

Thermal Detection of People with Mobility Restrictions for Barrier Reduction at Traffic Lights Controlled Intersections

Xiao Ni, Carsten Kühnel, and Xiaoyi Jiang, *Senior Member, IEEE*

Abstract—Rapid advances in deep learning for computer vision have driven the adoption of RGB camera-based adaptive traffic light systems to improve traffic safety and pedestrian comfort. However, these systems often overlook the needs of people with mobility restrictions. Moreover, the use of RGB cameras presents significant challenges, including limited detection performance under adverse weather or low-visibility conditions, as well as heightened privacy concerns. To address these issues, we propose a fully automated, thermal detector-based traffic light system that dynamically adjusts signal durations for individuals with walking impairments or mobility burden and triggers the auditory signal for visually impaired individuals, thereby advancing towards barrier-free intersection for all users. To this end, we build the thermal dataset for people with mobility restrictions (TD4PWMR), designed to capture diverse pedestrian scenarios, particularly focusing on individuals with mobility aids or mobility burden under varying environmental conditions, such as different lighting, weather, and crowded urban settings. While thermal imaging offers advantages in terms of privacy and robustness to adverse conditions, it also introduces inherent hurdles for object detection due to its lack of color and fine texture details and generally lower resolution of thermal images. To overcome these limitations, we develop YOLO-Thermal, a novel variant of the YOLO architecture that integrates advanced feature extraction and attention mechanisms for enhanced detection accuracy and robustness in thermal imaging. Experiments demonstrate that the proposed thermal detector outperforms existing detectors, while the proposed traffic light system effectively enhances barrier-free intersection. The source codes and dataset are available at <https://github.com/leon2014dresden/YOLO-THERMAL>.

Index Terms—Barrier-free intersection, thermal images, object detection, dataset.

I. INTRODUCTION

THE development of safe and accessible intelligent transportation system is crucial for promoting mobility and safety for all individuals [1], [2], [3], particularly those with disabilities [4], [5]. Urban intersections play a pivotal role in managing diverse traffic flows, including pedestrians, cyclists, drivers and public transportation users. However, conventional traffic systems are often inadequate in addressing the unique challenges faced by individuals with mobility restrictions due to fixed signal timings of traffic light and insufficient adaptive

features [6], [7]. Crossing streets poses a disproportionate risk for individuals with mobility restrictions. For instance, the mortality rate of wheelchair users is 36 percent higher in car-related accidents compared to standing pedestrians [8]. This underscores the need for innovative approaches to develop barrier-free intersections that ensure safety, comfort, and accessibility for everyone.

A barrier-free intersection is designed to ensure universal accessibility, providing safe and convenient passage for all users, including pedestrians, cyclists, and individuals with mobility restrictions. Barriers to achieving such intersections can be categorized into four broad groups: (1) attitudinal-related barriers, (2) physical barriers, (3) policy- and program-related barriers, and (4) information and communications technology (ICT)-related barriers [4]. This study focuses on the ICT-related category, where two of the most prevalent challenges are crossing signal times that are too short for people with mobility restrictions and the lack of automatically triggered auditory signals at crossings.

To address these challenges, this paper proposes a detector-based intersection control system capable of dynamically adjusting crossing signal times and providing auditory cues. When individuals with mobility restrictions cross the intersection, the system extends the crossing interval to ensure sufficient time for safe passage. Similarly, when visually impaired people attempt to cross the intersection, the system trigger the auditory signal to aid in orientation and guide them safely across the street. By tailoring signal durations and amplifying auditory guidance to accommodate individuals with mobility restrictions, the proposed system enhances safety and accessibility across urban road networks, particularly for people with mobility restrictions.

However, achieving barrier-free intersections with detectors is fraught with challenges. Conventional RGB cameras, used in various adaptive traffic systems, struggle in low-light and adverse weather conditions, limiting their reliability to detect pedestrians, especially at night or during harsh weather. Since these systems primarily depend on the RGB camera, the poor quality of the captured image under such conditions leads often to failures in key object detection tasks [9], [10], [11], [12]. Furthermore, privacy concerns associated with RGB camera systems discourage their deployment in sensitive urban spaces.

To overcome these limitations, we propose the incorporation of infrared sensors, which are merely affected by external illuminating and environmental conditions [13]. Although thermal imaging offers robustness in adverse conditions, its unique characteristics, such as the lack of color and fine texture details, as well as difficulties in distinguishing objects with similar heat signatures, still pose significant hurdles for

Manuscript received May 13, 2025. This work was funded in part by the "KIMONO-EF" Project by the Federal Ministry for Digital and Transport under Grant 45AVF3005A-E.

Xiao Ni is with the Faculty of Mathematics and Computer Science, University of Münster, 48149 Münster, Germany (e-mail: xiao.ni@uni-muenster.de), and also with University of Applied Sciences Erfurt, 99085 Erfurt, Germany.

Carsten Kühnel is with the Faculty of Business, Logistics and Transport, University of Applied Sciences Erfurt, 99085 Erfurt, Germany (e-mail: carsten.kuehnel@fh-erfurt.de).

Xiaoyi Jiang is with the Faculty of Mathematics and Computer Science, University of Münster, 48149 Münster, Germany (e-mail: xjiang@uni-muenster.de).

accurate detection [14]. In addition, while high-performance infrared cameras available for consumer applications are expensive, their resolution is still limited to a maximum of 640×480 pixels, significantly lower than RGB cameras, which can achieve resolutions of up to 4 megapixels.

To address these issues, we propose a lightweight real-time object detection framework built upon the YOLOv8 architecture, which is chosen for its optimal trade-off between speed and accuracy. This framework is specifically designed to tackle the inherent limitations of thermal imaging, such as low resolution, minimal texture details, and difficulties in distinguishing small objects. To achieve this, we integrate the SPPFCSPC module [15], the SPD-Conv module [16], and the Triplet-Attention mechanism [17], which enhance feature extraction, object representation, and differentiation, improving overall detection performance.

For an accurate detection of individuals with various mobility restrictions in intersections, we construct a thermal dataset focused on people with disabilities. This thermal dataset comprises 11196 manually annotated images, capturing a diverse range of people with mobility aids, assistive devices, and other characteristics specific to the mobility restrictions.

This paper offers three primary contributions:

- We build a specialized thermal infrared dataset focused on people with mobility restrictions, addressing critical gaps in existing datasets by incorporating previously underrepresented scenarios and environmental conditions.
- We propose a detector-based intersection control system that dynamically adjusts crossing signal times and increases the volume of auditory cues, specifically designed to improve accessibility for individuals with mobility and visual impairments.
- We propose a novel framework that addresses low resolution and limited texture in thermal imaging by integrating SPPFCSPC, SPD-Conv, and Triplet-Attention modules. Experiments validate its superior performance and robustness.

The remainder of this paper is structured as follows: Section II provides a brief overview of thermal infrared datasets, datasets specifically focused on people with mobility restrictions, real-time object detection methodologies. Section III introduces the proposed thermal dataset TD4PWMR, describing its collection process, annotation strategy, and dataset characteristics. Section IV presents a novel approach to intersection management that leverages thermal imaging-based pedestrian detection to dynamically adjust traffic lights, accommodating the specific needs of individuals with mobility restrictions. Section V elaborates on the proposed thermal object detection framework, detailing its key components. Section VI conducts a comprehensive performance evaluation, highlighting the effectiveness of our proposed model through extensive experiments. Finally, Section VII concludes the paper by summarizing key findings.

II. RELATED WORK

A. Thermal Datasets

There is currently no standardized and specialized thermal dataset for pedestrians in traffic environments. However,

TABLE I
COMPARISON OF THE PROPOSED THERMAL DATASET TD4PWMR TO EXISTING THERMAL DATASETS.

	OSU	FLIR	KAIST	DENSE	Ours
Resolution	320×240	640×512	640×480	640×480	640×512
Bit Depth	8	8	14	8	8
Total Frames	8544	9711	8970	11500	11196
Availability	✓	✓	✗	✗	✓

several existing datasets can be used to perform a simple evaluation of thermal detectors.

- 1) OSU Color-Thermal: The OSU dataset [18] is a thermal and RGB image fusion dataset designed for object detection and tracking. It focuses exclusively on pedestrians and all six videos are captured with a low-resolution fixed thermal sensor (320×240 pixels), the Raytheon PalmIR 250D, in static backgrounds.
- 2) FLIR: The FLIR dataset [19] is specifically designed for research on the visible and thermal sensor fusion. It contains 9711 thermal images across 15 different object categories, captured from a Teledyne FLIR Tau thermal camera with a resolution of 640×512 pixels.
- 3) KAIST: The KAIST dataset [20] integrates thermal imaging and other sensor modalities to address the challenges in autonomous and assisted driving for day and night. The dataset includes 8970 thermal images with the same resolution of 640×480 pixels, using a FLIR A655Sc thermal camera. However, The link provided in the paper for accessing the dataset and its associated toolkits is no longer functional.
- 4) DENSE. The DENSE dataset [21] is a multimodal adverse weather that includes data from a variety of sensors, such as cameras, LiDAR, radar and infrared sensors. The dataset was collected over a distance of 10000 km during driving in Northern Europe, providing a diverse range of challenging environmental conditions. Thermal images were captured from an Axis Q1922 FIR camera, which offers a resolution of 640×480 pixels, in dynamic and moving backgrounds. In total, the dataset includes 11500 thermal images. Although the evaluation tools remain available, the dataset's website is no longer accessible, restricting users from obtaining the dataset.

Table I provides a comparative overview of the proposed TD4PWMR dataset and existing thermal datasets based on various characteristics such as resolution, bit depth, total frames, and availability.

B. Datasets for People with Mobility Restrictions

Detecting people with mobility restrictions is a critical aspect of building barrier-free systems. Several research efforts have focused on datasets to advance this domain.

The Mobility Aid dataset [22], [23] is a specialized people with mobility restrictions in indoor environments, such as hospitals and public buildings. This dataset includes over 17,000 annotated RGB-D images, emphasizing privacy by leveraging depth data. This approach supports real-world applications

where RGB camera might be limited due to privacy concerns. This dataset includes five distinct classes: pedestrians, person in wheelchairs, pedestrian pushing a person in wheelchairs, person with crutches, and person using a walking frame. This dataset represents significant steps in addressing the challenges of detecting people with disabilities in structured environments. However, It still remains constrained to indoor scenarios, such as hospitals or controlled laboratory settings, limiting their applicability in outdoor or unstructured environments.

In [24], a custom dataset is built for disabled people with mobility aids like wheelchair, crutch, walking frame, walking stick, and mobility scooter. This dataset was manually annotated and contains 5,819 images. In this work, pedestrians and mobility aids are detected separately, and their association is determined based on their spatial proximity.

C. ICT-related Approaches for Barrier-free Intersection

Ensuring accessibility at urban intersections is critical to creating inclusive cities for individuals with mobility restrictions. This section will be devoted to describing research on ICT-related approaches to reducing barriers at intersections.

ICT-related barriers often pose more challenges for people whose disabilities affect hearing, speaking, reading, and who rely on alternative forms of communication [4]. Recent efforts have introduced smartphone-based accessibility solutions. For visually impaired people, 'smart' crosswalk systems have been designed with Bluetooth beacons that communicate with a custom smartphone app [25]. By measuring the strength of the received signal, these systems approximate the location of a user in relation to the crosswalk and provide audio or haptic guidance for safer navigation. However, many such applications require users to interact with complex interfaces, posing usability problems and limiting their reach [26].

An alternative approach proposed in [5] integrates stereo vision-based object detection with active Bluetooth beacons to localize people with disabilities in real time. In this system, active Bluetooth beacons not only support localization but also send disability-related information to the infrastructure, enabling the traffic light to deliver adaptive responses tailored to different impairment types. Nevertheless, this approach relies on people with disabilities obtaining and consistently carrying a dedicated disability identification device issued by local authorities, which may pose challenges in terms of convenience.

Inspired by these efforts, our solution incorporates thermal camera-based detection to dynamically adjust traffic signal durations for individuals with mobility restrictions and amplify auditory signals for visually impaired pedestrians. Unlike prior approaches, which rely on smartphones or tags, our solution requires no personal devices, thereby offering a privacy-preserving, and universally accessible framework for creating barrier-free intersections.

D. Real-time Object Detectors

YOLOv1 [27] marked a paradigm shift in object detection by introducing the first CNN-based one-stage detector

capable of achieving real-time performance. Over the years, the YOLO family has undergone significant improvement, consistently exceeding other one-stage detectors [28], [29] and establishing itself as the synonym for real-time object detection [30]. YOLO detectors are broadly categorized into two types: anchor-based [31], [32], [33], [34], [35] and anchor-free methods [36], [15]. Both categories effectively balance accuracy and speed, making them versatile tools for a wide array of practical applications.

In recent years, transformer-based detectors have attracted increasing attention. DETR [37], the first end-to-end detector built on Transformer, removes both hand-crafted anchors and non-maximum suppression. Although DETR offers clear advantages, it also faces challenges such as slow training convergence and high computational cost. Current DETR variants remain too compute-intensive for real-time detection. To address these limitations, RT-DETR [30] seeks to reduce computational cost and optimize query initialization, aiming to achieve performance comparable to real-time object detectors.

There are several specialized detectors designed for thermal images. However, [38], [39] are optimized for extremely low-resolution thermal images, such as 160×120 pixels, and rely on super-resolution techniques to enhance image quality. This approach is not suitable for our use case, as our thermal camera operates at a relatively high resolution compared to those scenarios. In [40], YOLOv3 model has been directly implemented for thermal object detection by retraining it on a thermal dataset. However, this approach does not adapt the model to the unique characteristics of thermal images, limiting its ability to fully leverage the potential of thermal data. Additionally, these papers of existing thermal detectors do not provide publicly available code implementations. For the reasons outlined above, we exclude them from the scope of this paper.

III. TD4PWMR: THERMAL DATASET FOR PEOPLE WITH MOBILITY RESTRICTIONS

To develop barrier-free intersections while protecting the privacy of pedestrians, it is crucial to establish a comprehensive thermal dataset that specifically focuses on pedestrians with mobility restrictions.

We cover various areas at intersections, including sidewalks, waiting zones, and intersection crossings, capturing each location during different time slots (sunrise, morning, afternoon, sunset, night, and dawn). Since the thermal contrast ratio is strongly influenced by ambient heating conditions, seasonal variations significantly affect the thermal signature of the scenes. To account for this, we captured images across all four seasons, ensuring a comprehensive dataset that reflects diverse environmental and thermal conditions. Furthermore, we deliberately balanced the number of images per season, aiming for an approximately equal distribution to minimize seasonal bias during model training and evaluation. Specifically, the dataset includes 3,086 images captured in spring, 2,584 in summer, 2,534 in autumn, and 2,992 in winter.

Our recording sensors are mounted on intersection poles, with a NVIDIA Jetson AGX Orin housed within the traffic

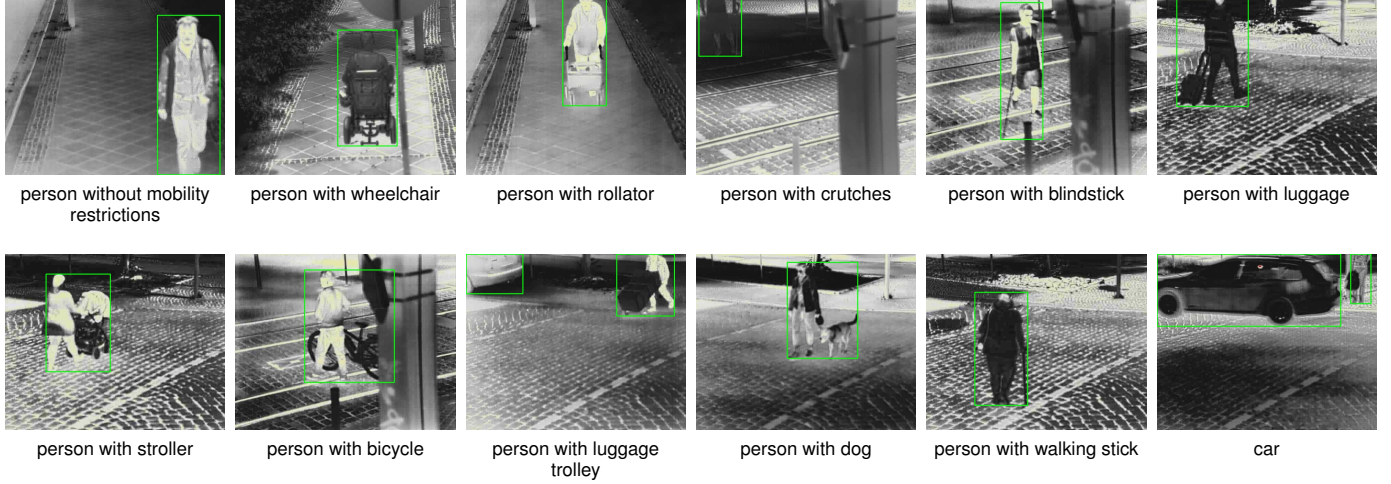


Fig. 1. One representative example from each class in the proposed TD4PWMR dataset.

control cabinet. To ensure stable and high-speed data collection, we utilized a solid-state drive (SSD), which offers significantly greater bandwidth compared to a traditional hard disk drive (HDD). In addition, we employ CAT7 cables between the edge computing device and the network switch, which connects to each sensor via CAT6 cables and aggregates the collected data. This setup reduces transfer latency and increases bandwidth. A total of eight infrared cameras (FLIR ThermiBot2) have been installed at the intersection. The camera operates within a spectral range of $7.5\text{--}13.5\ \mu\text{m}$ with a native resolution of 640×512 pixels and supports focal lengths ranging from $9\ \text{mm}$ to $35\ \text{mm}$ for versatile deployment scenarios. The infrared camera is barely affected by changes in illumination, making it a reliable choice for operation in adverse weather conditions.

To facilitate a variety of perception tasks, we manually annotated objects across all thermal video sequences. Prior to annotation, we established clear definitions for annotation targets and types. During the annotation process, the bounding boxes were drawn and the annotators assigned labels to objects by selecting from the following 12 predefined classes: person without mobility restrictions, person with wheelchair, person with rollator, person with crutches, person with blindstick, person with luggage, person with stroller, person with bicycle, person with shopping/luggage trolley, person with dog, person with walking stick, and car. For individuals assisted by mobility aids or restrained by mobility burden, the bounding box is drawn to encompass both the person and the associated object, treating them as a single object. Since these mobility aids and mobility burden are in constant use and often become partially occluded by the user's own body, making detection more challenging. By combining the person and the associated item into a single class, we leverage their relative positioning to enhance object detection accuracy and robustness. Figure 1 presents representative examples from each class, demonstrating the way we annotate the bounding boxes. In the images, facial features and specific identities remain difficult to discern due to the nature of thermal imaging. This characteristic ensures that surveillance and monitoring systems can operate

effectively without compromising personal privacy.

To facilitate effective model training and evaluation, the dataset is divided into training (80%) and evaluation (20%) subsets. The training subset is used to train detection models, ensuring that they learn to accurately identify and classify people with mobility restrictions across varied scenarios. The evaluation subset is reserved for testing, providing an unbiased measure of the model's performance in detecting and classifying people with mobility restrictions in unseen conditions.

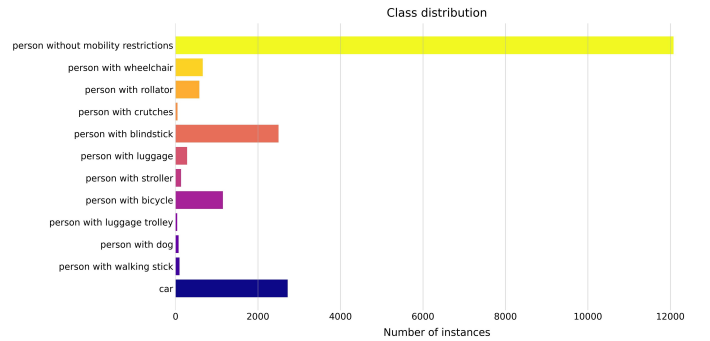


Fig. 2. Class distribution, highlighting the significant imbalance in the frequency of different classes.

Figure 2 reveals that the dataset exhibits a significant class imbalance, as the frequency of people with mobility restrictions is much lower compared to that of people without mobility restrictions. This imbalance poses a substantial challenge for training detection models, as they tend to focus disproportionately on the majority classes, leading to insufficient learning for the minority classes. Consequently, rare classes, such as people with various assistive devices, may not receive adequate training, resulting in poor detection performance and reduced model generalization for these underrepresented categories.

TABLE II
CATEGORIZATION OF PEDESTRIANS WITH MOBILITY RESTRICTIONS AND THEIR CORRESPONDING ADAPTATION STRATEGIES OF TRAFFIC LIGHT CONTROLLER.

Classes	Group	Strategy
Person with wheelchair Person with crutches Person with rollator Person with walking stick	Person with walking impairments	Extend the green time T_g of the current traffic direction, if person with walking impairments exists in the corresponding crossing area. The maximal extension of the green time $T_{g,ext}^{max} = 6 s$.
Person with blindstick	Person with visual impairments	Extend the green time T_g of the current traffic direction and increase the volume of audible guidance, if person with visual impairments exists in the corresponding crossing area. The maximal extension of the green time $T_{g,ext}^{max} = 8 s$.
Person with luggage Person with stroller Person with bicycle Person with luggage trolley Person with dog	Person with mobility burden	Extend the green time T_g of the current traffic direction, if person with walking impairment exists in the corresponding crossing area. The maximal extension of the green time $T_{g,ext}^{max} = 3 s$.

IV. DETECTOR-BASED SMART TRAFFIC LIGHT FOR PEOPLE WITH MOBILITY RESTRICTIONS

The traffic light controller is equipped with two cameras at each crossing area, one for each direction of pedestrian movement. These cameras are mounted atop the traffic light poles to provide a wide field of view and ensuring comprehensive coverage of the crosswalk. In the case of a four-way intersection, this configuration results in a total of eight cameras.

Table II categorizes pedestrians with mobility restrictions into three primary groups: people with walking impairments, people with visual impairments, and people with mobility burden. The first group includes individuals who use assistive devices such as wheelchairs, crutches, rollators and walking sticks. The second group consists of visually impaired pedestrians, identified by the use of blind sticks. The third group comprises individuals with mobility burden, such as those carrying luggage, pushing strollers, bicycle and luggage trolleys, or walking with dogs. Each class is assigned a corresponding group based on their needs and challenges at crosswalks.

To enhance intersection accessibility and safety, the traffic light controller applies tailored adaptation strategies for each group. The 95% design principle, widely used in ergonomic and infrastructure design, ensures accommodations for most users while preventing excessive customization needs [41]. Our measurements indicate that the sufficient crossing time for 95% of pedestrians with walking impairments is 5.8 seconds longer than that of the general population. For 95% of visually impaired pedestrians, crossing takes an average of 7.9 seconds longer, while 95% of pedestrians with mobility burden require an additional 2.7 seconds. Due to hardware constraints, traffic light controllers typically operate on integer-second timing intervals. As a result, green time extensions are rounded to the nearest whole second to ensure compatibility with existing infrastructure and to maintain synchronization across traffic control systems. Accordingly, the traffic light extends the green time by up to 6 seconds for individuals with walking impairments upon detection in the crossing area. When visually impaired pedestrians are detected in the crossing area, the maximum green phase extension is increased to 8 seconds, accompanied by an audible signals to aid navigation. Similarly,

pedestrians with mobility burden trigger a maximum green time extension of 3 seconds if they are detected in the crossing area. These adaptive strategies enhance safety and accessibility of the intersection while minimizing disruptions to overall traffic flow.

When multiple pedestrians with mobility restrictions from different categories are detected in the crosswalk simultaneously, the traffic light controller prioritizes the adaptation strategy of the category with the highest priority. Based on an evaluation of mobility capability and safety concerns, the priority order is as follows: people with visual impairments receive the highest priority, followed by people with walking impairments, and finally, people with mobility burden.

This prioritization is driven by the level of vulnerability and the ability of each group to navigate intersections safely. Visually impaired pedestrians face the greatest challenges, as they rely primarily on auditory cues and tactile guidance, making it difficult to judge traffic conditions, detect signal changes, or react quickly to unexpected obstacles. Their increased reliance on external assistance justifies the longest possible green light extension and enhanced audible guidance signals to ensure they cross safely. People with walking impairments face significant crossing challenges due to reduced movement speed and increased physical effort required to navigate curbs and uneven surfaces. However, they can still rely on visual cues and have some degree of maneuverability, making them less vulnerable than visually impaired pedestrians. Consequently, they receive a moderate extension of green time to accommodate their slower pace while ensuring smooth traffic flow. Finally, pedestrians with mobility restrictions move at a slightly reduced pace but do not face significant physical or sensory barriers as the other two groups. Thus, they receive the shortest green time extension to minimize traffic disruption while still allowing additional crossing time when necessary.

To detect pedestrians with mobility restrictions, we employ an efficient detection algorithm. When the detector identifies a pedestrian belonging to a target group, it classifies the individual and generates a bounding box around them. We define the middle bottom point of the bounding box as the pedestrian's position. The crosswalk area is predefined by outlining a polygon-shaped region on the street map, and the system checks whether the position of the detected pedestrian

falls within this designated area. This verification determines whether a person with mobility restrictions is present in the crosswalk, allowing the system to trigger appropriate adaptation strategies.

To increase the robustness of the traffic light control, we establish a reliable detection mechanism to determine the presence of pedestrians with mobility restrictions in the crosswalk. Once pedestrians with mobility restrictions are detected for the first time during the green phase, they are considered existing in the crossing area. However, accurately determining when they have fully exited the crosswalk remains a challenge. Due to occlusions and detection failures, the system can mistakenly conclude that they have already crossed when they are still in the intersection.

One potential solution to this issue is tracking-by-detection, which maintains pedestrian identities across frames and it could help to determine whether they have exited the crosswalk. However, this approach is not feasible in our system due to computational constraints. The detection algorithm is deployed on an edge computing device, the NVIDIA Jetson Orin, which has limited processing power. Given that the intersection is equipped with eight cameras, the system must process multiple video streams simultaneously. As a result, the interval between two processed frames for each camera is too large, making it impractical to apply tracking-by-detection algorithms that rely on high frame rate consistency.

To mitigate this, we implement a multi-frame validation approach, where the system only confirms the absence of pedestrians with mobility restrictions if they remain undetected for N consecutive frames after their initial detection in the crosswalk. This helps prevent premature termination of the extended green phase, ensuring sufficient crossing time. However, this approach also introduces a trade-off: If pedestrians with mobility restrictions have already exited the crosswalk and the system continues extending the green phase for the next N frames, potentially reducing traffic efficiency. Optimizing the parameter N is crucial to balancing robust protecting mechanism and smooth traffic flow, which we further analyze in section VI.

V. YOLO-THERMAL

Although thermal sensors provide distinct advantages over RGB cameras, they are also known for their limitations, such as the absence of color and fine texture details, lower resolution, and difficulty in distinguishing between objects with similar heat signatures. These limitations render conventional RGB-based detection methods less effective for deployment with thermal cameras. In this paper, we introduce a novel detection model built upon the YOLOv8 architecture [42], which is selected for its good balance between accuracy and speed. The model is specifically tailored for thermal imaging by incorporating several new modules.

A. Triplet-Attention

Triplet Attention [17] is an attention mechanism designed to enhance feature representation in deep learning models by capturing cross-dimensional interactions. Unlike conventional

attention modules that focus on either spatial or channel dependencies separately, Triplet-Attention integrates both by employing a three-branch structure.

Given an input tensor X of dimensions $C \times H \times W$, the module builds dependencies across spatial (H, W) and channel (C) dimensions through three key transformations:

1) *Cross-Dimension Interaction*: Traditional attention mechanisms like SENet [43] and CBAM [44] focus on independent spatial and channel-wise attention, potentially losing inter-dependencies. Conversely, Triplet Attention introduces a cross-dimension interaction strategy, ensuring that both spatial and channel dependencies are captured. The first branch models interactions between height H and channel C by permuting the input tensor, applying Z-Pool, and processing it through a convolutional layer. The second branch captures dependencies between width W and channel C similarly. The third branch follows CBAM's spatial attention paradigm, focusing on $H - W$ dependencies.

2) *Z-Pooling*: Instead of direct global average pooling (GAP), Triplet Attention employs a novel Z-Pool operation that concatenates both global max pooling (GMP) and GAP along a particular dimension:

$$Z - pool(X) = [\text{MaxPool}_{0d}(X), \text{AvgPool}_{0d}(X)], \quad (1)$$

where $0d$ denotes the dimension across which Z pooling operation is applied.

3) *Final Attention Weight Aggregation*: The refined feature representations from all three branches are aggregated as:

$$Y = \frac{1}{3}(\overline{X_1 W_1} + \overline{X_2 W_2} + X W_3) = \frac{1}{3}(\overline{Y_1} + \overline{Y_2} + Y_3), \quad (2)$$

where w_i denotes the attention weights computed in each branch and $\overline{Y_i}$ represents the 90° clockwise rotation applied to preserve the original input shape of $C \times H \times W$.

To maximize its effectiveness in the YOLO architecture, Triplet-Attention module should be strategically placed immediately after the backbone of YOLO, specifically the custom CSPDarknet53. The backbone is responsible for extracting low-level to mid-level features from the input image, forming the foundation for subsequent detection stages. By integrating the triplet-attention module at this point, the model achieves improved localization and classification performance in thermal images, effectively mitigating the challenges posed by low-resolution and ambiguous thermal object boundaries.

B. SPD-Conv

SPD-Conv, or Space-to-Depth Convolution [16], is a building block designed to replace traditional strided convolution and pooling operations in convolutional neural networks (CNN). It is particularly effective for tasks involving low-resolution images and small object detection, where preserving fine-grained details is crucial.

In our proposed thermal dataset, the image resolution is limited to 640×512 pixels, and because the dataset is collected in outdoor environments, pedestrians can often appear far from the camera, resulting in small representations within the images. These two characteristics (low resolution and small

object size) pose significant challenges for traditional down-sampling methods, which typically discard valuable fine-grained details.

SPD-Conv addresses these challenges by retaining fine-grained details during the space-to-depth transformation, re-distributing them into the channel dimension for enhanced feature representation. This ensures that even subtle features of small objects, such as distant pedestrians in thermal images, are effectively captured and processed. As a result, SPD-Conv is particularly well-suited for thermal datasets with these demanding characteristics, offering superior performance compared to conventional approaches.

The module consists of two primary components: a Space-to-Depth (SPD) transformation followed by a non-strided convolution layer.

1) *Space-to-Depth Transformation*: Let X be the input feature map with dimensions $S \times S \times C_1$, where S is the size of length and width, and C_1 is the number of channels. The SPD transformation slices X into submaps $f_{x,y}$ such that:

$$f_{x,y} = X[x : S : scale, y : S : scale], \quad (3)$$

where $scale$ is the downsampling factor.

This slicing generates $scale^2$ submaps, each of size $\frac{S}{scale} \times \frac{S}{scale} \times C_1$. These submaps are concatenated along the channel dimension, resulting in a transformed feature map X' with dimensions:

$$X' \in \mathbb{R}^{\frac{S}{scale} \times \frac{S}{scale} \times (scale^2 \cdot C_1)}. \quad (4)$$

2) *Non-Strided Convolution*: After SPD transformation, a non-strided (i.e., $stride = 1$) convolution layer with C_2 filters is applied, where $C_2 < scale^2 \cdot C_1$. This operation reduces the number of channels while preserving as much discriminative feature information as possible.

$$X'' = Conv(X'), \quad (5)$$

where $X'' \in \mathbb{R}^{\frac{S}{scale} \times \frac{S}{scale} \times C_2}$.

We replace all convolutional layers in the YOLO model with SPD-Conv layers. By leveraging space-to-depth transformation, SPD-Conv effectively captures fine-grained features. Furthermore, this replacement helps mitigate the limitations of low-resolution thermal imaging, improving the model's ability to detect and distinguish objects in thermal images.

C. SPPFCSPC

The Spatial Pyramid Pooling-Fast Cross Stage Partial Connections (SPPFCSPC) block [15] is a feature extraction component commonly utilized in deep learning models, particularly for object detection tasks. It combines Spatial Pyramid Pooling (SPP) [45], which captures multi-scale spatial features by applying pooling operations at various scales, with Cross Stage Partial Connections (CSPC) [46], which improves gradient flow and reduces computational redundancy by splitting feature maps into transformed and shortcut paths. This integration enables the module to extract both local and global contextual information effectively while maintaining lightweight computational requirements. By doing so, it significantly improves the ability to differentiate between objects

in thermal images, which is critical for accurately identifying objects in scenes with sparse texture details - a common challenge in imagery captured by thermal cameras. SPPFCSPC is particularly well-suited for handling objects of varying scales and optimizing feature reuse, making it ideal for real-time applications and scenarios with limited computational resources.

We replace SPPF (Spatial Pyramid Pooling Fast) with SPPFCSPC because SPPFCSPC integrates CSP connections into the SPPF block, improving gradient flow and computation efficiency. This typically leads to higher accuracy in object detection while maintaining competitive speed.

D. Quality Focal Loss

Additionally, there is still another issue that comes from our proposed thermal dataset: class imbalance. Traditional methods to address class imbalance include re-sampling, which involves oversampling the minority class or undersampling the majority class. However, both approaches come with significant drawbacks. Oversampling can lead to overfitting, while undersampling reduces available training data [47]. A more effective solution is Focal Loss [28], which adaptively reduces the loss contribution of well-classified samples and allows the model to focus on harder examples. The standard cross-entropy loss can be written as:

$$CE(p_t) = -\log(p_t), \quad p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases} \quad (6)$$

The loss contribution from well-classified samples (e.g., easy negatives with $p_t \approx 1$) dominates training, leading to under-representation of hard-to-classify samples, particularly the minority class. This can cause the model to bias heavily toward the majority class.

Focal Loss introduces a modulating factor $(1 - p_t)^\gamma$ to the cross-entropy loss, which dynamically reduces the loss contribution from well-classified examples ($p_t \rightarrow 1$) and focuses more on hard examples ($p_t \rightarrow 0$). The Focal Loss is defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t). \quad (7)$$

Quality Focal Loss (QFL) [48] is an extension of Focal Loss designed to improve object detection by jointly optimizing classification and localization quality. Unlike standard classification losses that treat class labels as binary values (0 or 1), QFL assigns a continuous quality score (e.g., Intersection over Union (IoU) between the predicted and ground truth bounding boxes) as the supervision target. This helps the model learn confidence-aware classification scores, reducing false positives.

QFL generalizes the standard Focal Loss to support continuous supervision labels instead of discrete ones. It is defined as:

$$L_{QFL} = -|y - \sigma|^\beta [y \log \sigma + (1 - y) \log(1 - \sigma)], \quad (8)$$

where $y \in [0, 1]$ is the soft target representing the IoU score between the predicted and ground truth bounding boxes, p denotes the predicted confidence score of the object, and β

TABLE III
COMPARISON WITH STATE-OF-THE-ART OBJECT DETECTION METHODS ON THE PROPOSED THERMAL DATASET FOR PEOPLE WITH MOBILITY RESTRICTIONS.

Model	Params (M) ↓	GFLOPs ↓	FPS ↑	AP^{val} ↑	AP_{50}^{val} ↑	AP_{75}^{val} ↑	AP_S^{val} ↑	AP_L^{val} ↑
YOLOv11-L	25.3	87.6	98.0	87.6	93.7	92.0	71.1	88.3
YOLOv11-X	56.9	196.0	51.3	87.5	94.4	90.8	73.9	88.1
YOLOv10-L	25.7	126.4	85.5	87.1	93.3	91.7	68.4	87.9
YOLOv10-X	31.6	169.9	53.8	86.1	91.4	90.0	69.1	86.9
YOLOv9-C	21.1	82.7	104.2	88.4	94.7	92.8	70.0	89.1
YOLOv9-E	54.0	173.4	47.6	88.8	94.2	92.3	75.5	89.5
YOLOv8-L	39.4	145.2	88.5	88.1	93.7	92.3	69.6	88.9
YOLOv8-X	61.6	226.8	55.9	88.4	94.7	92.3	72.2	89.3
YOLOv7-L	36.5	103.3	169.5	79.5	87.6	85.0	61.9	80.1
YOLOv7-X	70.9	188.2	105.3	82.9	90.7	88.6	64.3	83.4
YOLOv6-L	109.6	387.0	59.2	83.3	89.2	87.3	71.5	83.9
YOLOv6-X	171.0	603.4	38.2	81.7	87.8	86.0	69.3	82.3
RT-DETR (R50)	42.0	125.7	60.2	86.2	92.4	90.7	69.3	87.1
RT-DETR (R101)	60.9	186.3	45.7	88.6	94.0	92.6	69.1	89.5
Ours	42.1	150.1	90.1	89.1	95.1	93.0	71.5	89.9

is a modulating parameter that adjusts the weighting of hard examples. Unlike the original Focal Loss, which uses discrete labels ($y \in \{0, 1\}$), QFL enables a continuous label space, allowing more nuanced learning of classification confidence based on localization quality.

YOLOv8's loss function is composed of three key components: complete-IoU (cIoU) [49] for bounding box regression, Distribution Focal Loss (DFL) [48] for box refinement, and Binary Cross Entropy (BCE) [50] for classification and objectness. In this setup, we replace BCE with QFL to address class imbalance.

VI. EXPERIMENTS

A. Comparison with SOTA: Real-Time Object Detection

We conduct a comprehensive evaluation that includes both qualitative performance assessment and an analysis of computational complexity, measured in floating-point operations per second (FLOPs) and overall parameter count. To facilitate this evaluation, we benchmark state-of-the-art (SOTA) real-time object detection methods using our own thermal dataset TD4PWMR, training the models on its training partition and evaluating them on the test set. The corresponding quantitative results are presented in Table III, where the L and X models of YOLO detector are evaluated. These models have higher parameter counts, demand greater computational power, and result in longer inference times while achieving superior accuracy. To ensure a fair comparison, we retrain these models using the hyperparameters specified by the original authors. The detectors utilize a common input size of 640×640 pixels. The FPS is tested on GeForce RTX 2080 Ti GPU.

We evaluate model performance using COCO-style metrics. AP^{val} serves as the primary metric for object detection, computed by averaging the area under the precision-recall curve across multiple Intersection-over-Union (IoU) thresholds, ranging from 0.5 to 0.95 in increments of 0.05. AP_{50}^{val} and AP_{75}^{val} represent the average precision computed at fixed IoU thresholds of 0.50 and 0.75, respectively. Additionally, AP_L^{val}

and AP_S^{val} split the evaluation based on object size with the 96×96 pixel threshold, computing the AP for small and large objects, respectively.

The proposed model achieves the highest overall accuracy among all evaluated methods, attaining an AP^{val} of 89.1%. It also outperforms all competitors in terms of AP_{50}^{val} (95.1%) and AP_{75}^{val} (93.0%), highlighting its robustness across various IoU thresholds. Furthermore, in detecting large objects, our approach achieves the highest AP_L^{val} (89.9%).

In addition to its strong detection performance, the proposed model efficiently balances accuracy and computational complexity. With a moderate parameter count (42.1M) and computational cost (150.1 GFLOPs), it remains significantly more efficient than larger models such as YOLOv6-L (171.0M, 603.4 GFLOPs) while achieving superior accuracy. The model also achieves a high inference speed of 90.1 FPS, ranking among the fastest methods evaluated. It surpasses most competitors in inference efficiency, with only YOLOv7-L (169.5 FPS), YOLOv7-X (105.3 FPS), YOLOv9-C (104.2 FPS), and YOLOv11-L (98.0 FPS) achieving higher inference speed.

Compared to emerging real-time transformer-based models, our proposed approach demonstrates clear advantages. It consistently outperforms RT-DETR (R50) across all key evaluation metrics, achieving a higher AP^{val} (89.1% vs. 86.2%) and AP_{50}^{val} (95.1% vs. 92.4%), while delivering significantly improved inference speed (90.1 FPS vs. 60.2 FPS). Similarly, in comparison to RT-DETR (R101), our method attains a slight performance gain in AP^{val} (89.1% vs. 88.6%) and AP_{50}^{val} (95.1% vs. 94.0%), while nearly doubling the inference speed (90.1 FPS vs. 45.7 FPS).

Based on both qualitative and quantitative assessments, we conclude that the proposed object detection model outperforms SOTA performance while requiring fewer computational resources, achieving a lower parameter count and reduced computational complexity.

TABLE IV
ABLATION STUDY ON THE VALIDATION SET OF THE PROPOSED THERMAL DATASET FOR PEOPLE WITH MOBILITY RESTRICTIONS.

Method	Triplet-Attention	SPD-Conv	SPPFCSPC	QFL	$AP^{val} \uparrow$	$AP_{50}^{val} \uparrow$	$AP_{75}^{val} \uparrow$	$AP_S^{val} \uparrow$	$AP_L^{val} \uparrow$	FPS \uparrow
Baseline					88.1	93.7	92.3	69.6	88.9	88.5
YOLO-Thermalv1	✓				88.4	94.7	92.0	72.1	89.1	87.7
YOLO-Thermalv2	✓	✓			88.6	94.2	92.9	71.1	89.4	90.9
YOLO-Thermalv3	✓	✓	✓		88.7	94.7	92.4	71.0	89.4	90.1
YOLO-Thermalv4	✓	✓	✓	✓	89.1	95.1	93.0	71.5	89.9	90.1

TABLE V
ABLATION STUDY ON THE NUMBER OF CONSECUTIVE FRAMES IN THE MULTI-FRAME VALIDATION APPROACH.

Parameter N	Success rate (%) \uparrow	Latency (s) \downarrow	Real success rate (%) \uparrow	Average extended green time (s) for people with		
				walking impairment	visual impairment	mobility burden
1	13.4	0.8	32.3	1.6	2.2	0.8
2	77.2	1.2	95.4	2.9	3.9	1.3
3	92.7	1.6	96.6	3.1	4.2	1.6
4	94.6	1.9	96.9	3.3	4.3	1.7
5	94.8	2.3	97.0	3.5	4.5	1.9

B. Ablation Study

Table IV presents an ablation study conducted on the validation set of the proposed thermal dataset for people with mobility restrictions. The study evaluates the performance of different configurations of the YOLO-Thermal model, starting from YOLOv8, which serves as the baseline, and progressively adding different components.

1) *Triplet-Attention*: Adding Triplet-Attention improves feature representation, leading to a noticeable increase in AP^{val} from 88.1% to 88.4% and AP_{50}^{val} from 93.7% to 94.7%. Small object detection (AP_S^{val}) also improves from 69.6% to 72.1%, indicating better feature extraction for fine details. However, the FPS drops slightly to 87.7, suggesting a minor computational overhead.

2) *SPD-Conv*: Introducing SPD-Conv further enhances performance, boosting AP^{val} to 88.6% and AP_{75}^{val} to 92.9%, indicating improved precision at higher IoU thresholds. Additionally, the FPS is slightly improved to 90.9, demonstrating a more efficient inference process. Notably, SPD-Conv does not improve the detection accuracy of small objects (AP_S^{val}) and even degrades performance after integration into the model, contradicting the findings of the original paper. Nevertheless, it has been shown to enhance the overall detection accuracy of thermal images, aligning with the finding that it can improve detection in low-resolution images.

3) *SPPFCSPC*: The addition of SPPFCSPC continues the trend of improvement, improving AP^{val} to 88.7% and restoring AP_{50}^{val} to 94.7%. However, the FPS drops slightly to 90.1, reflecting an increased computational cost.

4) *QFL*: Finally, the incorporation of QFL significantly enhances the detection of hard-to-classify objects, pushing AP^{val} to 89.1% — the highest in this study. The improvement is evident across most metrics, with AP_{50}^{val} increasing to 95.1% and AP_L^{val} reaching 89.9%.

We deployed the proposed model on a local edge computing device (NVIDIA Jetson Orin), achieving an inference time of 40.4 milliseconds per frame. Given the parallel processing

of eight camera streams, the effective interval between consecutive frames from the same stream is 363.4 milliseconds, affecting the detection frequency and influencing the optimal configuration of the number of consecutive frames, denoted by parameter N , in the multi-frame validation approach.

To analyze the impact of N , we conducted an ablation study summarized in Table V. The table reports four key metrics: (1) success rate, defined as the proportion of events where the system successfully extends the green phase with sufficient additional time, under the assumption that the extension is terminated immediately at the start of the multi-frame validation approach, once no people with mobility restrictions are detected remaining in the crosswalk; (2) latency caused by the multi-frame validation approach and the integer-valued constraints of the traffic light controller; (3) real success rate, which is introduced to reflect the system's real performance under deployment conditions by adjusting the nominal success rate by accounting for latency; and (4) the average extended green time allocated to people with walking impairment, visual impairment, and mobility burden, respectively.

For $N = 1$, wasted time is minimal (0.8 s), but the success rate remains low (13.4 %) and the real success rate is only 32.3%, indicating frequent failures to adequately extend the green phase. At $N = 2$, performance improves significantly. The success rate rises to 77.2%, and the real success rate reaches 95.4%, effectively satisfying the design criterion of supporting at least 95% of people with mobility restrictions. The latency increases moderately to 1.2 s, which remains acceptable with respect to traffic efficiency. The average extended crossing times are 2.9 s for people with walking impairments, 3.9 s for people with visual impairments, and 1.3 s for people with mobility burdens. While increasing N to 3 marginally improves the success rate (92.7%) and the real success rate (96.6%), it introduces higher latency (1.6 s). Further increases to $N = 4$ and $N = 5$ result in diminishing returns, with the real success rate plateauing at approximately 97.0%, while latency rises significantly to 1.9 s and 2.3 s,

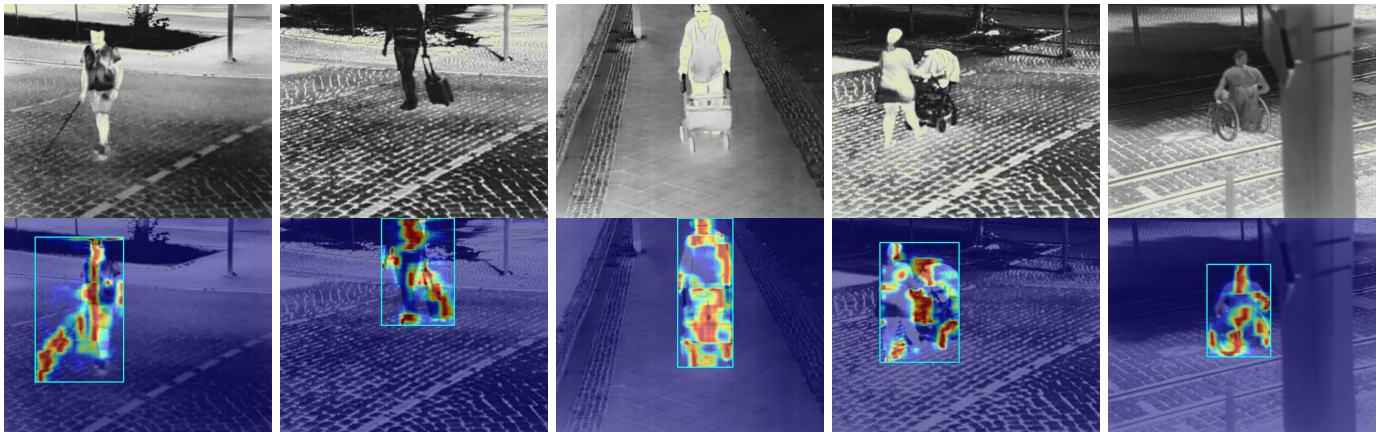


Fig. 3. Visualization of raw images (top) and corresponding LayerCAM heatmaps (bottom). Warmer colors indicate regions of higher activation.

respectively.

Notably, once $N > 2$, the observed increases in average extended green time are primarily driven by latency rather than substantive improvements in the reliability of the multi-frame validation approach. This outcome is undesirable, as it degrades traffic efficiency without yielding proportional gains in the real success rate, thereby offering no additional benefits in terms of safety and accessibility for pedestrians with mobility impairments. Furthermore, the latency introduced by the integer-valued constraints of the traffic light controller remains approximately constant at 0.5 s across different values of N . This indicates that the dominant source of increased latency at higher N originates from the delays intrinsic to the multi-frame validation process, rather than from the integer-valued constraints of the traffic light controller.

Therefore, setting N to 2 provides a balanced tradeoff between ensuring safety and accessibility of pedestrians with mobility restrictions and minimizing unnecessary latency, considering the constraints of our edge computing system and multi-stream processing setup.

C. Visualization

To further investigate the discriminative capacity of the learned features, we conduct a retrieval-based visualization experiment using LayerCAM [51], as shown in Fig. 3. LayerCAM is a gradient-based class activation mapping method that leverages gradient information from intermediate convolutional layers to generate spatially-aware heatmaps, highlighting the regions most influential to the model's prediction.

We visualize model attention in scenarios involving individuals using a mobility aid or with mobility burden. The top row of Fig. 3 shows the input thermal images, while the bottom row presents the corresponding LayerCAM heatmaps. The highlighted activations consistently localize around both the person and the associated object, particularly in the region that connects the two. This indicates that the model captures not only individual features but also the relational cues between the human subject and the associated object.

This observation supports our hypothesis that the relative position between the person and the associated object is critical for accurate detection. Moreover, the consistent activation patterns validate our annotation strategy, which treats the individual and the associated object as a single instance. This choice of annotation strategy effectively enables the detection model to better capture the semantic and spatial coherence of human-object interaction, thereby improving both detection accuracy and interpretability.

VII. CONCLUSION

This paper presents a novel thermal detector-based smart traffic light designed to enhance accessibility and safety for pedestrians with mobility restrictions. We introduce TD4PWMR, a specialized thermal dataset that captures diverse pedestrian scenarios across various environmental conditions, addressing critical gaps in existing datasets. To improve detection accuracy in thermal imaging, we propose YOLO-Thermal, an optimized object detection framework incorporating advanced feature extraction techniques such as Triplet-Attention, SPD-Conv, and SPPFCSPC. Experimental results demonstrate that YOLO-Thermal outperforms state-of-the-art models in both accuracy and efficiency. Finally, we implement an adaptive traffic light control strategy that dynamically adjusts green light durations and enhances auditory guidance based on real-time pedestrian detection. The proposed system significantly improves intersection accessibility while maintaining efficient traffic flow.

In future work, we aim to further enhance the system's real-time performance by integrating model compression techniques such as knowledge distillation and pruning. These approaches are expected to reduce computational overhead and accelerate inference, thereby mitigating delays introduced by the current multi-frame validation strategy. By minimizing unnecessary extensions of green light durations when no people with mobility restrictions still remain in the crosswalk, this optimization will contribute to a more responsive and efficient traffic control system that balances accessibility with urban mobility needs.

VIII. ACKNOWLEDGMENT

This work was funded by the Federal Ministry of Digital and Transport of Germany – 45AVF3005A-E.

REFERENCES

- [1] F.-Y. Wang, Y. Lin, P. A. Ioannou, L. Vlacic, X. Liu, A. Eskandarian, Y. Lv, X. Na, D. Cebon, J. Ma, L. Li, and C. Olaverri-Monreal, "Transportation 5.0: The dao to safe, secure, and sustainable intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 10, pp. 10262–10278, 2023.
- [2] N. Sharma and R. D. Garg, "Real-time IoT-based connected vehicle infrastructure for intelligent transportation safety," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8339–8347, 2023.
- [3] J. Roters, X. Jiang, and K. Rothaus, "Recognition of traffic lights in live video streams on mobile devices," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 10, pp. 1497–1511, 2011.
- [4] G. Asiedu-Ampem, A. Danso, J. Ayarkwa, D. Obeng-Atuah, E. Tudzi, and A. Afful, "Barriers to accessibility of urban roads by persons with disabilities: A review of the literature," *Journal of Transport & Health*, vol. 39, p. 101935, 2024.
- [5] D. Fernandez-Llorca, R. Quintero Minguez, I. Parra Alonso, C. Fernandez Lopez, I. Garcia Daza, M. A. Sotelo, and C. A. Cordero, "Assistive intelligent transportation systems: The need for user localization and anonymous disability identification," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 2, pp. 25–40, 2017.
- [6] R. Wunderlich, C. Liu, I. Elhanany, and T. Urbanik, "A novel signal-scheduling algorithm with quality-of-service provisioning for an isolated intersection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 536–547, 2008.
- [7] J. Wang, X. Guo, and X. Yang, "Efficient and safe strategies for intersection management: A review," *Sensors*, vol. 21, no. 9, 2021.
- [8] J. D. Kraemer and C. S. Benton, "Disparities in road crash mortality among pedestrians using wheelchairs in the usa: results of a capture-recapture analysis," *BMJ Open*, vol. 5, no. 11, 2015. [Online]. Available: <https://bmjopen.bmj.com/content/5/11/e008396>
- [9] K. Huang, L. Wang, T. Tan, and S. Maybank, "A real-time object detecting and tracking system for outdoor night surveillance," *Pattern Recognition*, vol. 41, no. 1, pp. 432–444, 2008.
- [10] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive YOLO for object detection in adverse weather conditions," in *AAAI Conference on Artificial Intelligence*, 2021.
- [11] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," in *European Conference on Computer Vision*, 2020, pp. 763–780.
- [12] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digital Signal Processing*, vol. 132, p. 103812, 2023.
- [13] R. Gade and T. Moeslund, "Thermal cameras and applications: A survey," *Machine Vision and Applications*, vol. 25, pp. 245–262, 01 2014.
- [14] N. Bustos, M. Mashhadi, S. K. Lai-Yuen, S. Sarkar, and T. K. Das, "A systematic literature review on object detection using near infrared and thermal images," *Neurocomputing*, vol. 560, p. 126804, 2023.
- [15] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu, "YOLOv6 v3.0: A full-scale reloading," *ArXiv*, vol. abs/2301.05586, 2023.
- [16] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Machine Learning and Knowledge Discovery in Databases*. Springer Nature Switzerland, 2023, pp. 443–459.
- [17] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 3138–3147.
- [18] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 162–182, 2007.
- [19] Flir thermal dataset. Accessed: 2025-02-06. [Online]. Available: <https://www.flir.com/oem/adas/adas-dataset-form/>
- [20] J. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.
- [21] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 679–11 689.
- [22] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard, "Deep detection of people and their mobility aids for a hospital robot," in *Proc. of the IEEE Eur. Conf. on Mobile Robotics*, 2017.
- [23] M. Kollmitz, A. Eitel, A. Vasquez, and W. Burgard, "Deep 3D perception of people and their mobility aids," *Robotics and Autonomous Systems*, vol. 114, pp. 29–40, 2019.
- [24] A. Mukhtar, M. J. Cree, J. B. Scott, and L. Streeter, "Mobility aids detection using convolution neural network (CNN)," in *International Conference on Image and Vision Computing New Zealand*, 2018, pp. 1–5.
- [25] F. Bustos, J. Gonçalves, and J. P. Coelho, "Improving pedestrian's cross-walk accessibility through digital fencing," in *Symposium of Applied Science for Young Researchers*, 2022.
- [26] T. Stefanov, S. Varbanova, and M. Stefanova, "Mobile devices supporting people with special needs," *International Journal of Advanced Computer Science and Applications*, vol. 13, 2022.
- [27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.
- [28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
- [29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.
- [30] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs beat YOLOs on real-time object detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2023.
- [31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [32] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 024–13 033.
- [33] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6517–6525, 2016.
- [34] J. Redmon, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [35] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.
- [36] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *ArXiv*, vol. abs/2107.08430, 2021.
- [37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [38] P. Shyam and H. Yoo, "Lightweight thermal super-resolution and object detection for robust perception in adverse weather conditions," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7456–7467.
- [39] Y. Jiang, Y. Liu, W. Zhan, and D. Zhu, "Improved thermal infrared image super-resolution reconstruction method base on multimodal sensor fusion," *Entropy*, vol. 25, no. 6, 2023.
- [40] M. Krišto, M. Ivacic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using YOLO," *IEEE Access*, vol. 8, pp. 125 459–125 476, 2020.
- [41] S. Pheasant and C. Haslegrave, *Bodyspace: Anthropometry, Ergonomics and the Design of Work*. CRC Press, 2018.
- [42] Ultralytics repository. Accessed: 2025-01-06. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision*, 2018, pp. 3–19.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

- [46] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, “CSPNet: A new backbone that can enhance learning capability of CNN,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1571–1580.
- [47] J. Johnson and T. Khoshgoftaar, “Survey on deep learning with class imbalance,” *Journal of Big Data*, vol. 6, p. 27, 03 2019.
- [48] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
- [49] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, “Enhancing geometric factors in model learning and inference for object detection and instance segmentation,” *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8574–8586, 2022.
- [50] D. R. Cox, “The regression analysis of binary sequences,” *Journal of the royal statistical society series b-methodological*, vol. 20, pp. 215–232, 1958.
- [51] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, “Layercam: Exploring hierarchical class activation maps for localization,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.



Xiaoyi Jiang received the bachelor’s degree from Peking University, Beijing, China, and the Ph.D. and Venia Docendi (Habilitation) degrees from the University of Bern, Bern, Switzerland, all in Computer Science. He was an Associate Professor with the Technical University of Berlin, Berlin, Germany. Since 2002, he has been a Full Professor of Computer Science with the University of Münster, Münster, Germany, where he was the Dean of the Faculty of Mathematics and Computer Science (2016–2023). His current research interests include pattern recognition, image analysis, and biomedical imaging. Dr. Jiang is an Editor-in-Chief of International Journal of Pattern Recognition and Artificial Intelligence. He also serves on the Advisory Board and Editorial Board of several journals, including International Journal of Neural Systems and Journal of Big Data. Previously, he has been Associate Editor for IEEE Trans. on Systems, Man, and Cybernetics - Part B / IEEE Trans. on Cybernetics, IEEE Trans. on Medical Imaging, and Pattern Recognition. He is a Senior Member of IEEE and a Fellow of IAPR.



Xiao Ni received the B.Eng. degree from Tongji University and is currently pursuing the Ph.D. degree from the University of Münster. His research interests include thermal image processing and object tracking.



Carsten Kühnel received his diploma in Industrial Engineering (Civil Engineering) from the Technical University of Darmstadt, Germany in 2004, followed by the Ph.D. degree in 2011 from the University of Kassel, Germany. He worked as the team leader for Innovative Technologies and Cooperative Systems at the Traffic Center Hessen and was responsible for the deployment of C-ITS in Germany within the C-ITS-Corridor project amongst others. Since 2017, he has been a Professor for Intelligent Transportation Systems at the University of Applied Sciences Erfurt

(FHE), Germany. Accompanied by two colleagues he leads the Institute for Transport and Spatial Planning at the FHE. He is a member of different working groups of the German Road and Transportation Research Association (FGSV), for example as the vice chair of the working group 3.1 Telematics.