International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Human Pose Estimation in Thermal Frames using Deep Learning Techniques

Dhananjay kumar Prasad*[1], Sonu Airen[2], Chandra Prakash Singar[3],Puja Gupta[3]

Submitted:07/06/2024 **Revised:**15/07/2024 **Accepted:**25/07/2024

Abstract: This paper presents a deep learning-based framework for human action recognition from thermal images, with a specific emphasis on pose estimation. The framework we proposed processes thermal images in stages. First, we extracted frames from the thermal video, followed by preprocessing the thermal frames, which included resizing, augmenting, and labelling action classes; labelling bounding boxes, and labelling 17 COCO-like keypoints. We developed a custom dataset with nine human actions including walking, sitting, lying, and an abnormal behaviour class. Lastly, we trained a YOLOv8-Pose model on the Thermal-IM dataset to both detect humans and estimate pose. Among the tested variants, the YOLOv8n-pose had the best accuracy-efficiency tradeoff. When evaluated on the Thermal-IM validation set, the YOLOv8n-pose achieved bounding box and pose mAP@0.5 average precision scores of 0.98 with mAP@0.5:0.95 scores of 0.96–0.97. It also achieved bounding box precision and recall values of 0.94 and 0.96, respectively, and pose precision and recall values of 0.93 and 0.96, respectively. The results show that the Deep Learning model can be effective for reliably detecting slight changes in human poses from thermal imagery in infinitely variable and difficult thermal conditions. Overall, the results confirm that pose-based analysis using thermal imagery is an appropriate, privacy-respecting and illumination independent, method for automated human behavior monitoring in complex indoor scenarios, with direct relevance for applications in surveillance, healthcare, and security fields of study.

Keywords: *Action Recognition, Human Pose Estimation, Thermal Imaging, Image Analysis, YOLOv8-Pose.*

1. Introduction

Thermal imaging, which captures variations in infrared radiation emitted from surfaces, has proven to be an essential tool for visual perception, especially in environments where traditional visible-light imaging fails. In contrast to color cameras that rely on ambient light, thermal cameras sense the heat that objects and living organisms radiate. This makes it possible to detect humans continuously in low-light conditions, through smoke and fog, or even in total darkness. Consequently, thermal imaging is particularly well-suited for applications that necessitate continuous observation in visually degraded environments. In the military, thermal imaging coupled with human pose estimation aids in enhancing situational awareness on the battlefield. It facilitates accurate tracking of individuals, enhances target localization, and assists in the automation of tactical assessment based on posture or movement. In search and rescue missions, such technology aids in the identification of victims obscured in rubble or out of sight, and also conveys information on their posture or condition, which can guide the level of urgency in response. Outside mission-critical uses, thermal-based human pose estimation (HPE) is also being introduced into civilian applications. In medicine, it supports passive, non-invasive monitoring of postures and physiological states in dark or low-visibility environments [1], offering valuable support in elder care and intensive care units. In sports and rehabilitation, thermal pose analysis aids in monitoring performance, detecting stress or fatigue early, and preventing injury. Smart homes also use thermal imaging for gesture recognition and natural user interaction, even in the absence of visible light, and while maintaining privacy and improving accessibility.

Present HPE systems are mainly based on deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models [2]. These have improved the field by providing high-accuracy pose detection of keypoints describing human joints. Pose estimation of multiple subjects, however, brings complexity, and models must detect multiple individuals and assign each joint to the appropriate person. To solve this, two main strategies are utilized: bottom-up and top-down. The bottom-up approach starts with the detection of all body joints within an image and then clusters them into

1Research Scholar, 2Associate Professor, 3Assistant Professor

1, 2,3Department of Information Technology , Shri Govindram Seksaria Institute of Technology and Science, Indore, M.P., India

Corresponding Author: Dhananjay kumar Prasad, dhananjay.prasad30121995@gmail.com

skeletons of individuals. Models such as OpenPose [3], HRNet [4], and CenterNet [5] adopt this approach. Top-down approaches, on the other hand, detect each person first and then estimate pose within the detected space. Though computationally more expensive, the approach tends to produce higher accuracy, with some examples being ViTPose [6], YOLOv8-Pose [7], and AlphaPose [8]. A closely related field is Human Action Recognition (HAR), whose aim is to automatically recognize and classify human activity recorded in visual data. Activities may be simple, such as sitting or walking, or more complex, involving several joints or synchronized movement. HAR is at the core of applications such as interactive media, autonomous systems, health behavior analysis, and smart living spaces. A lot of the advancement with HAR has been spurred by large annotated color datasets such as UCF101, KTH, and HMDB51 [11], which have made training and benchmarking increasingly advanced models possible.

Despite advancements in color-based pose estimation, applying these techniques to thermal or infrared (IR) images presents unique challenges. A significant limitation is the lack of large-scale, diverse, annotated IR datasets. Unlike color tasks supported by benchmarks like COCO [9] and MPII Human Pose [10], thermal datasets are often limited in scope and variability, restricting model generalization across different poses, body types, environments, and clothing conditions. Moreover, thermal imagery differs fundamentally from visible-light images. It lacks color and fine texture, instead encoding information based on heat patterns that can be influenced by environmental and physiological factors. These differences often result in low-contrast images with blurred outlines, making keypoint localization more difficult. Additionally, self-occlusion, a common issue in pose estimation, is exacerbated in thermal images. When body parts are close together, their thermal signatures may overlap, complicating joint assignment and increasing prediction ambiguity. Given these challenges, this study investigates the application of human pose estimation and action recognition techniques in thermal imaging contexts. The goal is to adapt and optimize deep learning models to operate effectively on thermal data, even in the absence of large-scale training sets. Through this research, we aim to advance the capabilities of thermal vision systems in domains ranging from public safety and healthcare monitoring to smart environments and defense analytics.

The rest of this paper is structured as follows. **Section II** reviews related work on thermal human detection, pose estimation, and action recognition. **Section III** explains the proposed method in detail. **Section IV** presents the experimental setup, results, and analysis. Finally, **Section V** concludes the paper and discusses potential future work.

2. LITERATURE REVIEW

Human action recognition and detection in thermal and infrared imagery has attracted growing research interest due to its potential in night-time surveillance and security applications. A variety of approaches leveraging deep learning, sensor fusion, and innovative feature extraction have been developed to address the inherent challenges of limited illumination and low-contrast thermal data.

Manssor et al. [13] solved the problems of pedestrian detection in thermal infrared images by enhancing the Tiny-YOLOv3 model. They augmented it with channel-wise contrast enforcement and paired it with a hybrid architecture consisting of PDM-Net and TIE-Net. Darknet-53, in this configuration, was tasked with extracting strong feature representations, while PDL-Net carried out classification operations. The approach effectively minimized loss of information during the early stages of processing, leading to more consistent detections, particularly in low-light environments where visible-spectrum detectors perform poorly. Imran et al. [14] proposed a four-stream deep learning architecture that pairs CNN and BiLSTM networks for detecting global and local motion patterns. Their approach infused dense optical flow-based features in the forms of SSDI and SDFDI, which allowed for encoding spatial and temporal information more holistically. By dividing video clips into segments and processing them through parallel CNN-BiLSTM streams, their system was able to fuse complementary features and deliver better action recognition across a wide range of activities. Krišto et al. [15] compared some of the top object detection models, including YOLOv3, by retraining them on thermal images recorded under different weather conditions such as rain, fog, and clear nights. Their findings suggested that YOLOv3 represented a good accuracy-speed trade-off and therefore was an efficient choice for real-time surveillance systems.

Batchuluun et al. [16] sought to recover skeletal keypoints from thermal video. They did so by converting single-channel thermal frames into three-channel inputs appropriate for a Joint-GAN model and subsequently generating joint and skeleton data with it. These were then passed through a CNN-LSTM architecture, allowing it to recognize complex human actions accurately.

Ding et al. [17] designed a thermal infrared system that was aimed at recognizing airport apron activities. Their pipeline began with the use of tracking algorithms to identify moving individuals from the background and subsequently extract spatiotemporal features within short time windows. These were input into a deep network with stacked LSTM layers to identify longer-term temporal patterns in order to aid the classification of walking, standing, and operational movements behaviors.

One of the most important works in this field is the work of Liu and Ostadabbas [18], who created the SLP dataset. The dataset is made up of thermal, visible, depth, and pressure images taken from 109 subjects who were lying in bed under three different conditions, namely uncovered, thinly covered, and fully covered conditions. Each person performed multiple poses across three main categories supine, left side, and right side leading to a total of 14,715 images. Their findings showed that visible images produced strong results when no cover was present, but performance dropped considerably when blankets were used. In such cases, thermal images were more effective for pose detection. Despite its usefulness, the SLP dataset has low variability since all data were recorded in the same environment with identical sensor settings, which limits its applicability to other contexts. Building on this dataset, Liu et al. [20] examined how combining different sensing

modalities including visible, thermal, depth, and pressure information could enhance pose estimation accuracy. Their experiments confirmed that multimodal input improves performance. Nevertheless, their work remained restricted to in-bed monitoring due to the dataset's narrow scope. In an effort to further improve thermal pose estimation, Chen et al. [21] compiled a large dataset containing 24,000 pairs of thermal and visible images recorded indoors. The visible images were high-resolution, whereas the thermal images had much lower resolution (80 × 60 pixels). To generate labels, OpenPose [3] was used on the visible images in the training set, while a subset of 2,000 test images was manually annotated. They proposed the ThermalPose model, which adapts OpenPose for thermal data. Experimental results showed that visible-based models achieved better accuracy under good lighting conditions, but in low-light or dark settings, ThermalPose outperformed all other approaches because visible cameras failed to detect people. However, the dataset's limited manually labeled subset and its indoor-only nature pose challenges for broader application.

To help address the lack of thermal data, Kniaz et al. [19] developed ThermalGAN, a generative adversarial network capable of translating visible images into thermal representations to support tasks like person re-identification. Their approach incorporated segmentation masks to estimate average temperatures for each object and to model variations within them. They introduced the ThermalWorld dataset, which includes over 15,000 visible–thermal image pairs with corresponding object annotations. Although this work shows promise in supplementing training data, its evaluation relied mostly on subjective judgments of image realism rather than objective performance metrics, making its practical impact on pose estimation uncertain.

Mehra et al. [22] also explored the benefits of fusing thermal and depth information for pose estimation. They developed a smaller dataset of 1,000 labeled images divided into training, validation, and testing sets. Using a modified version of the part affinity fields detector, they demonstrated that combining thermal and depth modalities resulted in more accurate detection than thermal input alone. However, the dataset's annotations covered only five keypoints per person, which limits its ability to support more detailed pose estimation. Several benchmark datasets have been introduced to facilitate research in thermal human detection and pose estimation. Each offers distinct characteristics suited for different application scenarios.

The CAMEL dataset [24] contains 26 sequences of paired color and thermal videos, totaling over 23,000 annotated frames, with around 7,775 precisely aligned pairs. Captured at 336 × 256 resolution and 30 fps in the LWIR spectrum, it features both indoor and outdoor urban settings under diverse lighting and weather conditions. The KAIST dataset [25] provides 95,000 aligned color–thermal image pairs and over 103,000 annotated pedestrian bounding boxes. Acquired at 640 × 480 resolution and 20 fps, it includes dynamic outdoor scenes captured from a moving vehicle across varying times of day and weather.The OTP dataset [27] offers 6,090 thermal images with bounding boxes and 17 keypoints per person, covering over 14,000 human instances in challenging outdoor conditions. It includes a range of activities, occlusions, scale variations, and environmental diversity. The LLVIP dataset [26] supports pedestrian detection and color–thermal fusion in low-light conditions, featuring 15,438 aligned image pairs from nighttime scenes. Its extension, LLVIP-POSE (LLVIP-P) [23], is the largest thermal pose estimation dataset to date, with over 26,000 annotated poses across training and test sets.

3. METHODOLOGY

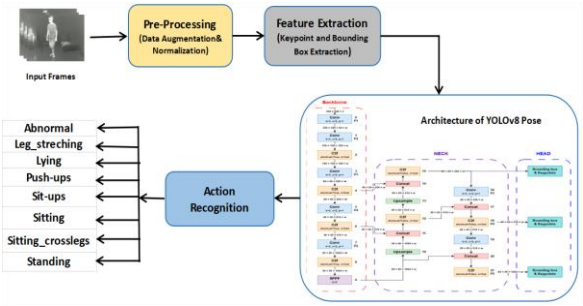


Fig.1. The Proposed Architecture of Human Action Recogniton

A. YOLOv8-Pose Overview

There are three parameters of YOLOv8-pose that determine the version: depth_multiple, width_multiple, and max_channels. The depth_multiple parameters decide how many bottleneck blocks are in the C2f block. The width_multiple and max_channels parameters define the output channels. The yolov8 stem is comprised of two convolution blocks with stride 2, kernel size 3. These two blocks create the origins of features and reduce the input resolution. The stage component in YOLOv8 is structured using the C2f block. The 8 stages are blocks no. 2, 4, 6, 8, 12, 15, 18, and 21. The stages in the backbone (blocks no. 2, 4, 6, and 8) utilize shortcuts while the neck (blocks 12, 15, 18, and 21) does not. Using shortcuts or not is based on seemingly sensible, valid results obtained from trial and error to try to achieve optimal. Downsampling for YOLOv8 is accomplished using a convolution block with a stride of 2 and a kernel size of 3. A stride of 2 will yield an output spatial resolution that is half the size. After the final block on the backbone, SPPF (Spatial Pyramid Pooling Fast) is used at the neck to give a multi-scale representation from the feature map. When pooling features at different scales, SPPF allows the model to capture features at different levels of abstraction. There are a few concat and upsample blocks on the neck. Upsampling increases the resolution of the feature map. YOLOv8 uses the nearest neighbor technique to conduct upsampling. This method fills the new pixels in a larger feature map by copying the value of neighboring pixels. Feature maps are concatenated with concat. The resolution does not change, however, the number of channels will increase when concatenating feature maps. YOLOv8 has three heads. The first head is connected to block No. 15 and detects small objects. The second head is connected to block No. 18 and detects medium objects. The third head is connected to block No. 21 and detects large objects. After these predictions, the model applies Non-

Maximum Suppression (NMS) to remove overlapping boxes and discard low-confidence results. This produces clean and reliable detections. The overall design of YOLOv8-Pose makes it well-suited for real-time human analysis in thermal imagery, especially in applications like activity monitoring, surveillance, and anomaly detection.

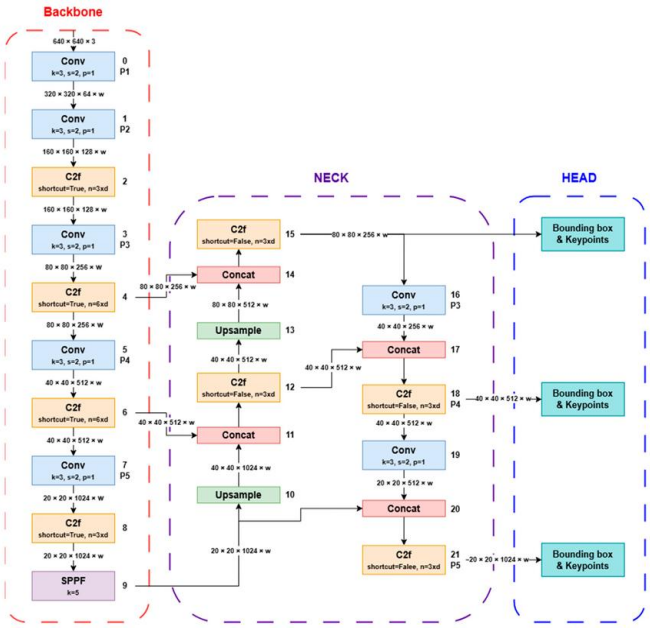


Fig. 2. YOLOv8Pose Architecture

B. Data Acquisition

This work used the thermal component of the Thermal Indoor Motion (Thermal-IM) dataset [12] to develop a human action detection system specifically designed for indoor environments. The thermal sequences were recorded with a Hikvision DS-2TD4237T-10 camera at a frame rate of 15 frames per second and a resolution of 288 × 384 pixels. To overcome challenges caused by low illumination and background clutter, only the thermal data were considered, even though the dataset also includes color and depth channels. The dataset comprises 783 video segments totaling over 560,000 thermal frames (approximately 10.4 hours) and captures actors performing everyday activities across various room configurations and camera positions. As seen in Fig. 3, this diverse and practical thermal-only dataset enables robust training and evaluation of models for pose estimation and action recognition

C. Data Extraction

53 thermal video samples were processed to construct a structured dataset for Human Action Recognition (HAR). Each video was accompanied by a corresponding JSON annotation file with temporal labels defining the start and end of various human actions. These actions were annotated into nine pre-defined classes: Abnormal, Leg_stretching, Lying, Push-ups, Sitting, Sitting_crosslegs, Sit-ups, Standing, and Walking. A Python script was designed to extract frames and corresponding short video samples automatically from the annotated areas. The source videos were captured using a thermal infrared camera in MPEG-4 (.mp4) format, resolution 288 × 384 pixels, and captured at 15 frames per second (FPS) in the 7.5–14 µm spectral band. All the extracted frames were resized to a standard 640 × 640 pixels to ensure consistency for model input. The output was organized into class-specific directories to ensure a clean and well-annotated dataset appropriate for keypoint extraction and action classification. This processed dataset was used as the foundation for training and testing the YOLOv8-Pose model on thermal human action sequences. The video input was broken down into separate frames based on its frame rate, which is the number of frames recorded per second.

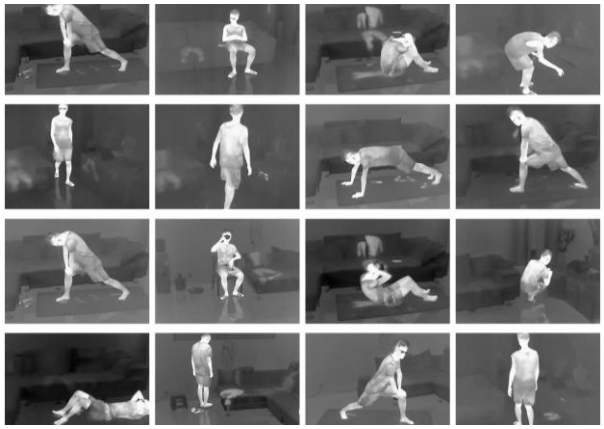


Fig. 3. Sample Images From Thermal-IM Dataset [12].

D. Data Preprocessing

To standardize the input for the YOLOv8-Pose model, all thermal images were resized to a fixed resolution of 640×640 pixels using a custom script built with the Pillow library. This resizing ensured uniform spatial dimensions across the dataset, facilitating efficient training and inference. To enhance model generalizability, standard data augmentation techniques are applied to the thermal images. Horizontal flipping simulates mirrored movement, while Gaussian noise is introduced to mimic real-world thermal sensor distortions as seen in Fig 4. These augmentations were implemented using the PyTorch and torchvision libraries.

(a) (b) (c)

Fig. 4. Preprocessed image (a) Original (b) Horizontal Filp (c) Gaussian Noise

After preprocessing, the final dataset consisted of 9,414 thermal images categorized into nine distinct human action classes as shown in table 1, including both simple and complex movements. The dataset was organized into class-specific directories, making it suitable for supervised learning tasks. The preprocessing steps helped reduce overfitting, increased robustness to real-world conditions, and provided a consistent and diverse foundation for pose estimation and action recognition in thermal environments.

TABLE 1. Summary of Image Distribution Across Action Classes

Class	Number of Images
Abnormal	873
Leg_streching	1,749
Lying	414
Push-ups	777
Sit-ups	1692
Sitting	1350
Sitting_crosslegs	468
Standing	312
Walking	1761
Total	9414

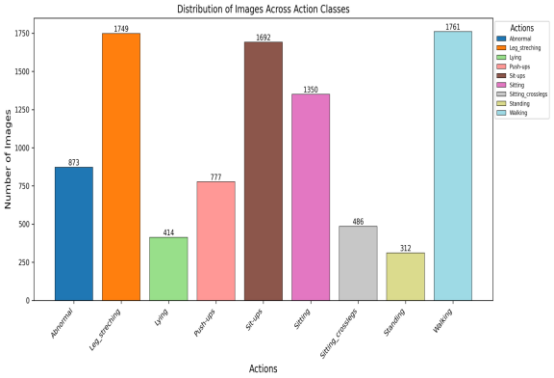


Fig.5. Distribution of Images Across Action Class

E. Keypoint and Bounding Box Annotation

In this work, we manually prepared a dataset consisting of 9,414 thermal images, each annotated with bounding boxes and 17 human keypoints according to the widely adopted COCO keypoint format. These keypoints capture critical anatomical landmarks such as the nose, eyes, shoulders, elbows, wrists, hips, knees, and ankles. Due to the nature of thermal imagery, where body outlines and joint positions are often less distinct, precise annotation proved to be a challenging and time-intensive task. For initial annotations, we utilized a YOLOv8m-pose model pre-trained on the COCO dataset to automatically detect bounding boxes and estimate keypoint positions. These initial predictions were then carefully reviewed and corrected through manual refinement to ensure accuracy, especially in cases where keypoints were missed or incorrectly positioned. A confidence threshold of 0.7 was used to filter reliable detections. The results were converted into normalized YOLO format, with annotations saved as .txt files corresponding to each image. Only frames containing valid detections were retained, and the dataset was organized into separate images and labels directories. This structured format enabled effective training and evaluation of the pose estimation model.

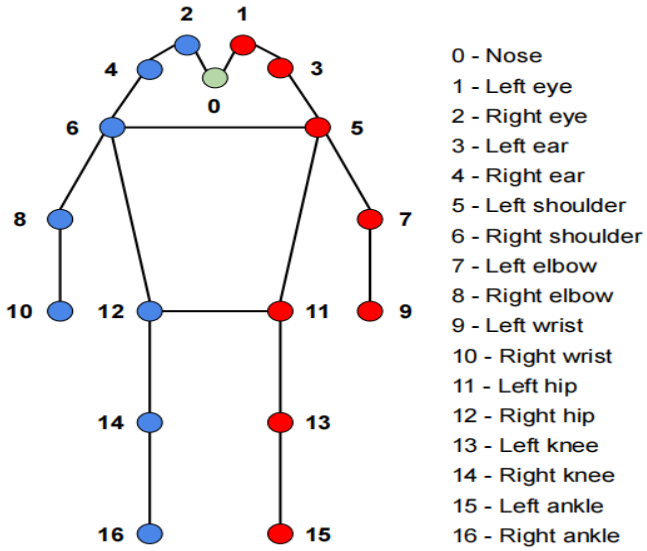


Fig.6. Keypoints Annotation Format used for Thermal-IM Dataset

Algorithm 1: Algorithm for Thermal Image YOLOPose Annotation

Input: YOLOPose results on 640×640 image
Output: YOLO-style .txt file with bbox + 17 keypoints

Steps:
Start
Set image width W = 640, **height** H = 640
For each detection in results:
Extract bounding box → (x_center, y_center, width, height)
Normalize:
• x_center = x_center / 640
• y_center = y_center / 640
• width = width / 640
• height = height / 640
For each of 17 keypoints (x_i, y_i):
• Normalize $x_i = x_i$ / 640, $y_i = y_i$ / 640
• Set visibility $v_i = 2$
Format output line as:
→ class_id x_center y_center width height x_1 y_1 v_1 x_2 y_2 v_2 ... x_{17} y_{17} v_{17}
Save the line to .txt file
Stop

After training, the models are employed to predict keypoints from test samples. The coordinates of keypoint are skeletal descriptors of the underlying human action. Through spatial configuration and temporal evolution of keypoints analysis, actions are classified by posture and movement patterns. This completes the HAR pipeline to allow automatic action recognition from thermal video streams.

F. Evaluation Metrics

The quantitative performance of the suggested YOLOv8-Pose model was evaluated with typical object detection and human pose estimation metrics. These metrics give a general idea about the accuracy of the model to detect people and predict anatomical keypoints from thermal images and the computational complexity in real-time applications.

- **Object Detection Evaluation**

The performance of the suggested YOLOv8-Pose model was evaluated based on typical object detection and human pose estimation metrics. The metrics provide a comprehensive assessment of the model's accuracy in detecting humans and estimating anatomical keypoints from

thermal images and its computational cost for potential real-time application.

Precision (P) = the number of true positive detections among all the predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall (R) is the ratio of correct positive detections out of all actual ground-truth instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Where:

- **TP:** True Positives (correctly detected persons),
- **FP:** False Positives (incorrect detections),
- **FN:** False Negatives (miss detections).

In order to evaluate detection performance at various levels of localization accuracy, mean Average Precision (mAP) was computed over various levels of Intersection over Union (IoU) ranging from 0.50 to 0.95 with step size 0.05:

$$\text{mAP}^{@[0.50:0.95]} = \frac{1}{10} \sum_{i=0.50}^{0.95} AP_i \tag{3}$$

- **Pose Estimation Evaluation**

The accuracy of 17-keypoint human pose estimation was evaluated using the **Object Keypoint Similarity (OKS)** metric, which measures the similarity between predicted and ground-truth keypoints while accounting for object scale and keypoint visibility:

$$\text{OKS} = \frac{\sum_i \exp(-\frac{d^2}{2s^2 k_i^2}).\delta(v_i) > 0)}{\sum_i \delta(v_i > 0)} \tag{4}$$

Based on the Object Keypoint Similarity (OKS), the following evaluation metrics were computed to assess the model's performance in human pose estimation: **AP_{50}** and **AP_{95}**, which represent the Average Precision at OKS thresholds of 0.50 and 0.95, respectively; **Mean Average Precision (mAP)**, computed across multiple OKS thresholds; and **Average Recall (AR)**, which measures the proportion of visible keypoints correctly predicted by the model. In the OKS formulation, d_i denotes the Euclidean distance between the predicted and ground-truth keypoints, ***s*** represents the object scale (bounding box area), k_i is a keypoint-specific falloff constant, v_i indicates the visibility of the keypoint, and **δ(.)** is the indicator function. These metrics collectively provide a comprehensive evaluation of the model's accuracy and consistency in detecting human keypoints from thermal images.

- **Loss Function**

The total loss function combined bounding box loss, classification loss, distribution focal loss, pose keypoint loss, and keypoint objectness loss. The final loss was computed as:

$$TotalLoss = \lambda_{box}.BoxLoss + \lambda_{cls}.ClsLoss + \lambda_{dfl}.dflLoss + \lambda_{pose}.PoseLoss + \lambda_{kobj}.kobjLoss \tag{5}$$

Here, λ represents the loss weights used to balance each component in the total loss: λ_{box} for bounding box loss, λ_{cls} for classification loss, λ_{dfl} for distribution focal loss, λ_{pose} pose keypoint loss, and λ_{kobj} for keypoint objectness loss.

4. RESULT AND DISCUSSION

A. Implementation and parameter Settings

Before starting the training process, it was necessary to adjust properly the model hyperparameters to provide stable performance. During this work, an input resolution of 640 × 640 pixels was used to find a balance between accuracy of detection and computational load in all experiments. A mini-batch size of 32 was used, which is a good balance between the utilization of GPU memory and model convergence. Stochastic Gradient Descent (SGD) was used as the optimizer, with momentum equal to 0.9 to help speed up learning and weight decay equal to 0.0005 to promote generalization. An initial learning rate of 0.01 was used, and each model was trained for 100 epochs until the validation metrics converged. To make the models more robust and limit overfitting, various data augmentation strategies were employed, including random horizontal flip, scaling, mosaic augmentation, and color changes. Automatic Mixed Precision (AMP) was activated during training to reduce memory usage and speed up computation.

All the experiments were conducted on the Lightning AI cloud platform with NVIDIA Tesla T4 and NVIDIA L4 GPUs and CUDA acceleration. In this research, the performance of four YOLOv8 Pose models were evaluated: YOLOv8n-pose (nano), YOLOv8s-pose (small), YOLOv8m-pose (medium), and YOLOv8l-pose (large). The models were executed in Python 3.10.10 with PyTorch 2.7.0 using the Ultralytics YOLOv8 framework (version 8.3.133). The computational environment consisted of 64-bit Intel Xeon-class processors and 32 GB of RAM. Automated checkpointing and validation were conducted during training to track key metrics, including precision,

recall, mean Average Precision (mAP), and pose estimation accuracy.

B. Training and Validation on IM-Thermal Dataset

The IM-Thermal dataset was carefully partitioned to ensure robust model evaluation and fair comparison across different YOLOv8-pose variants. Specifically, the dataset was split into three subsets: 70% for training (6,589 images), 10% for validation (941 images), and 20% for testing (1,884 images). This stratified division maintained an even representation of all nine human action classes, including Abnormal, Leg Stretching, Lying, Push-ups, Sitting, Sitting Cross-legged, Sit-ups, Standing, and Walking. During training, the models learned to detect bounding boxes and estimate human pose keypoints in thermal images. The validation set was employed to monitor learning progress, fine-tune hyperparameters, and assess intermediate performance after each epoch. This approach was essential for minimizing overfitting and ensuring the models retained good generalization ability on unseen data. The reserved test set was only used in the final evaluation to measure accuracy and robustness in real-world scenarios.

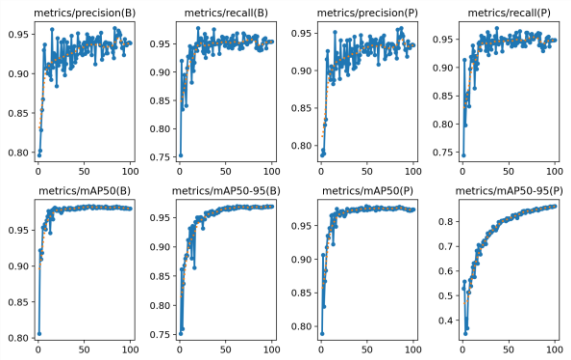


Fig.7. Model Performance Metrics for Bounding box (B) and Pose (P) Estimation over 100 Epochs.

Figure 7 illustrates the model's performance trends across key evaluation metrics. The results exhibit an initial phase of rapid improvement, followed by stabilization at consistently high values. Final performance metrics demonstrate excellent accuracy, with precision, recall, and mAP@0.5 achieving approximately 0.94, 0.96, and 0.98, respectively. Furthermore, the more stringent mAP@0.5:0.95 for pose estimation shows steady growth throughout training, ultimately reaching around 0.85.

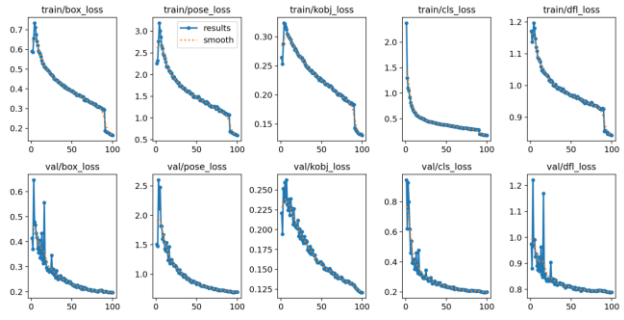


Fig.8. Training and Validation Loss Curves over 100 Epochs.

As shown in Figure 8 shows all Traning and validation loss curves display a sharp decrease in the initial epochs before converging steadily. This pattern indicates that the model learned the task effectively and generalized well to the validation data without significant overfitting.

Fig. 9. Bounding Box Precision vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

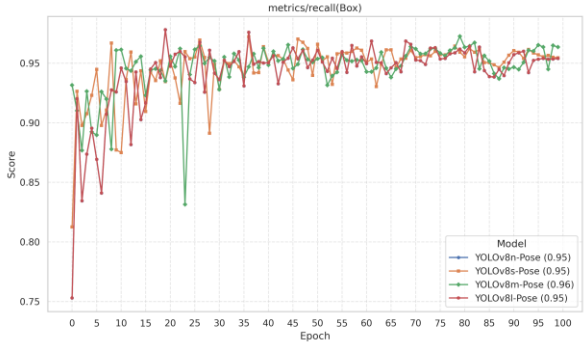


Fig. 10. Bounding Box Recall vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

Fig.11. Bounding Box mAP@0.5 vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

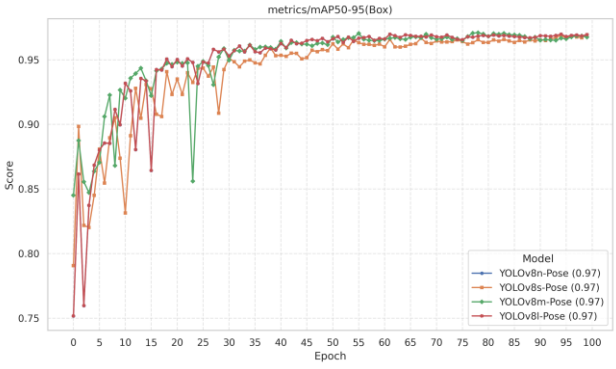


Fig.12. Bounding Box mAP@0.5:0.95 vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

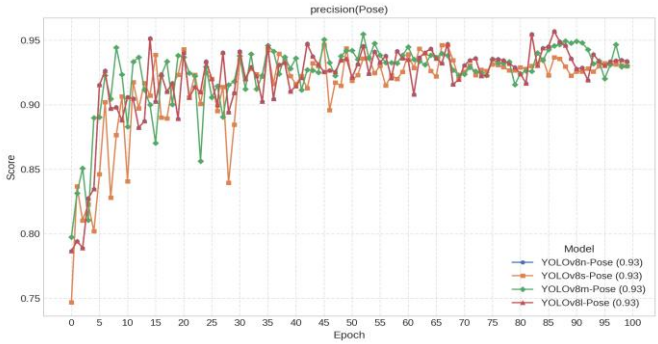


Fig.13. Pose Keypoints Precision vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

Fig.14. Pose Keypoints Recall vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

Fig.15. Pose Keypoints mAP@0.5 vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset

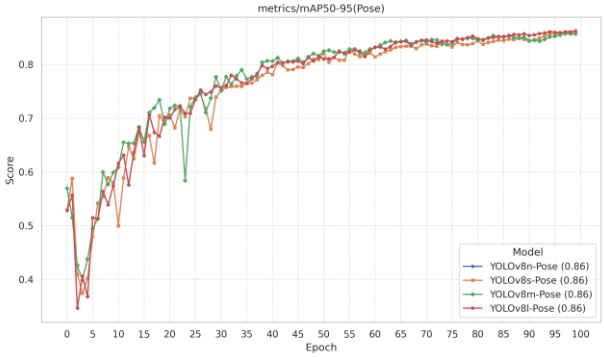


Fig.16. Pose Keypoints mAP@0.5:0.95 vs Epochs for YOLOv8-Pose models on Thermal-IM Dataset.

Figures 9 through 16 provide a detailed overview of how the YOLOv8-Pose models performed during training on the thermal human pose dataset. Throughout the experiments, all models showed gradual improvement in both object detection and pose estimation metrics. Among the four versions, the YOLOv8l-Pose consistently delivered stronger results, achieving higher precision, recall, and mean Average Precision (mAP) for both bounding box detection and keypoint localization. These outcomes suggest that the larger model is better equipped to handle the complexities of human detection and pose estimation in thermal imagery. Overall, YOLOv8l-Pose proved to be the most accurate and dependable across the tasks evaluated in this work.

TABLE 2. Comparison of Detection and Pose Estimation Performance Across Different YOLOv8-Pose Models on Thermal-IM Dataset

Model	**Box**				**Pose**			
	Precision	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**
YOLOv8n-pose	0.94	0.96	0.98	0.97	0.93	0.96	0.98	0.85
YOLOv8s-pose	0.94	0.96	0.98	0.97	0.93	0.95	0.97	0.86
YOLOv8m-pose	0.93	0.97	0.98	0.97	0.93	0.96	0.98	0.86
YOLOv8l-pose	0.94	0.95	0.98	0.97	0.93	0.94	0.97	0.86

TABLE 3. Validation Performance Metrics of the YOLOv8n-pose model on theThermal-IM Dataset.

Class	**Box**				**Pose**			
	Precision	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**
All	0.936	0.960	0.982	0.966	0.931	0.955	0.975	0.852
Abnormal	0.936	0.96	0.991	0.969	0.936	0.96	0.991	0.877

Leg strechi ng	0.93 8	0.9 59	0.98 9	0.983	0.93 8	0.9 59	0.98 9	0.922
Lying	0.91 3	0.9 60	0.98 1	0.972	0.91 3	0.9 60	0.98 1	0.899
Push-ups	0.90 3	0.9 09	0.94 3	0.897	0.90 3	0.9 09	0.94 0	0.734
Sitting	0.99 3	0.9 94	0.99 5	0.985	0.97 8	0.9 79	0.98 8	0.914
Sitting crossle gs	0.91 4	0.9 44	0.98 5	0.985	0.91 4	0.9 44	0.97 4	0.850
Sit-ups	0.94 8	0.9 20	0.97 8	0.955	0.94 8	0.9 20	0.97 8	0.693
Standi ng	0.92 4	1.0 00	0.99 2	0.976	0.92 4	1.0 00	0.99 2	0.957
Walki ng	0.95 7	0.9 56	0.98 6	0.971	0.92 8	0.9 27	0.94 6	0.822

TABLE 4. Validation Performance Metrics of the YOLOv8s-pose model on theThermal-IM Dataset.

Class	**Box**				**Pose**			
	Precision	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**
All	0.935	0.955	0.981	0.969	0.930	0.950	0.974	0.857
Abnormal	0.936	0.986	0.992	0.978	0.936	0.986	0.992	0.895
Leg streching	0.936	0.968	0.989	0.982	0.936	0.968	0.989	0.923
Lying	0.908	0.960	0.980	0.972	0.908	0.960	0.980	0.878
Push-ups	0.922	0.861	0.934	0.907	0.922	0.861	0.924	0.733
Sitting	0.994	0.993	0.995	0.989	0.979	0.978	0.988	0.915
Sitting crosslegs	0.915	0.944	0.984	0.984	0.915	0.944	0.976	0.879
Sit-ups	0.936	0.915	0.973	0.951	0.936	0.915	0.973	0.699
Standing	0.920	1.000	0.995	0.985	0.920	1.000	0.995	0.956
Walking	0.945	0.964	0.987	0.975	0.916	0.935	0.952	0.834

TABLE 5. Validation Performance Metrics of the YOLOv8m-pose model on theThermal-IM Dataset.

Class	**Box**				**Pose**			
	Precision	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**
All	0.934	0.965	0.983	0.969	0.929	0.961	0.976	0.859
Abnormal	0.913	1.000	0.991	0.976	0.913	1.000	0.991	0.901
Leg streching	0.933	0.968	0.988	0.981	0.933	0.968	0.988	0.929
Lying	0.925	0.985	0.990	0.981	0.925	0.985	0.990	0.897
Push-ups	0.919	0.902	0.941	0.907	0.919	0.902	0.930	0.720
Sitting	0.996	0.993	0.995	0.987	0.989	0.985	0.991	0.923
Sitting crosslegs	0.913	0.944	0.984	0.984	0.913	0.944	0.976	0.874
Sit-ups	0.931	0.915	0.975	0.950	0.931	0.915	0.975	0.704
Standing	0.920	1.000	0.995	0.983	0.920	1.000	0.995	0.955
Walking	0.951	0.976	0.983	0.968	0.922	0.947	0.945	0.831

TABLE 6. Validation Performance Metrics of the YOLOv8l-pose model on theThermal-IM Dataset.

Class	Box				Pose			
	Precision	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5–0.95**
All	0.939	0.954	0.980	0.969	0.934	0.949	0.974	0.862
Abnormal	0.935	0.983	0.991	0.976	0.935	0.983	0.991	0.887
Leg streching	0.940	0.975	0.989	0.985	0.940	0.975	0.989	0.929
Lying	0.920	0.960	0.981	0.973	0.920	0.960	0.981	0.915
Push-ups	0.922	0.861	0.934	0.912	0.922	0.861	0.934	0.735
Sitting	0.996	0.993	0.995	0.987	0.989	0.985	0.993	0.919
Sitting crosslegs	0.912	0.944	0.982	0.982	0.911	0.944	0.975	0.869
Sit-ups	0.952	0.907	0.970	0.946	0.942	0.898	0.967	0.698
Standing	0.922	1.000	0.992	0.988	0.922	1.000	0.992	0.961
Walking	0.950	0.964	0.986	0.974	0.921	0.935	0.946	0.845

A comparative analysis of four YOLOv8-pose model variants (nano, small, medium, and large) was conducted. The results reveal a uniformly high level of performance with negligible differences between the model sizes. For the initial task of subject localization, all variants achieved an identical and excellent Box mAP@0.5 of 0.98. Similarly, for the final Pose based action classification, accuracy remained outstanding, with a mAP@0.5 score between 0.97 and 0.98 for all models. A marginal advantage for the larger models was only observed in the stricter Pose mAP@0.5–0.95 metric. These results demonstrate the model's robustness and accuracy in detecting human poses from thermal imagery under challenging indoor conditions.

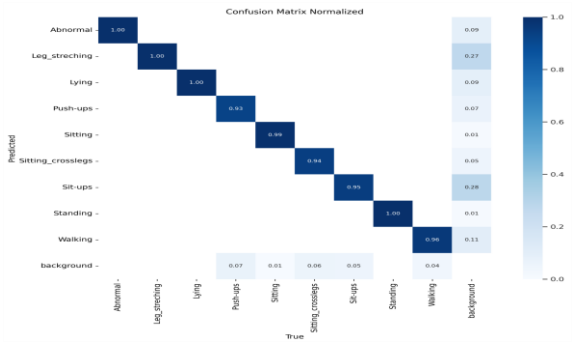
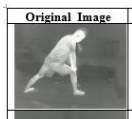
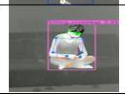


Fig.17. Confusion Matrix for Human Action Recognition

TABLE. 3. Class-wise best pose detection results showing original and predicted images with corresponding confidence scores

Original Image	Predicted Image	Class	Confidence
		Lying	0.9902
		Sitting	0.9806
		Walking	0.9833

Original Image	Predicted Image	Class	Confidence
		Leg streching	0.9808
		Abnormal	0.9878
		Sitting crosslegs	0.9863

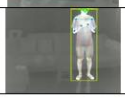
Original Image	Predicted Image	Class	Confidence
		Push-ups	0.9895
		Standing	0.9868

Table 3 presents predicted actions and confidence scores for a set of thermal images. The model shows strong performance across all classes, with confidence values mostly above 0.97. It achieves the highest confidence for Lying (0.9902), followed by Push-ups (0.9895) and Abnormal (0.9878). Even visually similar actions like Sitting_crosslegs and Leg_streching are classified accurately. These results indicate the model's effectiveness in recognizing various actions from thermal images, despite limited texture and visual cues.

5. CONCLUSION AND FUTURE WORK

In this work, an extensive analysis of several YOLOv8-Pose models, such as YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l, was performed on a thermal image dataset specially prepared for human action recognition. The models exhibited high detection precision, with box mAP@0.5 being more than 98% in all configurations and pose mAP@0.5 between 97.4% and 97.6%. Of particular

interest was YOLOv8l-pose, which posted the highest pose mAP@0.5–0.95 of 86.2%, indicating its best ability to correctly localize keypoints under adverse thermal imaging scenarios. The experimental results validate the efficacy of state-of-the-art pose estimation architectures in learning discriminative spatial representations from thermal data. In summary, this work proves that light models such as YOLOv8n-pose can still achieve competitive performance with faster inference speeds, making them appropriate for real-time applications where computational resources are scarce.

In future research, this work can be continued by acquiring larger, more varied thermal datasets of complex scenes and unusual actions to better enhance model robustness. Further, modeling temporal dynamics or graph-based approaches on pose sequences may further enhance action recognition on continuous video. Investigating domain adaptation from color to thermal data and model optimization for real-time edge deployment are also potential avenues to continue this research further.

Author contributions

Dhananjay Kumar Prasad: Conceptualization, Methodology, Software, Field study, Data curation, Writing-Original draft preparation, Validation, Field study. **Sonu Airen, Dr. Chandra Prakash Singar, Dr. Puja Gupta:** Visualization, Investigation, Writing-Reviewing, and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Shuangjun Liu and Sarah Ostadabbas, “Seeing under the cover: A physics guided learning approach for inbed pose estimation,” 2019

[2] E. Samkari, M. Arif, M. Alghamdi, and M. A. Al Ghamdi. Human pose estimation using deep learning: A systematic literature review. Machine Learning and Knowledge Extraction, 5(4):1612–1659, 2023.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1):172–186, 2021

[4] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. Proc. of the EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5686–5696, 2019.

[5] Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. arXiv preprint arXiv:1904.07850.

[6] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In Proc. of the Advances in Neural Information Processing Systems, 2022.

[7] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics yolov8, 2023.

[8] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6):7157–7173, 2023.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollar, ´ “Microsoft coco: Common objects in context,” 2015.

[10] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686–3693

[11] Srihari, P., 2022. Spatio-Temporal Information for Action Recognition in Thermal Video Using Deep Learning Model. International journal of electrical and computer engineering systems, 13(8), pp.669-680.

[12] Tang, Z., Ye, W., Ma, W.C. and Zhao, H., 2023. What happened 3 seconds ago? inferring the past with thermal imaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17111-17120).

[13] S. A. Manssor, S. Sun, M. Abdalmajed, and S. Ali, "Real-time human detection in thermal infrared imaging at night using enhanced Tiny-YOLOv3 network," Journal of Real-Time Image Processing, vol. 19, pp. 261–274, 2022.

[14] J. Imran and B. Raman, "Deep residual infrared action recognition by integrating local and global spatio-temporal cues," Infrared Physics & Technology, vol. 102, p. 103014, 2019.

[15] M. Krišto, M. Ivašić-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using YOLO," IEEE Access, vol. 8, pp. 125459–125476, 2020.

[16] G. Batchuluun, J. K. Kang, D. T. Nguyen, T. D. Pham, M. Arsalan, and K. R. Park, "Action recognition from thermal videos using joint and skeleton information," IEEE Access, vol. 9, pp. 11716–11733, 2021.

[17] M. Ding, Y. Y. Ding, X. Z. Wu, X. H. Wang, and Y. B. Xu, "Action recognition of individuals on an airport apron based on tracking bounding boxes of the thermal

infrared target," Infrared Physics & Technology, vol. 117, p. 103859, 2021.

[18] Gupta, P. and Kulkarni, N., 2013. An introduction of soft computing approach over hard computing. International Journal of Latest Trends in Engineering and Technology (IJLTET), 3(1), pp.254-258

[19] Liu, Y., & Ostadabbas, S. SLP: A Dataset for In-Bed Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.

[20] Gupta, P., Sharma, V. and Varma, S., 2022. A novel algorithm for mask detection and recognizing actions of human. Expert Systems with Applications, p.116823.

[21] Kniaz, V., Mizginov, R., & Afonin, S. ThermalGAN: Multimodal RGB-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW), 2018.

[22] Liu, Y., Shao, Z., & Ostadabbas, S. Multimodal In-Bed Human Pose Estimation under Blankets. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2249–2258, 2021.

[23] Chen, C., Xie, W., Yang, Y., & Liu, Y. ThermalPose: Estimating Human Pose from Thermal Images Using Self-Supervised Multi-Modal Learning. arXiv preprint arXiv:2109.10199, 2021.

[24] Singh, U., Gupta, P. and Shukla, M., 2022. Activity detection and counting people using Mask-RCNN with bidirectional ConvLSTM. Journal of Intelligent & Fuzzy Systems, 43(5), pp.6505-6520

[25] Mehra, D., Suri, S., & Gupta, R. Fusion of Thermal and Depth Data for Enhanced Human Pose Estimation. In: International Conference on Computer Vision Systems (ICVS), 2022.

[26] Singh, U, Gupta, P., Shukla, M., Sharma, V., Varma, S. and Sharma, S.K., 2023. Acknowledgment of patient in sense behaviors using bidirectional ConvLSTM. Concurrency and Computation: Practice and Experience, 35(28), p.e7819.

[27] Mickael Cormier, Caleb Ng Zhi Yi, Andreas Specker, Benjamin Blaß, Michael Heizmann, and Jurgen Beyerer. Lever- ¨ aging thermal imaging for robust human pose estimation in low-light vision. In Proceedings of the Asian Conference on Computer Vision, pages 67–83, 2024.

[28] Evan Gebhardt and Marilyn Wolf. Camel dataset for visual and thermal infrared multiple object detection and tracking.

[29] Gupta, P., Sharma, V. and Varma, S., 2022, September. An Algorithm for Counting People using Dense Nets and Feature Fusion. In 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1248-1253). IEEE.

[30] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[31] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3496–3504, 2021.

[32] Askat Kuzdeuov, Darya Taratynova, Alim Tleuliyev, and Huseyin Atakan Varol. Openthermalpose: An open-source annotated thermal human pose dataset and initial yolov8-pose baselines. In 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), pages 1–8. IEEE, 2024.