IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 February 2023, accepted 31 March 2023, date of publication 5 April 2023, date of current version 12 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264714

RESEARCH ARTICLE

Human Pose Estimation Using Thermal Images

JAVIER SMITH[1], PATRICIO LONCOMILLA[2], AND JAVIER RUIZ-DEL-SOLAR[1,2], (Senior Member, IEEE)
[1]Department of Electrical Engineering, Universidad de Chile, Santiago 8370451, Chile
[2]Advanced Mining Technology Center, Universidad de Chile, Santiago 8370451, Chile

Corresponding author: Patricio Loncomilla (ploncomi@gmail.com)

This work was supported in part by the Agencia Nacional de Investigación y Desarrollo (ANID) through PIA under Grant AFB18004, in part by FONDECYT under Grant 1201170, and in part by FONDEQUIP under Grant EQM170041.

ABSTRACT This study addresses the human pose estimation problem on thermal images using Convolutional Neural Networks and Vision Transformer architectures. To do this, eight human pose estimation methods designed for visible images were extended to be applied in the thermal domain. Due to the lack of large, representative datasets containing labeled thermal images, this extension requires transfer learning between the visible and the thermal domain, and a database for fine-tuning the networks in the thermal domain. Thus, it is proposed to train the networks using a grayscale version of the COCO dataset, and then fine-tune them in the thermal domain. Fine-tuning is carried out using the new UCH-Thermal-Pose database presented in this work. This database includes 600 thermal images for training, 200 for validation, and 104 for testing, all of them fully labeled. Moreover, in the paper, a comparative study of the eight extended deep-based methods for human pose detection is carried out. The UCH-Thermal-Pose database is available at https://datos.uchile.cl/dataset.xhtml?persistentId=doi%3A10.34691%2FUCHILE%2F4B6NA3, and the source code of all the methods is available at https://github.com/jsmithdlc/Thermal-Human-Pose-Estimation.

INDEX TERMS Convolutional neural networks, vision transformer, deep neural networks, human pose estimation, thermal images.

I. INTRODUCTION

Human pose estimation consists of predicting the location of the parts of a person's body in an image. The parts of the body are represented by an articulated skeleton, which is composed of a set of joints (also named keypoints). Then, an algorithm that performs human pose estimation receives an image as input, on which it computes a set of keypoints, grouping them into a skeleton. An example of a skeleton is shown in Figure 1.

Human pose estimation is normally carried out using visible images. However, the estimation of human pose in thermal images is an important research topic, since (i) thermal images can be captured in darkness as they are not affected by illumination, which enables applications that must work at night or inside environments with no lighting, like those related to surveillance; (ii) the use of thermal images allows estimating the pose of people under covers, which is important for medical applications used for monitoring bedridden patients; (iii) thermal cameras are able to detect persons despite the environment containing heavy smoke or dust, which enables their use in applications like search and rescue; and (iv) pose estimation in thermal images is also used in applications that require that the identity of the people is preserved. Some examples of human pose detection using thermal images are shown in Figure 2.

As an example of the increasing interest in human pose estimation in thermal images, the IEEE VIP Cup 2021 "Privacy-Preserving In-Bed Human Pose Estimation"[1] looks for "computer vision-based solutions for in-bed pose estimation under the covers". The appropriate solution for this challenge requires the use of thermal cameras, among other sensors.

However, since thermal images are not as popular as standard visible images, there are almost no datasets of thermal images with annotated human poses. This is an important drawback for the development of human pose detectors based

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

[1]https://signalprocessingsociety.org/community-involvement/video-image-processing-cup

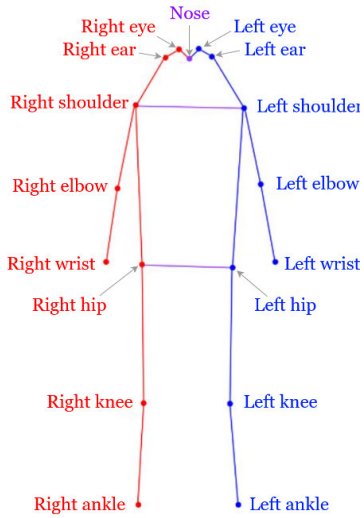

FIGURE 1. Human pose represented as a 17-keypoint skeleton.

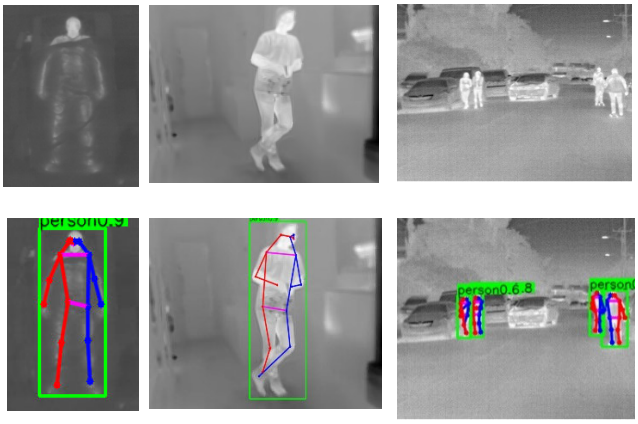

FIGURE 2. Examples of applications of human pose estimation in thermal images. Left column: pose detection of an in-bed person covered by a blanket. Center column: pose detection in an application that requires anonymity; facial landmarks are mostly undistinguishable. Right column: pose detection in a scene with poor illumination.

on deep/machine learning, which require having annotated images for their training.

In this context, this paper addresses the human pose estimation problem using various deep learning methods. Several human pose estimation methods for visible images, based on the use of Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), are extended to be applied in the thermal domain. This extension requires transfer learning between the visible and the thermal domain, as well as having a dataset for fine-tuning the networks in the thermal domain. The work reported in this paper proposes a transfer learning methodology, which considers training first the human pose detectors using grayscale images of the COCO dataset [4] before fine-tuning them in the thermal domain. The use of grayscale images is justified because thermal images are also monochromatic, and therefore the domain transfer is more direct than the case when color images are used. Fine-tuning is carried out using the new UCH-Thermal-Pose database, presented in this work, which contains 600 images for training, 200 for validation, and 104 for testing. All database images are labeled.

The paper presents a comparative study of eight different methods for human pose detection, all of them built by using the proposed training methodology: four top-down methods: Simple Baselines [5], AlphaPose [6], RSN [7] and ViTPose [8]; and four bottom-up methods: Bottom-Up-HRNet [9], OpenPose [10], CenterNet [11], and PoseAE [12]. It is worth mentioning that RSN [7] and ViTPose [8] are state-of-the-art methods; they obtained the first and third position in the *pose estimation test-dev* of the COCO (Common Objects in Context) dataset as of December 2022.[2] In addition, YOLOv7 [13], a state-of-the-art object detector, is used for generating the bounding boxes required by the top-down detectors.

Thus, the main contributions of the paper are:

- The extension of eight methods for human pose detection from the visible domain to the thermal domain, using a *transfer learning methodology* that considers pre-training the human pose detectors using *grayscale images* of the COCO dataset before the fine-tuning in the thermal domain.

- A comparative study of eight different methods for human pose detection, four top-down and four bottom-up; all of them built by using the proposed *transfer learning* methodology. The comparison is useful for understanding which network parameters and design decisions increase accuracy. The source code of these approaches is made public for research purposes.

- The proposal of UCH-Thermal-Pose, a new dataset for human pose estimation in the thermal domain, whose usefulness is validated by training several detectors, including two state-of-the-art methods: RSN [7] and ViTPose [8]. This database is made public for research purposes.

This paper is organized as follows: In Section II, a general background regarding human pose estimation, and its applications in the thermal domain, is given. Section III continues with a description of the methods and datasets used in this work, as well as the training methodology. Section IV covers the experimental results obtained from the evaluation of thermal images, including the precision and speed of the trained pose detection systems. A comparison of different training strategies and detection configurations is also provided, along with the corresponding analysis in Section V. Challenges on human pose estimation in the thermal domain are discussed in Section VI. Finally, in Section VII the conclusions obtained in this work are summarized.

II. HUMAN POSE ESTIMATION USING THERMAL IMAGES

State-of-the-art methods for human pose detection from visible images are based on using convolutional neural networks (CNNs), or vision transformers (ViTs). These methods can also be applied to detect human poses using thermal images.

[2]https://paperswithcode.com/sota/pose-estimation-on-coco-test-dev

However, thermal images present different characteristics compared to visible images. The invariance of thermal images with regard to environmental illumination makes them one of the best approaches for dealing with dark environments. Also, people are visible in thermal images even when they are covered, which allows for viewing bedridden people. However, detecting people becomes more difficult when their temperatures are close to those of their environments.

An important limitation for developing performant detectors of people's poses in the thermal domain is that labeled datasets for thermal images are fewer by far than those of visible images, and the available datasets have low variability in the kind of images contained in them. This is because visible images are abundant on the Internet, which enables a great diversity of such datasets needed for training high-performing detectors based on CNNs or ViTs. Thus, the lack of abundant and diverse annotated datasets in the thermal domain results in the detectors not being able to generalize in the task of detecting poses of people in environments different from the specific ones contained in these datasets.

In the following subsections, several methods for detecting people's poses, both on thermal and visible images, will be described and analyzed, including the main two families of detectors: top-down detectors, which work by first detecting people, and then detecting keypoints for each person, and bottom-up detectors, which detect the keypoints of all the people present in the image, and then group them together into various individual poses.

A. TOP-DOWN HUMAN POSE DETECTION

Top-down human pose detection works by separating the pose detection process into two steps: detection of people, and per-person pose estimation [9].

For detecting people, an object detector trained for the detection of people is used. Common people detectors are based on two-stage object detectors, such as Faster-RCNN [14], or single-stage object detectors, such as the YOLO object detector variants [15], [16], [17], [18], [13].

In a second step, a pose detector is used on each of the detected bounding boxes. Popular pose detectors include Mask-RCNN [19], Simple Baselines [5], AlphaPose [6], HRNet [20], Residual Steps Network [7] and ViTPose [8].

Mask R-CNN detects bounding boxes of people, and then computes a segmentation mask. This mask can be used for representing locations for each of the joints.

Simple Baselines uses a ResNet-based architecture as a feature extractor, modified by the addition of deconvolutional layers in the upsampling phase.

AlphaPose is a more complex human pose estimation approach. It comprises the use of a Symmetric Spatial Transformer network to detect the region of a single person in an inaccurate bounding box detection. A Parametric Pose Non-Maximum Suppression is used to solve redundant detections, and a Pose-Guided Proposals Generator to augment the training data [21].

HRNet is a person pose estimator that works by generating heat maps for each of the joints. It includes two key characteristics: (i) Connecting the high-to-low resolution convolution streams in parallel, and (ii) exchanging the information across resolutions repeatedly.

Residual Steps Network (RSN) [7] is a state-of-the-art person pose estimator based on CNNs, but it uses an attention mechanism. It uses Residual Step Blocks (RSBs) that contain four branches, each with a different number of 3×3 convolutional layers. The outputs from these layers are concatenated, and then a 1×1 convolution is applied and followed by a residual connection.

According to *pose estimation test-dev* of the COCO dataset, the best pose detector, as of December 2022, is ViTPose [8]. A plain, non-hierarchical visual transformer ViT [22] is used as a backbone to extract feature maps for the given person instances. Then, a lightweight decoder, composed of two deconvolution layers and one prediction layer, is used for generating the height maps for the keypoints.

Despite top-down methods being highly performant, their runtime depends on the number of persons present in the images. This is explained by a pose detector having to be applied independently on each detected person. Thus, this family of methods is not recommendable when the images to be processed contain many people.

B. BOTTOM-UP HUMAN POSE DETECTION

In this approach, the keypoints of all the people present in the image are detected in one step and then grouped together into skeletons corresponding with each of the persons present in the image. For detecting the keypoints, detectors that compute heat maps per keypoint are used, and then these keypoints are grouped based on various criteria, which depend on the specific method being used.

Popular bottom-up human pose detection methods are Bottom-Up-HRNet [9], OpenPose [10], and CenterNet [11] trained for human keypoint detection. Another popular keypoint detector is PoseAE [12].

OpenPose [23] is one of the most popular Bottom-Up detectors in the visible domain. It distinguishes itself from the other methods explored here mainly by its use of Part Affinity Fields (PAF) for grouping keypoint coordinates, detected via heat maps, into the different person instances, while retaining real-time inference.

CenterNet [11] is generally used as an object detection technique. Nevertheless, as the authors showed, the model can also be extended to solve other tasks, such as human pose estimation. As a Bottom-Up human pose detector, it presents a unique method for grouping the keypoints into final poses, in which a regression from the person's center is used to localize keypoints initially. Then, the keypoints are further corrected by regressing an offset for their position.

PoseAE [12] also detects keypoints through heat maps and so-called associative embeddings for each predicted keypoint. These embeddings serve as grouping cues, where

keypoints that belong to the same person should have a small distance between them.

Bottom-Up HRNet [9] uses the same keypoint grouping technique as CenterNet, but, at the regression head, combines the predicted keypoint heat maps with the feature map generated by the backbone architecture to produce better results. The keypoint coordinate predictions are further refined using a Spatial Transformer Network (SPN).

In general, bottom-up human pose detectors are faster than top-down ones. This runtime difference becomes larger when several persons are present in the images. However, bottom-up detectors in general are less performant than top-down ones.

C. HUMAN POSE DETECTION USING THERMAL IMAGES

Methods for detecting human pose over thermal images hardly appear in the literature [25], [26], [2], [27]. One of the main reasons is the lack of databases of thermal images with annotated poses of humans.

A relevant application of human pose detection in thermal images is detecting bedridden people, which has the potential to be used in healthcare applications. A system that can recover poses of covered human bodies from thermal images is proposed in [1]. It is stated that the main difficulty to deal with people under covers is heat diffusion, which causes the heat on the cover to be different from that of the uncovered person. In addition, a dataset named SLP was created. It contains visible, thermal, depth, and pressure map images, captured from 109 in-bed participants. The participants change their poses randomly from the three main categories of supine, left side, and right side. For each category, 15 poses are collected. Also, three cover categories (uncovered, thin cover, and thick cover) are included. The result is a total number of 14,715 images. This dataset is used for training stacked hourglass networks [24] for detecting poses of people, both in visible and in thermal images. The results show that, while pose detection using visible images from uncovered persons is very performant, it has a large performance drop on covered persons, making use of thermal images most suited to this condition. Note that despite the dataset used being large, its images have low variability since all the people are captured on the same bed, with the same sensor configuration. Therefore, this dataset is useful only for training pose detectors of bedridden people.

The use of multiple sensing modalities benefits from the advantages related to each of the sensors. In [25], the previous work reported in [1] is extended. Several person pose detectors are trained on the SLP dataset [1], and different combinations of input modalities (visible, thermal, depth, and pressure maps) are tested and compared. That work is aimed only at detecting the poses of people in bed because the SLP dataset has low variability, as mentioned previously. Thus, it cannot be extended to other applications in which thermal images can be used.

The lack of large datasets containing thermal images of people is an obstacle to achieving high performance in detectors. A dataset containing paired visible and thermal images is introduced in [26], as well as a new network architecture for detecting people in the thermal domain. The dataset contains 24,000 pairs of thermal and visible images, captured in indoor environments. The visible images have a resolution of 1920×1080, while the thermal images have a much lower resolution of 80×60. Some of the visible images were captured with low illumination, which impairs person detectors based on visible images. The ground truth for the training subset is based on the detection of people on the visible images, using OpenPose [10]. The test subset is formed by 1,000 pairs captured with good illumination, and 1,000 pairs captured in darkness. The ground truth for the test subset is labeled manually by using visible images in scenes with good illumination, but by using thermal images in settings in darkness. Here, a network named ThermalPose, based on OpenPose is introduced. This network is trained by using thermal images from the training subset. It is shown from the results of this work that, on images with good illumination, person detectors based on visible images beat ThermalPose. However, in scenes with predominant darkness, ThermalPose behaves better than all the networks tested based on visible images, as the latter are unable to detect people. While the dataset used in this work is large, only 2,000 image pairs were labeled manually (i.e. the test subset). Furthermore, all the images were captured in indoor environments with very low variability.

Datasets containing labeled visible images are by far more abundant and diverse than those of thermal images, but methods that can translate visible images into thermal images have the potential of helping to overcome this situation. The development and testing of an algorithm named ThermalGAN is reported in [2]. It works by training a GAN for translating visible images into thermal images, and it is aimed at the re-identification of people in thermal images. For improving the predicted temperature of the objects, segmentation masks of objects are used for computing per object average temperatures, and then the temperature variations inside each object are predicted. A dataset named ThermalWorld is generated, which contains 15,118 pairs of visible + thermal images, including annotated masks per object. This dataset is composed of two splits: the first one includes cropped persons (used for reidentification) and the second one contains images containing ten object categories (used for object detection). Despite this work being able to translate images, the quantitative evaluation of thermal object detection is based on humans indicating whether or not the translated images are real. Thus, the dataset is not used for evaluating the performance of the translation quantitatively.

The use of both thermal and visible images has the potential of improving the performance obtained by using only one of them. The creation of a dataset containing both thermal and depth images is reported in [27]. It contains 700 labeled thermal + depth images for training, 100 for validation, and 200 for testing. A variant of a part affinity fields detector [20] is used for testing pose detection on this dataset, and the

use of thermal + depth images improves when using only thermal images. Despite this dataset being labeled manually, it considers only five keypoints, which is a number much smaller than that used in most pose detectors.

III. EXTENDING HUMAN POSE ESTIMATION METHODS FROM THE VISIBLE TO THE THERMAL DOMAIN

Selected pose detection methods are extended to estimate human pose in thermal images, taking successful top-down and bottom-up methods used for human pose estimation in visible images as a base. In addition, the performance of these methods is analyzed under different criteria using a new database, UCH-Thermal-Pose. Each method is trained to detect 17 keypoints of the human body: nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles, the same ones specified in the Microsoft COCO Keypoints challenge [4].

A. DATASETS

The Microsoft COCO 2017 dataset [4] was used for the pre-training of the models. This dataset is regarded as a common training and evaluation benchmark for human pose estimation methods in the visible domain. It provides nearly 200k visible images for training, with almost 250k person instances labeled with their keypoints in various types of contexts. There is a total of 17 types of human body keypoints in this dataset, as was mentioned above. In this work, images from this dataset were transformed to grayscale, so that the models were trained using samples that better resemble the target domain of thermal images; both domains, the grayscale and the thermal, are monochromatic. A comparison between training with the original RGB images and their grayscale conversions is made for one of the models (CenterNet) to support this hypothesis (see Appendix A2). The training of all models using both types of images would have been a massive undertaking, given the limited time and resources, so the results obtained for CenterNet are generalized for the other models.

In order to train and evaluate the extended methods a new dataset, UCH-Thermal-Pose, was built and made public for research purposes. The database is composed of two sets, A and B.

UCH-Thermal-Pose Set-A was built by collecting thermal images from different public sources and then annotating them. The dataset is composed of 800 images including both indoor and outdoor settings, each one containing at least one person. In each image, the human keypoints were annotated manually using the LabelMe [28] tool. 600 of these images are used for training the methods, and the remaining 200 are used as a validation set. Table 1 shows a more detailed description of the different image sources and the number of images extracted from each one for training and validation.

Variability in thermal image datasets is hard to find, since most of the existent datasets are captured in a limited amount of environments, using static camera setups [27], [30]. Consequently, the images selected for UCH-Thermal-Pose Set-A were chosen while trying to maximize various settings, weather conditions, and poses.

TABLE 1. Different thermal image sources used in the new UCH-Thermal-Pose set-A database, which consists of 600 training images and 200 validation images.

Source	Type	Environment	Train	Validation
FLIR [3]	Video	Outdoor	271	83
ThermalWorld [2]	Images	Indoor + Outdoor	161	32
LSIFIR [31]	Video	Outdoor	36	11
BUTIV [32]	Video	Indoor + Outdoor	30	10
Terravic [33]	Images	Outdoor	2	1
infAR [30]	Video	Outdoor	12	3
SLP [25]	Images	Indoor	23	7
OSU [34]	Video	Outdoor	1	2
Biloudet [29]	Video	Indoor	9	3
CSIR [35]	Images	Outdoor	2	1
Kaist [36]	Video	Outdoor	13	4
Thermal Dogs and People [42]	Images	Outdoor	11	6
Thermal Soccer [38]	Images	Outdoor	20	4
TUFTS Face [37]	Images	Indoor	9	3
CVC [39]	Video	Outdoor	0	20
AAU VAP Trimodal [40]	Video	Interior	0	8
Web search	Images	Outdoor	0	2

UCH-Thermal-Pose Set-B is composed of thermal images acquired in our laboratory, which were annotated using the same procedure used for annotating the images in set A. The dataset is composed of 104 thermal images captured using a FLIR FC-690 S thermal camera, at a resolution of 640 × 480, and comprising a total of 3 different camera angles. All these images, except three, which are devoid of people, contain between 1 and 4 people in different poses. Examples of these images are shown in Figure 6.

Figures 3, 4, and 5 provide an overview of the UCH-Thermal-Pose dataset. A histogram showing the frequency of images at different sizes is provided in Figure 3. The distribution of the number of keypoints labeled for each person is shown in Figure 4. Finally, the distribution of the labeled bounding box areas is shown in Figure 5, shown both as pixels (left) and relative to the size of the image (right).

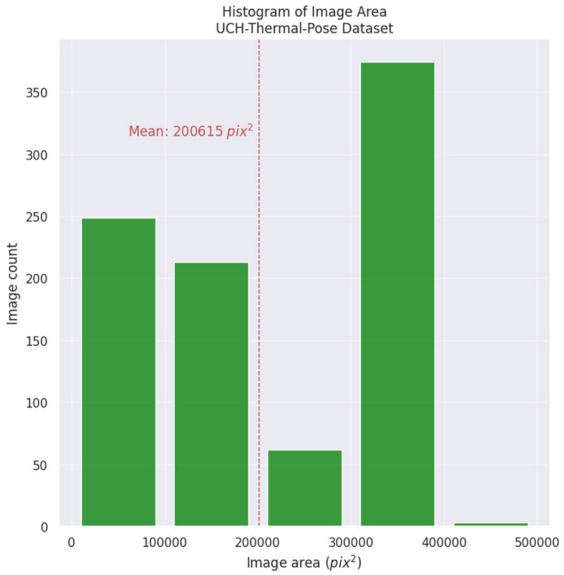

FIGURE 3. Histogram of image area, in pixels, from the UCH-Thermal-Pose dataset.

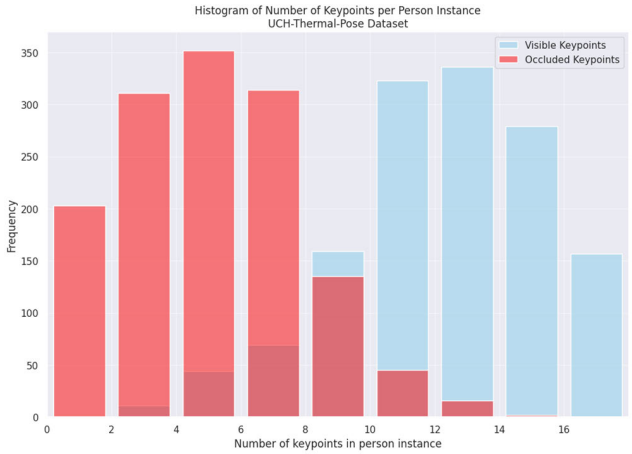

FIGURE 4. Histogram of the number of visible and occluded labeled keypoints per person instance, from the UCH-Thermal-Pose dataset.

B. THERMAL HUMAN POSE ESTIMATION METHODS

Top-down and bottom-up methods are used in this work for human pose estimation in the thermal domain.

The following top-down methods are analyzed: Simple Baselines [5], AlphaPose [6], RSN [7] and ViTPose [8].

These methods were chosen because of their use on common benchmarks like Microsoft COCO and MPII [41], and their ubiquity in the field of human pose estimation. Since top-down methods depend largely on off-the-shelf object detectors for identifying humans in images, the choice of detector heavily influences the final results. The same object detector, YOLOv7 [17], therefore, was used in each method for predicting bounding boxes for humans.

The analyzed bottom-up methods are CenterNet [11], OpenPose [10], Bottom-Up HRNet [9], and PoseAE [12].

TABLE 2. Human pose estimation methods considered in this study.

Method	Backbone	Bottom-Up / Top-down	Person Detector
ViTPose [8]	ViT	Top-down	YOLOv7
RSN [7]	RSN	Top-down	YOLOv7
AlphaPose [6]	ResNet-152	Top-down	YOLOv7
Simple Baselines [5]	ResNet-50	Top-down	YOLOv7
CenterNet [11]	DLA-34	Bottom-Up	---
CenterNet [11]	Hourglass-104	Bottom-Up	---
CenterNet [11]	Hourglass-52	Bottom-Up	---
CenterNet [11]	HRNet-W32	Bottom-Up	---
Bottom-Up HRNet [9]	HRNet-W32	Bottom-Up	---
PoseAE [12]	Hourglass-52	Bottom-Up	---
OpenPose [10]	VGG-19	Bottom-Up	---

CenterNet has a simple architecture in which the backbone can be changed easily. In consequence, four backbones are considered for CenterNet: DLA-34, Hourglass-104, Hourglass-52 and HRNet-W32. These backbones are also used to analyze freezing of layers in Appendix A.3. In the case of the other methods, the backbones from the original papers are used.

The methods to be compared are summarized in Table 2.

C. EVALUATION METRICS

Evaluation of human pose estimation models done in this work, is based on a metric named Object Keypoint Similarity (OKS), that was introduced in the COCO challenge [4]. OKS is computed as shown in (1), and its possible values are in the range of 0-1. In equation (1), d_i is the distance between each predicted keypoint and their respective ground truth, s is the object scale computed from the bounding box size, and k_i a fall-off constant for each type of keypoint that quantifies the variance in the annotation process for that keypoint. Only keypoints that are inside the image boundaries ($v_i > 0$) impact the final value.

$$OKS = \frac{\sum_i e^{-\frac{d_i^2}{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i \delta(v_i > 0))} \tag{1}$$

OKS can thus quantify how close a human pose prediction is to its actual ground-truth values, and a prediction can be considered correct if this metric is over a certain threshold. In this way, detections are gathered as True Positives whose OKSs are over a defined threshold (TP), detections whose

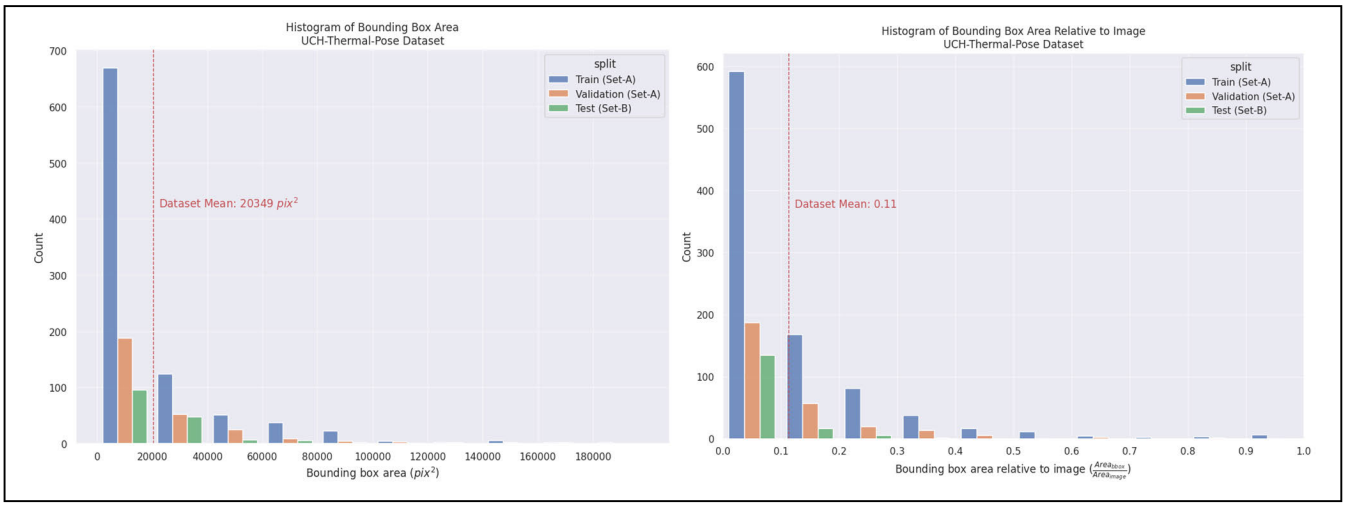

FIGURE 5. Histogram of bounding box size per split from UCH-Thermal-Pose Dataset, in pixels (left) and relative to image size (right).

FIGURE 6. Examples of thermal images included in UCH-Thermal-Pose set-B.

OKSs are under the threshold as False Positives (FP), and ground-truth human poses that were not properly detected as False Negatives (FN). Then, precision and recall can be calculated easily using (2) and (3), in which TP are true positives, FP are false positives, and FN are false negatives. In this case, precision defines the ability of the model to detect only the correct poses (a fraction of correctly detected poses of the total detections), whereas recall defines the ability of the model to detect the poses of the humans that are present in the images (a fraction of detected poses from all the poses to be detected).

$$Precision = \frac{TP}{FP + TP} \tag{2}$$

$$Recall = \frac{TP}{FN + TP} \tag{3}$$

Following common practices, average precision (AP), and average recall (AR) are calculated by averaging over 11 different OKS thresholds, from 0.5 to 0.95 with steps of 0.05. These correspond to the main metrics used in the experiments reported. In addition, AP and AR at OKS thresholds of 0.5 and 0.75, as well as for medium (M), and large (L) objects, are also calculated. They correspond to AP0.5, AP0.75, APM, and APL, respectively.

The inference speed of the trained models is estimated using the Frames per second (FPS) metric. This metric is important as a reference for real-time applications. It defines the number of images (frames) that a model can detect in the lapse of one second. Models that score FPS over 15 are commonly considered to be real-time detectors.

D. TRAINING METHODOLOGY

All the methods, except OpenPose, are first pretrained on the COCO grayscale dataset, i.e., using visible images from COCO converted to grayscale, and then fine-tuned using the UCH-Thermal-Pose A dataset. OpenPose is not pretrained using COCO grayscale images given that a long training is required by this method. Instead, available pretrained weights, obtained after training the network in the pose estimation task using color images, are used as a base for fine-tuning this method. The training of all methods was carried out using 1 or 2 Tesla V100 GPUs contained in a NVIDIA DGX-1 deep-learning server.[3] This server includes 8 V100 GPUs, each having 32 GB ram, and a dual 20-core Intel Xeon processor.

1) PRETRAINING

The pretraining of the networks includes two steps: (i) first, using the weights resulting from training the network on the ImageNet dataset (image classification task), and/or the COCO color dataset (object detection task), and then (ii) pretraining using COCO grayscale on the pose estimation task, based on the previous weights. Grayscale images are used

[3]https://www.amax.com/products/nvidia-products/nvidia-dgx-1/

TABLE 3. Base weights used as starting points for pretraining human pose estimation methods on COCO.

Method	Base training weights	Base training objective task
ViTPose [8]	ImageNet	Masked Image Modeling
RSN [7]	ImageNet	Masked Image Modeling
AlphaPose [6]	ImageNet	Classification
Simple Baselines [5]	ImageNet	Classification
CenterNet DLA-34 [11]	ImageNet + COCO (color)	Classification + Object Detection
CenterNet Hourglass-104 [11]	ImageNet + COCO (color)	Classification + Object Detection
CenterNet Hourglass-52 [11]	ImageNet	Classification
CenterNet HRNet-W32 [11]	ImageNet + COCO (color)	Classification + Object Detection
Bottom-Up HRNet [9]	ImageNet	Classification
PoseAE [12]	ImageNet	Classification
OpenPose* [10]	ImageNet	Classification

because thermal images are also monochromatic. The results of this procedure are pre-trained networks. The only exception is OpenPose, which, as mentioned before, is pretrained by using only color images.

With respect to the first step, the base weights available in the original code repositories were used. Some networks included in this work use base weights obtained by training directly on ImageNet. Others use base weights obtained by training on ImageNet first and then training in COCO color for object detection. For instance, using CenterNet with backbone DLA-34, network parameters are initialized from base weights trained over COCO color for object detection, which were, themselves, trained over ImageNet weights. Meanwhile, for other models, such as PoseAE, training is resumed from the base weights obtained directly from using ImageNet. Table 3 shows the information on the base weights used as a starting point for the pretraining of the methods.

The second pretraining step is performed on the task of person pose estimation. The training hyperparameters were chosen guided by the references of each work but accommodating for available computational resources when necessary. Table 4 shows a fuller description of the hyperparameters used in each case.

TABLE 4. Learning hyperparameters evaluated in this work for pretraining the networks on the COCO human keypoint dataset. *OpenPose was not pretrained on COCO in this work, but pretraining done by the authors includes values indicated in the table.

Model	Epochs	Image type	Initial learning rate	Learning rate steps	Batch size GPU 1	Batch size GPU 2
ViTPose [8]	210	Gray scale	0.0005	[170, 200]	64	-
RSN [7]	200	Gray scale	0.0005	-	32	-
AlphaPose [6]	200	Gray scale	0.001	pre-DPG: [90,120] DPG: [160,190]	32	-
Simple Baselines [5]	140	Gray scale	0.001	[90, 12]	32	32
CenterNet DLA-34 [11]	320	Gray scale	0.0005	[270, 300]	8	32
CenterNet Hourglass-104 [11]	150	Gray scale	0.00025	[130]	4	24
CenterNet Hourglass-52 [11]	200	Gray scale	0.0002	[170]	32	32
CenterNet HRNet-W32 [11]	320	Gray scale	0.001	[270, 300]	12	12
Bottom-Up HRNet [9]	140	Gray scale	0.001	[90, 12]	12	12
PoseAE [12]	326	Gray scale	0.0002	[200]	32	32
OpenPose* [10]	40	color	0.0001	-	10	10

2) FINETUNING

After pretraining, all methods are fine-tuned using the 600 annotated images from UCH-Thermal-Pose Set-A. Given the fact that this training set is small, various combinations of learning rate and batch size are used, deviating from the reference values. Additionally, different learning rate schedules are explored. Table 5 shows a summary of the best training parameters found in each case.

For CenterNet DLA-34 and CenterNet Hourglass-104, additional experiments were done by freezing different numbers of layers of the backbone network during fine-tuning (see Appendix A3 for details). These freezing regimes were labeled [Freeze 1 . . . Freeze N], where 1 corresponds to freezing just the first convolutional block right after the image

TABLE 5. Learning hyperparameters used for fine-tuning the networks evaluated in this work.

Methods	Epochs	Initial learning rate	Learning rate steps	Batch size GPU 1	Batch size GPU 2
ViTPose [8]	50	5.00E-04	-	16	-
RSN [7]	50	5.00E-04	-	16	-
AlphaPose [6]	50	1.50E-03	[35,45]	64	-
Simple Baselines [5]	50	1.00E-03	[35,45]	32	32
CenterNet DLA-34 [11]	50	5.00E-04	[35,45]	8	16
CenterNet Hourglass-104 [11]	20	3.50E-04	[15,18]	4	8
CenterNet Hourglass-52 [11]	20	2.00E-04	[15,18]	8	8
CenterNet HRNet-W32 [11]	50	1.00E-03	-	8	8
Bottom-Up HRNet [9]	50	1.00E-03	[35,45]	8	8
PoseAE [12]	50	2.00E-04	-	16	16
OpenPose [10]	7	5.00E-05	-	16	-

input, and N corresponds to freezing the entire backbone network. In the case of DLA, N = 6 was used, while in Hourglass, N = 4 was selected. Diagrams of Figures 7 and 8 illustrate which extensions of the backbone network each freeze regime includes, with the names of the weight parameters being frozen in each case.

The experiments showed that the best alternative for CenterNet DLA-34 and for CenterNet Hourglass-104 is to use N = 1 (see Figure 16 in Appendix A3), i.e., to freeze the first block. Therefore, in these two cases this freezing regime was selected for the fine-tuning process.

IV. EXPERIMENTAL RESULTS

A. PRECISION AND RECALL ON THE UCH-THERMAL-POSE SET-A DATASET

Tables 6 and 7 show the AP and AR for each model, evaluated on the validation subset consisting of 200 thermal images. These results are reported after pretraining the models over the COCO dataset, and then fine-tuning using 600 thermal images of the UCH-Thermal-Pose set-A, that is, using the best hyperparameters found for each model in this last stage (see Table 5). Also, as already mentioned, the variants used for the networks CenterNet DLA and CenterNet Hourglass-104 were trained with the first convolutional block frozen, since this proved to be the most effective (see Appendix A3).

The results obtained show that the highest precision among all models was obtained by ViTPose, with 85.7% average precision. Meanwhile, the highest recall was achieved also by the ViTPose model, with 88.7% average recall. A comparison between precision and recall scores shows that the problem of false positives is more serious than that of false negatives.

The results also show that top-down models obtain a higher performance than bottom-up models, both in precision and recall.

The results for CenterNet variants show that they perform well when using backbones DLA-34, and 2-stack Hourglass-104. Meanwhile, the system shows poor results when using backbones of 4-stack Hourglass-52, and HRNet-W32 networks. Using the latter as part of the Bottom-Up HRNet system yields the worst result in precision, and second-to-worst result in recall. Therefore, a conclusion can be made that this specific backbone architecture is not suited for human pose estimation on thermal images. The best-performing bottom-up thermal pose detector is PoseAE. Compared to ViTPose, it achieves a 6.7% lower AP and a 10.3% lower AR. Also, HRNet and OpenPose perform worse than the best CenterNet model.

B. INFERENCE SPEEDS

Inference times for each model and for all the test images were gathered to obtain the FPS. The results are shown in Table 8, on which the fastest top-down models are slower than the fastest bottom-up ones. Additionally, only ViTPose, RSN and the CenterNet variants with backbone DLA-34 and with backbone Hourglass-52 can be considered real-time detectors ($\geq$15 FPS). FPS was measured using two different GPUs: the Tesla V100 (see technical specifications in Section III-B), and the GTX 1660.[4] The latter GPU has 6 GB ram, and 1408 CUDA cores, 28% of those available on the Tesla V100.

In Figure 9, a diagram shows the FPS of the various methods when different numbers of people are included in the analysis. These results reflect the dependence of top-down model speeds on the number of people present in the image. For the four models belonging to this paradigm, speed decreases dramatically as more people are present in the image. Meanwhile, models following the bottom-up paradigm present nearly flat curves when being evaluated for images with different numbers of people. This is an important factor to consider when deciding what applications these types of models might have.

Figure 10 shows FPS versus AP for the various methods under comparison. ViTPose and RSN achieve both good AP and high FPS. The fastest method is CenterNet DLA-34, but its AP is considerably lower (around −20% AP) than the previous two methods.

[4]https://www.nvidia.com/en-us/geforce/graphics-cards/gtx-1660-ti/

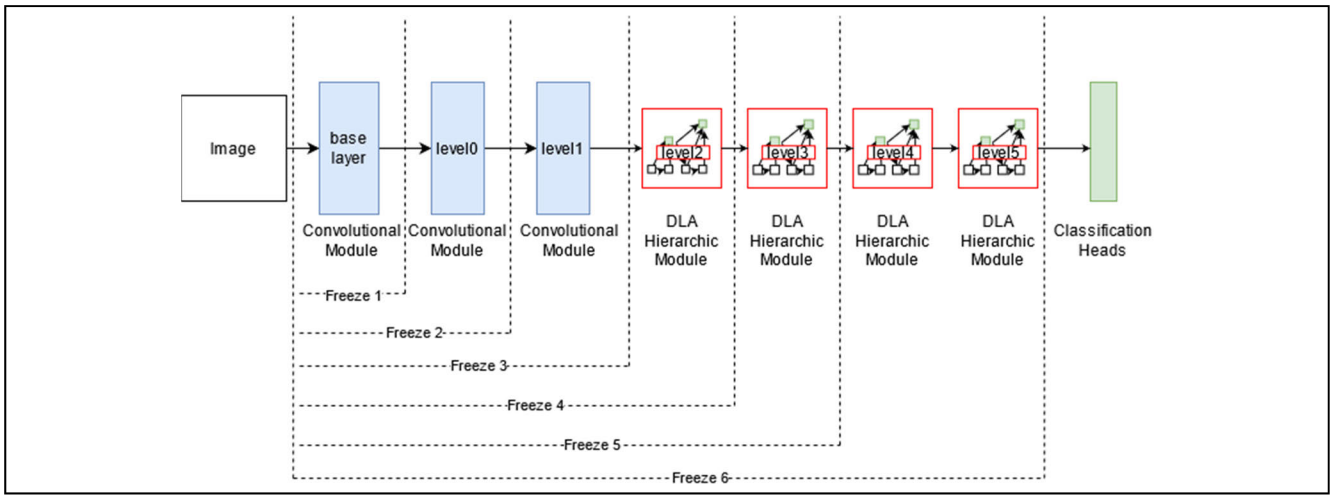

FIGURE 7. Convolutional blocks frozen for a CenterNet DLA-34 network.

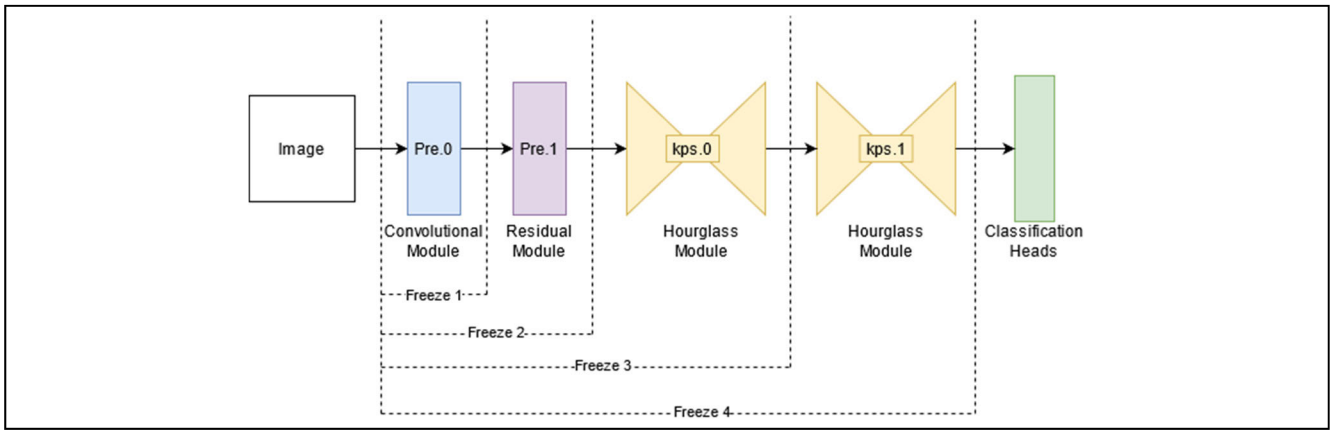

FIGURE 8. Convolutional blocks frozen for a CenterNet Hourglass-104 network.

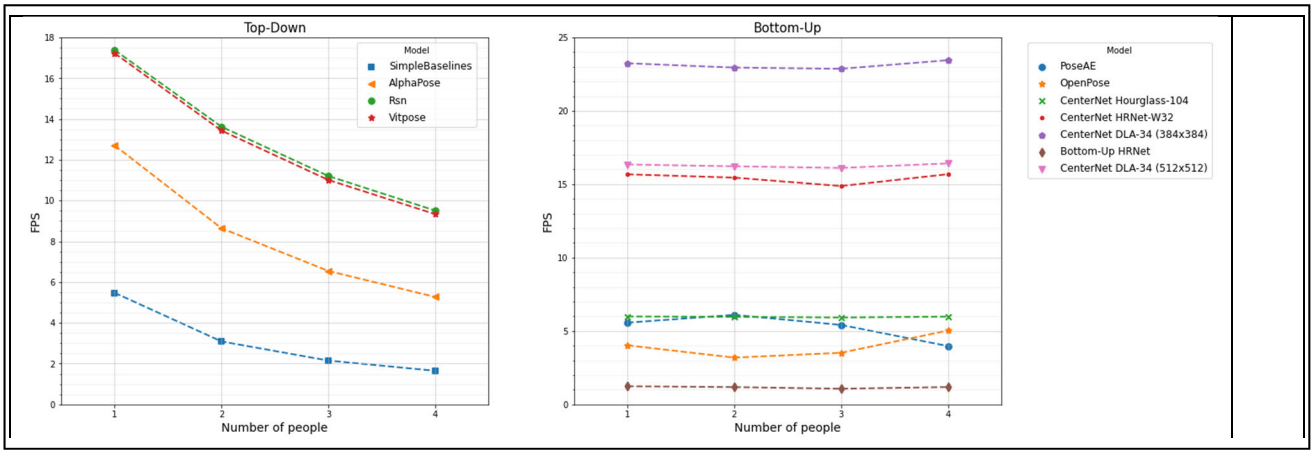

FIGURE 9. FPS when analyzing images containing different numbers of people for the methods under comparison.

C. PRECISION AND RECALL ON THERMAL IMAGES FROM UCH-THERMAL-POSE SET-B

Tables 9 and 10 show the AP and AR for each model, evaluated using the Set-B of the UCH-Thermal-Pose dataset. If these results are compared with those displayed in the previous section, the results obtained for this set are superior for every trained model. Most of the bottom-up models, with the exception of CenterNet with Hourglass-52 backbone, and Bottom-Up HRNet, score an AP over 70%, and an AR over 80%. Likewise, results for top-down models indicate an AP

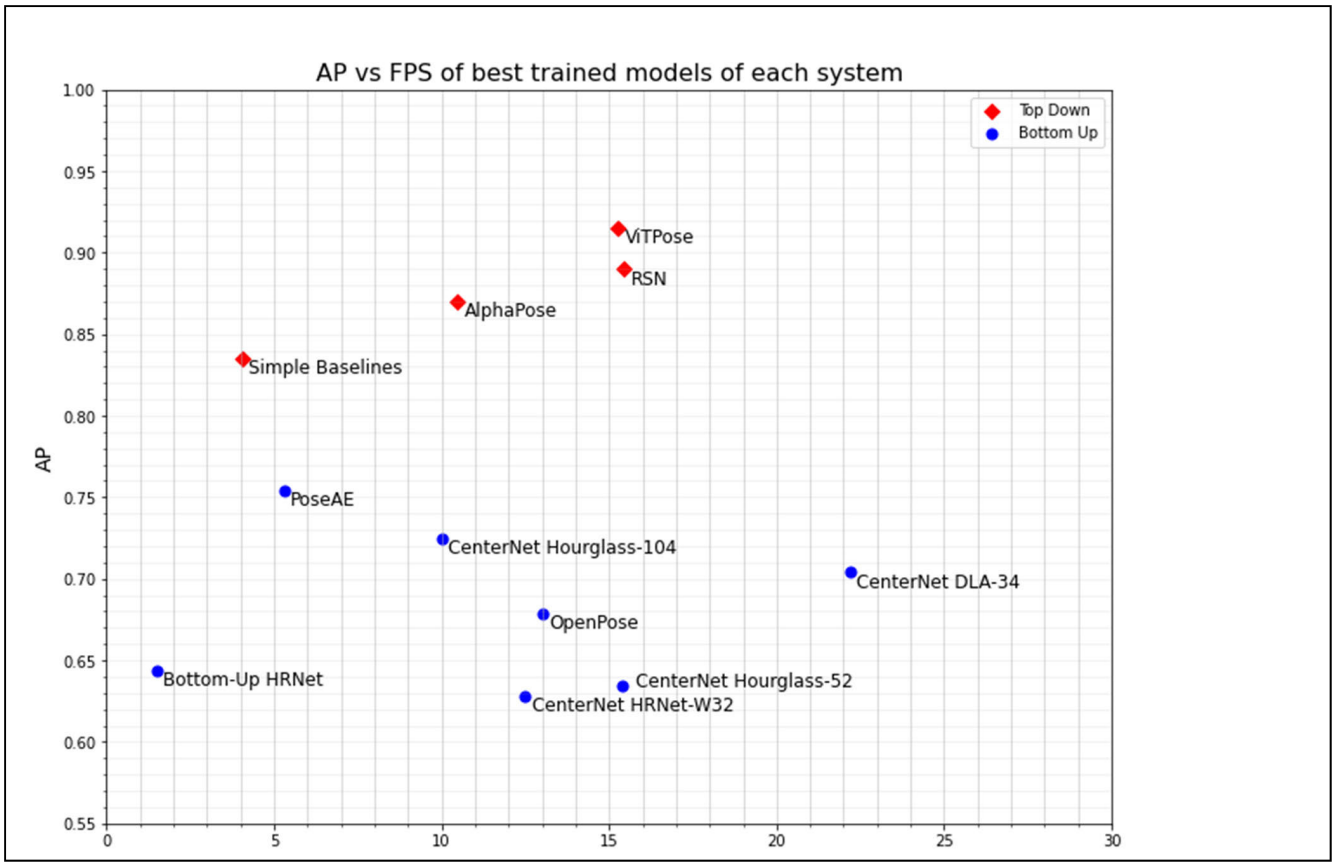

FIGURE 10. FPS versus AP for the different methods evaluated in this work.

over 83%, and an AR over 86%, reaching an impressive 91.5% average precision, and 94.1% average recall with the ViTPose method.

Nevertheless, the results displayed here should be taken with caution. Images from this set are quite closely correlated since they were all captured in the same indoor laboratory environment. On the other hand, images that belong to Set-A of UCH-Thermal-Pose Database are more varied and contain a greater array of poses and contexts. Therefore, it is expected that detection over that set would be a more challenging task.

By analyzing the results for bottom-up models, it can be observed that CenterNet with backbone HRNet-W32, and OpenPose score more favorably compared with other models from the same paradigm. This is a drastic change from what was evidenced in the experiments with UCH-Thermal-Pose set A, in which these models were among the least precise. PoseAE, which was highly performant in the UCH-Thermal-Pose Set A, achieves a low performance on UCH-Thermal-Pose Set B, but remains the best on large objects, compared to the highest scoring methods in this set. Also, Bottom-Up HRNet performs noticeably less well than CenterNet HRNet-W32, and OpenPose.

To see the kind of detection obtained by the methods, examples of human pose detection in some UCH-Thermal-Pose Set-B images using CenterNet DLA-34 are shown in Figure 11. Notice how the method correctly places keypoints on the different subjects.

V. ANALYSIS

Exhaustive experiments were performed to evaluate the behavior of different keypoint detectors on the thermal domain. Four top-down, and four bottom-up network architectures were compared, while the best training regimes for six of them were studied exhaustively. (The extra experiments are reported in the Appendix). The comparison of the networks considers runtime, AP, and AR, for different intersection-over-union thresholds, and for two different sizes of people (medium and large).

The main measures of performance for the detectors, i.e., AP, and speed, are summarized in Figure 10. It can be seen in that figure that ViTPose is the most highly performing method. On the other hand, CenterNet DLA-34 performs better than all the other architectures regarding inference speed, while maintaining an acceptable precision, and recall. Therefore, when high precision is needed for detecting keypoints of people in the thermal domain, ViTPose is recommended. On the other hand, when real time processing is the main requirement, CenterNet DLA-34 is the method of choice for detecting human skeletons. It must be noted that the best ViTPose model obtained was pretrained with a grayscale

TABLE 6. Average precision for various network architectures on the UCH-Thermal-Pose set-A validation split, pre trained on the COCO dataset and then fine-tuned. With APM/APL: AP for Medium/Large size objects, OD: Object Detection, and PE: Pose Estimation.

Experiment	Input resolution	AP	AP0.5	AP0.75	APM	APL
Top-Down						
ViTPose [8] + YOLOv7	OD: [640x640] PE: [256x192]	**0.857**	**0.978**	**0.932**	**0.813**	**0.901**
RSN [7] + YOLOv7	OD: [640x640] PE: [256x192]	0.809	0.949	0.903	0.773	0.840
AlphaPose [6] + YOLOv7	OD: [640x640] PE: [256x192]	0.813	0.962	0.908	0.778	0.840
Simple Baselines [5] + YOLOv7	OD: [640x640] PE: [384x288]	0.797	0.974	0.877	0.763	0.824
Bottom-Up						
CenterNet DLA-34 [11]	[384x384]	0.704	0.931	0.834	0.651	0.755
CenterNet DLA-34 [11]	[512x512]	0.692	0.935	0.804	0.622	0.769
CenterNet 2-Stack Hourglass-104 [11]	[512x512]	0.725	0.947	0.831	0.677	0.786
CenterNet 4-Stack Hourglass-52 [11]	[256x256]	0.634	0.917	0.729	0.578	0.702
CenterNet HRNet-W32 [11]	[512x512]	0.628	0.898	0.738	0.601	0.689
Bottom-Up HRNet [9]	[512x512]	0.644	0.910	0.805	0.634	0.717
PoseAE [12]	[256x256]	**0.754**	**0.959**	**0.882**	**0.728**	**0.817**
OpenPose [10]	[368x368]	0.679	0.877	0.768	0.615	0.745

TABLE 7. Average recall for various network architectures on UCH-Thermal-Pose set-A validation split, trained on the COCO dataset and then fine-tuned. With ARM/ARL: AR for Medium/Large size objects, OD: Object Detection and PE: Pose Estimation.

Experiment	Input resolution	AR	AR0.5	AR0.75	ARM	ARL
Top-Down						
ViTPose [8] + YOLOv7	OD: [640x640] PE: [256x192]	**0.887**	**0.986**	**0.948**	**0.850**	**0.916**
RSN [7] + YOLOv7	OD: [640x640] PE: [256x192]	0.845	0.969	0.921	0.822	0.866
AlphaPose [6] + YOLOv7	OD: [640x640] PE: [256x192]	0.846	0.976	0.924	0.823	0.867
Simple Baselines [5] + YOLOv7	OD: [640x640] PE: [384x288]	0.830	0.983	0.897	0.808	0.845
Bottom-Up						
CenterNet DLA-34 [11]	[384x384]	0.780	0.975	0.889	0.746	0.818
CenterNet DLA-34 [11]	[512x512]	0.772	0.971	0.868	0.717	0.837
CenterNet 2-Stack Hourglass-104 [11]	[512x512]	0.801	**0.979**	0.893	0.770	**0.848**
CenterNet 4-Stack Hourglass-52 [11]	[256x256]	0.722	0.964	0.814	0.676	0.771
CenterNet HRNet-W32 [11]	[512x512]	0.737	0.961	0.839	0.708	0.785
Bottom-Up HRNet [9]	[512x512]	0.724	0.950	0.857	0.715	0.743
PoseAE [12]	[256x256]	**0.820**	**0.979**	**0.911**	**0.797**	0.843
OpenPose [10]	[368x368]	0.735	0.900	0.811	0.688	0.780

version of COCO. This is also true for CenterNet DLA-34, although the latter was fine-tuned with the first convolutional blocks frozen (see Appendix A3).

It can also be noted that the accuracy of some of the trained models benefits from large resolution images. This is evidenced in Tables 11-16 in Appendix A1, in which,

TABLE 8. Frame rates for various network architectures. With OD: Object detection, and PE: Pose estimation.

Experiment	Input Resolution	FPS TESLA V100	FPS GTX 1660
Top-Down			
ViTPose [8]	OD: [640x640] PE: [256x192]	15.3	13.9
RSN [7]	OD: [640x640] PE: [256x192]	**15.4**	**15.0**
AlphaPose [6]	OD: [640x640] PE: [256x192]	10.5	9.9
Simple Baselines [5]	OD: [640x640] PE: [384x288]	4.1	3.9
Bottom-Up			
CenterNet DLA-34 [11]	[384x384]	**22.2**	**22.2**
CenterNet DLA-34 [11]	[512x512]	18.2	15.9
CenterNet 2-Stack Hourglass-104 [11]	[512x512]	10.0	9.1
CenterNet 4-Stack Hourglass-52 [11]	[256x256]	15.4	15.0
CenterNet HRNet-W32 [11]	[512x512]	12.5	16.9
Bottom-Up HRNet [9]	[512x512]	1.5	1.4
PoseAE [12]	[256x256]	5.3	5.77
OpenPose [10]	[368x368]	13.0	6.8

TABLE 9. Average precision for different network architectures on UCH-Thermal-Pose Set-B, trained on the COCO dataset, and then fine-tuned on the UCH- Thermal-Pose Set-A. With APM/APL: AP for Medium/Large size objects, OD: Object Detection, and PE: Pose Estimation. Detection of keypoints of people in the UCH-Thermal-Pose Set-B using CenterNet DLA-34.

Experiment	Input Resolution	AP	AP0.5	AP0.75	APM	APL
Top-Down						
ViTPose [8] + YOLOv7	OD: [640x640] PE: [256x192]	**0.915**	**0.983**	**0.950**	0.844	**0.934**
RSN [7] + YOLOv7	OD: [640x640] PE: [256x192]	0.890	0.967	0.936	0.833	0.904
AlphaPose [6] + YOLOv7	OD: [640x640] PE: [256x192]	0.870	0.950	0.941	**0.847**	0.874
Simple Baselines [5] + YOLOv7	OD: [640x640] PE: [384x288]	0.835	0.941	0.889	0.708	0.873
Bottom-Up						
CenterNet DLA-34 [11]	[384x384]	0.739	0.901	0.848	0.667	0.760
CenterNet DLA-34 [11]	[512x512]	0.767	0.937	0.858	0.712	0.786
CenterNet 2-Stack Hourglass-104 [11]	[512x512]	0.735	0.900	0.832	0.722	0.749
CenterNet 4-Stack Hourglass-52 [11]	[256x256]	0.678	0.894	0.769	0.628	0.704
CenterNet HRNet-W32 [11]	[512x512]	0.795	**0.967**	**0.928**	**0.771**	0.808
Bottom-Up HRNet [9]	[512x512]	0.662	0.882	0.755	0.578	0.748
PoseAE [12]	[256x256]	0.759	0.876	0.847	0.526	**0.856**
OpenPose [10]	[368x368]	**0.806**	0.955	0.890	0.692	0.844

in some cases, precision and recall were reduced to half or less by using input images half the size of the images used for training. These tables also show that bottom-up detectors are most affected by lowering input resolution, compared with top-down detectors. This may be due to the fact that top-down detectors separate person detection and pose estimation into

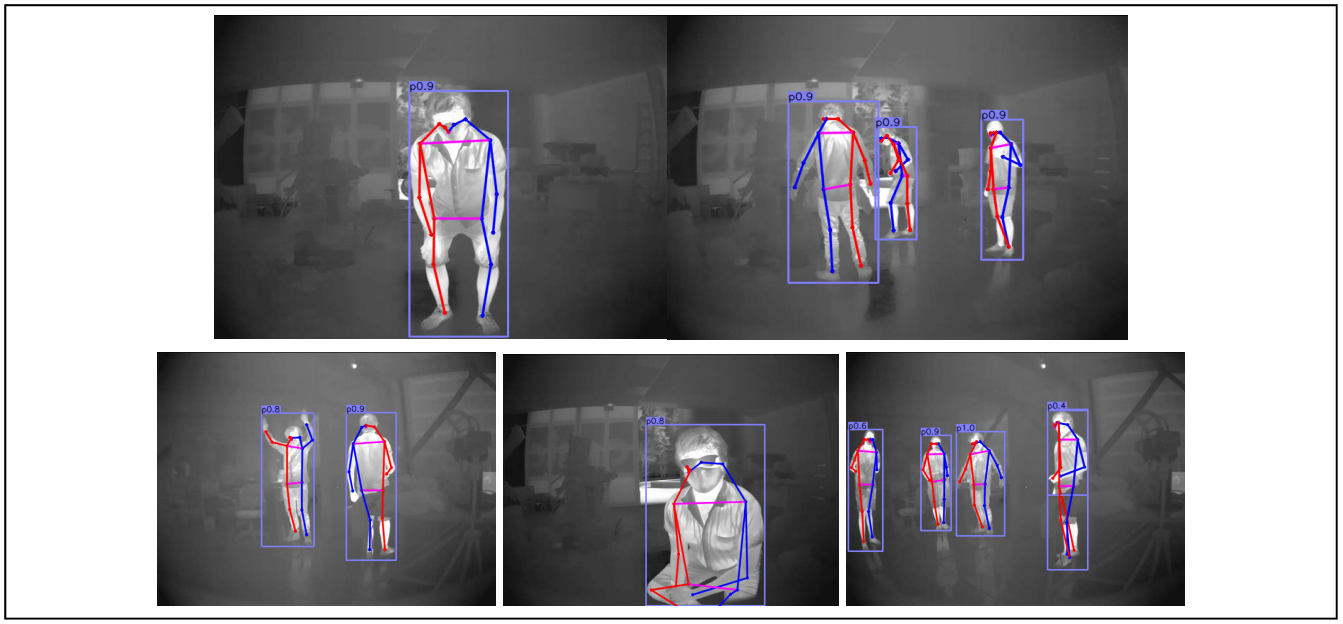

FIGURE 11. Detection of keypoints of people in the UCH-Thermal-Pose Set-B using CenterNet DLA-34.

TABLE 10. Average recall for different network architectures on UCH-Thermal-Pose Set-B, trained over the COCO dataset, and then fine-tuned on the UCH-Thermal-Pose set-A. With ARM/ARL: AR for Medium/Large size objects, OD: Object detection, and PE: Pose estimation.

Experiment	Input Resolution	AR	AR0.5	AR0.75	ARM	ARL
Top-Down						
ViTPose [8] + YOLOv7	OD: [640x640] PE: [256x192]	**0.941**	**0.994**	**0.969**	0.893	**0.951**
RSN [7] + YOLOv7	OD: [640x640] PE: [256x192]	0.919	0.988	0.956	0.893	0.925
AlphaPose [6] + YOLOv7	OD: [640x640] PE: [256x192]	0.909	0.975	0.969	0.893	0.912
Simple Baselines [5] + YOLOv7	OD: [640x640] PE: [384x288]	0.874	0.969	0.919	0.810	0.889
Bottom-Up						
CenterNet DLA-34 [11]	[384x384]	0.825	0.944	0.906	0.762	0.839
CenterNet DLA-34 [11]	[512x512]	0.846	0.969	0.919	0.779	0.861
CenterNet 2-Stack Hourglass-104 [11]	[512x512]	0.846	0.963	0.925	0.793	0.857
CenterNet 4-Stack Hourglass-52 [11]	[256x256]	0.779	0.963	0.850	0.752	0.785
CenterNet HRNet-W32 [11]	[512x512]	**0.864**	**0.988**	**0.969**	**0.821**	0.874
Bottom-Up HRNet [9]	[512x512]	0.762	0.963	0.831	0.693	0.778
PoseAE [12]	[256x256]	0.840	0.938	0.906	0.659	**0.880**
OpenPose [10]	[368x368]	0.849	0.975	0.912	0.776	0.866

two parts, while bottom-up detectors perform both detections simultaneously, so they can be more affected by changes in the input resolution.

VI. FUTURE CHALLENGES

After analyzing the results obtained in the experiments reported, some future challenges were identified. The first

challenge is the need for having architectures whose structure is adapted to thermal images. The architectures explored in this work are aimed at processing color images, which implies that the same thermal image is fed into three channels at the input of the network, which is redundant for thermal images. Exploring architectures that consider only one input image (a thermal one) could improve the runtime of the methods.

A second challenge is related to the possibility of using RGB and thermal images in an integrated way. A detection system could benefit from switching between both modalities depending on the environmental conditions, such as using RGB images for well-illuminated surroundings, and thermal images for low-light surroundings. However, additional efforts would be needed to capture the state of the environment and to automate the switching between both modalities.

Another alternative is feeding both RGB and thermal images into the detection networks simultaneously, to benefit from the complementary nature of both sensors. This can be achieved by adding the thermal image as an extra channel in the network inputs. However, thermal and color images need to be aligned first for performing this step, which imposes an extra challenge on the annotation procedure, since most of the datasets available do not have intrinsic and extrinsic parameters for the cameras. In addition, calibrating thermal cameras is not as straightforward as it is with RGB cameras because normal checkerboards are not useful for this task. Also, the error of the mapped keypoints between the two images increases with the baseline between the cameras. Color images can be aligned by using local descriptors; however, this procedure is not straightforward for aligning images from different domains. Thus, the automatic alignment of images from the two domains could be an interesting topic to be explored in the future, as this could enable the use of both thermal and color images in a unified and straightforward way for tasks like person detection in environments containing dust, or detection of people in hospital beds, in which they might be either covered or uncovered.

VII. CONCLUSION

Several network architectures for detecting keypoints of people in the thermal domain are explored and compared in this work. Since the different networks were developed for color images, their performance on thermal images cannot be deduced from studies in the existing literature. Procedures for training the networks were explored, such as using images from the COCO dataset transformed into grayscales when pretraining the networks. This procedure was able to increase the accuracy of the networks for most of them. Also, freezing different numbers of convolutional blocks when pretraining was also explored, and freezing only the first convolutional block proved to increase the accuracy of the networks when trained in the thermal domain.

A dataset labeled with human keypoints in the thermal domain, UCH-Thermal-Pose, was created for enabling a comparison between the different networks. UCH-Thermal-Pose Set-A is composed of images from various existing datasets, in which keypoints on the selected images were annotated manually, while the UCH-Thermal-Pose Set-B is composed of thermal images acquired in our laboratory, and annotated using the same procedure used for annotating the images in set A.

The results presented in this paper show that pretraining with grayscale images improves the performance of most of the network architectures, while freezing the first layer improves the results on the CenterNet DLA-34, and CenterNet Hourglass-104 architectures.

The best architectures obtained for detecting human keypoints on thermal images are ViTPose when high accuracy is required, and CenterNet DLA-34 when real time processing is the main requirement.

TABLE 11. Metrics for bottom-up detectors for input resolutions 128 × 128 and 256 × 256.

	128x128			256x256		
Model	**AP**	**AR**	**FPS**	**AP**	**AR**	**FPS**
CenterNet DLA-34	0.288	0.374	**34.5**	0.576	0.664	**31.3**
CenterNet DLA-34 (train 384x384)	0.377	0.447	33.3	0.572	0.667	**31.3**
CenterNet 2-Stack Hourglass -104	0.307	0.382	15.9	**0.600**	**0.700**	10.5
CenterNet 4-Stack Hourglass -52	**0.380**	**0.484**	20.8	0.583	0.684	11.9
CenterNet HRNet-W32	0.268	0.325	17.5	0.475	0.595	17.2
Bottom-Up HRNet	0.218	0.282	2.6	0.475	0.581	1.8
PoseAE	0.000	0.000	27.0	0.000	0.000	20.0
OpenPose	0.224	0.285	37.0	0.525	0.582	20.0

APPENDIX A
SELECTION OF BEST TRAINING PROCEDURES AND HYPERPARAMETERS FOR THE NETWORKS

A.1 TESTING DIFFERENT INPUT RESOLUTIONS

AP, AR, and FPS were tested for bottom-up models using different input resolutions. The results are shown in Tables 11 and 12. Using a low input resolution, such as 128 × 128, increases speed significantly for all models. Each one of them, except for Bottom-Up HRNet, performs at real-time speeds with that resolution, but with very low precision.

TABLE 12. Metrics for bottom-up detectors for input resolutions 384 × 384 and 512 × 512.

	384x384			512x512		
Model	**AP**	**AR**	**FPS**	**AP**	**AR**	**FPS**
CenterNet DLA-34	0.641	0.738	23.8	0.714	0.791	16.1
CenterNet DLA-34 (train 384x384)	0.654	0.743	**24.4**	0.694	0.779	**16.9**
CenterNet 2-Stack Hourglass-104	**0.674**	**0.766**	7.1	0.717	0.797	6.0
CenterNet 4-Stack Hourglass-52	0.612	0.689	7.6	0.557	0.640	6.5
CenterNet HRNet-W32	0.532	0.674	17.2	0.629	0.737	**16.9**
Bottom-Up HRNet	0.548	0.657	1.6	0.619	0.708	1.5
PoseAE	0.667	0.763	10.0	**0.738**	**0.803**	5.3
OpenPose	0.641	0.699	12.5	0.672	0.726	6.6

TABLE 13. Metrics for top-down detectors. The person detector is trained for input resolutions 128 × 128 and 256 × 256.

	YOLOv3 DETECTOR INPUT RESOLUTION					
	128x128			**256x256**		
Model	**AP**	**AR**	**FPS**	**AP**	**AR**	**FPS**
AlphaPose (detected bboxes)	0.591	0.636	12.2	0.707	0.747	12.8
Simple Baselines (detected bboxes)	0.547	0.595	3.5	0.64	0.686	3.9

As resolution increases, precision and recall improve but are accompanied by a decrease in speed. Therefore, it is easy to establish a tradeoff between speed and precision/recall for these models.

With top-down models, changing input resolution is more complex since there are two networks with different input resolutions: the person detector, and the keypoint detector. Nevertheless, experiments were made by changing the input size of the YOLOv3 person detector while maintaining the input resolution of the pose estimation network. These results are shown in Tables 13 and 14. Note that YOLOv3 was used in these experiments because of it being used in the AlphaPose paper [6].

TABLE 14. Metrics for top-down detectors. The person detector is trained for input resolutions 384 × 384 and 512 × 512.

	YOLOv3 DETECTOR INPUT RESOLUTION					
	384x384			**512x512**		
Model	**AP**	**AR**	**FPS**	**AP**	**AR**	**FPS**
AlphaPose (detected bboxes)	0.741	0.781	12.3	0.731	0.762	12.4
Simple Baselines (detected bboxes)	0.689	0.731	4.0	0.685	0.721	4.1

TABLE 15. Metrics for top-down detectors. The keypoint detector is trained for input resolutions 128 × 128 and 256 × 256.

	POSE ESTIMATOR INPUT RESOLUTION					
	128x96			**256x192**		
Model	**AP**	**AR**	**FPS**	**AP**	**AR**	**FPS**
AlphaPose (ground truth bboxes)	0.050	0.003	24.1	0.791	0.825	20.9
AlphaPose (detected bboxes)	0.003	0.002	8.8	0.088	0.118	7.5

TABLE 16. Metrics for top-down detectors. The person detector is trained for input resolutions 384 × 384 and 512 × 512.

	POSE ESTIMATOR INPUT RESOLUTION					
	384x288			**512x384**		
Model	**AP**	**AR**	**FPS**	**AP**	**AR**	**FPS**
AlphaPose (ground truth bboxes)	0.716	0.756	18.6	0.456	0.537	20.7
AlphaPose (detected bboxes)	0.777	0.815	5.1	0.675	0.710	4.5

Remarkably, the AP and AR of top-down detectors are less affected by a decrease in image resolution when compared to decreases in resolution for bottom-up detectors. As can be seen, both keep an AP over 50%, and an AR over 59% when the detecting was done on 128 × 128 images. And even more, at 256 × 256 images, AP and AR are very close to those obtained at high resolution images. This fact can be explained

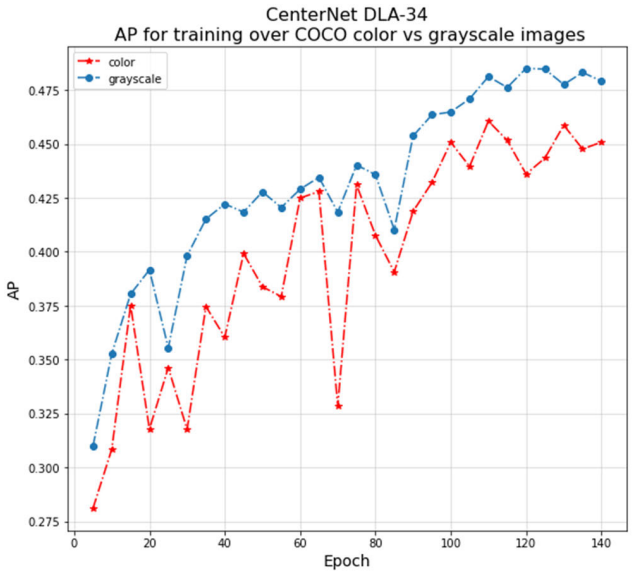

FIGURE 12. Average precision for CenterNet DLA-34 trained on grayscale and color images, for different training epochs.

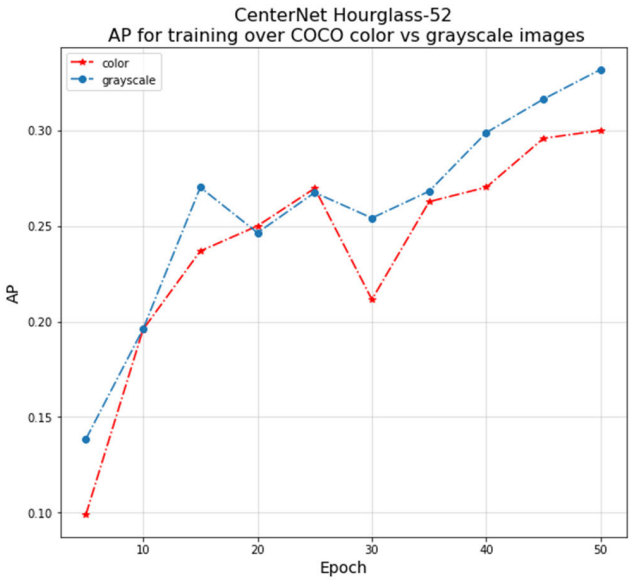

FIGURE 14. Average precision for CenterNet Hourglass-52 trained on grayscale and color images, for different training epochs.

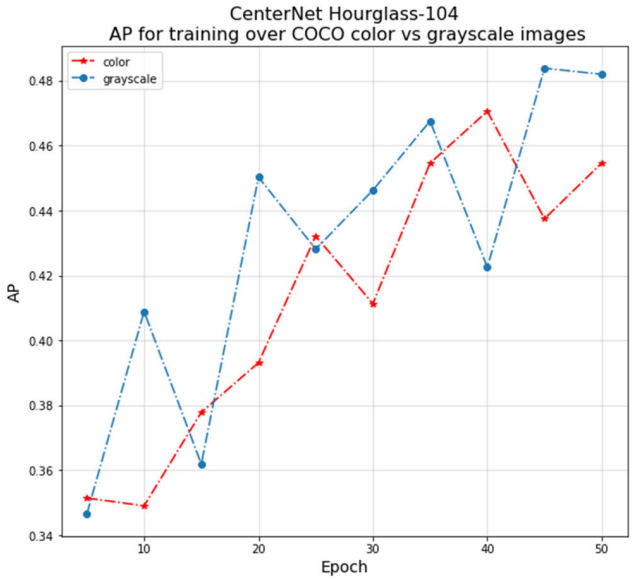

FIGURE 13. Average precision for CenterNet Hourglass-104 trained on grayscale and color images, for different training epochs. Average precision for CenterNet DLA-34 trained on grayscale and color images, for different training epochs.

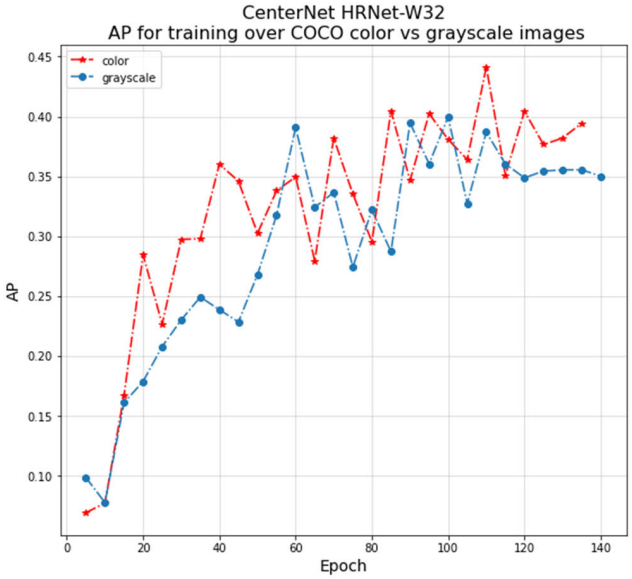

FIGURE 15. Average precision for CenterNet HRNet-W32 trained on grayscale and color images, for different training epochs.

by object detectors being robust when detecting people over low-resolution images.

However, no gains in speed are evidenced when detecting on smaller images. FPS remains almost the same for the different input sizes considered. A plausible reason behind this is that, even though the object detector receives lower resolution images, the pose estimation network keeps operating at the same resolution. This constitutes a bottleneck since the actual speed of the full detector is determined mainly by the pose estimation network. Additionally, we observed that the YOLOv3 model predicts a larger number of false positives when receiving low-resolution images, which means that a greater number of bounding boxes would be fed to the pose estimation network.

On the other hand, when changing the input resolution of the pose estimation network but preserving the resolution of the whole images, the results shown in Tables 15 and 16 are obtained. The models are evaluated using both ground truth bounding boxes, and detected ones. By observing the results, it can be seen that, when using input resolutions lower than the default ones of the models (256×192 for AlphaPose,

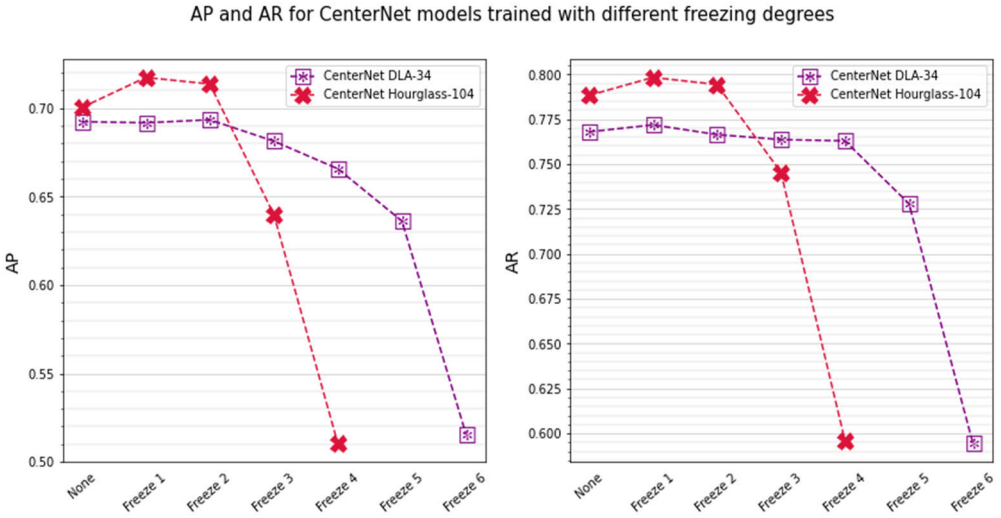

FIGURE 16. Average precision and average recall for two CenterNet models, for different numbers of frozen convolutional blocks.

and 384 × 288 for Simple Baselines), precision and recall are negatively impacted, reaching values below 6% for all the metrics. Moreover, when increasing resolution from default values, precision and recall are also negatively affected by large margins. Speed is not substantially modified when using different resolutions.

A.2 COMPARISON BETWEEN PRETRAINING ON GRAY SCALE AND COLOR IMAGES

Experiments carried out with the CenterNet architecture show that models pretrained over gray images exhibit, in general, a greater affinity for detection in the thermal domain with respect to models pretrained over color images. Figure 12 shows some good examples of the previous statement, exhibiting two identical CenterNet DLA-34 networks trained with grayscale images, and with color images, respectively, over 140 epochs. The AP of both variants was evaluated on the thermal images test set in each epoch (without fine-tuning).

Figures 12-15 show the results of using the DLA-34, Hourglass-104, Hourglass-52, and HRNet-W32 backbones. With networks using the first three backbones, training using grayscale images provides better results compared to those when using color images, as is shown in Figures 12 to 14. Nevertheless, this statement does not hold true when using an HRNet-W32 backbone, where the curves corresponding to training with RGB and grayscale images are more intertwined as can be seen in Figure 15.

A.3 FINE-TUNING CENTERNET WITH FROZEN LAYERS

During the training process, freezing layers promote obtaining better results when little training data is available, since a lower number of parameters from the network need to be learned. For some architectures, freezing individual convolutional layers can be difficult to implement, but in that case convolutional blocks, which are subsets of convolutional layers, can be frozen. Results from experiments on freezing different numbers of convolutional blocks (freeze 1 to freeze 6) for two CenterNet models, DLA-34, and Hourglass-104, are shown in Figure 16. The convolutional blocks used for the freezing experiments are shown in Figures 7 and 8.

As is shown in Figure 16, precision and recall lower significantly when the entire backbones are frozen (freeze 4 in Hourglass and freeze 6 in DLA). However, when using the freeze 1 regime, both networks show gains in both precision and recall. Recalling Figures 7 and 8, this regime corresponds to freezing the first convolutional block for each backbone. The freeze 1 regime was included as part of the training of the final models reported for these variants of CenterNet.

REFERENCES

[1] S. Liu and S. Ostadabbas, "Seeing under the cover: A physics guided learning approach for in-bed pose estimation," in *Proc. MICCAI*, 2019, pp. 236–245.

[2] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov, "ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2019, pp. 606–624.

[3] *FREE FLIR Thermal Dataset for Algorithm Training*, FLIR Syst., Orlando, FL, USA, 2020.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[5] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2019, pp. 466–481.

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[7] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2020, pp. 455–472.

[8] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," 2022, *arXiv:2204.12484*.

[9] K. Sun, Z. Geng, D. Meng, B. Xiao, D. Liu, Z. Zhang, and J. Wang, "Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates," 2020, *arXiv:2006.15480*.
[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[11] X. Zhou, D. Wang, and P. Krahenbuhl, "Objects as points," 2019, *arXiv:1904.07850*.
[12] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 2277–2287.
[13] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[15] J. Redmon, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016, pp. 779–788.
[16] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[18] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[19] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 2980–2988.
[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.
[21] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digit. Signal Process.*, vol. 126, Jun. 2022, Art. no. 103514.
[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–22.
[23] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.
[24] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, vol. 9912, 2016, pp. 483–499.
[25] S. Liu, X. Huang, N. Fu, C. Li, Z. Su, and S. Ostadabbas, "Simultaneously-collected multimodal lying pose dataset: Towards in-bed human pose monitoring under adverse vision conditions," 2020, *arXiv:2008.08735*.
[26] I.-C. Chen, C.-J. Wang, C.-K. Wen, and S.-J. Tzou, "Multi-person pose estimation using thermal images," *IEEE Access*, vol. 8, pp. 174964–174971, 2020.
[27] R. Mehra, M. Chetty, and J. K. Kamalu, "Multiperson pose estimation using thermal and depth modalities," Stanford, CA, USA, Tech. Rep., 2017.
[28] W. Kentaro, "Labelme: Image polygonal annotation with Python," GitHub Repository, 2016.
[29] R. Bergevin, P.-L. St-Charles, and G.-A. Bilodeau, "Mutual foreground segmentation with multispectral stereo pairs," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 375–384.
[30] Y. Liu, Z. Lu, J. Li, C. Yao, and Y. Deng, "Transferable feature representation for visible-to-infrared cross-dataset human action recognition," *Complexity*, vol. 2018, pp. 1–20, Jan. 2018.
[31] D. Olmeda, U. Nunes, J. Armingol, and A. La Escalera, "LSI far infrared pedestrian dataset," Intell. Syst. Lab, e-cienciaDatos, 2013.
[32] Z. Wu, N. Fuller, D. Theriault, and M. Betke, "A thermal infrared video benchmark for visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 201–208.
[33] R. Miezianko, "Terravic research infrared database," IEEE OTCBVS WS Series Bench.
[34] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Proc. 7th IEEE Workshops Appl. Comput. Vis.*, Jan. 2005, pp. 364–369.
[35] A. Akula, R. Ghosh, S. Kumar, and H. K. Sardana, "Moving target detection in thermal infrared imagery using spatiotemporal information," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 30, no. 8, pp. 1492–1501, 2013.
[36] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.
[37] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.
[38] R. Gade and T. B. Moeslund, "Constrained multi-target tracking for team sports activities," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–11, Dec. 2018.
[39] Y. Socarras, S. Ramos, D. Vazquez, A. Lopez, and T. Gevers, "Adapting pedestrian detection from synthetic to far infrared images," in *Proc. ICCV*, 2013, pp. 1–3.
[40] C. Palmero, A. Clapés, C. Bahnsen, A. Mogelmose, T. B. Moeslund, and S. Escalera, "Multi-modal RGB–depth–thermal human body segmentation," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 217–239, Jun. 2016.
[41] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

JAVIER SMITH was born in Concepción, Biobío Region, Chile, in 1996. He received the B.S. and Engineer's degrees in electrical engineering from Universidad de Chile, Santiago, Metropolitan Region, Chile, in 2021. Since 2021, he has been a Machine Learning Engineer with SoyMomo, Santiago. His research interests include computer vision, reinforcement learning, and natural language processing.

PATRICIO LONCOMILLA was born in Santiago, Chile, in 1980. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Universidad de Chile, Santiago, in 2004, 2006, and 2011, respectively. Since 2011, he has been a Research Assistant with the Advanced Mining Technology Center, Universidad de Chile, and a full-time Researcher, since 2019. His research interests include signal and image processing, pattern recognition, visual object recognition and representation, and robotics, with more than 20 publications in these fields.

JAVIER RUIZ-DEL-SOLAR (Senior Member, IEEE) received the degree in electrical engineering from Universidad Técnica Federico Santa María, Valparaíso, Chile, in 1991, and the Doctor–Engineer degree from the Technical University of Berlin, Berlin, Germany, in 1997. Since 2009, he has been the Executive Director of the Advanced Mining Technology Center, Universidad de Chile, Santiago, Chile. His main research interests include fundamental research in perception and learning and applications of robotics technology in the real world. In recent years, his research has focused on the application of deep reinforcement learning to mobile robotics applications.