Anomaly Detection for People with Visual Impairments Using an Egocentric 360-Degree Camera

Inpyo Song[1], Sanghyeon Lee[2], Minjun Joo[1], Jangwon Lee[1]
[1]Department of Immersive Media Engineering, Sungkyunkwan University
[2]School of Electronics and Information Engineering, Korea Aerospace University
{songinpyo, jmjs1526, leejang}@skku.edu, tkdgus4693@kau.kr

Abstract

*Recent advancements in computer vision have led to a renewed interest in developing assistive technologies for individuals with visual impairments. Although extensive research has been conducted in the field of computer vision-based assistive technologies, most of the focus has been on understanding contexts in images, rather than addressing their physical safety and security concerns. To address this challenge, we propose the first step towards detecting anomalous situations for visually impaired people by observing their entire surroundings using an egocentric 360-degree camera. We first introduce a novel egocentric 360-degree video dataset called **VIEW360** (Visually Impaired Equipped with Wearable 360-degree camera), which contains abnormal activities that visually impaired individuals may encounter, such as shoulder surfing and pickpocketing. Furthermore, we propose a new architecture called the **FDPN** (Frame and Direction Prediction Network), which facilitates frame-level prediction of abnormal events and identifying of their directions. Finally, we evaluate our approach on our VIEW360 dataset and the publicly available UCF-Crime and Shanghaitech datasets, demonstrating state-of-the-art performance. Code and dataset are available at https://github.com/Songinpyo/VIEW360.*

1. Introduction

People with visual impairments encounter various challenges in their daily lives, especially related to their physical safety and security risks, as they may not perceive their surroundings as easily as sighted individuals [2]. Traditionally, white canes and guide dogs have been widely used to help them navigate their environment and understand their surroundings. However, these traditional assistive systems suffer from certain limitations. For example, white canes only provide limited information about the surroundings [9], and training guide dogs is both time-consuming and expensive [24]. As a result, there has been a significant increase in developing a visual aid system to create new "eyes" for visually impaired individuals using wearable devices and Artificial Intelligence (AI) technologies [12, 26]. However, much of the research up to now has primarily focused on tasks like the image captioning and visual question answering rather than addressing real-world issues such as physical safety and security concerns [11–13]. Therefore, this paper proposes a first step towards to address the physical safety and security concerns of visually impaired people by employing a 360-degree wearable camera. In particular, we have two primary objectives: (1) detecting suspicious or abnormal activities within a 360-degree video stream, (2) identifying the direction of these activities.

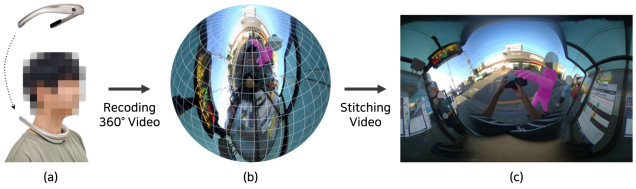

Figure 1. This paper aims to tackle safety and security concerns faced by visually impaired individuals. To tackle these concerns, we introduce a new dataset, **VIEW360**, specifically designed for detecting unusual activities by observing their entire surroundings using an egocentric 360-degree camera. The dataset is collected through a process involving (a) capturing footage with a wearable 360-degree camera worn around the neck, (b) recording egocentric 360-degree videos to encompass the wearer's surroundings, and (c) stitching these videos into panoramic views for comprehensive analysis. In the depicted scene, the individual highlighted in magenta is attempting a wallet theft.

To accomplish this, we first introduce a new dataset called **VIEW360**, which has been designed for detecting anomalies in a camera wearer's entire surroundings. This dataset includes egocentric 360-degree videos captured at several public locations, such as ATM booths, parks, and cafes. VIEW360 is comprised of 575 videos that illustrate real-life situations, which were obtained through interviews

with visually impaired individuals [2]. Our dataset can be categorized as one of the datasets for the Video Anomaly Detection (VAD) task in the field, which aims to identify frames in a video where unusual or anomalous events occur. There are generally two approaches to this task: semi-supervised Video Anomaly Detection (**sVAD**) and weakly-supervised Video Anomaly Detection (**wVAD**). The sVAD approach trains networks using only normal videos without any annotations [25, 32], whereas the wVAD approach uses video-level labels for training, but with limited annotation [28, 29].

Among the two lines of approaches, lately, researchers have shown an increased interest in wVAD methods due to their promising performance on public VAD benchmark datasets [7, 10]. However, these methods face challenges in identifying abnormal activities in rapidly changing scenarios. This limitation is inherent to recent wVAD approaches, which tend to predict anomaly scores at the snippet-level (a brief segment extracted from a video). Consequently, these approaches assign identical anomaly scores to a fixed number of frames (referred to as snippets), often resulting in overly generalized predictions. Therefore, these techniques could face difficulties in identifying sudden real-world anomalies, such as short and unexpected activities like shoulder surfing at an ATM booth.

To address these limitations, we introduce a framework called Frame and Direction Prediction Network (**FDPN**). It is designed to predict anomaly scores at the frame level, extending beyond mere snippet-level scores. We achieve this through our innovative coarse-to-fine learning approach. By utilizing existing snippet-level predictions [7, 29] as pseudo-supervision, our FDPN is trained to predict frame-level anomalies, eliminating the requirement for extra annotations. This frame-level prediction proves particularly advantageous in identifying abrupt abnormal events occurring within a brief timeframe, as depicted in Figure 2.

Furthermore, we employ an off-the-shelf saliency detection model as our pre-processing step, denoted as saliency-driven image masking, to further enhance the process. This technique identifies visually striking regions in a frame that may contain anomalies. Since anomalies often appear in salient regions, this method allows us to narrow down the search space for anomaly detection. This is especially useful for handling 360-degree images due to their extensive visual coverage. We also incorporate direction classification from a 360-degree egocentric perspective, offering practical guidance for visually impaired individuals during anomalous events. This is achieved through a dedicated subnetwork utilizing saliency heatmaps.

Finally, we evaluate the proposed approach on our VIEW360 and the publicly available UCF-Crime, Shanghaitech datasets. Our approach achieves state-of-the-art performance on the VIEW360, UCF-Crime and Shanghaitech datasets. In summary, this paper makes the following contributions:

- To our knowledge, this is the first study to address the physical safety and security concerns of people with visually impairments by detecting anomalous events and identifying their direction in the surroundings.
- We introduce **VIEW360**, a novel egocentric 360-degree video-based dataset to address safety and security concerns of visually impaired people in real-world scenarios.
- We propose a novel architecture called **FDPN** that can predict more precise anomaly scores at the frame-level based on rich scene representation without the need for additional frame-level annotation.

2. Related Work

2.1. AI for People with Visual Impairments

In the last decade, there has been a dramatic increase in developing a visual aid system aimed at creating new "eyes" for the visually impaired people using AI technologies. Gurari et al. have introduced the VizWiz-VQA dataset, a collection of images and questions gathered from blind individuals [12]. These researchers have also created another VQA dataset, named the VizWiz-Priv dataset, which focuses on identifying unintended leaks of personal information through VQA for visually impaired users [11]. Furthermore, diverse approaches have emerged in recent years to offer various perspectives within this VQA task. These include delving into the reasons behind variations in responses to identical visual questions among distinct individuals [3], addressing the domain gap between images captured by visually impaired individuals and sighted individuals [13], as well as tackling poor image quality in VQA systems by creating a dataset and task to predict reasons for low-quality images [8].

While existing efforts have made substantial progress in comprehending the contextual aspects of images, less attention has been directed towards video analysis, especially in addressing the security and physical concerns that visually impaired individuals might encounter in their daily lives [27]. Therefore, this study bridges a research gap in AI-based visual aid systems for people with visual impairments by introducing a novel 360-degree egocentric video anomaly detection task along with a new dataset.

2.2. Anomaly Detection in Videos

Video anomaly detection, a crucial computer vision task, identifies frames with abnormal events in videos. Approaches are categorized as semi-supervised and weakly-supervised. Semi-supervised methods, assuming most data is normal, detect abnormalities by identifying image frames

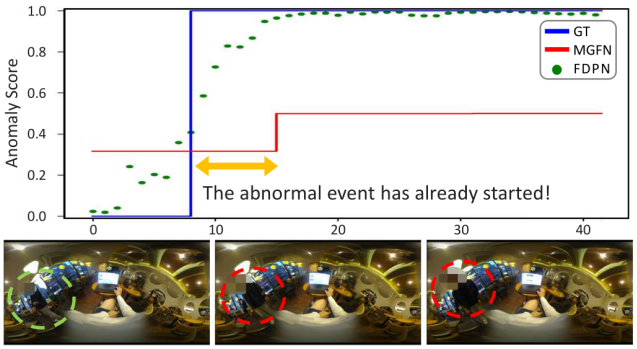

Figure 2. This figure contrasts anomaly scores at event start and end boundaries for state-of-the-art method MGFN and our FDPN on VIEW360 dataset. MGFN often makes false predictions at event boundaries because it predicts at the snippet-level, whereas our proposed method makes better predictions at the event boundaries since it can make frame-level predictions.

that significantly differ from previously observed data [23, 25]. This approach makes sense since abnormal events "unusually" occur in the real world and it offers a clear advantage as it does not require annotation costs. However, their performance often declines when classifying unseen data.

In contrast, weakly-supervised approaches aim to enhance detection performance by utilizing minimal annotations, such as video-level labels, during training. Recent advancements in this field include the use of graph networks to handle noisy labels in abnormal videos [38], and the integration of motion information to improve detection accuracy [40]. Further innovations, such as inter-class distancing, sequence learning, self-supervised techniques, and magnitude-based methods, have significantly contributed to the progress in anomaly detection [7, 16, 29, 30, 33]. Additionally, the adoption of vision-language models and prompt-enhanced techniques has advanced capabilities in this domain [6, 15, 35, 36]. However, existing methods, which are primarily snippet-level anomaly detection approaches, often struggle with abrupt events due to uniform scoring. Our proposed frame-level approach leverages snippet-level predictions as pseudo-supervision, thereby improving detection accuracy for short-lived anomalies.

3. VIEW360:Dataset for the Visually Impaired

This section presents the "**V**isually **I**mpaired **E**quipping **W**earable **360**-degree camera" (**VIEW360**) dataset, designed to advance AI-assisted technology for individuals with visual impairments through a 360-degree wearable camera [19]. This is the first dataset that contains 360-degree egocentric videos of anomaly activities that visually impaired people commonly encounter in their daily lives. We selected the locations and anomalies for our dataset through interviews with visually impaired individuals [2]. We focused on three types of abnormal scenarios: *Glance, Stealing*, and *Teasing*, as illustrated in Figure 3. Data was collected from eight different public venues: *Cafes, Restaurants, Bus stops, Elevators, Parks, Libraries, Offices*, and *Automated Teller Machine (ATM) booths*.

Dataset	Videos	Avg. anomaly duration (s)	Video source
UCSD Ped1 [17]	70	11.2	CCTV
UCSD Ped2 [17]	28	13.7	CCTV
Avenue [21]	37	9.2	CCTV
UBnormal [1]	543	10.7	Virtual scene
Shanghaitech [20]	437	6.7	CCTV
NWPU Campus [4]	547	10.8	CCTV
UCF-Crime [28]	1,900	20.1	CCTV
XD-violence [34]	4,754	37.5	Movie, Game, etc.
VIEW360 (Ours)	575	3.5	Ego 360° camera

Table 1. Comparison of anomaly detection datasets, distinguishing between sVAD (top) and wVAD (bottom). Our VIEW360 dataset, stands out by exclusively featuring egocentric 360-degree videos. Notably, VIEW360 focuses on shorter average anomaly duration, emphasizing the detection of quick, transient anomalies.

Video Collection All videos in the dataset were captured using a 360-degree wearable camera mounted on the neck of an actor who simulated being visually impaired. To ensure diversity, 11 participants wore the camera and performed abnormal behaviors. We aimed to construct a dataset with a range of situations, so we mostly collected short videos ranging from 10 to 60 seconds instead of long videos typical of most video anomaly datasets. In total, we collected 575 videos consisting of 484,364 frames.

Annotations There are two types of annotations available: temporal annotations and directional annotations. Temporal annotations are created in accordance with the conventions of existing anomaly detection datasets [28, 34]. Specifically, we annotated video-level labels for the training set, while the testing set includes frame-level labels for evaluation purposes. A challenge in labeling the anomaly detection dataset is determining the boundary between the beginning and end of the anomaly. To address this, a total of five annotators were engaged, cross-validating each other's labels to maintain consistency and enhance accuracy. For direction prediction, we deliberately simplified the annotation to three categories, *Left back, Center*, and *Right back*. This sparse but practical directional information can aid visually impaired individuals in swiftly identifying potential threat directions, enabling quicker decision-making and more effective responses in real-world scenarios.

Dataset Statistics The training set comprises 375 videos: 181 normal and 194 abnormal. The testing set includes 200 videos: 95 normal and 105 abnormal. Our dataset features 3 directional labels, *Left back* (106 videos), *Center* (101 videos), and *Right back* (105 videos). Details of training/testing sets, locations, and abnormal directions are in Figure 4. Abnormal situations are evenly distributed between the training and testing sets. Video length (in seconds) and abnormal class distribution are in Figure 5.

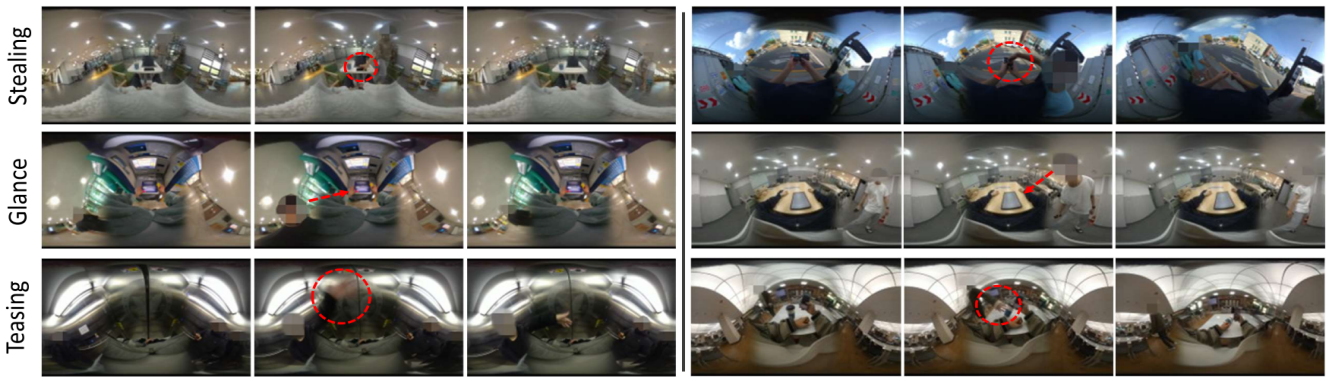

Figure 3. Here are some abnormal instances in our VIEW360 dataset. The first row shows theft of personal belongings from the camera-wearer. The second row depicts shoulder-surfing attacks: someone covertly observing the camera wearer's ATM use and smartphone without their awareness. The third row portrays a person with visual impairments being mocked or harassed.

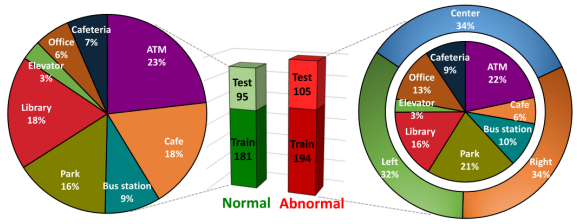

Figure 4. Distribution of the VIEW360 dataset, illustrating training/testing splits, video locations, and abnormal event orientations. Includes a bar chart of normal and abnormal video counts, a pie chart of video location distribution, and a donut chart of abnormal event directions.

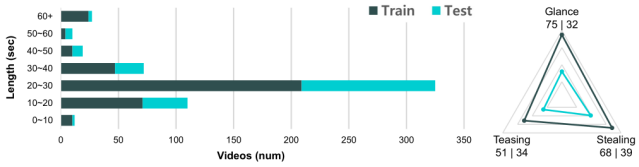

Figure 5. Video duration and abnormal classes in VIEW360.

Privacy and Ethics The dataset collection for the VIEW360 was rigorously conducted under the approval of the Institutional Review Board (IRB), ensuring that all research activities adhered to the highest ethical standards and guidelines. Informed consent was obtained from all participants involved in the simulated scenarios, thereby guaranteeing their full awareness of the data collection's purpose and scope, and affirming their rights as participants. During the collection process, special attention was given to respecting the rights of others in private spaces and diligently avoiding the capture of sensitive areas or activities. For non-consenting individuals appearing in public areas within the videos, privacy measures like blurring and facial masking were applied to uphold their anonymity and ensure the dataset adhered to ethical research practices.

4. Proposed Approach

Building upon our VIEW360 dataset, our goal is to identify short-lived abnormal activities in a 360-degree video stream and determine their direction. In this section, we introduce our framework, the Frame and Direction Prediction Network (FDPN), designed to achieve these aims.

4.1. Overview

Our FDPN begins by identifying salient regions within the input frame that may contain anomalies given 360-degree panoramic image frames as input. Once the salient regions in the image frames are identified, FDPN internally employs two types of input images: 1) masked images, retaining only the salient regions while masking out other portions, and 2) original images, the unaltered 360-degree panorama input images. FDPN then proceeds to extract snippet-level features from the original images using Inflated 3D ConvNet (I3D) [5], and it extracts frame-level (image-level) features from the masked images using ResNet [14]. After that, we employ the snippet-level features for prediction at the snippet level and to generate pseudo-labels for training our Frame Prediction Subnetwork (FPS). Finally, these snippet-level features are combined with the frame-level features, and this concatenated feature set is utilized to compute anomaly scores at the frame level.

For identifying the direction of abnormal events, we construct the Direction Prediction Subnetwork (DPS). It operates on concatenated features of snippet-level and image-level features, along with the saliency maps used in the initial step. Figure 6 depicts the overall architecture of the FDPN. We present the details of each module within this process in the following subsections.

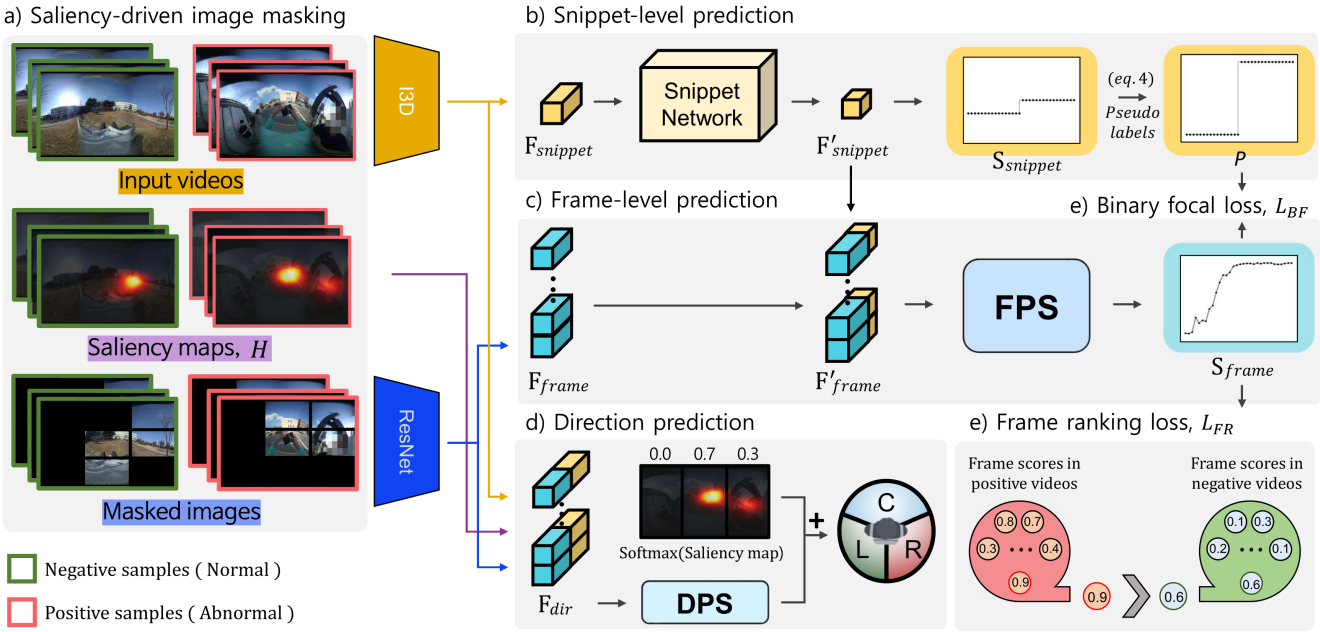

Figure 6. Overview of our FDPN. During training, positive and negative video pairs are fed into the framework. **(a)** Saliency map-based masking is applied to these pairs, identifying salient regions within input image frames. Snippet-level original frames are processed using an I3D feature extractor, while frame-level masked images are handled by ResNet. **(b)** Following feature extraction, the Snippet Network generates F'_{snippet} and computes snippet-level anomaly score S_{snippet}, used to create a pseudo label ($\mathcal{P}$) for training the FPS. **(c)** FPS employs the concatenated feature F'_{frame}, combining F_{frame} from masked images and F'_{snippet}, to compute frame-level anomaly score S_{frame}. **(d)** For direction estimation, we construct the DPS. DPS initially estimates direction using concatenated feature F_{dir}, refining output with softmax-applied saliency values. **(e)** To train the model, we use binary focal loss with S_{frame} and the pseudo labels ($\mathcal{P}$). In addition, Frame Ranking Loss (L_{FR}) ensures higher anomaly scores for positive videos than for negative ones.

4.2. Saliency-driven Image Masking (a)

We first employ TASED-Net [22] to derive saliency heatmaps H as shown in Figure 6-(a). These heatmaps emphasize visual significance within the frame, with each pixel value $H_{x,y}$ representing the saliency score at the corresponding coordinates (x,y) within the image. Subsequently, we divide these heatmaps into an $n \times n$ grid. For each cell denoted by integers (i, j), we then compute the importance score by summing the pixel saliency values within it:

$$G_{i,j} = \sum_{x,y \in \text{cell}_{i,j}} H_{x,y} \tag{1}$$

After this step, we pinpoint the *top-K* salient regions with the highest scores from the grid scores $G_{i,j}$. Following this, we generate a binary mask M where the *top-K* cells are assigned a value of 1, and the remainder are assigned a value of 0:

$$M_{i,j} = \begin{cases} 1 & \text{if } G_{i,j} \text{ is in } \textit{top-K} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Finally, the masked image I_{masked} is obtained by multiplying the original image I by the binary mask M:

$$I_{\text{masked}} = I \odot M \tag{3}$$

This approach offers a significant advantage: it guides focus toward event-specific regions by utilizing the saliency map to pinpoint critical areas while concealing others. This emphasis improves the analysis of potential anomaly locations, as shown in Figure 7.

From I_{masked}, frame-level features are extracted using ResNet, yielding $F_{\text{frame}} \in \mathbb{R}^{B \times T \times N \times C}$. Simultaneously, snippet-level features are derived from the original frames using I3D, resulting in $F_{\text{snippet}} \in \mathbb{R}^{B \times T \times C}$. Here, B represents the number of video pairs, T the number of snippets, each snippet consists of N frames and C the channel.

4.3. Coarse-to-Fine Learning (b - c)

Snippet-level Prediction (b) To enable our FDPN to produce frame-level anomaly scores, the network needs frame-level ground truth labels for training. Consequently, we initially train the Snippet Network for snippet-level prediction. This involves using pre-trained networks [7,29] to create feature F'_{snippet} and snippet-level anomaly scores S_{snippet}. Subsequently, the snippet-level anomaly score S_{snippet} is duplicated N times, following the specified rule below, resulting in pseudo labels $\mathcal{P}$ crafted for training our FPS at the

Figure 7. This figure shows that our saliency-driven image masking preprocess applied to events at an ATM booth. It is divided into two rows, representing normal and abnormal events, with each row consisting of the original image, saliency map, and masked image, respectively. In the bottom row representing abnormal events, we marked the grid cells where the anomaly activity happened in red.

frame level.

$$\mathcal{P} = \begin{cases} 0 & \text{if } S^{+}_{\text{snippet}} < 0.5 \text{ or } S^{-}_{\text{snippet}} \\ 1 & \text{if } S^{+}_{\text{snippet}} \geq 0.5 \end{cases}, \tag{4}$$

S^{+}_{snippet} and S^{-}_{snippet} denote snippet-level anomaly scores for positive and negative videos, respectively. Positive videos receive pseudo labels of 0 or 1 according to their anomaly scores, while negative videos (without anomalies) are assigned a pseudo label of 0.

Frame-level Prediction (c) With the acquired pseudo labels, we proceed to train the FPS. By combining F_{frame} and F'_{snippet}, we create F'_{frame} as FPS input. This design aims to encompass both frame-level and snippet-level information. Following training, FPS is primed to estimate frame-level anomaly scores, denoted as $S_{\text{frame}} \in \mathbb{R}^{B\times T\times N}$. We designed this FPS to handle sequences, inspired by PoolFormer [37]. It employs average pooling and 1D convolutions to capture frame connections and compute frame-level anomaly scores. This improves the network's capability to understand relationships among neighboring frames.

4.4. Direction Prediction (d)

We also developed the DPS to determine abnormal event directions by utilizing F_{dir}, a concatenation of F_{frame} and F_{snippet}. DPS employs the same PoolFormer-based architecture as FPS and computes direction scores for three orientations (Center, Left back, and Right back). It integrates saliency heatmap values across a 1x3 grid and applies softmax to enhance anomaly direction prediction. This approach leverages spatial information from the saliency heatmap and our PoolFormer-based architecture to improve the accuracy of abnormal event direction prediction, reducing spatial ambiguity in video frames.

4.5. Loss Functions (e)

We utilize three loss functions for frame-level prediction and a single loss function for direction prediction. The first loss, Binary Focal Loss [18], L_{BF}, compares the pseudo labels $\mathcal{P}$ to the frame-level anomaly scores S_{frame} with a focusing parameter γ that emphasizes hard-to-classify examples.

$$\begin{aligned} L_{BF} = &-\mathcal{P}(1 - S_{\text{frame}})^{\gamma} \log(S_{\text{frame}}) \\ &-(1-\mathcal{P}) S^{\gamma}_{\text{frame}} \log(1 - S_{\text{frame}}) \end{aligned} \tag{5}$$

The second loss function, Frame Ranking Loss L_{FR}, prioritizes the top $\mathcal{R}$ frames with the highest anomaly scores in both positive and negative videos. It drives the anomaly scores of positive videos towards 1 and those of negative videos towards 0.

$$L_{FR} = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left(1 - S^{+}_{\text{frame},r} + S^{-}_{\text{frame},r}\right) \tag{6}$$

The third loss, Smoothness Loss, L_{smooth} [28], penalizes rapid changes in the predicted anomaly scores to ensure smooth transitions between adjacent frames. It is calculated for all $\mathcal{F}$ frames as follows:

$$L_{smooth} = \frac{1}{\mathcal{F}} \sum_{f=1}^{\mathcal{F}} (S^{f}_{\text{frame}} - S^{f-1}_{\text{frame}})^2 \tag{7}$$

For direction prediction, we utilize Directional Focal Loss L_{DF}, employing the identical focusing parameter γ.

$$L_{DF} = -\sum_{k=1}^{C} y_k (1 - p_k)^{\gamma} \log(p_k) \tag{8}$$

Here, C represents the three directions (Center, Left back, Right back) in our VIEW360 dataset. y_k is the ground-truth label for direction and p_k is the predicted probability for that direction. The parameter γ is shared with Binary Focal Loss and adjusts the class contribution to the loss.

The overall loss is computed as follows, where λ_1, λ_2, and λ_3 represent weight factors for each respective loss function.

$$L = L_{BF} + \lambda_1 L_{FR} + \lambda_2 L_{smooth} + \lambda_3 L_{DF} \tag{9}$$

5. Experiments

5.1. Datasets and Evaluation Metrics

We evaluate our method on three anomaly detection datasets: VIEW360, UCF-Crime, and Shanghaitech [20, 28]. We used Area Under the Receiver Operating Characteristic (AUC-ROC) evaluation metric same with established works [7, 29]. For VIEW360, we additionally employed AUC-ROC and Area Under the Precision-Recall (AUC-PR) metrics to evaluate model, addressing both false positives and false negatives. This approach ensures a balanced assessment of the model's ability to accurately detect anomalies while minimizing unnecessary alerts for visually impaired users, enhancing the system's practical reliability.

Method	Publication	Feature	AUC-ROC	AUC-PR
MIST [10]	CVPR"21	I3D	79.30	19.58
RTFM [29]	ICCV"21	I3D	83.92	24.94
S3R [33]	ECCV"22	I3D	83.96	25.17
MGFN [7]	AAAI"23	I3D	80.43	20.16
DMU [39]	AAAI"23	I3D	83.88	25.11
CLIP-TSA [15]	ICIP"23	CLIP	80.03	17.41
VadCLIP [35]	AAAI"24	CLIP	79.92	21.28
FDPN (Ours)	-	I3D	**86.00**	**26.97**

Table 2. Comparison of frame-level AUC-ROC and AUC-PR performances on VIEW360 dataset.

Method	Publication	Feature	UCF-Crime	Shanghaitech
MIST [10]	CVPR"21	I3D	82.30	94.83
RTFM [29]	ICCV"21	VidSwin	83.31	96.76
RTFM [29]	ICCV"21	I3D	84.03	97.21
MSL [16]	AAAI"22	I3D	85.30	95.45
MSL [16]	AAAI"22	VidSwin	85.62	96.93
S3R [33]	ECCV"22	I3D	85.99	97.48
MGFN [7]	AAAI"23	I3D	86.98	-
DMU [39]	AAAI"23	I3D	86.97	-
CLIP-TSA [15]	ICIP"23	CLIP	87.58	98.32
PE-MIL [6]	CVPR"24	I3D	86.83	98.35
*TPWNG [36]	CVPR"24	CLIP	87.79	-
*VadCLIP [35]	AAAI"24	CLIP	88.02	-
FDPN (Ours)	-	I3D / CLIP	**88.03**	**98.51**

Table 3. Comparison of frame-level AUC-ROC performance on UCF-Crime and Shanghaitech datasets. Models marked with "*" use multi-modal text information, which was not available for evaluation on the Shanghaitech dataset.

5.2. Implementation Details

Following existing methods [7,33], each video is divided into 32 snippets with 16 frames in each snippet during the training stage. For the hyperparameters, we set $B = 16$, $T = 32$, $N = 16$, $n = 3$, $K = 4$, $\gamma = 2$, $\mathcal{R} = 48$, $\lambda_1 = 1$, $\lambda_2 = 1.6e^{-3}$, $\lambda_3 = 0.3$.

5.3. Evaluation

VIEW360 As shown in Table 2, our approach outperforms state-of-the-art methods in both AUC-ROC and AUC-PR on the VIEW360 dataset. We achieved improvements of 2.08% and 2.03% in AUC-ROC and AUC-PR, respectively, compared to RTFM (used as the snippet network). These results highlight the efficacy of our method in detecting subtle anomalies and short-life anomalies in the dataset, important for applications with visually impaired users.

UCF-Crime and Shanghaitech Our FDPN achieved state-of-the-art performance on UCF-Crime and Shanghaitech as shown in Table 3. These results highlight the robustness of our frame-level prediction method across different anomaly detection tasks.

Analysis of Existing Methods We observed an intriguing performance discrepancy between recent state-of-the-art methods like MGFN and VadCLIP across different datasets. While these methods excelled on UCF-Crime, they underperformed on VIEW360. This variance primarily stems from the inherent differences in dataset characteristics. VIEW360 contains more subtle and shorter-duration anomalies compared to UCF-Crime. Therefore, MGFN, optimized for high-magnitude snippet training, tends to struggle with accurate clip selection in datasets featuring subtler anomalies like VIEW360. Similarly, VadCLIP's reliance on abnormal category classification for training proves less effective for VIEW360, where abnormal events are more challenging to distinguish than in UCF-Crime.

DPS	Saliency	Direction acc.
✓		56.50
	✓	67.80
✓	✓	**75.04**

(a)

Dataset	SD	OD	Diff.
VIEW360	86.00	84.78	-1.24
UCF-Crime	88.03	86.37	-1.66

(b)

Table 4. (a) Ablation study for direction prediction on VIEW360, (b) Effectiveness comparison of image masking process with Saliency Detection (SD) or Object Detection (OD)

Grid	3x3 (9)		4x4 (16)		5x5 (25)	
	Top-K	ROC	Top-K	ROC	Top-K	ROC
Unmasked 85.02	5 / 9	85.19	10 / 16	85.19	15 / 25	85.02
	4 / 9	**86.00**	8 / 16	85.50	12 / 25	85.41
	3 / 9	85.54	6 / 16	85.26	9 / 25	85.30

Table 5. Comparison of different grid sizes and Top-K salient regions for Saliency-driven Image Masking on VIEW360

Snippet Network Selection. We tailored our snippet network selection to each dataset's characteristics. For UCF-Crime, which features prominent anomalies, we opted for MGFN due to its proven effectiveness in such scenarios. In contrast, for VIEW360, which contains more subtle and shorter-duration anomalies, we selected RTFM. RTFM's approach of training on entire videos minimizes the risk of missing critical frames, making it more suitable for datasets with nuanced anomalies. For Shanghaitech, which primarily comprises abnormal object appearances, we chose CLIP-TSA to leverage the strong image feature extraction capabilities of CLIP. This customized strategy enables our FDPN to effectively adapt to diverse anomaly detection scenarios, contributing to its robust performance across different datasets.

Direction Prediction on VIEW360 Our FDPN also aims to identify the direction of detected abnormal activities to assist visually impaired individuals. With our DPS, we achieved 75.04% directional prediction accuracy by leveraging salient area information from the saliency detector. We further conduct ablation experiments to understand the impact of each module. Results in Table 4-a demonstrate the essential role of both image features and salient area information in estimating abnormal event directions, which highlights their combined effectiveness on VIEW360.

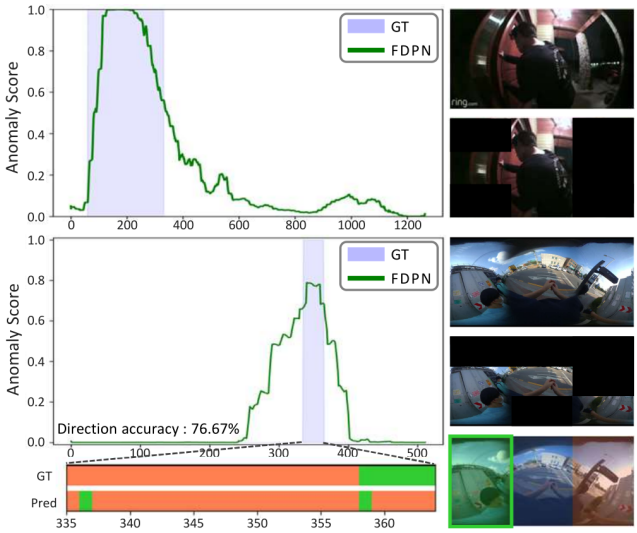

Figure 8. The figure shows FDPN's anomaly scores on the UCF-Crime (top) and VIEW360 datasets (bottom). Ground truth abnormal frames are highlighted with blue boxes on the graph. Beside it, we present original and saliency-driven masked images of frames. In the VIEW360, direction prediction results and ground truth direction are additionally displayed as colored bars: green for left, blue for center, and orange for right.

5.4. Ablation Study

Saliency Detection vs Object Detection As part of our ablation study, we also analyzed the benefits of using a saliency detector versus a general object detector (YOLO v7 [31]) in our pre-processing step. To do this, we replaced the saliency detector as object detector which is trained on COCO dataset. Results in Table 4-b confirmed that the saliency detector is more effective for anomaly detection. The primary advantage of saliency detection is its focus on active region rather than numerous inactive objects present in video. Moreover, being class-agnostic, it can detect any active image objects, enhancing flexibility and robustness.

Optimizing Image Masking Our saliency-driven image masking's effectiveness is influenced by grid size and top-K salient regions retained within the grid, while others are masked. We thus examined various grid sizes and top-K regions. Table 5 illustrates this experiment, emphasizing the critical role of proper masking in highlighting key image areas for more precise detection. In the table, in a $n \times n$ grid, "4 / 9" indicates retaining 4 salient regions out of a total of 9, with 5 areas masked.

Verifying Frame-level Prediction's Benefit Our FDPN excels in identifying abrupt abnormal events occurring within brief timeframes, thanks to frame-level prediction as illustrated in Figure 2. To validate this capability, we conducted a deeper analysis comparing the accuracy improvement of our FDPN against MGFN across varying anomaly durations. We evaluated anomaly detection performance using various score thresholds: 0.6, 0.7, 0.8, and 0.9. The results, presented in Figure 9, clearly demonstrate that FDPN achieves more substantial improvements for videos containing shorter-duration anomalous events.

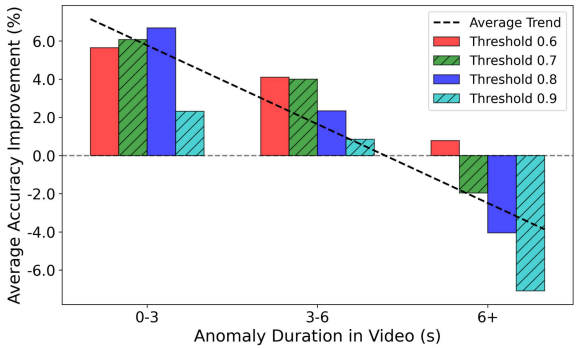

Figure 9. Accuracy improvement of FDPN over MGFN across anomaly durations on UCF-Crime dataset. Thresholds range from 0.6 to 0.9. Representative abnormal events: 0-3s (RoadAccidents), 3-6s (Abuse, Shooting), 6+s (Assault, Explosion).

6. Conclusion

The aim of this research was to tackle safety and security challenges for individuals with visual impairments. To achieve this, we introduced a novel problem and dataset, **VIEW360**, aimed at detecting anomalous events and determining their direction by observing the entire surroundings through an egocentric 360-degree camera. Additionally, we present **FDPN**, a new weakly-supervised video anomaly detection method for frame-level detection and direction prediction. Experimental results show that our FDPN enables more precise detection of anomalous events, as evidenced by achieving state-of-the-art results on the UCF-Crime, VIEW360 and Shanghaitech datasets. This method shows promise in enhancing safety for individuals with visual impairments. We believe our research offers valuable insights for developing AI-based assistive systems, contributing to their safety and security.

Limitations While our dataset was developed through consultation with blind individuals to reflect real-world situations, it may not encompass all challenges faced by the visually impaired community. Additionally, the model's processing speed of 1.7 FPS presents a limitation, as it falls short of the real-time performance needed for practical assistive technology. Future work will focus on optimizing processing speed while maintaining analytical accuracy.

Acknowledgement: This work was supported by the National Research Foundation of Korea(NRF) grant (No. RS-2024-00458696) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grants (No.RS-2023-00254129, No.RS-2024-00459638) funded by the Korean government.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *CVPR*, 2022. 3

[2] Tousif Ahmed, Kurt Andersen, Patrick Shaffer, Dave Crocker, Saptarshi Ghosh, Kay Connelly, Krishna P Gummadi, David Crandall, Aniket Kate, Apu Kapadia, et al. Addressing physical safety, security, and privacy for people with visual impairments. In *Twelfth Symposium on Usable Privacy and Security (SOUPS)*, 2016. 1, 2, 3

[3] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *ICCV*, 2019. 2

[4] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *CVPR*, 2023. 3

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 4

[6] Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, et al. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, 2024. 3, 7

[7] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI*, 2023. 2, 3, 5, 6, 7

[8] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *CVPR*, 2020. 2

[9] Aline Darc Piculo dos Santos, Fausto Orsi Medola, Milton José Cinelli, Alejandro Rafael Garcia Ramirez, and Frode Eika Sandnes. Are electronic white canes better than traditional canes? a comparative study with blind and blindfolded participants. *Universal Access in the Information Society*, 2021. 1

[10] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 2021. 2, 7

[11] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *CVPR*, 2019. 1, 2

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 1, 2

[13] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*. Springer, 2020. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[15] Hyekang Kevin Joo et al. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *ICIP*. IEEE, 2023. 3, 7

[16] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, 2022. 3, 7

[17] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 2013. 3

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 6

[19] LINKFLOW. FITT360. https://www.ftt360.us/, 2021. (online). 3

[20] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, 2018. 3, 6

[21] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013. 3

[22] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *ICCV*, pages 2394–2403, 2019. 5

[23] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *ICCV*, 2019. 3

[24] Clarence J Pfaffenberger, JP Scott, JL Fuller, BE Ginsburg, SW Biefelt, et al. *Guide dogs for the blind: their selection, development, and training.* Elsevier Scientific Publishing Company., 1976. 1

[25] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, 2018. 2, 3

[26] Shankar Sivan and Gopu Darsan. Computer vision based assistive technology for blind and visually impaired people. In *Proceedings of the 7th International Conference on Computing Communication and Networking Technologies*, 2016. 1

[27] Inpyo Song, Minjun Joo, Joonhyung Kwon, and Jangwon Lee. Video question answering for people with visual impairments using an egocentric 360-degree camera. *arXiv preprint arXiv:2405.19794*, 2024. 2

[28] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 2, 3, 6

[29] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, 2021. 2, 3, 5, 6, 7

[30] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020. 3

[31] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, 2023. 8

[32] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *ECCV*. Springer, 2022. 2

[33] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *ECCV*. Springer, 2022. 3, 7

[34] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*. Springer, 2020. 3

[35] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *AAAI*, volume 38, 2024. 3, 7

[36] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *CVPR*, 2024. 3, 7

[37] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022. 6

[38] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, 2019. 3

[39] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI*, 2023. 7

[40] Yi Zhu and Shawn D. Newsam. Motion-aware feature for improved video anomaly detection. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 270. BMVA Press, 2019. 3