## Supplementary Material

| Section | Content |
|---|---|
| 1 | Additional Ablation Study |
| 2 | Additional Qualitative Analysis |
| 3 | Added Discussion on Performance Comparison |
| 4 | Real-world Challenging Situation for Human subNet |
| 5 | Necessity of 5-Fold Test Evaluation |

**Table 1. Overview of Supplementary Material**

### 1. Additional Ablation Study

In this section, we provide more ablation studies of our method to showcase the need of soft-selection coupler (SSC) in entangling the feature representation learned by scene and human subNets. Since our human-scene network (HSN) dissociates the local and global feature representations to capture human and scene centric spatio-temporal cues, it is required to couple both representations to obtain overall anomaly detection score. Table 2 presents an experimental ablation study on various feature coupling techniques to combine the local and global representations learned in HSN. We start with conventional feature combination strategies *early and late fusion* and found that *late fusion* performs significantly better (+4.17%) than *early fusion*. In *early fusion*, we `concatenate` the feature map from multi-granularity temporal modeling (MGTM) block of scene subNet, tracklet selection and relation modeling (TSRM) block of human subNet to learn the fused feature in a MLP ranker model. In contrast, the *late fusion* is achieved by computing the `average` score obtained from scene and tracklet rankers of scene and human subNets. From *early and late fusion* experiments, we infer that feature combination at later stage of HSN is more meaningful than early stage. Further, we consider popular *cross-attention* Chen et al. (2021) mechanism to generate an enhanced feature map to be considered for overall anomaly detection. In the designed *cross-attention* block, the intermediate representations of tracklet and scene rankers are considered as `key` and `query` respectively. The `value` is considered by concatenating the intermediate representations of tracklet and scene rankers. Finally, the `key`, `query` and `value` are used to generate the enhanced feature map by highlighting the salient temporal segments for anomaly detection. From experimentation we find that, although the designed *cross-attention* block for feature combination is superior (+3.81%) to *early fusion* but it is worse (0.36%) than *late fusion*.

| Methods | AUC(%) |
|---|---|
| Early Fusion | 79.52 |
| Late Fusion | 83.69 |
| Cross-Attention Chen et al. (2021) | 83.33 |
| **Soft-selection Coupler** | **84.33** |

**Table 2. Experimental Ablation for various feature coupling techniques in HSN using I3D-Inc backbone on UCF-Crime dataset.**

From the above three experiments we find that, (i) representation coupling at the later stage of HSN has superiority than early stage, (ii) features from intermediate layer of ranker models can learn better attention weights for coupling. For this, we design a soft-selection coupler by considering the above findings. It combines the representations learned by human and Scene subNets at the score level by generating selection factors *i.e.* $S_{HsN}$ and $S_{SsN}$ respectively which are computed by considering the intermediate representations of tracklet and scene rankers. Thus, the soft-selection coupler achieves (+4.81%), (+0.64%) (+1.0%) performance boost compared to *early fusion*, *late fusion* and *cross-attention* mechanism respectively.

### 2. Additional Qualitative Analysis

In this section we present a detailed qualitative analysis of the proposed method as shown in Figure 1. The analysis is performed along two aspects *i.e.* correct and partially correct detection cases. For each case the plot "Ground truth Vs. prediction" is shown in row-1 and the corresponding selection factors, indicating on which human or scene subNet to focus, are shown in row-2. The prediction is obtained by applying a threshold on the prediction scores and the threshold value is chosen where it has maximum true positive rate and minimum false positive rate.

**Correct Cases :** To showcase the correct detection cases of HSN, we take two samples from UCF-Crime Sultani et al. (2018) dataset *i.e.*"Shooting18" and "Robbery102"videos. "Shooting18" portrays a typical real-world human-centric anomaly scenario where a person is shooting another one with a weapon as shown in the sample frame. In this case HSN detects the anomaly precisely by assigning a higher selection factor to human subNet. Similarly, in "Robbery102", two people try to rob a women as shown in the sample frames. HSN also correctly detects this human-centric anomaly by focusing on human-subNet. In this video, HSN failed at few temporal locations since the anomaly action is fully occluded by a trash bin as shown in the middle sample frame.

**Partially Correct Cases :**We define partially correct detection case when the model correctly detects the anomaly localized areas but produces some acceptable false positives. To showcase such cases of HSN, we find two samples from UCF-Crime Sultani et al. (2018) dataset *i.e.* "Shoplifting15" and "Arrest01" videos. Unlike earlier examples ("Shooting18" and "Robbery102") where anomalies are characterized by strong motion cue, the "Shoplifting15" video has a subtle motion. Despite, HSN detects the anomaly in "Shoplifting15" across precise temporal locations by assigning a high selection factor to human subNet. On top of that, in this video HSN also reports few false positives where the scenario is quite similar to an argument and fighting like situation which makes the prediction partially correct. Similarly in "Arrest01", we see that HSN is correctly able to detect the start of anomaly but it produces false positives even after the ground truth region. To further analyse the case, we visualize the video and find that the false positive region corresponds to the partially occluded arrest situation where some `body parts` of the abnormal human are partially visible as shown in the sample frame.

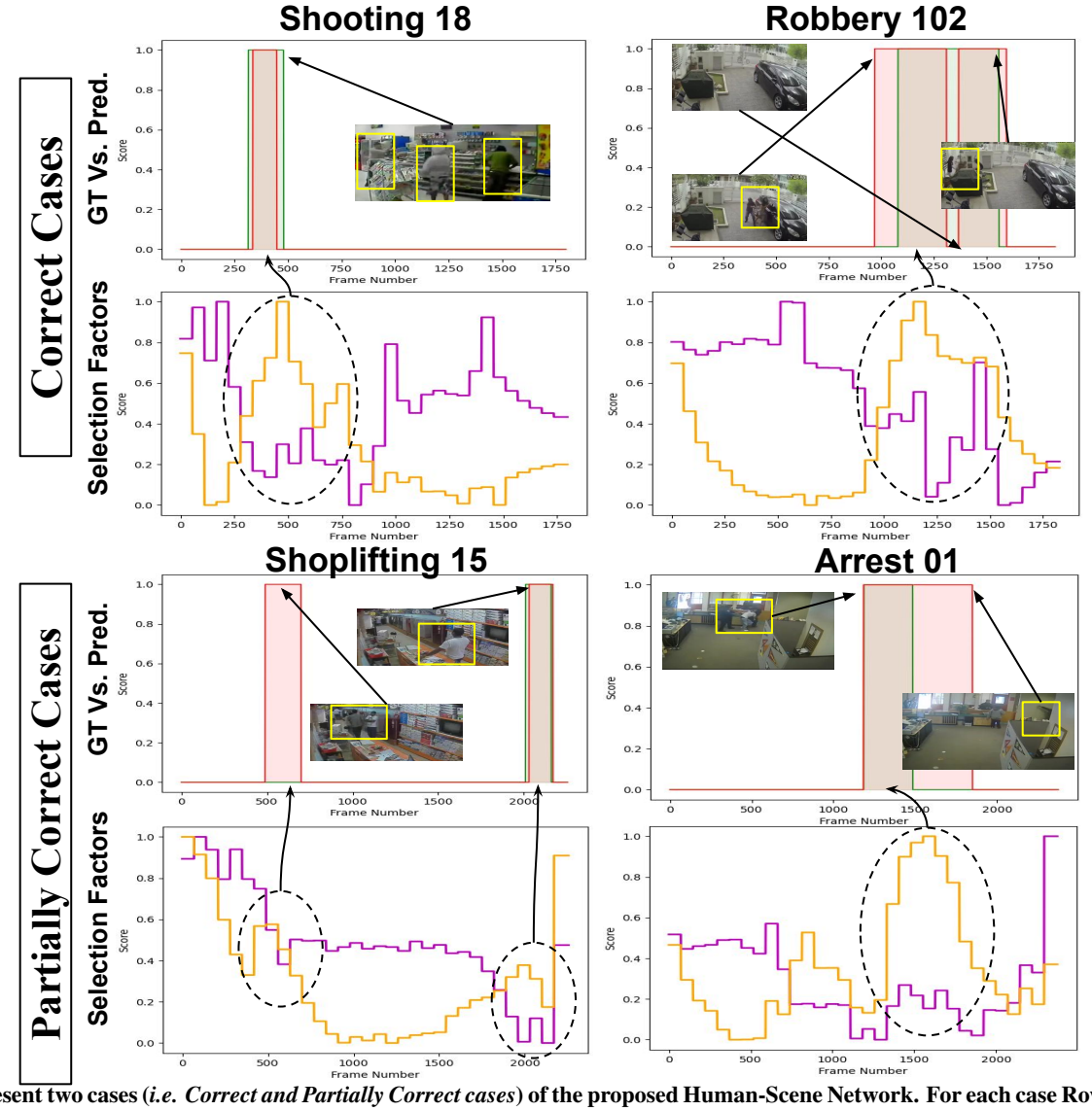


**Fig. 1. We present two cases (*i.e. Correct and Partially Correct cases*) of the proposed Human-Scene Network. For each case Row-1 shows the Ground Truth (Green shed) Vs. Prediction (Red shed) to showcase the achieved preciseness, Row-2 presents the corresponding selection factors for HsN (Yellow plot) and SsN (Violet plot) to highlight the robustness of soft-selection coupler in focusing on the appropriate subNet. The dotted circles in Row-2 highlight the selection factors in anomaly localized areas.**

## 3. Added Discussion on Performance Comparison

**Overall performance Comparison with Unsupervised Methods :** Here, we provide the overall performance comparison of our method with the state-of-the-art unsupervised and weakly-supervised methods as shown in Table 3. Since our method follows weakly-supervised learning paradigm, so it gives a performance boost of 14.41% and 17.29% compared to the competitive unsupervised method Zaheer et al. (2022) on UCF-Crime (similar to real-world scenarios) and ShanghaiTech datasets respectively. This is due to weakly-supervised methods have higher robustness and generalization capabilities than that of unsupervised methods. Similarly, compared to recent weakly-supervised methods Majhi et al. (2021b) and Tian et al. (2021), our human-scene network outperforms on five out of six scenarios considered.

**Observation from ShanghaiTech Dataset :** Unlike UCF-Crime Sultani et al. (2018) that has a real-world complex and diverse data distribution, ShanghaiTech has only 65 anomaly videos for training with simple human anomalies (like fall down, riding cycle on pedestrian road) of lower diversity. The reason behind less competitive performance of our method on ShanghaiTech dataset is the difficulties involved in training human-scene network (HSN) with the soft-selection coupler (SSC) in such skewed data distribution (*i.e* limited number of samples from human anomalies and no scene anomalies). Since the SSC block of HSN learns to focus on either human or scene subnet and the overall efficacy of the HSN depends upon the effective selection of SSC, with such human-only distribution the SSC could not be optimized effectively, resulting in lower performance.

## 4. Real-world challenging situation for Human subNet

The objective of human subNet is to model fine-grained representation in human-centric anomaly (*i.e.* Abuse, Assault,

| | Methods | Feature | UCF-C | ST AUC(%) | IITB-C |
|---|---|---|---|---|---|
| Unsupervised | SVM Baseline | - | 50.00 | - | - |
| | Hasan et al. (2016) | AE | 50.60 | 60.85 | - |
| | Lu et al. (2013) | C3D | 65.51 | - | - |
| | Liu et al. (2018) | C3D | - | 72.80 | 64.65 |
| | Wang (2019) | I3D | 70.46 | - | - |
| | Sun et al. (2020) | TCN | 70.70 | - | - |
| | Zaheer et al. (2022) | ResNext | 71.04 | 78.93 | - |
| Weakly-supervised | Sultani et al. (2018) | C3D | 75.41 | - | - |
| | | I3D-Inc | 77.42 | 80.02 | 74.59 |
| | Zhang et al. (2019) | C3D | 78.66 | - | - |
| | Lin et al. (2019) | C3D | 78.28 | - | - |
| | Zhu and Newsam (2019) | C3D | 79.00 | - | - |
| | Zaheer et al. (2020) | C3D | 79.54 | 84.16 | - |
| | Zhong et al. (2019) | C3D | 81.08 | 76.44 | - |
| | | TSN-Inc | 82.12 | 84.44 | - |
| | Feng et al. (2021) | C3D | 81.40 | - | - |
| | | 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) | C3D | 83.28 | 91.51 | - |
| | | I3D-Res | 84.30 | **97.21** | 81.12* |
| | **Our** | **I3D-Inc** | **84.33** | **93.72** | **84.12** |
| | | **I3D-Res** | **85.45** | 96.22 | **86.98** |

**Table 3. 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. Kindly note that * marked AUC is our implementation.**

Robbery, Shoplift) detection. But, there exists few real-world challenging situations where human subNet face difficulties in detecting human anomalies. First, due to heavy occlusion, anomaly partially out of camera field of view (FOV), lower illumination condition, noisy image quality, subtle motion, the human detection and tracking are inconsistent and flickering for anomaly causing humans (mostly in Assault, Robbery, Shoplift categories). Thus, the human subNet fails to select the abnormal human trajectories and results in slightly lower performance. Second, with no human involvement in the anomaly the human subNet lags in detection performance. Figure 3 shows the sample anomalies which are challenging for human subNet.

## 5. Necessity of 5-Fold Test Evaluation

The anomaly detection performance obtained from official test-split Sultani et al. (2018) of UCF-Crime has a strong bias towards the easy anomalies (contains sharp changes in the appearance and motion cues) like *explosion, road accidents, shooting* as they have significantly more samples for testing. Due to this, global scene-based methods Tian et al. (2021); Sultani et al. (2018); Wu et al. (2020) which are capable of capturing sharp changes tend to perform well on the official test split. From Figure 2, it can be seen that anomalies like *abuse, arrest,*

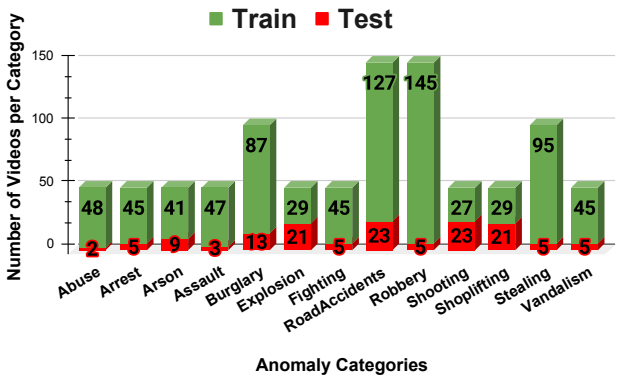


**Fig. 2. Visualization of number of videos used for training and testing for each abnormal category in the official split of UCF-Crime dataset.**

*arson, assault, burglary, fighting, robbery, stealing vandalism* have very few test samples compared to their training counterparts and these anomalies are characterized by both subtle and smooth spatio-temporal cues which are difficult to detect in a real-world complex scenario. For this, we adopt the 5-fold test evaluation in our work which covers the entire dataset for an unbiased evaluation of both simple and complex anomalies.

## References


Chen, C.F.R., Fan, Q., Panda, R., 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 357–366.

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).

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.

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

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.

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.

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.

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.

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.

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, 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.

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).

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


**Fig. 3. Visualization of sample video anomalies where human-subNet can results in slightly lower performance. The red and green shadings show the anomaly and normal ground truth across temporal locations. The white-dotted-circles highlight anomaly regions where peeple detection tracking methods has failed due to various factors mentioned in the left side.**