Leveraging Thermal Imaging for Robust Human Pose Estimation in Low-Light Vision

Mickael Cormier[1,3,4], Caleb Ng Zhi Yi[1], Andreas Specker[1,4], Benjamin Blaß[2], Michael Heizmann[3,1,4], and Jürgen Beyerer[3,1,4]

[1] Fraunhofer IOSB, Germany, {firstname.lastname}@iosb.fraunhofer.de
[2] Stahl-Holding-Saar, Germany, benjamin.blass@stahl-holding-saar.de
[3] Karlsruhe Institute of Technology, Germany, {firstname.lastname}@kit.edu
[4] Fraunhofer Center for Machine Learning, Germany

Abstract. Human Pose Estimation (HPE) is becoming increasingly ubiquitous, finding applications in diverse fields such as surveillance and worker safety, healthcare, sport and entertainment. Despite substantial research in HPE within the visible domain, there is limited focus on thermal imaging for HPE, primarily due to the scarcity and annotation difficulty of thermal data. Thermal imaging offers significant advantages, including better performance in low-light conditions and enhanced privacy, which can lead to greater acceptance of monitoring systems. In this work, we introduce LLVIP-Pose, an extension of the existing LLVIP dataset, to include 2D single-image pose estimation for aligned night-time RGB and thermal images, containing approximately 26k annotated skeletons. We detail our annotation process and propose a novel metric for identifying and correcting poorly annotated skeletons. Furthermore, we present a comprehensive benchmark of top-down, bottom-up, and single-stage pose estimation models evaluated on both RGB and thermal images. Our evaluations demonstrate how pre-training on grayscale COCO data with data augmentation can benefit thermal pose estimation. The LLVIP-Pose dataset addresses the lack of thermal HPE datasets, providing a valuable resource for future research in this area. The pose annotations and baseline code are available on github: https://github.com/MickaelCormier/llvip-pose.

1 Introduction

Human pose estimation (HPE) constitutes a critical task within the domain of computer vision, aiming to predict human poses from image or video sources. The ability to extract poses from visual data facilitates enhanced scene understanding, thereby garnering significant interest across various academic and industrial fields, including autonomous driving [28, 36], fall detection in healthcare [1], human-computer interaction [16], robotics [45], and surveillance [9]. Specifically, in surveillance, HPE enables skeleton-based activity recognition, autonomously triggering alerts that necessitate immediate human intervention. This proves highly advantageous as it replaces vision-based activity recognition that may otherwise contain sensitive biometric data [9]. Furthermore, HPE may

Fig. 1: Thermal image of a worker lying unconscious in a maintenance hall. Colleagues are not able to see him to call for help.

also be leveraged for worker safety through privacy-preserving monitoring systems, which ensure that the collected data cannot be misused against employees, thereby fostering a secure and ethical work environment. As shown in Fig. 1, detecting an injured worker or preventing collision with larger machines may prevent heavy injuries or fatalities. However, with enough context information, the identity of a person in thermal imaging may be inferred. Thus, HPE provides a second layer of anonymization. However, prevailing HPE techniques predominantly rely on images from the visible light spectrum, thereby rendering them susceptible to low-light conditions and obstructive weather phenomena [8, 18].

To address these limitations, thermal infrared cameras present a promising alternative. These cameras exhibit substantial robustness against the aforementioned conditions. This robustness is evident in tasks such as person detection, where thermal-based predictions demonstrate superior accuracy compared to their vision-based counterparts. Nonetheless, the potential of thermal imagery in the context of HPE remains largely underexplored.

The primary impediment to progress in this area is the scarcity of thermal-based HPE datasets. The acquisition of such datasets comes at considerable costs due to the challenges associated with manually annotating poses, where annotating a single pose can take up to 60 seconds [10]. Furthermore, the absence of a benchmark to compare thermal-based with vision-based HPE complicates the assessment of the respective advantages and disadvantages of these modalities. Thus, the LLVIP dataset [18] emerges as a dataset of particular interest, offering aligned images from both modalities and designed to investigate the impact of low-light conditions on pedestrian detection. However, it lacks pose annotations and necessitates modifications to serve as a suitable benchmark for HPE.

In this work, we propose to extend the aformentioned dataset into a new benchmark dataset called LLVIP-Pose. To this aim, pose annotations are incorporated, where poses are shared between the two modalities. Since manual pose annotation demand high costs, we propose a semi-autonomous workflow to accelerate the pose annotation and validation processes, where previously each

pose was annotated and validated by hand. For this purpose, pose predictions from HPE models are corrected by the annotators, eliminating the need to label a pose from scratch. To minimize the validation costs, an outlier detector based on the principle of anatomically correct skeleton structure leverages the relations between bone segments to highlight poorly annotated poses. This results in the contribution of 26k pose annotations for thermal imaging, which is by far the largest dataset for HPE in thermal imaging. Alongside the acquisition of this dataset, two studies are conducted to examine the viability and capabilities of thermal-based HPE models. First, our benchmark experiments compare the performance of thermal-based models against traditional RGB-based models featuring different multi-person pose estimation methodologies, e.g., top-down, bottom-up, and single-stage approaches. Next, generalization experiments are conducted on the thermal images of the LLVIP-Pose dataset with models pretrained on the COCO dataset to provide insight into the generalization capabilities of RGB-trained models for thermal applications.

The main contributions of this paper are summarized as follows:

- We propose LLVIP-Pose, the first large scale visible-infrared paired dataset for HPE in thermal imaging and low-light vision.
- We propose a semi-automated workflow to label and validate skeletons in still-images in LLVIP-Pose.
- We evaluate the experimental results of recent methods for HPE on LLVIP-Pose, and find that the dataset is a challenge for both HPE in RGB-low-light vision and in thermal imaging, although the latter shows impressive results.

2 Related Work

2.1 2D Human Pose Estimation

In HPE, human poses are predicted from images or videos by localizing keypoints (e.g., body parts or joints) and visualizing their connections based on a skeletal topology. Keypoint estimation is traditionally categorized into regression-based [22,27,33,40] or heatmap-based [7,35,41] methods. Regression-based methods predict keypoint coordinates directly, while heatmap-based methods infer them through multiple heatmaps, each representing the likelihood of a specific keypoint type. However, recents methods also use classification to perform discrete regression [23, 25]. Top-down approaches [23, 27, 35, 40, 41, 43] localize individuals first and then estimate their poses. In contrast, bottom-up approaches [5, 7, 15, 31] localize keypoints first and then group them. Top-down methods offer better accuracy but are slower with more individuals, while bottom-up methods are faster but less accurate. One-stage pipeline approaches [25, 26, 42] generate pose candidates without intermediate steps, processing both global and local context simultaneously. They are computationally efficient but suffer from the same accuracy issues as bottom-up approaches. HPE is predominantly applied to RGB images, leveraging rich color and texture information. However, its application to infrared thermal images remains largely unexplored. A major reason for this gap is the absence of thermal image datasets for HPE.

2.2 RGB Datasets

With growing interest in applications for HPE [38], numerous large-scale datasets have gained prominence [2, 3, 11, 21, 24, 34, 37, 44]. The COCO dataset [24], one of the most widely utilized, comprises over 200,000 images and 250,000 annotated poses. Similar to the MPII dataset [3], COCO features non-continuous images with common poses and frontal views. PoseTrack18 [2], based on the MPII dataset, includes continuous video frames capturing more complex real-life scenarios in controlled environments, such as sports events. COCO defines its own topology with 17 keypoints, including five on the head (nose, eyes, ears), which are challenging to detect in realistic scenarios with steeper camera angles. Consequently, the MPII and PoseTrack18 topologies simplify this by reducing head keypoints to two and three, respectively. In this work, we prefer this representation to facilitate manual annotation. These datasets primarily represent human poses in common and straightforward situations with favorable camera angles. OCHuman [44] and OCPose [44] address (self-)occlusion with frontal views in single, non-continuous images with two subjects. CrowdPose [21] features crowded scenarios in controlled environments like group photos or sports events. However, training on these datasets often transfers poorly to real-world surveillance scenarios, which involve steep camera angles, heavy (self-)occlusion, and dense crowds. Occlusion in crowded environments and complex poses remains a significant challenge in HPE [9] even for annotators labeling manually [10].

2.3 Thermal Infrared Datasets

The CAMEL dataset [14], modeled after the MOT challenge [11,20,29], includes 26 video sequences of RGB-thermal pairs, totaling around 23,437 annotated pairs, with 7,775 aligned. Captured at 336 $\times$ 256 pixels resolution and 30 fps in the LWIR band, it features indoor and outdoor urban environments under varying conditions and times. The KAIST dataset [17] offers 95,000 aligned RGB-thermal image pairs with 103,128 annotated bounding boxes for pedestrians, captured at 640 $\times$ 480 resolution and 20 fps in the LWIR band. It features outdoor scenarios with varying conditions and day times from a moving vehicle's perspective. The ThermalIM dataset [32] aims to infer past human motion via thermal residuals on objects. It consists of 783 video clips with thermal images at 384 $\times$ 288 pixels resolution and RGB at 1,920 $\times$ 1,080 pixels. It lacks bounding boxes, but provides 2D and 3D pose labels for RGB images, captured in three rooms with different actors, angles, and layouts. The OpenThermalPose dataset [19] includes 6,090 images with 14,315 annotated human instances. It provides bounding boxes and 17 keypoints, following COCO standards. The dataset spans various activities such as fitness exercises, multi-person interactions, and outdoor walking in different locations and weather conditions. However, the point of view is not suitable for surveillance contexts. The LLVIP dataset [18] addresses image fusion and low-light pedestrian detection, featuring 15,438 aligned image pairs in the LWIR band at a resolution of 1,280 $\times$ 1,024 pixels across 26 nighttime scenarios. It is possible to transfer labels from thermal

images to RGB images due to alignment, delivering bounding boxes usable for both modalities.

3 LLVIP-Pose

We propose LLVIP-Pose, an HPE extension of the LLVIP visible-infrared paired dataset for low-light vision. The train-test split is shared with the original LLVIP dataset. After annotation and dataset cleaning, 26,135 bounding boxes with poses are provided. The final training set includes 6,853 images pairs (18,633 persons) and the test set contains 3,462 image pairs (7,502 persons).

In the remainder of this section, the annotation and validation of the human poses is described. First, a semi-automated annotation process is proposed. Second, a distance metric is proposed in order to identify outliers in the manual annotation which may contain annotation errors. Finally, the validation workflow is described.

3.1 Annotation Process

A semi-autonomous pipeline is proposed that uses an HPE model to reduce the labor-intensive process of labeling each keypoint individually. Instead, the focus is on correcting the predicted poses from the HPE model. Predicted poses from existing RGB-based pose estimators delivered suboptimal results due to the low visibility in the RGB images. Therefore, paid-annotators manually labeled poses until a sufficient number of poses was reached to train an HPE model. As a compromise between high quality predicted poses and fast training, HRNetw48-udp model [35] is used, which is a top-down HPE model pre-trained on the COCO dataset. Prior to prediction, the initial width and height of the bounding boxes were increased by a tenth of their width and height to ensure that all body parts are enclosed within the bounding box. These RGB-based pose predictions were visualized on thermal images and corrected by the annotators for their 2D image coordinates and visibility score. Subsequently, a thermal-based pose estimator is trained to provide predictions for the remaining more challenging sequences.

3.2 Validation Process

After the initial pose annotation phase, a validation phase is carried out for quality assurance. Validation often incurs high time costs, as each individual pose is reviewed manually.

Anthropometric Detection of Outliers. Based on anthropometric measurements, we propose to evaluate the viability of poses based on the proportions of various body segments. Drillis *et al.* [12] provided estimates of different body lengths as a percentage of the body height H, as depicted in Fig. 2. In this work,

we re-contextualized the body segment lengths as ratios represented as the percentage of the shortest over the longest body segment. This new representation is necessary as the availability of body height H is not guaranteed in the poses, e.g., poses found at the edges of the image or occlusions blocking either the upper or lower body.

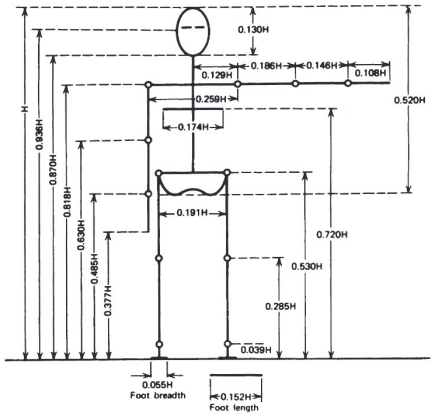


Fig. 2: Body segments with respect to body height H. (Source: [12, 39])

As the anatomical structure of a human skeleton consists of multiple body parts interconnected together at the joints (keypoints), the length of a body part affects the length of its neighbors. Therefore, the use of the hip-shoulder, thigh-torso and torso-leg ratios is proposed as these rations share a common connection with the keypoints of the hip. Ratios including the arms were omitted, as their high degrees of freedom in 3D space are not fully captured in 2D images. In addition to the three ratios, the hip-shoulder-angle is included as an additional term to evaluate the viability of the poses, where the angle difference between both body segments is calculated. In the case of LLVIP, the angle difference is approximately zero, as the three main actions performed are standing, walking, and riding.

The evaluator features a distance metric d_i to score the individual poses, where a lower score indicates a low difference between the values derived from the poses x_i and the groundtruth values y_i found in Tab. 1.

The pose and distance metric are depicted in Eq. (1) and Eq. (2).

$$\text{Metric} = \frac{\sum_{i=1}^{n_{\text{values}}} d_i(x_i, y_i)}{n_{\text{values}}} \tag{1}$$

$$d_i(x_i, y_i) = \begin{cases} 1.0 - f(|x_i - y_i| \mid 0, \sigma_{|x_i - y_i|}) \cdot \sigma_{|x_i - y_i|}\sqrt{2\pi}, & \text{if } x_i \text{ is present} \\ 1.0 - f(|x_{\text{approx},i} - y_i| \mid 0, \sigma_{|x_i - y_i|}) \cdot \sigma_{|x_i - y_i|}\sqrt{2\pi}, & \text{if } x_i \text{ is not present} \end{cases} \tag{2}$$

Table 1: Groundtruth values used in the pose validator.

Ratio	**Value**
Hip-Shoulder	0.737
Hip-Shoulder-Angle	0.0
Thigh-Torso	0.851
Torso-Leg	0.587

The variables and parameters found in the metrics are:

- x_i: calculated values
- y_i: groundtruth values
- $x_{\text{approx},i}$: approximated values
- $\mu_{|x_i-y_i|}$: mean of the error terms between calculated and groundtruth values
- μ_{x_i}: mean of calculated values
- $\sigma_{|x_i-y_i|}$: standard deviation of the error terms between calculated and groundtruth values
- σ_{x_i}: standard deviation of calculated values
- n_{values}: number of values
- n_x : number of available values $(1 - n_{\text{values}})$

Through the distance metric d_i each calculated value x_i is evaluated against its respective groundtruth value y_i. Their error term z_i is calculated based on the Manhattan Distance and is set into an unnormalized Gaussian with the mean $\mu_{|x_i-y_i|}$ and the standard deviation $\sigma_{|x_i-y_i|}$ of the error term, as depicted by Eq. 3. The resulting pose metric score is the average of the distance metric for each value.

$$f(z \mid \mu_{|x_i-y_i|}, \sigma_{|x_i-y_i|}) = \frac{1}{\sigma_{|x_i-y_i|}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{z-\mu_{|x_i-y_i|}}{\sigma_{|x_i-y_i|}}\right)^2} \tag{3}$$

Poses are classified as an outlier, if the calculated pose metric score exceeds a threshold, which is calculated based on the percentage of the standard deviation σ_{x_i} over the mean μ_{x_i} for the different values. This representation reflects the variability of the values, where a high variability indicates an inconsistent quality of pose annotations resulting in a lower threshold. For the hip-shoulder-angle, the threshold is calculated as the percentage of the mean μ_{x_i} over the standard deviation σ_{x_i}, as $\mu_{x_i} >> \sigma_{x_i}$. Formally, the threshold is computed as

$$\text{Threshold} = \frac{\sum_{i=1}^{n_{\text{ratios}}}(1 - \frac{\sigma_i}{\mu_i})}{n_{\text{ratios}}} \tag{4}$$

The calculation of the values relies on the availability of keypoints. If a ratio is excluded from the overall pose metric, multiple different thresholds would need to be assigned for each pose within a sequence. Thus, the omission of any value is undesirable as it will introduce inconsistency in the pose metric. For this reason, we propose to approximate the missing ratios or angle from

the distribution of known n_x values, where the error term of the missing ratio or angle is approximated based on the n_x error terms with respect to their respective standard deviation $\sigma_{|x_i - y_i|}$, as depicted in Eq. 5.

$$\frac{|x_{\text{approx},i} - y_i|}{\sigma_{|x_i - y_i|}} = \frac{(\sum_{n=1}^{n_x} \frac{|x_i - y_i|}{\sigma_{|x_i - y_i|}})}{n_x} \tag{5}$$

Validation Workflow. After applying the outlier detector on the manually annotated sequences, a ranking based on the metric is used to prioritize the work of the human validators. Those are tasked with the correction of the detected outliers for each sequence in the dataset. The mean and the standard deviation of the error terms z_i are summarized in Fig. 3. We report the mean and standard deviation before and after the validation phase of the entire dataset to ascertain the effectiveness of outlier detection and improve the consistency of poses.

It is observed that the mean and standard deviation decrease for each value, showcasing the increase of quality as the individual error terms are converging towards zero.

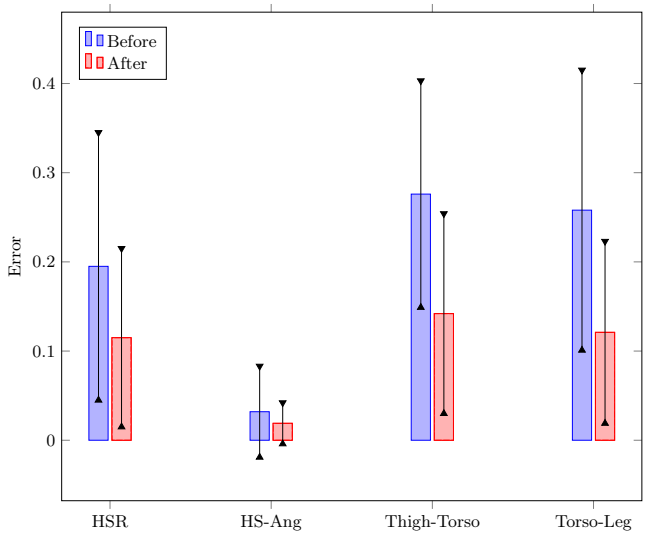


Fig. 3: Means and standard deviations of the error term before and after the validation phase. One can observe that both the mean and standard deviation decrease which proves the effectiveness of the proposed pose validator.

In summary, instead of manually reviewing each pose and probably generating similar errors during the annotation process, the proposed distance metric is used to prioritize and objectively quantify the potential error from each pose estimation. Through multiple review iterations, the correction of a quarter to

half of a sequence optimizes the overall metric by 0.135 points. In certain sequences, an optimization of 0.277 points is observed, strongly reducing the costs of validation while guaranteeing appropriate quality of annotation.

4 Experiments

This section details the experiments for pedestrian detection required for top-down models, the HPE experiments on LLVIP-Pose dataset for RGB and thermal images, and evaluates the results. Further experiments on grayscale pre-training and data-augmentation are also presented. The experiments are conducted on NVIDIA RTX Quatro 6000 24GB or Tesla V100-SXM2-32GB depending on their availability.

4.1 Pedestrian Detection

For our experiments, a YOLOX [13] object detection model is trained for person detection for each modality. These models are used to generate the detections for the evaluation of the top-down HPE models in the benchmark experiments. The YOLOX-l predictions with the same configuration setting as the LLVIP paper report worse predictions for both modalities as the baseline. The YOLOX-x versions achieved better detection results on RGB images but slightly lower detection performance on thermal images. The detailed results are presented in Tab. 2.

Table 2: Person detection results for LLVIP datasets. YOLOv5-l results are taken from the LLVIP paper. The YOLOX-l and YOLOX-x models are trained with the mmdetection toolbox [6]. Best model is highlighted in red. Second best is highlighted in blue.

Models	**Thermal**		**RGB**		Epoch	Lr
	AP	AP_{50}	AP	AP_{50}		
YOLOv5-l [18]	0.670	0.965	0.527	0.908	200	3.2×10^{-3}
YOLOX-l (ours)	0.652	0.959	0.529	0.906	200	3.2×10^{-3}
YOLOX-x (ours)	0.664	0.962	0.536	0.908	100	1.0×10^{-2}

4.2 Evaluation Setup

The Average Precision (AP) based on the object keypoint similarity (OKS) metric is used for evaluation on the LLVIP-Pose. In addition to AP and AP_{50}, the AP_M and AP_L are reported for medium and large bounding boxes as well as the AP_E^C, AP_M^C and AP_H^C for easy, medium and hard crowdedness levels based on [21].

4.3 Human Pose Estimation

Seven SOTA 2D HPE models with representatives for top-down (TD), bottom-up (BU) and one stage (OS) models are trained and evaluated on the newly acquired LLVIP-Pose dataset. To this aim, the mmpose toolbox [30] is used. See the supplementary materials for detailed training configurations. The models are trained with the same configuration with a batch size of 16 and weights from ImageNet, if provided. Otherwise, the weights are randomly initialized. To assess the effectiveness of thermal-based pose predictions against low-light RGB, each model is trained twice: once with thermal images and once with RGB images independently.

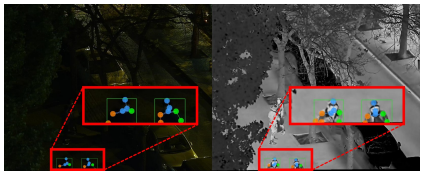
The evaluation results on the test set of the LLVIP-Pose dataset are reported in Tab. 3 and Tab. 4. As can be observed, the thermal-based models outperform the RGB-based models in all AP categories. The best model for both modalities is the ViTPose-h, for both groundtruth and predicted bounding boxes.

Table 3: Benchmarking results with AP for the different bounding box sizes (AP_M, AP_L). Best model is highlighted in red. Second best is higlighted in blue.

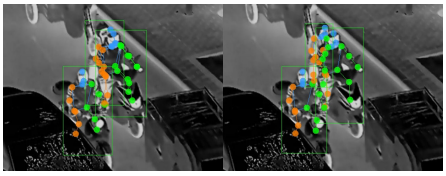
Thermal									
Models	**Category**	**Groundtruth BBoxes**				**Predicted BBoxes** (Thermal, AP: 0.664)			
		AP	AP_{50}	AP_M	AP_L	AP	AP_{50}	AP_M	AP_L
HRNetw48-udp [35]	TD	0.900	0.980	0.605	0.902	0.886	0.965	0.248	0.887
ViTPose-h [41]	TD	0.916	0.990	0.634	0.917	0.899	0.965	0.270	0.902
DeepPose-r50 [33]	TD	0.852	0.979	0.350	0.854	0.841	0.964	0.176	0.842
SimCC [23]	TD	0.877	0.979	0.457	0.879	0.867	0.964	0.217	0.869
DEKR [15]	BU					0.845	0.953	0.065	0.850
YOLOX-Pose-l [26]	OS					0.848	0.962	0.223	0.850
RTMO-l [25]	OS					0.855	0.961	0.243	0.858
RGB									
Models	**Category**	**Groundtruth BBoxes**				**Predicted BBoxes** (RGB, AP: 0.536)			
		AP	AP_{50}	AP_M	AP_L	AP	AP_{50}	AP_M	AP_L
HRNetw48-udp [35]	TD	0.643	0.914	0.176	0.647	0.591	0.866	0.087	0.594
ViTPose-h [41]	TD	0.681	0.937	0.253	0.684	0.632	0.879	0.142	0.634
DeepPose-r50 [33]	TD	0.596	0.906	0.153	0.599	0.547	0.854	0.064	0.549
SimCC [23]	TD	0.619	0.906	0.199	0.621	0.572	0.856	0.074	0.574
DEKR [15]	BU					0.562	0.854	0.010	0.566
YOLOX-Pose-l [26]	OS					0.569	0.867	0.128	0.571
RTMO-l [25]	OS					0.561	0.853	0.096	0.563

While considering the top-down approaches for both modalities, one notices a larger AP gap in the results between groundtruth and predicted bounding boxes for the RGB-based models. The switch from groundtruth to predicted

bounding boxes results in a drop of approximately 0.015 AP and 0.05 AP for thermal- and RGB-based models, respectively. The lower drop of APs showcases the higher detection rate of individuals in thermal images and the difficulty of person detection in RGB images due to low-illumination. The thermal-based models exhibit higher AP_M, showcasing the effectiveness of thermal images for medium bounding boxes. However, the reported AP_M does not highlight the ability of thermal images for medium scale human instances due to the fact that the scales of the person for medium and large bounding boxes are similar for both, as depicted in Fig. 4a.



(a) Medium bounding boxes for image pair with pose predictions from ViTPose-h. Medium bounding boxes contain human instances with similar scale to instances of large bounding boxes but are cropped due to image boundary.



(b) Comparison of SimCC with groundtruth and predicted bounding boxes. Poses with groundtruth bounding boxes (left) report lower AP_H^C than predicted bounding boxes (right).

Fig. 4: Visualization of pose predictions for both medium bounding boxes (a) and hard crowding level (b).

Bottom-up and single-stage models exhibit similar trends, with higher APs reported for thermal images, comparable to results of the top-down models.

As shown in Tab. 4, thermal models report higher AP^C than RGB models for both predicted and groundtruth bounding boxes. Single-stage models with the exeption of ViTPose-h have higher AP_H^C than top-down approaches for both modalities. For the top-down approaches, two overlapping bounding boxes result in sub-optimal pose predictions where the background bounding box receives the pose for the foreground bounding box, leaving the background bounding box without any poses and the foreground bounding box with two poses. This occurrence is relevant to the thermal images on SimCC, where a higher AP_H^C is reported for predicted bounding boxes, as shown in Fig. 4b.

Qualitative results are shown in Fig. 5, showcasing pose predictions from the ViTPose models for RGB and thermal images, respectively. Qualitatively, thermal- and RGB-based predictions prove similar results. However, for scenarios with low-illumination (e.g., top right) and high illumination changes (e.g., top-left, bottom-middle, and bottom-right) thermal-based predictions exhibit higher accuracy as indicated by the higher heatmap values. On the top right image pairs, thermal-based pose predictions capture the body orientations of the pedestrians situated in the dark, whereas RGB-based predictions failed to do so.

Table 4: Benchmarking results with AP for the different crowding level(AP^C_E, AP^C_M, AP^C_H). Best model is highlighted in red. Second best is higlighted in blue.

Thermal											
Models	**Category**	**Groundtruth BBoxes**					**Predicted BBoxes** (Thermal, AP: 0.664)				
		AP	AP_{50}	AP^C_E	AP^C_M	AP^C_H	AP	AP_{50}	AP^C_E	AP^C_M	AP^C_H
HRNetw48-udp [35]	TD	0.900	0.980	0.906	0.886	0.833	0.886	0.965	0.893	0.866	0.760
ViTPose-h [41]	TD	0.916	0.990	0.921	0.901	0.837	0.899	0.965	0.907	0.878	0.797
DeepPose-r50 [33]	TD	0.852	0.979	0.859	0.826	0.751	0.841	0.964	0.849	0.809	0.728
SimCC [23]	TD	0.877	0.979	0.884	0.853	0.721	0.867	0.964	0.876	0.839	0.724
DEKR [15]	BU						0.845	0.953	0.852	0.818	0.707
YOLOX-Pose-l [26]	OS						0.848	0.962	0.851	0.838	0.775
RTMO-l [25]	OS						0.855	0.961	0.859	0.845	0.797
RGB											
Models	**Category**	**Groundtruth BBoxes**					**Predicted BBoxes** (RGB, AP: 0.536)				
		AP	AP_{50}	AP^C_E	AP^C_M	AP^C_H	AP	AP_{50}	AP^C_E	AP^C_M	AP^C_H
HRNetw48-udp [35]	TD	0.643	0.914	0.651	0.626	0.546	0.591	0.866	0.598	0.576	0.497
ViTPose-h [41]	TD	0.681	0.937	0.683	0.678	0.714	0.632	0.879	0.635	0.626	0.567
DeepPose-r50 [33]	TD	0.596	0.906	0.601	0.584	0.560	0.547	0.854	0.551	0.529	0.480
SimCC [23]	TD	0.619	0.906	0.626	0.601	0.575	0.572	0.856	0.578	0.556	0.508
DEKR [15]	BU						0.562	0.854	0.570	0.536	0.519
YOLOX-Pose-l [26]	OS						0.569	0.867	0.571	0.565	0.504
RTMO-l [25]	OS						0.561	0.853	0.561	0.565	0.527

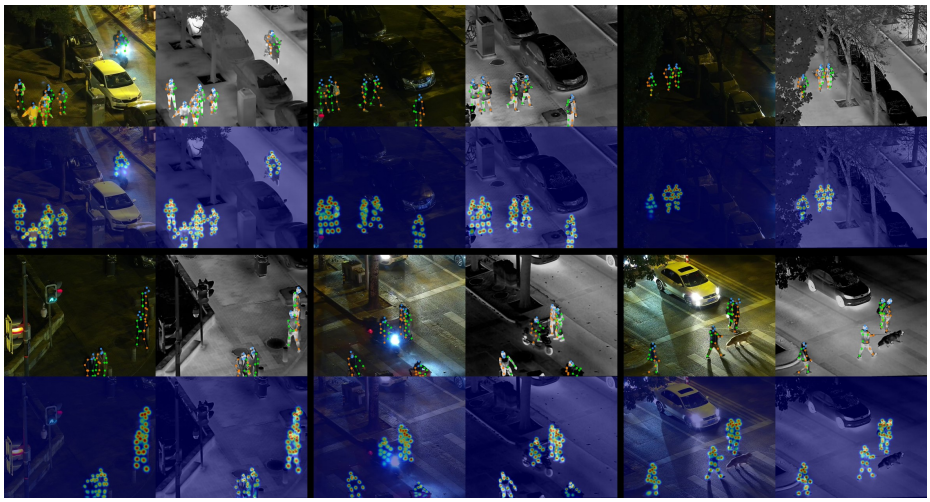


Fig. 5: Qualitative results of ViTPose for RGB and thermal image pairs. For scenarios with low-illumination (e.g. top right) and high illumination changes (e.g. top-left, bottom-middle and bottom-right) thermal-based predictions exhibit higher accuracy as indicated by the higher pose heatmap values.

4.4 Further experiments

Due to the scarce availability of annotated thermal data for HPE, leveraging RGB data is investigated. More precisely, experiments with COCO RGB pretraining and COCO reduced to grayscale for pretraining are conducted. The pre-trained models are first evaluated directly on LLVIP-Pose on the 13 common keypoints between the COCO topology and the LLVIP (Posetrack18) topology. Each model is then fine-tuned on the LLVIP-Pose dataset and once more evaluated.

Further variants are trained using the Albumentation library [4] to apply pixel-wise augmentation techniques on the COCO training data. In this case, HSV augmentation and image inversion with a probability of $p = 1$ are used to further simulate thermal data. HSV augmentation is applied prior to grayscaling to provide more variations, followed by image inversion to imitate the appearance of thermal images. As a control, the same augmentations are applied to RGB images and an experiment combining both RGB and grayscale images is conducted.

For a representative overview, three models are trained, representing the three categories of HPE models. The results are reported in Tab. 5. Prior to fine-tuning, two main observations are reported. The first is the benefit of the HSV and image inversion augmentations to both RGB and grayscale images and the second is the lower APs of models trained with grayscale images. This observation showcases the domain gap between RGB and thermal images, in which grayscale images lack the characteristics of thermal images, but the loss of two distinct color channels hinders keypoint localization. The APs reported for "RGB + Grayscale" for DEKR and YOLOX-Pose further support this with the lowest APs recorded. HRNetw48-udp acts, however, as an outlier, where the mixture of RGB and grayscale images resulted in the higher AP.

After the fine-tuning using the LLVIP-Pose train set, the grayscale (monochromatic) models report higher AP compared to the RGB models. This observation shows how learning monochromatic features can assist the fine-tuning of thermal data, as the grayscale models even outperforms the "RGB + Augmentations" models. The sub-optimal performance of models trained with RGB and grayscale COCO images report the highest AP for YOLOX-Pose and the second highest AP for HRNetw48-udp and DEKR, trailing behind the APs for "Grayscale + Augmentations". This demonstrates the importance of RGB features alongside monochromatic features, as it outperforms the APs of the grayscale models. "Grayscale + Augmentations" exhibits the highest APs for HRNetw48-udp and DEKR, however, the lowest AP for YOLOX-Pose, despite showcasing the highest AP without fine-tuning.

5 Conclusions

In conclusion, this work addresses the limitations in HPE under low-light conditions by introducing the LLVIP-Pose dataset, a modification of the existing

Table 5: Pre-training and fine-tuning results of three SOTA models with different augmentations. The tests are conducted on the LLVIP-Pose test dataset with the 13 mutual keypoints between the COCO and PoseTrack18 skeleton topology.

Models	**Augmentations**	**Thermal**							
		COCO Pre-Training				**LLVIP Fine-Tuning**			
		AP	AP_{50}	AP_M	AP_L	AP	AP_{50}	AP_M	AP_L
HRNetw48-udp [35]	RGB	0.626	0.813	0.435	0.626	0.930	0.990	0.720	0.930
	Grayscale	0.611	0.789	0.431	0.611	0.938	0.990	0.705	0.938
	RGB + Grayscale	0.631	0.814	0.466	0.633	0.939	0.990	0.744	0.939
	RGB + Aug	0.705	0.872	0.411	0.707	0.937	0.989	0.764	0.938
	Grayscale + Aug	0.739	0.896	0.384	0.742	0.941	0.990	0.786	0.942
	LLVIP-Pose (Scratch)	-	-	-	-	0.917	0.979	0.685	0.917
DEKR [15]	RGB	0.427	0.606	0.016	0.440	0.888	0.964	0.071	0.892
	Grayscale	0.400	0.573	0.012	0.417	0.894	0.964	0.067	0.898
	RGB + Grayscale	0.383	0.550	0.010	0.398	0.896	0.965	0.062	0.899
	RGB + Aug	0.540	0.719	0.007	0.553	0.892	0.963	0.061	0.897
	Grayscale + Aug	0.601	0.782	0.009	0.616	0.897	0.965	0.066	0.900
	LLVIP-Pose (Scratch)	-	-	-	-	0.862	0.952	0.070	0.866
YOLOX-Pose-l [26]	RGB	0.511	0.707	0.061	0.513	0.880	0.964	0.295	0.881
	Grayscale	0.496	0.684	0.045	0.499	0.881	0.972	0.287	0.882
	RGB + Grayscale	0.491	0.677	0.055	0.494	0.882	0.971	0.219	0.883
	RGB + Aug	0.598	0.799	0.049	0.600	0.881	0.972	0.245	0.883
	Grayscale + Aug	0.607	0.814	0.042	0.610	0.878	0.971	0.235	0.879
	LLVIP-Pose (Scratch)	-	-	-	-	0.863	0.962	0.222	0.864

LLVIP dataset to include pose annotations for both visible and thermal imagery. Our dataset includes an extensive contribution of 26,135 pose annotations, significantly enhancing the resources available for thermal-based HPE research. By implementing a semi-autonomous workflow for pose annotation and validation, we have substantially reduced the manual effort and associated costs. Our benchmark experiments demonstrate that thermal-based HPE models can outperform traditional RGB-based models in challenging conditions, while our generalization experiments provide valuable insights into the transferability of pre-trained RGB models to thermal applications.

These findings underscore the potential of thermal imagery for robust and privacy-preserving human pose estimation, paving the way for future advancements in various applications such as surveillance, worker safety, and beyond. Future work may address the problem of generalization between different sensors, cameras, and perspectives. Ensuring that HPE models can robustly adapt to varying hardware and viewpoints will be crucial for the widespread adoption and effectiveness of these technologies in real-world scenarios.

References

1. Alam, E., Sufian, A., Dutta, P., Leo, M.: Real-time human fall detection using a lightweight pose estimation technique. In: International Conference on Computa-

tional Intelligence in Communications and Business Analytics. pp. 30–40. Springer (2023)
2. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
4. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2) (2020). https://doi.org/10.3390/info11020125, https://www.mdpi.com/2078-2489/11/2/125
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
6. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
7. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020)
8. Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K., Kweon, I.S.: Kaist multi-spectral day/night data set for autonomous and assisted driving. IEEE Transactions on Intelligent Transportation Systems **19**(3), 934–948 (2018)
9. Cormier, M., Clepe, A., Specker, A., Beyerer, J.: Where are we with human pose estimation in real-world surveillance? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 591–601 (January 2022)
10. Cormier, M., Röpke, F., Golda, T., Beyerer, J.: Interactive labeling for human pose estimation in surveillance videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1649–1658 (October 2021)
11. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv:2003.09003[cs] (Mar 2020), http://arxiv.org/abs/1906.04567, arXiv: 2003.09003
12. Drillis, R., Contini, R.: Body segment parameters, new york university. Tech. rep., NY, Technical Report (1966)
13. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
14. Gebhardt, E., Wolf, M.: Camel dataset for visual and thermal infrared multiple object detection and tracking. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2018)
15. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14676–14686 (2021)

16. Heindl, C., Ikeda, M., Stübl, G., Pichler, A., Scharinger, J.: Metric pose estimation for human-machine interaction using monocular vision. arXiv preprint arXiv:1910.03239 (2019)
17. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baselines. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
18. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3496–3504 (2021)
19. Kuzdeuov, A., Taratynova, D., Tleuliyev, A., Varol, H.A.: 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). pp. 1–8 (2024). https://doi.org/10.1109/FG59268.2024.10581992
20. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942 [cs] (Apr 2015), http://arxiv.org/abs/1504.01942, arXiv: 1504.01942
21. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
22. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1944–1953 (2021)
23. Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., Xia, S.T.: Simcc: A simple coordinate classification perspective for human pose estimation. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
25. Lu, P., Jiang, T., Li, Y., Li, X., Chen, K., Yang, W.: Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. arXiv preprint arXiv:2312.07526 (2023)
26. Maji, D., Nagori, S., Mathew, M., Poddar, D.: Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2637–2646 (2022)
27. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., den Hengel, A.v.: Poseur: Direct human pose regression with transformers. In: European conference on computer vision. pp. 72–88. Springer (2022)
28. Martin, M., Popp, J., Anneken, M., Voit, M., Stiefelhagen, R.: Body pose and context information for driver secondary task detection. In: 2018 IEEE Intelligent Vehicles Symposium (IV). pp. 2015–2021 (2018). https://doi.org/10.1109/IVS.2018.8500523
29. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (Mar 2016), http://arxiv.org/abs/1603.00831, arXiv: 1603.00831
30. MMPose-Contributors: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)

31. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. Advances in neural information processing systems **30** (2017)
32. Tang, Z., Ye, W., Ma, W.C., Zhao, H.: What happened 3 seconds ago? inferring the past with thermal imaging. In: CVPR (2023)
33. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)
34. Vendrow, E., Le, D.T., Cai, J., Rezatofighi, H.: Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4811–4820 (June 2023)
35. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. TPAMI (2019)
36. Wang, S., Jiang, K., Chen, J., Yang, M., Fu, Z., Wen, T., Yang, D.: Skeleton-based traffic command recognition at road intersections for intelligent vehicles. Neurocomputing **501**, 123–134 (2022). https://doi.org/https://doi.org/10.1016/j.neucom.2022.05.107, https://www.sciencedirect.com/science/article/pii/S0925231222006944
37. Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., Fang, L.: Panda: A gigapixel-level human-centric video dataset. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3265–3275 (2020). https://doi.org/10.1109/CVPR42600.2020.00333
38. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
39. Winter, D.A.: Biomechanics and motor control of human movement. John wiley & sons (2009)
40. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 466–481 (2018)
41. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
42. Yang, J., Zeng, A., Liu, S., Li, F., Zhang, R., Zhang, L.: Explicit box detection unifies end-to-end multi-person pose estimation. arXiv preprint arXiv:2302.01593 (2023)
43. Yang, S., Quan, Z., Nie, M., Yang, W.: Transpose: Keypoint localization via transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11802–11812 (2021)
44. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
45. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3d human pose estimation in rgbd images for robotic task learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1986–1992. IEEE (2018)