Journal of Thermal Analysis and Calorimetry
https://doi.org/10.1007/s10973-024-13343-w

# StereoThermoLegs: label propagation with multimodal stereo cameras for automated annotation of posterior legs during running at different velocities

**Daniel Andrés López[1] · Barlo Hillen[2] · Markus Nägele[3] · Perikles Simon[2] · Elmar Schömer[1]**

Received: 31 August 2023 / Accepted: 24 May 2024


**Abstract**
In sports science, thermal imaging is applied to investigate various questions related to exercise-induced stress response, muscle fatigue, anomalies, and diseases. Infrared thermography monitors thermal radiation from the skin's surface over time. For further analysis, regions of interest are extracted and statistically analyzed. Although computer vision algorithms have grown in recent years due to data-driven approaches, this is not the case for detailed segmentation in thermal images. In a supervised manner, machine learning optimizations require a large amount of training data with input and ground truth output data. Unfortunately, obtaining annotated data are a costly problem that increases with the complexity of the task. For semantic segmentation, pixel-wise label masks must be created by experts. Few datasets meet the needs of sports scientists and physicians to perform advanced applications of thermal computer vision during physical activity and generate new insights in their fields. In this paper, a new method is introduced to transfer segmentation masks from the vision domain to the thermal domain with a stereo-calibrated time-of-flight camera and high-resolution mid-wave infrared camera. A post-processing procedure is then utilized to obtain dense pixel masks for the posterior legs during walking and running on a treadmill. The developed StereoThermoLegs dataset is based on 14 participants and includes 11 subjects for training with 12,826 thermograms and the remaining three individuals for testing with 3433 images. A deep neural network was trained with the DeepLabv3+ architecture, the AdaBelief optimizer, and Dice loss as a benchmark. After 29 epochs, the test set achieved an average intersection over union of 0.66. The analysis of the posterior leg region, specifically the left and right calf, offered the most insights, with values of 0.83 and 0.83, respectively. The first multimodal stereo dataset containing synchronized visual and thermal images of a runner's back provides a starting point for data-driven segmentation tasks in sports science and medicine. Our technique allows for automatic production of customized datasets for deep learning, accelerating the implementation of baseline outcomes for newly identified areas of interest in thermal imaging, while bypassing the requirement for extensive manual annotation. The approach is not exclusive to stereo rig and segmentation tasks utilizing RGBD and thermal cameras, but can be applied to other imaging tasks and modalities.

**Keywords** Artificial neural networks · Incremental exercise testing · Semantic segmentation · Thermal imaging

✉ Daniel Andrés López
daniel.andres@uni-mainz.de

Barlo Hillen
b.hillen@uni-mainz.de

Markus Nägele
markus.naegele@optoprecision.de

Perikles Simon
simonpe@uni-mainz.de

Elmar Schömer
schoemer@uni-mainz.de

[1] Research Group of Computational Geometry, Institute of Computer Science, Johannes Gutenberg University Mainz, Staudinger Weg 9, 55128 Mainz, Germany

[2] Department of Sports Medicine, Disease Prevention and Rehabilitation, Institute of Sports Science, Johannes Gutenberg University Mainz, Albert-Schweitzer-Straße 22, 55128 Mainz, Germany

[3] OptoPrecision GmbH, Auf der Höhe 15, 28357 Bremen, Germany

Published online: 18 June 2024

## Introduction

In sports science and medicine, infrared thermography (IRT) is applied to assess the surface radiation of human skin. Applications include inflammation detection [1], symmetry in muscle activity [2], tumor detection (breast cancer) [3], internal load assessment [4, 5]. Magalhaes et al. have shown a huge amount of applications that employ thermal image data [6], but they focused on generating insights from processed thermal data and not on how to process thermograms. The region of interest (ROI) and information extraction is still often done manually [7], which limits the amount of data to a few examples of a large possible data stream. Perpetuini et al. report that the studies either select the ROIs manually or do not mention how the ROI is extracted. There are only a few studies that describe an automated ROI selection process.

When an IRT study involves automatic ROI selection, a problem specific algorithm is developed for a particular type of thermal image, as in [8], or geometric shapes are tracked, as in [9]. Data-driven algorithms for real shape extraction of human body parts are rarely discussed in sports science and medicine related publications (e.g., [10]). Together with the need for manual work for data labeling and data policies, no publicly available datasets have yet been published that meet all the needs of these disciplines, such as thermograms of resting or exhausted humans, high resolution on specific body parts, or stereo pairs for visual and thermal fusion.

We present a new dataset *StereoThermoLegs* containing image pairs of thermograms and visual data together with corresponding label masks for both domains for the segmentation of posterior legs in thermograms acquired during standing, walking and running. The dataset is automatically generated with data from a known domain, vision images, and transforming the information to the unknown domain, thermal radiation. In Fig. 1, examples of the resulting segmentation masks are shown next to the thermograms. The masks, along with the thermograms, can be employed for supervised learning in the thermal image domain, with no need for the visual modality.

## Related work

Thermal datasets with humans have already been created for various applications. He et al. reviewed the algorithmic application of thermal cameras [11] and discussed deep learning techniques in this area. Deep learning requires prepared datasets for optimization, but none of the existing ones focus on medical applications with exercising humans. In the work of Hillen et al., they reviewed several necessities for high quality thermograms [12], which also are not available in current datasets.

With ThermalFaceDB [13], the authors published a dataset of thermal facial expressions along with facial landmarks for face recognition. This includes a variety of head positions and different people and expressions, but its application in sports science is limited due to the lack of images of physically exhausted individuals. Kniaz et al. showed with ThermalGAN a novel work on color to thermal image translation to improve person re-identification [14]. Along with the dataset, ThermalWorld was released, which contains more than 5000 annotated images with 10 classes of outdoor scenes such as people, cars, or buildings. The purpose of the dataset is to help people re-identification in low-light scenes with thermal cameras.

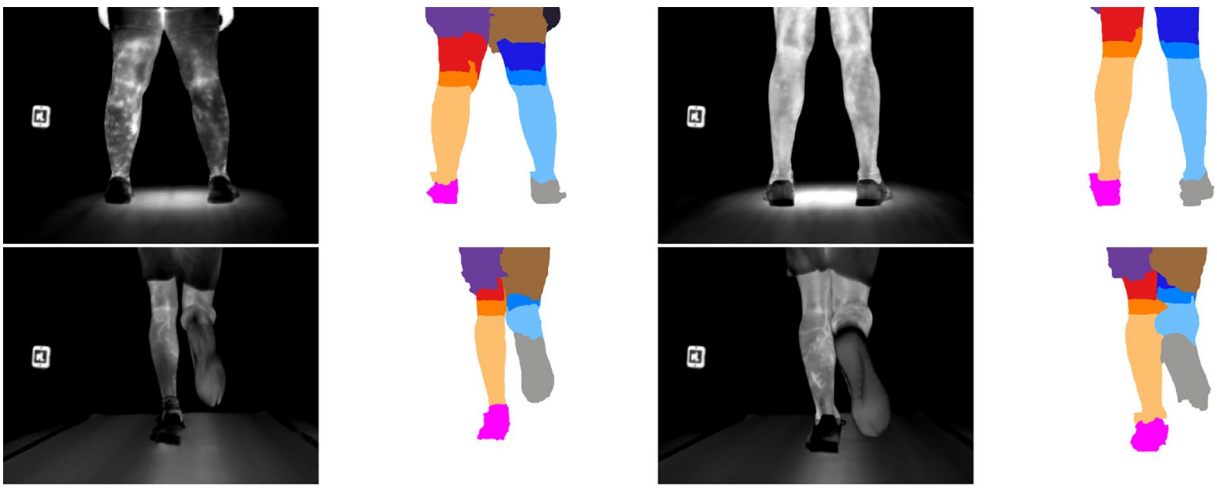


**Fig. 1** 4 Samples of proposed *StereoThermoLegs* dataset: thermogram and its corresponding segmentation mask with each side from top to down: clothes, thigh, knee, calf, shoe. Top left and right are resting after the experiment trial, lower left and right are running.

In their paper [15], Soldan et al. describe a system to fuse a thermal camera with a time-of-flight (ToF) camera by registering them in the coordinate system. The calibration process is key to obtaining a stable system and has been thoroughly analyzed by the authors. This includes considerations for building a calibration target that is visible in each modality: visible light, depth, and thermal spectrum. They suggest that a heated cardboard circle-based pattern works best in all three modalities. Each camera itself must be calibrated, and stereo calibration requires an image in all three fields of view simultaneously.

In the work of Richter et al., they present a method for fusing a ToF camera with a thermal camera to directly assess regions of interest by finding a skeletal pose in the ToF data and transforming it to the thermal domain [16]. The skin temperature is measured along the skeletal region of interest. The transformation process is performed with a calibrated stereo system of both systems. The idea of transforming data from one domain to another must be applied to any measurement, as it leads directly to the region of interest to be analyzed. Our approach is based on the same principles, but we focus on dense mapping instead of skeletal parts, and we show how to generate a dataset for training a single camera.

Knyaz and Mashkantsev [17] demonstrated the fusion of thermal and vision cameras to produce multimodal 3D reconstructed scenes. They equipped an unmanned aerial vehicle (UAV) with a stereo rig and captured outdoor scenes and buildings. Calibration was performed by finding calibration patterns and known building features such as straight lines, as well as additional distances measured with a laser scanner to obtain true dimensions. A structure-from-motion 3D model was created for each camera, and the 3D models were fused together.

Bultman et al. proposed in [18] a sensor fusion system mounted on a UAV including a vision and a thermal camera together with a LiDAR sensor. In the work, segmentations from multiple domains are merged to generate a fused segmentation map and 3D point cloud. Furthermore, the training of individual segmentation algorithms is improved by propagating labels from one modality to another (vision to LiDAR) together with similar existing datasets and retraining the algorithm.

Hillen et al. described a deep learning method to segment the posterior legs in the movements of running persons and analyze the region over a whole exercise [5]. Additionally, they compared manual and automatic methods of ROI selection and found that the automatic method was superior, especially the analysis with several thousands thermograms. Their method was refined in [4] by separating left and right and focusing on the most interesting part in the posterior leg in movement, the calves. Although for both methods, the training dataset was annotated manually and is limited in the number of thermograms.

## Methods

In this section, we describe how a stereo rig is leveraged to create a dense annotated segmentation dataset, and we present a benchmark deep neural network (DNN) training procedure for future comparisons. The resulting dataset can be deployed in a single domain to train a deep neural network to perform on the single modality. The stereo system is not needed afterward, so the technical setup and calibration routine can be omitted in inference and is only required for data generation.

The task for the algorithm is to segment multiple body parts in the posterior legs of a running individual thermal images. We apply semantic segmentation to identify ROIs and also to distinguish between left and right leg. For each leg, we aim to distinguish between: shoe, lower leg/calf, knee, upper leg/thigh, clothing, and background.

### Dataset generation

We propose a method of label propagation from one image in visual domain to IRT domain. In visual domain, labels can be gathered more easily than in thermograms. With a stereo registration, both domains are coupled and information can be projected from visual into the thermal image plane. The visual systems needs an external source of depth in the same coordinate system (RGBD) for backprojection. The calibration process is necessary once, before capturing stereo pairs. For each image pair, we generate the labels in visual domain, transform them to thermal domain and smooth them in post-processing steps to obtain a final segmentation mask for thermograms.

#### Calibration

The first step is the geometric calibration of both camera systems to be ready for stereo applications. For this purpose, we generate point correspondences for the intrinsic and distortion as well as for the stereo calibration with a calibration pattern that is visible in all modalities. The RGBD and IRT cameras have different image sizes. Therefore, each intrinsic parameter is calibrated individually and resized to same size as the RGBD image. Consequently, the stereo registration $[\mathbf{R}|\mathbf{t}]_{\mathrm{IRT,RGBD}}$ can be estimated. As a calibration pattern, we applied a symmetric circle grid pattern made of black painted aluminum (emissivity 0.98) and a backplate made of white expanded polystyrene at room temperature. The metal plate is cooled to 8°C to increase the thermal radiation difference with the expanded polystyrene and make the circles

easier to see in the thermal camera. Black and white colors also allow for easy detection in visual space. An example pair of corresponding images with the calibration pattern is shown in Fig. 2. For better detection, the temperature scale is set to a temperature higher than the metal plate and lower than the backplate. The depth camera has already been fused to the vision camera by the manufacturer.

### Label generation for vision data

To transform images and labels from the visual spectrum to the thermal spectrum, we first need to generate labels in the visual spectrum. Our goal is to generate a dense segmentation mask for body parts, including the separation of whether a part is pure skin or wearing clothing. In medical applications, the temperature distribution is measured for skin parts. To obtain these labels, we have a naive approach by generating a mask for the whole body with a DeepLabv3+ trained network [19] (Fig. 3b: gray mask) and a mask for the skin by applying a FCNResnet101 [20] (Fig. 3b: white mask). We also need the skeletal pose of the human, which can be obtained from Yolo-Pose [21, 22]. The model recognizes up to 17 keypoints related to skeletal joints and connections between them. For dense labels, the Watershed algorithm [23] is applied, which fills similar areas with the same value. As initial markers, the skeleton vertices are drawn, at the end of each a larger ellipse is drawn for better boundary detection. Additionally, for the special class *Knee*, the joints are introduced as ellipses to separate the thigh from the calf (Fig. 3 c: skeleton overlay). Outside the body part, the segmentation is background. Afterward, the

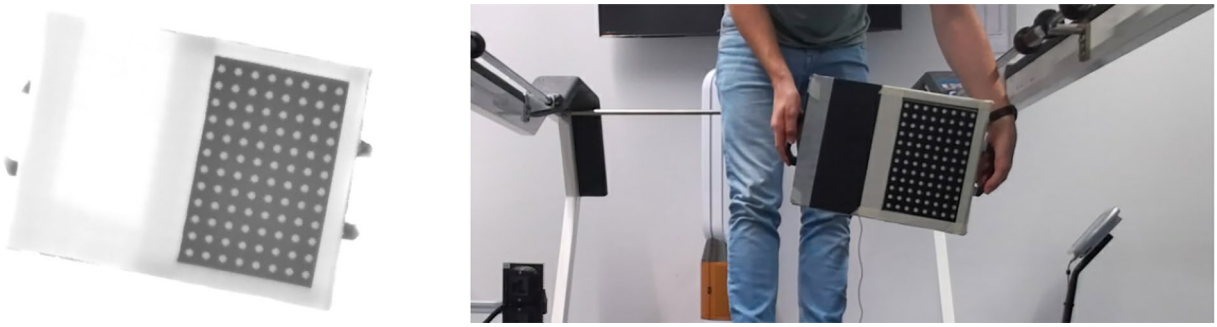


**Fig. 2** Cropped images in both domains with captured calibration board. Left: thermal (temperature scale 10–20°C), right: vision.

**Fig. 3** From left to right: Vision image; body mask: human skin (white), body mask including skin and clothing (gray); vision image together with detected human skeleton pose; label mask for the vision image based on the skeleton and the body mask.

Watershed algorithm is applied. This process is repeated once for the label values of the skin and once for the label values of the clothing. Both watershed masks are merged into one: for parts with skin, from the skin watershed mask, otherwise from the non-skin watershed. Small clothing labels are removed. Figure 3 shows the main steps: input image, skin and body masks, skeleton with respect to body and skin mask and the resulting dense label mask. As the field of view for the vision data is larger than for the thermal image, and we only require accurate segmentation from hips to feet, we do not optimize segmentation in body parts above the hips, as these regions will not be transformed in the next step.

### Image transformation

With the depth data from RGBD images and a calibrated stereo system, it is possible to transform information from one image plane to another as generally described in [24]. Rectified image planes of both cameras have been implemented in the transformation process as fundamental step for further developments. Thereby, the rectified images $\mathbf{p}_{\mathrm{rRGBD}}$ and $\mathbf{p}_{\mathrm{rIRT}}$ are computed with the rectification rotations and the calibration matrices: $\mathbf{R}_{\mathrm{rRGBD,RGBD}}$, $\mathbf{R}_{\mathrm{rIRT,IRT}}$, $\mathbf{K}_{\mathrm{rRGBD}}$ and $\mathbf{K}_{\mathrm{rIRT}}$. Lens distortion is removed as part of this image rectification. Next, we reconstruct the 3D points from the (homogeneous) 2D point, the camera calibration matrix $\mathbf{K}$ and the depth $d$ with (1).

$$\mathbf{P}_{\mathrm{rRGBD}} = \mathbf{K}^{-1}_{\mathrm{rRGBD}} \cdot d_{\mathrm{rRGBD}} \cdot \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}_{\mathrm{rRGBD}} \tag{1}$$

In 3D space we perform a coordinate system change from rRGBD over RGBD and IRT to rIRT by applying the relative rotation and translation received from the stereo calibration $[\mathbf{R}|\mathbf{t}]_{\mathrm{IRT,RGBD}}$ (2).

$$\begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix}_{\mathrm{rIRT}} = \begin{bmatrix} \mathbf{R} & 0 \\ 0 & 1 \end{bmatrix}_{\mathrm{rIRT,IRT}} \cdot \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix}_{\mathrm{IRT,RGBD}} \cdot \begin{bmatrix} \mathbf{R} & 0 \\ 0 & 1 \end{bmatrix}^{-1}_{\mathrm{rRGBD,RGBD}} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}_{\mathrm{rRGBD}} \tag{2}$$

The transformed points are projected to rIRT image plane by camera projection (3).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}_{\mathrm{rIRT}} = \frac{1}{z_{\mathrm{rIRT}}} \cdot \mathbf{K}_{\mathrm{rIRT}} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{\mathrm{rIRT}} \tag{3}$$

The RGBD points are located on the IRT image plane by inverse rectification of $\mathbf{p}_{\mathrm{rIRT}}$. For each point, the intensity value must be copied from the vision image and placed at the new position in the IRT image plane. With ToF technology, depth acquisition for round surfaces is inaccurate at certain degrees as reflections won't registered correctly by the sensor, which leads to an invalid measurement. Missing values in the depth map are interpolated to avoid invalid transformations [25].

The texture to be transformed can be any image with the same image plane as the vision image plane. Therefore, the labels generated from the vision domain are transformed. Figure 4 shows the transformation of the vision data into the IRT image plane. The superposition of the two images shows the high accuracy of the transformation.

The projection and transformation processes operate on real numbers, but RGBD and IRT images are stored with discrete integer positions. If a transformed point does not match the fixed positions, the closest value is selected.

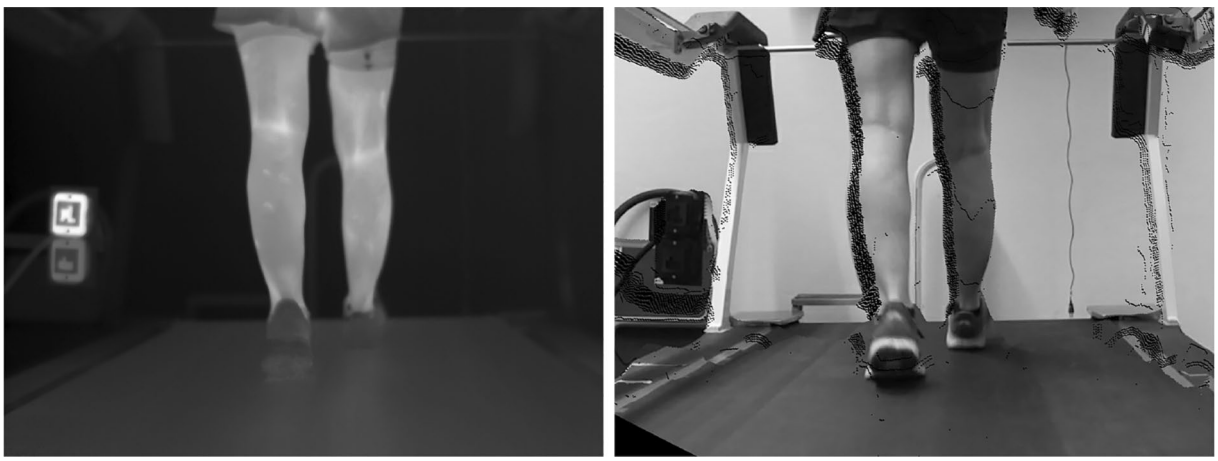


**Fig. 4** Left: thermogram in the IRT image plane; right: grayscale vision data transformed in the IRT image plane.

### Label smoothing in IRT

Due to the transformation process, not all pixels in transformed labels have a corresponding label, especially at object boundaries where pixels could not be transformed due to invalid depth measurements (see Fig. 4 black contours of legs). Additional distortions, perspective shifts due to stereo baseline offsets, and the discrete transformation space lead to an incomplete label image in the IRT domain. As a result, several steps are taken to fill in all information gaps in the segmentation map and form an initial image to apply a Watershed algorithm for a fully dense mask. The initial markers for the watershed algorithm consider 0 as an unknown pixel to be determined.

We process the transformed labels with image operations together with the original thermogram. Figure 5 shows the main steps described in the following paragraph. The first step is to enlarge the knee area of each side by drawing a horizontal line on the sides of each knee if the transformed label is not large enough. At this stage, the shoe labels are shrunk by erosion for the first time.

During the transformation process, some pixels were misplaced and are now without any connection to other pixels, to remove them, all connected components smaller than 40 pixels are removed for the body segmentation mask. Label pixels should be inside the body segmentation mask. If not, they will be reset to 0.

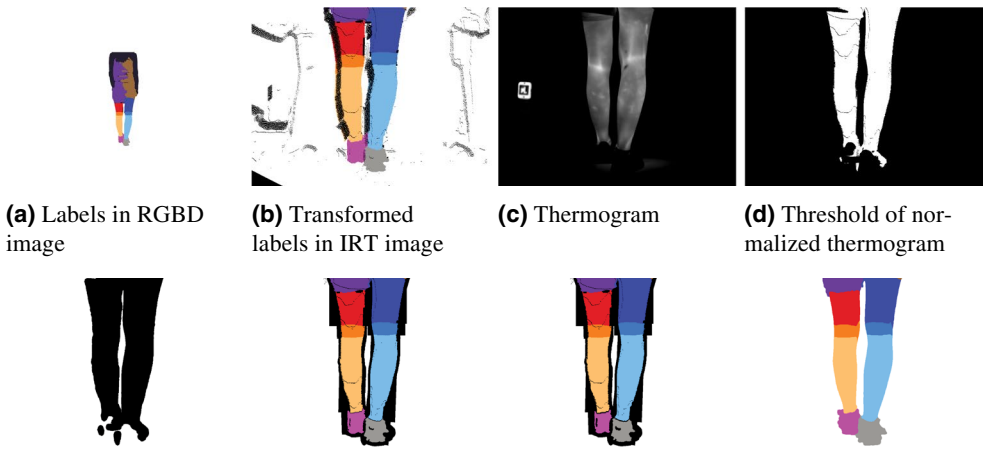
In the 8-bit images, the range of the temperature scale in our thermograms is set to 25 to 35°C. When thresholding this thermogram, it is not always possible to determine the correct contour because some areas are too similar to the background (see Fig. 6). Therefore, the original 16-bit thermogram is normalized to an 8-bit image by setting the minimum to 0 and the maximum to 255 and scaling in between. A threshold of this normalized 16-bit thermogram with Otsu's algorithm [26] determines the contour of the body. All existing labels outside the mask are removed, as this area is likely to be a background pixel. From the inverse thresholded image, a distance map is computed to measure the distance of each likely background pixel to a given foreground pixel. If the distance is greater than 20 pixels, we assume it is definitely background and set the pixel to that label, otherwise keep it as unknown.

Our model consists of individual connected components for each class. We assume that our process will produce at least one large component that will be kept and others that will be discarded.

The shoe and clothing labels tend to be underestimated in the distance map mask because they do not have high temperatures and are often barely visible in thermograms. For them, the distance map is calculated around their convex hull.

To ensure true neighborhood between classes on a body side, background is removed for each connecting class in the common rotated rectangle. In addition, there is no background above clothing because people are only captured from the waist down. The area above existing clothing markers is also cleared and a clothing marker line is drawn.

A large area is not well covered by the transformation and post-processing steps, the inner area between the legs. This area is not always connected to the outer background area, and the watershed will not fill it as a background if there is no initial label. To overcome this problem, we introduce a way to generate initial background labels there. For each of the classes shoe, calf, knee, and thigh, we find the center of the connected component on the left and right side. A virtual



**(a)** Labels in RGBD image
**(b)** Transformed labels in IRT image
**(c)** Thermogram
**(d)** Threshold of normalized thermogram
**(e)** Distance map mask
**(f)** Distance map mask and extra space applied
**(g)** Erosion of shoes
**(h)** Applied Watershed

**Fig. 5** Major post-processing steps from label transformation to final dense segmentation mask in IRT image plane.

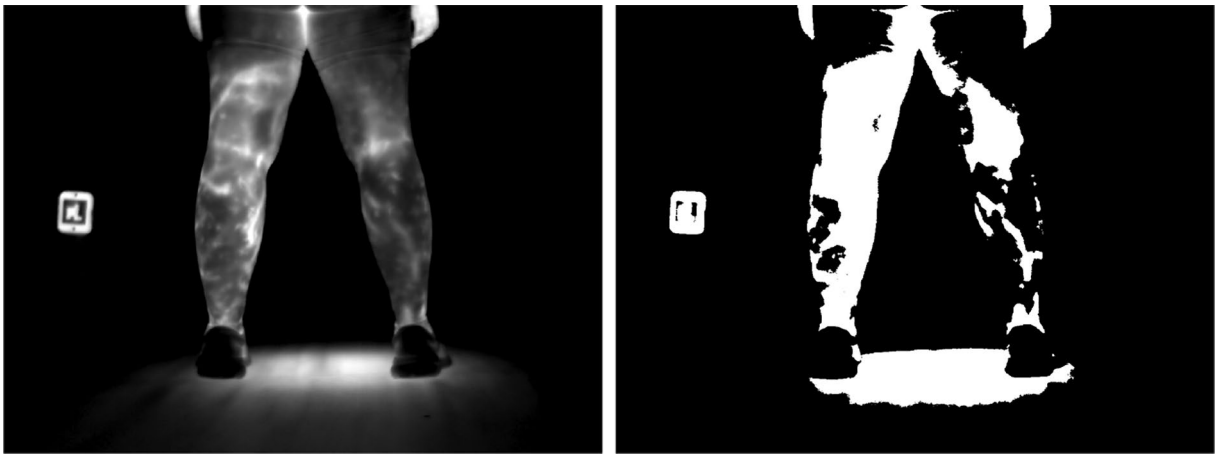


**Fig. 6** Thresholding a thermogram results in an unclosed mask when applied to an exhausted subject. Left: thermogram; right: threshold with Otsu's algorithm.

line is drawn between these two centers of the same entity. If the line crosses two edges of the thermogram (based on the Canny edge detector [27]), we assume that there is a background area between them. A line of thickness 1 is drawn from the intersections with the edges as background. Additionally, a circle of radius 2 is placed between the shoes to focus more on the background.

To avoid too large initial markers for the shoes, which would lead to too large watersheds, large markers (over 5000 pixels) are slightly eroded.

Finally, the Watershed algorithm with the prepared labels and the thermogram is applied. Morphological closing eliminates the boundaries estimated by the watershed algorithm. As a last step, the label contours are smoothed by a median filter.

### Benchmark

A benchmark for training and testing the proposed dataset ensures that others can compare their results with a baseline. Therefore, an actual semantic segmentation network (DeepLabv3+) is trained. Training procedure includes also randomly applied data augmentation: elastic grid and artificial optical deformations, small rotations and brightness changes and coarse dropouts. For training images, it will be randomly zoomed and resized to $640 \times 480$ pixels. As target objects could be possible not inside the region of interest, the cropping process will take care of including a certain amount of non-background class pixels.

As a reference metric, we choose the well-known segmentation metric *Intersection over Union* (IoU, Jaccard index). It is defined as the ratio between the intersection of the prediction and ground truth labels with the union of both labels. We report an average IoU and also the individual values to show the strengths and weaknesses of the model with respect to the different classes.

The batch size is set to 6. Dice [28] is selected as the loss function because it performs well for segmentation tasks and is related to the IoU metric. The network weights are updated by backpropagation and the advanced gradient descent optimizer AdaBelief [29] with a learning rate of 0.001. The performance of the network is monitored during training with a validation set. Images are normalized by subtracting the mean and dividing by the standard deviation before being fed into the network.

## Experimental

The proposed dataset is based on an experiment with running individuals recorded by a multimodal stereo system. Each participant had to perform several speed increments on a treadmill while standing, walking, and running: 0.5 min of standing, followed by 2 min of 4 and 6 km $h^{-1}$ walking, 8, 10, and 12 km $h^{-1}$ running and finally, a 3 min standing phase.

### Participants

The study includes data of 14 individuals (6 female, 8 male, age $31.79 \pm 9.3$). All participants signed an informed consent form and agreed to have their data included in the dataset and in this publication. In addition, a questionnaire on medical issues and personal abilities ensured the safety of the trial. Furthermore, RPE was assessed at each stage with the Borg scale (6–20) [30] to maintain a safe procedure. All participants, except one, completed all phases as prescribed. Participant 5 skipped the last running stage, so we added a 2 min walking stage at 4 km $h^{-1}$ at the end. The experiments

were performed in the Department of Sports Medicine, Disease Prevention, and Rehabilitation of the Institute of Sports Science in Mainz, Germany. All procedures were approved by the Human Ethics Committee Rhineland-Palatinate (IRB Number: 2021-15713 1) and are conform to the World Medical Association Code of Ethics (Declaration of Helsinki). Written informed consent was obtained from all participants.

### Experimental setup

We propose an experimental setup involving a multimodal stereo imaging system. First, we capture visible light (VIS) images (1920 × 1080 pixels, 30 fps) and the corresponding depth map from the integrated time-of-flight (ToF, 30 fps) camera with a Microsoft Azure Kinect. Both cameras are registered by the manufacturer, and the depth map is internally matched to the 2D coordinate system of the VIS camera and build a RGBD system. Additionally, a thermal camera (VarioCam HD head, JENOPTIK AG, Jena, GER: Microbolometer FPA detector with 1024 × 768 IRT-pixels, spectral range: 7.5–14 µm, 30 fps) was applied to measure thermal radiation. Both camera housings are rigidly mounted with the Kinect above the thermal camera. The system is mounted on a tripod behind an indoor treadmill (Saturn, HP cosmos, Nussdorf-Traunstein, GER) with a distance of 1.8 m. The assembled setup is shown in Fig. 7 from behind during an experiment of this study. To build the stereo system, we synchronized the image capture process. Therefore, the Azure Kinect acts as the hardware trigger source and the thermal camera receives the signal to start its own capturing. Although all camera components reach 30 fps in free run mode, we needed to reduce the frame rate to 15 fps in synchronized mode to avoid running out of synchronization. In addition, the thermal camera periodically recalibrates its sensor with a Non-Uniformity Correction (NUC), so we skipped 0.5 s of processing afterward.

## Results

The results of the applied methodology are a final dataset and the training results of the DNN.

### Dataset content

We processed 14 assigned trials and selected 10% of the images for each person to form the dataset. The selection of images was done randomly, including all phases of standing, walking, and running. We also split the set into a training set (11 persons, 12,826 thermograms) and a test set (3 persons, 3433 thermograms). In our benchmark, we additionally split the training set into training (9 persons) and validation (2 persons). The mean and the standard deviation for image normalization based on training and validation set is mean = 0.084 and SD = 0.165.

**Fig. 7** Experimental setup with stereo camera setup in the foreground and a runner on a treadmill (left) and the operator of the treadmill (right) in the background. The upper camera is the Microsoft Azure Kinect, and the lower camera is the JENOPTIK VarioCam HD head.

The dataset contains the IRT images with the generated labels and the corresponding RGBD images (visual domain and depth values). We also provide the camera calibrations and the stereo registration.

### Benchmark

The benchmark was trained for 34 epochs before early stopping, with the best validation IoU achieved after the 29th epoch. It achieves an overall mean IoU of 0.66 on the test set. Table 1 shows the IoU per class and the overall results. The results are similar from left to right. But there are big differences depending on the body part. The best single class IoU, besides the background, are the lower leg classes. The lower legs are the easiest class to recognize because they are most visible in the images, being in the center of the image. However, in later stages, they may be obscured by the opposite side or by a raised shoe. For the thighs, there is also an interaction with moving clothing. These are not the same size for everyone and can be loose or tight around the body. In the tight case, thermal

**Table 1** The proposed deep neural network's overall IoU and per class IoU for the test set

| Class | IoU left side | IoU right side |
|---|---|---|
| Mean | 0.66 (no side) | |
| Background | 0.98 (no side) | |
| Clothes | 0.51 | 0.53 |
| Thigh | 0.60 | 0.54 |
| Knee | 0.59 | 0.61 |
| Calf | 0.83 | 0.83 |
| Shoe | 0.64 | 0.61 |

radiation can pass through the clothing. The knee is not detected as well as the legs. In special cases, such as the raised shoe, it disappears from view. Shoes and clothing are more difficult to detect, as described in label transformation, they are sometimes indistinguishable from the background.

The test set includes a person without shorts in the images, which also needs to be treated, but with the current training routine, the network is not able to segment all parts correctly, as can be seen in the first row of Fig. 8. In the second row, the participant's knee will be detected more precisely. The generated label contains incorrect labels for overlapping leg parts. The third person's result also improves over the ground truth data, as the background class between the legs is correctly identified.

## Discussion

The data presented show that labels in thermal images can be automatically generated with a stereo system of thermal, visual, and depth domains by propagating labels from existing segmentation algorithms to thermal images.

Three different networks are implemented to generate dense target labels in the vision image: a body segmentation,

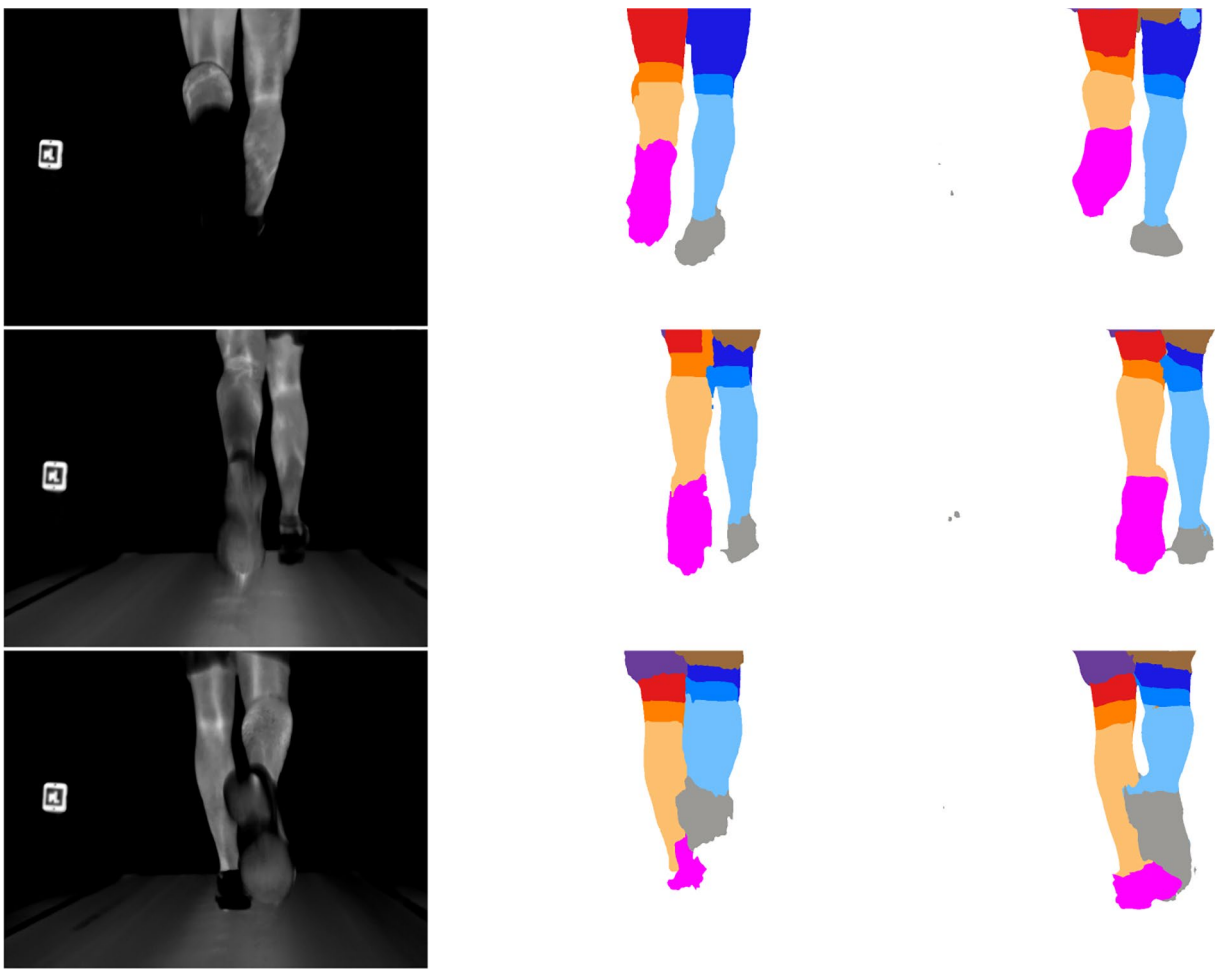


**Fig. 8** Network results compared to ground truth for a sample of each person in the test set. From left to right: thermogram, ground truth, and prediction.

a skin segmentation, and a skeleton detection. Instead of generating segmentation masks by three networks, a direct approach fits better and reduces inconsistencies like alignment issues of body and skin mask shown in Fig. 3 (second). The detected edges of the legs are different in the body mask (gray) and the skin mask (white). But our label smoothing approach overcomes this alignment problem with Watershed to find the real edges.

For stereo transformation, the depth image will be interpolated to ensure that there is always a valid depth value. But these values can be incorrect and lead to erroneous 3D point reconstructions. Also, the transformed image is not fully dense in the IRT image plane due to the different views of the two cameras. Therefore, post-processing must be performed to obtain a fully dense segmentation mask. The steps involve removing artifacts, filling the affected pixels with a neighboring label, and adjusting them for contour-based edges.

In the presented dataset, we focus on the movement of the posterior calves. However, the nature of running is a cyclic movement of the legs. This will lead to a huge redundancy in the images with same leg positions and angles. Therefore, we do not take all of the thermograms from the entire set of measurements, but rather select a small number to remove redundancy while still preserving the variations in velocity and thermal patterns of the vascular system. The post-processing methods were optimized to deal with the motion and thermal patterns. This approach can be extended to generate labels for additional ROIs.

One of the challenges of creating a thermal dataset is dealing with objects that have low contrast to the background. The dataset has a temperature scale of 25–35°C in 8-bit images. Therefore, pixels with temperatures below 25°C get a value of 0 and are indistinguishable from surrounding objects if they are also below 25°C or if background temperature radiation has similar values as the object. There are two areas where these effects are most common: first, the foot sole or other parts that are not visible or difficult to distinguish from background in the thermal image, hamper segmentation. Second, the posterior leg surface radiation at later stages (e.g., Figure 6). As reported in [5], the surface radiation of runners' legs decreases over time with increasing load. This results in leg surface temperatures close to 25°C. Both phenomena complicate the post-processing after the label transformation and training with inaccurate labels will lead to incorrect predictions. One should further optimize the labels by executing an advanced post-processing routine with varying temperature scales of the thermal image. A high computational cost of label generation steps and DNN training is affordable because it is computed only once and not with each inference. Another approach is to apply normalized full 12-bit dynamics, as these full-range thermograms have better contrast characteristics. The disadvantage of normalization is the different thermal information distribution in each subsequent image, which reduces the accuracy of the evaluated thermal statistics. In addition, with fixed temperature scales, we obtain comparable image properties for DNN training.

The benchmark shows the principle of training with the newly generated dataset. The IoU for each class is completely different, which can be explained by the classes themselves. The worst classes are shoes and clothes. We have already explained the problem of generating good labels for these classes, which is why they are difficult to train. As the quality of the label increases, so do the test results, as in the other parts. Calves are superior. They are often visible, have no interactions with the clothing, and the connection with the shoes has a clear border. As mentioned in [5], it is not necessary to match the shape perfectly, as small variations do not have a big effect on statistical properties like mean intensity. Although the connector class *knee* is positioned by the skeleton recognition, it is not separated by intensities from the other leg parts and therefore has a worse recognition. The thigh recognition is related to the clothing detection, as these are faulty and lead to either background or wrong thigh prediction. Improvements to the DNN training can be made by manually refining the generated labels, including more training samples, or employing an active learning approach that refines the labels with the predictions from a previous training run.

With the benchmark, it will be possible to compare future work with this dataset. Hence in this work it is not possible to compare the benchmark result with previous work. The dataset was completely generated automatically. Our previous work in [5] rely on manual annotated data. As the data was captured with different environment settings it is not comparable directly.

The presented work is optimized for runners running on a treadmill while their calves are detected (sagittal view). The pre- and post-processing steps of the label generation are strongly adapted to the calf detection task. Our label generation method does not include the depth data from the RGBD camera. In new scenarios with different ROIs, it may not be possible to clearly distinguish between sides based on the visual image. Therefore, depth data can also be included to improve label generation. This also allows for non-stationary tasks, such as people walking on a track instead of a treadmill. The recording mode must be selected according to the new task, taking into account that the focus of both cameras, especially the IRT camera, is fixed and cannot be changed during the experiment, as this would destroy the extrinsic stereo calibration. The focus range of the IRT camera is small and does not allow large shifts. Therefore, the depth shift of the images should be limited by the depth information of the RGBD camera.

## Conclusions

In this paper, we present a method for automatic label generation for thermograms of moving humans. The novel application combines several existing technologies for images in the visual domain: stereo camera systems, human skeletal pose estimation, skin detection, and combines them to generate dense segmentation masks in a corresponding thermal image. The StereoThermoLegs dataset improves the automatic image analysis of IRT applications of humans during exercise. The developed method including a stereo rig with a common modality (vision) and our target modality (thermal radiation) helps to automatically generate a dataset that can support deep learning. Researchers who employ IRT as an assessment tool to understand physiological processes have mostly worked with static and manually labeled scenes and are now able to analyze activity in thermograms. In addition, our approach allows for rapid adaptation to new environments, activities, and ROIs by creating specialized datasets with less effort than manual labeling. The task is not limited to semantic segmentation: object detection, instance segmentation, classification, and other tasks would also benefit from the presented approach. In the future, our method can be extended to transform video tracking information from the visual to the thermal domain to generate thermal video tracking datasets.

**Acknowledgements** We would like to thank all participants for their voluntary participation in this study.

**Author contribution** Daniel Andrés López contributed in study design, implementation of algorithms, data labeling and selection, and stereo camera setup and calibration. He also wrote the first manuscript draft. Barlo Hillen contributed in study design, data interpretation and manuscript editing. Markus Nägele contributed in stereo camera instrumentation, calibration and reviewing the manuscript. Perikles Simon helped in study design, interpretation and manuscript reviewing. Elmar Schömer contributed in algorithm design, interpretation and manuscript reviewing.

**Funding** Open Access funding enabled and organized by Projekt DEAL. The authors gratefully announce that this research study was funded by the Johannes Gutenberg University research program "Inneruniversitäre Forschungs förderung (Stufe I)" with title: "Extend DeepSpoMed: Computer-aided domain adoption for human skin body segmentation in thermal images".

**Code and data availability** The proposed dataset is publicly available [31]. Essential code may be made available upon reasonable request and with permission of the institutional boards.

## Declarations

**Conflict of interest** The authors declare that they have no Conflict of interest.

**Ethical approval** All authors read and approved the final manuscript.



## References

1. Das K, Bhowmik MK, Prasad Mukherjee D. Segmentation of Knee Thermograms for Detecting Inflammation. In: 2019 IEEE international conference on image processing (ICIP). Taipei: IEEE; 2019. pp. 1550–1554. https://doi.org/10.1109/ICIP.2019.8803094.
2. Szurko A, Kasprzyk-Kucewicz T, Cholewka A, Kazior M, Sieroń K, Stanek A, Morawiec T. Thermovision as a tool for athletes to verify the symmetry of work of individual muscle segments. Int J Environ Res Public Health. 2022. https://doi.org/10.3390/ijerph19148490.
3. Tayel MB, Elbagoury AM. An efficient and reliable method for regional analysis of breast thermographic images. Glob Sci J. 2020;8(9):1508–18.
4. Hillen B, Andrés López D, Pfirrmann D, Neuberger EW, Mertinat K, Nägele M, Schömer E, Simon P. An exploratory, intra- and interindividual comparison of the deep neural network automatically measured calf surface radiation temperature during cardiopulmonary running and cycling exercise testing: A preliminary study. J Therm Biol. 2023;113: 103498. https://doi.org/10.1016/j.jtherbio.2023.103498.
5. Hillen B, Andrés López D, Schömer E, Nägele M, Simon P. Towards exercise radiomics: deep neural network-based automatic analysis of thermal images captured during exercise. IEEE J Biomed Health Inform. 2022;26(9):4530–40. https://doi.org/10.1109/JBHI.2022.3186530.
6. Magalhaes C, Mendes J, Vardasca R. Meta-Analysis and systematic review of the application of machine learning classifiers in biomedical applications of infrared thermography. Appl Sci. 2021;11(2):842. https://doi.org/10.3390/app11020842.
7. Perpetuini D, Formenti D, Cardone D, Filippini C, Merla A. Regions of interest selection and thermal imaging data analysis in sports and exercise science: a narrative review. Physiol Meas. 2021;42(8):08–01. https://doi.org/10.1088/1361-6579/ac0fbd.
8. Cañada-Soriano M, Bovaira M, García-Vitoria C, Salvador-Palmer R, Ortiz Cibrián, de Anda R, Moratal D, Priego-Quesada JI. Application of machine learning algorithms in thermal images for an automatic classification of lumbar sympathetic blocks. J Therm Biol. 2023;113: 103523. https://doi.org/10.1016/j.jtherbio.2023.103523.
9. Perpetuini D, Formenti D, Cardone D, Trecroci A, Rossi A, Di Credico A, Merati G, Alberti G, Di Baldassarre A, Merla A. Can data-driven supervised machine learning approaches applied to infrared thermal imaging data estimate muscular activity and fatigue? Sensors. 2023;23(2):832. https://doi.org/10.3390/s23020832.
10. Lou A, Guan S, Kamona N, Loew M. Segmentation of infrared breast images using multiresunet neural networks. In: 2019 IEEE applied imagery pattern recognition workshop (AIPR).

Washington: IEEE; 2019. pp. 1–6. https://doi.org/10.1109/AIPR47015.2019.9316541.
11. He Y, Deng B, Wang H, Cheng L, Zhou K, Cai S, Ciampa F. Infrared machine vision and infrared thermography with deep learning: a review. Infrared Phys Technol. 2021. https://doi.org/10.1016/j.infrared.2021.103754.
12. Hillen B, Pfirrmann D, Nägele M, Simon P. Infrared thermography in exercise physiology: the dawning of exercise radiomics. Sports Med. 2020;50(2):263–82. https://doi.org/10.1007/s40279-019-01210-w.
13. Kopaczka M, Kolk R, Merhof D. A fully annotated thermal face database and its application for thermal facial expression recognition. In: 2018 IEEE international instrumentation and measurement technology conference (I2MTC). Houston: IEEE; 2018. pp. 1–6. https://doi.org/10.1109/I2MTC.2018.8409768.
14. Kniaz VV, Knyaz VA, Hladůvka J, Kropatsch WG, Mizginov V. Thermalgan: multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Leal-Taixé L, Roth S (eds) Computer vision—ECCV 2018 workshops. Cham: Springer; 2018. pp. 606–624. https://doi.org/10.1007/978-3-030-11024-6_46.
15. Rangel J, Soldan S, Kroll A. 3D thermal imaging: fusion of thermography and depth cameras. In: Proceedings of the 2014 international conference on quantitative InfraRed thermography. Bordeaux: QIRT Council; 2014. https://doi.org/10.21611/qirt.2014.035.
16. Richter J, Wiede C, Kaden S, Weigert M, Hirtz G. Skin temperature measurement based on human skeleton extraction and infra-red thermography—an application of sensor fusion methods in the field of physical training. In: Proceedings of the 12th international joint conference on computer vision, imaging and computer graphics theory and applications, vol 6. Porto: SCITEPRESS—Science and Technology Publications; 2017. pp. 59–66. https://doi.org/10.5220/0006095100590066.
17. Knyaz VA, Moshkantsev PV. Joint geometric calibration of color and thermal cameras fopr synchronized multimodal dataset creating. Int. Arch. Photogr., Remote Sens. Spat. Inf. Sci. XLII-2/W18, 2019. https://doi.org/10.5194/isprs-archives-XLII-2-W18-79-2019.
18. Bultmann S, Quenzel J, Behnke S. Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation. Robot Auton Syst. 2023;159: 104286. https://doi.org/10.1016/j.robot.2022.104286.
19. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision—ECCV 2018, vol 11211 LNCS. Cham: Springer; 2018. pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
20. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR). Boston: IEEE; 2015. pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
21. Maji D, Nagori S, Mathew M, Poddar D. YOLO-Pose: enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In: 2022 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). New Orleans: IEEE; 2022. pp. 2636–2645. https://doi.org/10.1109/CVPRW56347.2022.00297.
22. Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), vol abs/2207.0. Vancouver: IEEE; 2023. pp. 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721.
23. Meyer F. Color image segmentation. In: 1992 international conference on image processing and its applications. Maastricht: IET; 1992. pp. 303–306.
24. Hartley R, Zisserman A. Multiple view geometry in computer vision. 2nd ed. New York: Cambridge University Press; 2004. https://doi.org/10.1017/CBO9780511811685.
25. Telea A. An image inpainting technique based on the fast marching method. J Graph Tools. 2004;9(1):23–34. https://doi.org/10.1080/10867651.2004.10487596.
26. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9(1):62–6. https://doi.org/10.1109/TSMC.1979.4310076.
27. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell PAMI. 1986;8(6):679–98. https://doi.org/10.1109/TPAMI.1986.4767851.
28. Bertels J, Eelbode T, Berman M, Vandermeulen D, Maes F, Bisschops R, Blaschko MB. Optimizing the dice score and Jaccard index for medical image segmentation: theory and practice. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, editors. Medical image computing and computer assisted intervention–MICCAI 2019, vol. 11765 LNCS. Cham: Springer; 2019. p. 92–100. https://doi.org/10.1007/978-3-030-32245-8_11.
29. Zhuang J, Tang T, Ding Y, Tatikonda S, Dvornek N, Papademetris X, Duncan JS. AdaBelief optimizer: adapting stepsizes by the belief in observed gradients. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems, vol. 33. Virtual: Curran Associates Inc; 2020. p. 18795–806.
30. Borg GAV. Perceived exertion: a note on "history'' and methods. Med Sci Sports. 1973;5(2):90–3. https://doi.org/10.1249/00005768-197300520-00017.
31. Andrés López D, Hillen B, Nägele M, Simon P, Schömer E. StereoThermoLegs dataset. Zenodo. 2024. https://doi.org/10.5281/zenodo.8289870.