J., Hur, M., Khoo, D.A.A., 2023a. Tevad: Improved video anomaly detection with captions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5548–5558.

Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., Wu, Y.C., 2023b. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 387–395.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE. pp. 539–546.

Cong, Y., Yuan, J., Liu, J., 2013. Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition 46, 1851–1864.

Fan, Y., Yu, Y., Lu, W., Han, Y., 2024. Weakly-supervised video anomaly detection with snippet anomalous attention. IEEE Transactions on Circuits and Systems for Video Technology .

Feng, J.C., Hong, F.T., Zheng, W.S., 2021. Mist: Multiple instance self-training framework for video anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14009–14018.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S., 2016. Learning temporal regularity in video sequences, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hinami, R., Mei, T., Satoh, S., 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge, in: Proceedings of the IEEE international conference on computer vision, pp. 3619–3627.

Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L., 2019. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7842–7851.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al., 2017. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 .

Kim, J., Grauman, K., 2009. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2921–2928.

Lea, C., Vidal, R., Reiter, A., Hager, G.D., 2016. Temporal convolutional networks: A unified approach to action segmentation, in: European conference on computer vision, Springer. pp. 47–54.

Li, S., Liu, F., Jiao, L., 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1395–1403.

Lin, S., Yang, H., Tang, X., Shi, T., Chen, L., 2019. Social mil: Interaction-aware for crowd anomaly detection, in: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE. pp. 1–8.

Liu, W., Luo, W., Lian, D., Gao, S., 2018. Future frame prediction for anomaly detection–a new baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6536–6545.

Liu, Y., Liu, J., Yang, K., Ju, B., Liu, S., Wang, Y., Yang, D., Sun, P., Song, L., 2023. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. IEEE Transactions on Industrial Informatics .

Lu, C., Shi, J., Jia, J., 2013a. Abnormal event detection at 150 fps in matlab, in: Proceedings of the IEEE international conference on computer vision, pp. 2720–2727.

Lu, C., Shi, J., Jia, J., 2013b. Abnormal event detection at 150 fps in matlab, in: Proceedings of the IEEE international conference on computer vision, pp. 2720–2727.

Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J., 2021. Localizing anomalies from weakly-labeled videos. IEEE transactions on image processing 30, 4505–4515.

Majhi, S., Das, S., Brémond, F., Dash, R., Sa, P.K., 2021a. Weakly-supervised joint anomaly detection and classification, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE. pp. 1–7.

Majhi, S., Das, S., Brémond, F., 2021b. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection, in: 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. doi:10.1109/AVSS52988.2021.9663810.

Purwanto, D., Chen, Y.T., Fang, W.H., 2021. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 173–183.

Ramachandra, B., Jones, M., 2020. Street scene: A new dataset and evaluation protocol for video anomaly detection, in: The IEEE Winter Conference on Applications of Computer Vision, pp. 2569–2578.

Rodrigues, R., Bhargava, N., Velmurugan, R., Chaudhuri, S., 2020. Multi-timescale trajectory prediction for abnormal human activity detection, in: The IEEE Winter Conference on Applications of Computer Vision (WACV).

Roy, P.R., Bilodeau, G.A., Seoud, L., 2021. Local anomaly detection in videos using object-centric adversarial learning, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV, Springer. pp. 219–234.

Sapkota, H., Yu, Q., 2022. Bayesian nonparametric submodular video partition for robust anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3212–3221.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823.

Sultani, W., Chen, C., Shah, M., 2018. Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488.

Sun, C., Jia, Y., Hu, Y., Wu, Y., 2020. Scene-aware context reasoning for unsupervised abnormal event detection in videos, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 184–192.

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

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G., 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4975–4986.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: The IEEE International Conference on Computer Vision (ICCV).

Wan, B., Fang, Y., Xia, X., Mei, J., 2020. Weakly supervised video anomaly detection via center-guided discriminative learning, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1–6.

Wan, B., Jiang, W., Fang, Y., Luo, Z., Ding, G., 2021. Anomaly detection in video sequences: A benchmark and computational model. IET Image Processing 15, 3454–3465.

Wang, Jue Cherian, A., 2019. Gods: Generalized one-class discriminative subspaces for anomaly detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 8201–8211.

Wang, Z., Zou, Y., Zhang, Z., 2020. Cluster attention contrast for video anomaly detection, in: Proceedings of the 28th ACM international conference on multimedia, pp. 2463–2471.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z., 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: European Conference on Computer Vision, Springer. pp. 322–339.

Yu, J., Lee, Y., Yow, K.C., Jeon, M., Pedrycz, W., 2021. Abnormal event detection and localization via adversarial event prediction. IEEE Transactions on Neural Networks and Learning Systems .

Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I., 2022. Generative cooperative learning for unsupervised video anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14744–14754.

Zaheer, M.Z., Mahmood, A., Shin, H., Lee, S.I., 2020. A self-reasoning framework for anomaly detection using video-level labels. IEEE Signal Processing Letters 27, 1705–1709.

Zhang, D., Huang, C., Liu, C., Xu, Y., 2022a. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. IEEE Signal Processing Letters 29, 1197–1201.

Zhang, J., Qing, L., Miao, J., 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 4030–4034.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022b. Bytetrack: Multi-object tracking by associating every detection box, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII, Springer. pp. 1–21.

Zhao, B., Fei-Fei, L., Xing, E.P., 2011. Online detection of unusual events in videos via dynamic sparse coding, in: CVPR 2011, pp. 3313–3320. doi:10.1109/CVPR.2011.5995524.

Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G., 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhou, H., Yu, J., Yang, W., 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. arXiv preprint arXiv:2302.05160 .

Zhu, Y., Newsam, S., 2019. Motion-aware feature for improved video anomaly detection. arXiv preprint arXiv:1907.10211 .