A Foundation for 3D Human Behavior Detection in Privacy-Sensitive Domains

Thomas Heitzinger
https://cvl.tuwien.ac.at/staff/thomas-heitzinger/
Martin Kampel
https://cvl.tuwien.ac.at/staff/martin-kampel/

Computer Vision Lab
TU Wien
Vienna, Austria

Abstract

Human behavioral analysis applications in the fields of ambient assisted living (AAL) and human security monitoring require continuous video analysis of individuals. Although intelligent systems deployed in these areas are intended to have a positive impact on the persons involved, subsequent continuous monitoring naturally raises ethical concerns and questions about privacy implications. To address these issues, we present a foundation for identity-preserving 3D human behavior analysis. Our main contributions are a fast 3D detection system and a public multi-modal dataset. The introduced detection system uses an innovative target assignment scheme to significantly improve performance, especially in challenging scenes with a large number of person-person and person-object occlusions. On our dataset, the system shows performance on par with the state-of-the-art in 3D object detection while being lightweight enough for configuration and deployment to edge devices. The dataset is large, at a total of 85k annotated frames. To reduce privacy intrusion, it consists entirely of spatio-temporally aligned depth and thermal sequences. Annotation is provided as 3D bounding boxes, along with pose labels and consistent person IDs for use in tracking. The dataset is designed to be flexible. Data representation in either image view or point clouds and the option for projected 2D bounding boxes, allows use in a variety of 2D or 3D tasks. Target applications of our work are privacy-sensitive domains that currently require continuous monitoring using RGB-based systems, including ambient assisted living tasks (e.g., motion rehabilitation, fall detection, vital sign detection) and human security monitoring applications, such as construction safety, critical care and correctional facility monitoring.

1 Introduction

Developments in machine vision have led to the emergence of intelligent systems in areas key to human health and well-being. Notable examples include applications in ambient assisted living (AAL), such as motion rehabilitation [11, 62], fall detection [1, 14, 56], contact-less heart rate and respiratory rate detection [9, 63]. Such systems, especially if designed for medical emergency detection, require continuous video surveillance. Similar circumstances are found in domains monitoring human safety, including construction safety [58, 64], intensive care [26, 32] and supervision in correctional facilities [8].

The overarching commonality among the addressed domains is the strict requirement for continuous video supervision to ensure participant safety. While intelligent systems in such areas are designed to have a positive impact on involved persons, continuous surveillance, especially in the age of IoT devices, naturally leads to ethical concerns and questions about loss of privacy [2, 15, 41]. This is

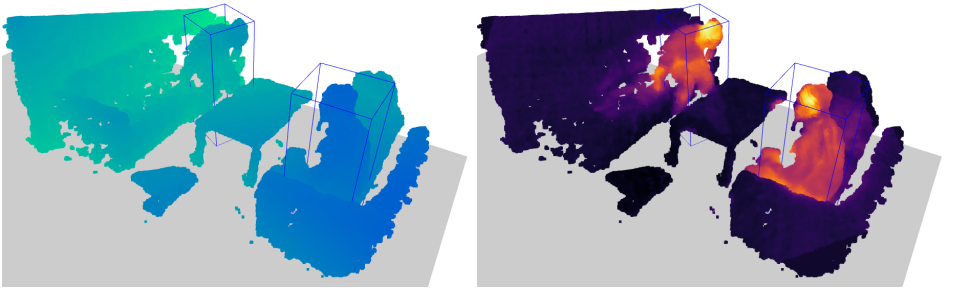


Figure 1: Sample image with annotations viewed as a point cloud. False color based on depth (left) and based on temperature (right).

of particular concern if participants have limited control over their environment, either due to mental and/or physical impairments or because they are deprived of their liberty. We argue that the identity of individuals is in many cases irrelevant. The central questions is not *who*, but rather *what* – it is the action that matters, not the person. The main aim of our paper is thus to address a dichotomy between the *utility* of machine vision-based human behavior analysis, and *impact on privacy*. It is our goal to lay the foundations for a robust human behavior detection (HBA) system that can serve as a replacement for current RGB-based surveillance, and at the same time, protect the privacy of individuals - thus enabling future research on assistance systems that support vulnerable persons in a more dignified manner. It is an explicit *non-objective* of our work to enable the expansion of surveillance into previously unmonitored areas, such as toilets or bathrooms, for which specialized solutions exist [42, 45].

The ability to reliably detect and track individuals is the underlying requirement for human behavior analysis [10, 39]. It is key to all the addressed applications and a necessity for the development of semantically more abstract action and behavior recognition tasks. Motivated by this, we introduce a dataset captured solely with depth and thermal sensors (Figure 1). As opposed to other postprocessing-based anonymization techniques [6], these imaging modalities preserve sensitive information by design - the individuals identity is already protected during data acquisition. Our presented detection system is highly configurable, and even allows for deployment to resource constrained edge devices. This can serve as an additional safety measure for the protection of personal data since a sensor-near inference setup keeps potentially sensitive imagery local to the sensing unit.

The main contributions of our paper are:

- A fast and highly configurable 3D detection model that performs on par with state-of-the-art models while being suitable for inference on edge devices.
- An innovative broadcasted target assignment scheme to replace traditional methods at negligible additional processing cost, and an optional birds eye view module for boosted performance.
- The *MIPT* dataset[1]: A large multimodal dataset for privacy-preserved 3D human detection and tracking consisting of spatio-temporally aligned depth and thermal sequences.

In Section 2 we give a summary over related work on multimodal datasets and detection in the context of human behavior analysis. Section 3 presents our multimodal dataset and its key features. Our proposed detection system and innovations are introduced in Section 4, while Section 5 presents conducted experiments and their results. Finally, Section 6 concludes the paper.

[1] URL: https://cvl.tuwien.ac.at/research/cvl-databases/mipt-dataset/, DOI: 10.5281/zenodo.5592323

2 Related Work

We address related work both in terms of publicly available multimodal data, as well as object detection in human behavior analysis. Additionally, we categorize our system among general 3D object detection methods.

Multimodal data in the context of HBA Currently available public datasets most similar to ours are *IPHD* [13] and *IPT* [22]. Like our dataset, IPHD is large at 100k frames and consists of depth and thermal data. Key differences are annotations that are only provided in 2D image space and its structure as an image dataset instead of sequences - making it unsuitable for tracking. The IPT dataset features 10 sequences of depth-only data and 3D location annotations (not 3D bounding boxes). Especially missing labeling of person orientations make this dataset unsuitable for behavior and interaction detection. Other datasets only providing 2D annotation are the *AAU VAP Trimodal People Segmentation Dataset* [44] and *RGB-D People Dataset* [37, 51] which – in addition – are small in size with 3 and 1 sequences respectively. Synthetic datasets use 3D models to generate imagery and perfectly accurate annotations at the cost of an introduced synthetic data bias, which may harm the generalization ability of models on real world data. In this category Shotton *et al*. [59] present a dataset for human pose recognition in depth data. The multimodal *SDT* dataset [48] features artificial depth and thermal frames intended for 2D object detection. For interaction detection, the RGB-D *SBU Kinect Interaction Dataset* [67] provides sequences of depth data and mapped motion sequences which include both neutral ("Exchanging", "Shaking Hands") and violent actions ("Punching", "Kicking"). Similarly, the *Cornell Activity Dataset* [52] presents a collections of human actions in RGB-D. Both are recorded in idealized environments (no other objects in the background, no occlusions) that do not reflect activity found in real world applications. Privacy concerns in the collection of large datasets for human fall detection are addressed in the work of Asif *et al*. [3] which presents a large-scale synthetic database using human motion captures of fall and related events.

General 3D Object Detection Our detection framework is structured as a single stage detector similar to *YOLO* [53], but extended to the three-dimensional case. As such, it does not rely on a region proposal network (RPN) [20]. The previous YOLO extension *Complex-YOLO* [60] uses predefined heights based on each class as additional information source for the transition from 2D to 3D. Our method processes data entirely in image view and is thus distinct to all methods operating directly on pointclouds, such as *PointNet*-based methods [49, 50, 54]. If used in configuration with the presented birds-eye-view module, our system is a fusion method with components similar in purpose to *MV3D* [12] and *AVOD* [29].

Detection and Privacy in HBA A publication by Ijjina and Chalavadi [25] presents a deep learning approach for the detection of motion sequences in RGB-D data, but does not address localization. Using the same modalities, Liu *et al*. [35] present a method that focuses solely on recognition of person orientations and is robust to occlusions and complex environments. In a work on multimodal behavioral analysis, Clapés *et al*. [40] explore the feasibility of person re-identification with a combination of RGB, depth and thermal imaging. Work of Callemein *et al*. [9] addresses privacy concern by downscaling RGB input images to low resolution. This is however associated with lower detection accuracy. Pittaluga *et al*. [47] propose an adversarial training approach to prevent extraction of sensitive information at the encoder stage. Other work focuses on anonymization methods at the sensor level. Using a line sensor to measure a one-dimensional brightness distribution, Nakashima *et al*. [42] demonstrate person localization in simple environments. Existing sensors can be modified with privacy-preserving optics [45] that preserve privacy through a defocusing blur.

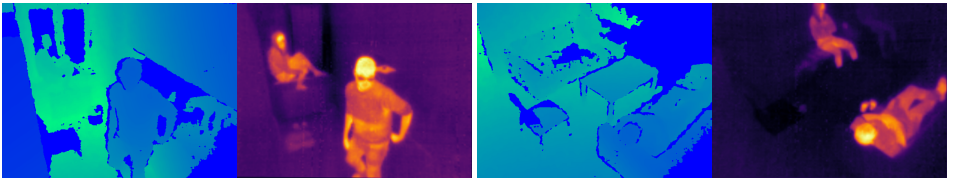


Figure 2: Spatio-temporally aligned depth (left) and thermal (right) frames in our dataset.

Resorting to a low-dimensional input stream or blurring/blocking image content has the disadvantage of limiting the performance and accuracy of derived assistance systems. The aforementioned methods target applications for which a high-resolution image sensor (RGB, thermal, depth or other) is not acceptable, such as assistance systems for bathrooms or restrooms. As such, they are therefore orthogonal to our work, since our main goal is to replace existing RGB-based monitoring systems.

The degree of anonymity provided by depth and thermal sensors is an actively discussed topic in existing literature [23, 28, 30], as face recognition on close-up portrait images has been demonstrated for both modalities. Furthermore, gait analysis [43] allows for some degree of identification independent of the imaging modality. Regardless, the goal of this paper is not to make identification impossible from an algorithmic perspective, but rather to provide a basis for a less intrusive experience for subjects compared to existing RGB-based systems. The modalities provided by our dataset can be used in a variety of configurations. The feasibility of systems based solely on depth modality is demonstrated, for example, by privacy-preserving fall detection solutions [27, 55]. Specifically for thermal sensors, it has been proposed to obfuscate human imagery by blocking sensor measurements in human temperature ranges before the digitization process [46]. If desired, such a step can be implemented as a preprocessing step for optional reduced use of thermal data. In this way, the dataset can be tailored to the needs of the application, to balance participant comfort with system performance.

3 Dataset

Overview We present a public , multimodal dataset for identity preserved detection and tracking (*MIPT*). The dataset features human activity in indoor sequences (Figure 2) and is suitable for privacy-sensitive 2D and 3D tasks. Its main distinguishing features are:

- **Size:** A total of *85k* frames showing indoor scenes spread over 20 densely labeled sequences – 9 of which are multimodal, with spatio-temporally aligned depth and thermal frames. The remaining sequences are depth only.
- **3D Labeling:** For all frames we provide labeling of person instances as *3D bounding boxes* (also projected 2D bounding boxes), basic pose labels for the classes "Standing", "Sitting" and "Lying", and a consistent *person ID* for use in tracking. Additionally we provide tracking ground truth files in the MOT20 Challenge [16] format.

Activities in the MIPT dataset are mostly scripted to guarantee a high degree of movement, poses and behavioral patterns in a short time frame. The sensor is static, mimicking a surveying setup as would be found in AAL or security applications. Nine sequences are recorded with a depth/thermal sensor setup similar to the one presented in [48], using a modified Orbbec Astra[2] and a FLIR Lepton 3.5[3] thermal camera module. These sequences are recorded at a resolution of 640×480 (thermal

[2] https://orbbec3d.com/product-astra-pro/ (last accessed on June 16th 2021)
[3] https://www.flir.com/products/lepton/ (last accessed on June 16th 2021)

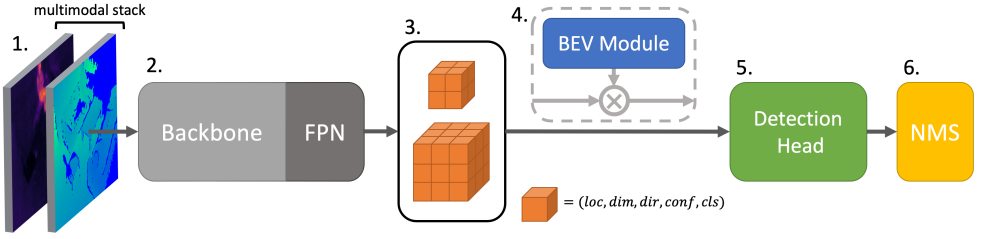


Figure 3: Basic structure of the detection system: 1. Input frames, 2. Backbone and FPN for feature extraction, 3. Reshape at each scale to a 3D grid of *cells*, 4. (Optional) augmentation of each cell with information captured in the birds-eye-view module (BEV), 5. Transformation into 3D bounding boxes, 6. Filtering and fusion of bounding boxes.

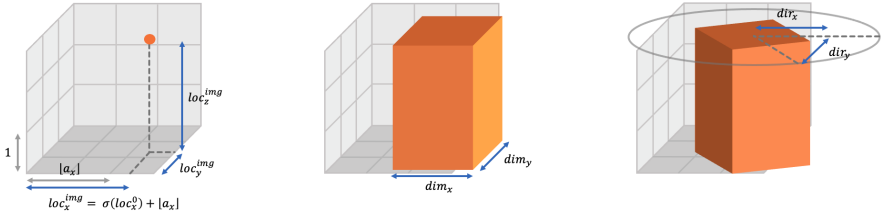


Figure 4: Transformation of raw neural network output in a *cell* to a bounding box.

frames upscaled from 160×120) at a frame rate of 8.7 (the upper limit for the thermal module). The remaining eleven sequences are depth-only – ten of which reuse raw data from the IPT dataset [22] but are entirely relabeled. Annotations in the original IPT dataset only feature positions of people in a birds-eye-view coordinate frame. Labeling of box dimensions, direction and a refinement of locations are our contribution.

Bounding Box Format The coordinate system is chosen such that the positive x-axis is rightward, positive y-axis is forward and positive z-axis is upward. Our dataset consists of indoor scenes, as such the ground plane is flat and all bounding boxes are assumed to be at ground level. Bounding boxes are expressed by: Their 3D location at head level $loc = (loc_x, loc_y, loc_z)$ (i.e. loc_z is the bounding boxes height), their width and length dimensions $dim = (dim_x, dim_y)$ and their rotation around the z-axis (rotation is 0 pointing towards the positive x-axis and approaches $\pi/2$ turning towards the positive y-axis). In practice we express rotation as a directional vector $dir = (dir_x, dir_y)$ with the constraint $\|dir\|_2 = 1$. In case classification is required on top of detection, a bounding box specification further includes a confidence score $conf$ and an n_{cls}-dimensional probability vector cls. Thus a 3D bounding box is fully expressed by the tuple $box = (loc, dim, dir, conf, cls)$ and can be written as a vector in $\mathbb{R}^{8+n_{cls}}$. For ground truth bounding boxes we define $conf := 1$ and cls to be the one-hot encoded ground truth class.

4 Proposed Method

Network Structure and Detection Head The input of our detection system is a two-channel stack of spatio-temporally aligned depth and thermal frames (optionally a single channel if only the depth modality is used). It uses a flexible backbone (ResNet [21] or Mobilenet v2 [37]) and is configured as a feature pyramid network (FPN) [34], for feature extraction at multiple scale levels (see step 2 in Figure 3). Although we intend to solve a 3D detection task, processing is entirely performed in image space (with an added depth dimension) using 2D convolutions.

For each scale level of the FPN the feature tensor of size (n_x, n_y, n_{ch}) is reshaped to shape $(n_x, n_y, n_z, 8+n_{cls})$ before it is passed to the detection head (step 3 in Figure 3), which requires the number of FPN output channels n_{ch} to be set to $n_z(8+n_{cls})$. Here n_z is a hyperparameter that sets the number of *depth layers* we predict at.

A bounded cuboid shaped subregion of the 3D image space is rescaled to size (n_x, n_y, n_z) and subdivided into $1 \times 1 \times 1$ sized subregions, referred to as *cells*. This allows each cell to be assigned to a feature vector of length $8+n_{cls}$, representing a possible bounding box prediction. The main objective of the detection head is the transformation of a raw prediction $box^0 \in \mathbb{R}^{8+n_{cls}}$ to a bounding box in 3D world space (see Figure 4):

$$\begin{aligned} loc_i^{img} &= \text{sigmoid}(loc_i^0) + \lfloor a_i \rfloor, & i &\in \{x,y,z\}, & conf &= \text{sigmoid}(conf^0), \\ dim_i &= m_i \exp(dim_i^0), & i &\in \{x,y\}, & cls &= \text{softmax}(cls^0), \\ dir_i &= dir_i^0 \frac{\text{sigmoid}(\|dir^0\|_2)}{\|dir^0\|_2}, & i &\in \{x,y\}. \end{aligned} \tag{1}$$

Here, $a \in \mathbb{R}^3$ is an *anchor* and serves as an offset to the center of the cell that box^0 is assigned to ($\lfloor a \rfloor$ rounds down to nearest integer coordinates). Furthermore, $m = (m_x, m_y)$ is an *anchor box*, obtained as the mean bounding box dimension over the training set.

Assume there is an invertible function $\mathcal{T}$ that transforms locations in image space to corresponding (3D) locations in world space. A final bounding box prediction is obtained as $loc^{wld} = \mathcal{T}[loc^{img}]$ and $box^{wld} = (loc^{wld}, dim, dir, conf, cls)$. Each box component is optimized during model training to match the values of a ground truth (= target) box. Thus, the output loc^0 corresponds to a scaled residual between an anchor location and target location and the dimensions dim^0 are scaling factors with respect to mean box dimensions.

Broadcasted Target Assignment One of the critical steps in anchor-based object detection is target assignment, i.e. which cell is responsible for the prediction of a given ground truth bounding box. In 2D detection models such as YOLO [53], this problem is solved by assigning the cell and anchor box with highest overlap to the target box. As our system uses only a single anchor box m, such an approach in most cases selects the anchor a^* closest to the ground truth location in image space $\widehat{loc}^{img}$. For the set of anchors $\mathcal{G}$ we get:

$$a^* = \underset{a \in \mathcal{G}}{\arg\min} \left\| \widehat{loc}^{img} - a \right\|_\infty . \tag{2}$$

The residual $\widehat{loc}^{img} - \lfloor a^* \rfloor$ is used as a regression target for the output loc^{img} in the cell corresponding to a^*. We argue that in applications where high occlusions between objects are expected, this process can be improved upon. Instead of a single cell, we assign a *set* of cells to a single target box, a process we call broadcasted target assignment (BTA). To do so, instead of a target location we parameterize a target *line*: $\widehat{loc}^{img}(s)$, $s \in [0,1]$, where $\widehat{loc}^{img}(0)$ corresponds to the location of the target box at ground level and $\widehat{loc}^{img}(1)$ corresponds to the location at head level (see Figure 5).

The set of anchors $\mathcal{A}$ assigned to the target line corresponds to all cells intersected by it:

$$\mathcal{A} := \left\{ a \in \mathcal{G} : \left\| \widehat{loc}^{img}(s^*(a)) - a \right\|_\infty < 1/2 \right\}, \qquad s^*(a) := \underset{s \in [0,1]}{\arg\min} \left\| \widehat{loc}^{img}(s) - a \right\|_2 . \tag{3}$$

The individual regression targets are defined as the closest points $\widehat{loc}^{img}(s^*(a))$ to the target line. Using BTA, each cell additionally predicts a parameter $s(a)$, which is optimized to match $s^*(a)$ and indicates the location of the prediction along the target line. In the final transformation of the predicted

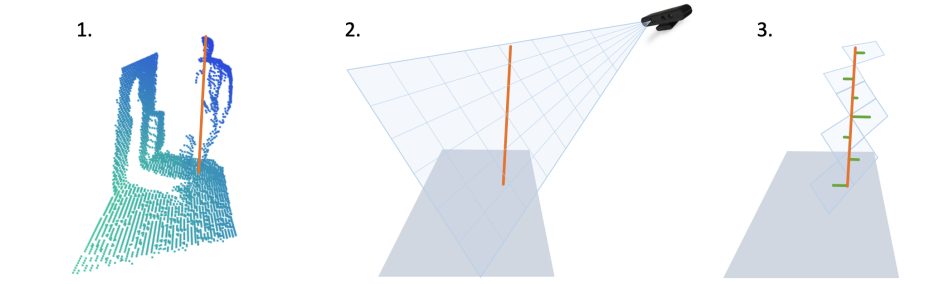


Figure 5: Depiction of broadcasted target assignment: 1. Parametrization of a target *line*, 2. Subdivision of the 3D image space into *cells*, 3. Assignment of regression targets.

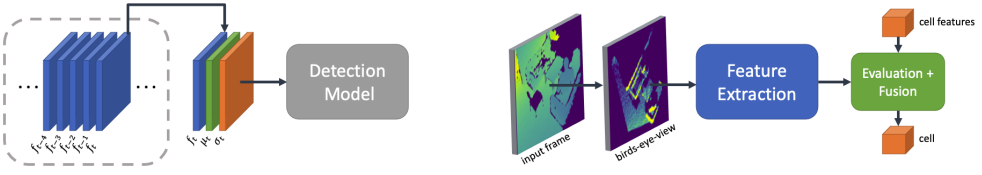


Figure 6: Optional components for our system. The background model (left) and cell feature augmentation using a secondary birds-eye-view branch.

location to world space, the computation of the height coordinate z is replaced with $loc_z^{wld} = \frac{\mathcal{T}[loc^{img}]_z}{s(a)+\varepsilon}$, where $\varepsilon > 0$ is a small regularization constant. In this way the predicted relative height $s(a)$ is used to reconstruct the full height of the bounding box.

Background Model In order to utilize temporal information available due to the sequential format of our dataset, we employ a background model as presented by Wren *et al.* [56] (left side in Figure 6). Given a sequence of input frames $(f_i)_{i=0}^{N}$, instead of a single-channel input of the most recent frame f_t, the model is given a stack (f_t, μ_t, σ_t) of the current frame, and geometrically weighed mean and standard deviation over previous frames:

$$\begin{aligned} \mu_t &= \alpha f_t + (1-\alpha)\mu_{t-1}, & \mu_0 &= f_0, \\ \sigma_t^2 &= \alpha(f_t - \mu_t)^2 + (1-\alpha)\sigma_{t-1}^2, & \sigma_0^2 &= 0. \end{aligned} \tag{4}$$

The scalar parameter α determines how strongly previous frames are weighed. In the multimodal settings this operation is performed for every modality.

Birds-Eye-View Module We present an optional birds-eye-view module (BEV), which uses a secondary view point and feature extraction branch for augmentation of cell features. Distance information of depth frames is used to transform input frames to a birds-eye-view point density representation (right side of Figure 6), from which features are extracted using a reduced Mobilenet v2 [57] backbone. The resulting feature map of shape $(n_x^{bev}, n_y^{bev}, n_c^{bev})$ (width, height and number of channels) is used to augment intermediate features at each cell. Let $\mathcal{S}$ be a coordinate transformation from world space to the 2D birds-eye-view. For each anchor a we evaluate the birds-eye-view feature map at location $a^{bev} = \mathcal{S}[\mathcal{T}[a]]$ (rounded integer coordinates) to obtain a feature vector of length n_c^{bev}. For each cell, intermediate feature before the detection head are concatenated with the new features, and fused using an inverted residual block [57].

Postprocessing To filter box predictions of insufficient confidence and fuse overlapping bounding bounding boxes we use non-maximum suppression with soft voting [7]. Overlapping bounding boxes

of the form $box^{wld} = (loc^{wld}, dim, dir, conf, cls)$ are averaged based on prediction confidence using a simple arithmetic mean of each attribute. If broadcasted target assignment is active, it is advisable to instead use a harmonic mean for the box height box_z^{wld}, as it is formed as the ratio of two predictions.

5 Experiments and Results

Evaluation For evaluation we use *average precision (AP)* [17] and *average heading similarity (AHS)* [29]. The AHS metric is a stricter extension of AP, where differences in orientation between predicted and ground truth bounding boxes are taken into account. If not stated otherwise, detection threshold is 0.25 based on 3D IoU. Tests are performed using two base configurations:

- **Res50**: Based on a ResNet50 [21] backbone at an image resolution of 416×416, with detection heads at three scale levels. Its intent is a demonstration of the best performance our proposed model can reach.
- **Mob2**: Based on a Mobilenet v2 [57] backbone. It uses an image resolution of 320×320 and detection heads at two scale levels. It main motivation is to be fast, so it can be used on edge devices.

More recent backbone architectures such as MobileNet v3 [24] did not show improved performance, which may suggest that architectures found trough neural architecture search (NAS) [59] overfit on either the RGB modality, or datasets used for its evaluation.

Ablation Study To demonstrate the effectiveness of our detection system and its extensions we perform an ablation study. Both the **Res50** and the **Mob2** configuration are tested with various combinations of the proposed extension modules (see Table 1). Our evaluations shows a clear positive impact of each introduced module. The baseline Res50 evaluation is provided both with anchor assignment based on the closest cell (equation 2), as well as based on largest IoU overlap (marked Res50_{IoU}) – a strategy found in existing architectures such as Faster-RCNN [54]. Differences between the two methods are within the margin of error. To validate the choice of our background model it is compared against a 10 channel input stack of the current, and previous frames sampled at a $0.5s$ interval (denoted BG_{Stack}), yielding a roughly 8% performance drop. Our Mob2 configuration, with utilization of a background model, shows detection quality on par with the Pointpillars [31] architecture – at less than a tenth of the inference time (on an Nvidia GeForce GTX TITAN X) and more than 11 fps on the Nvidia Jetson Nano edge device.

Comparison with state-of-the-art methods We compare our method against the state-of-the-art in 3D object detection [31, 36, 51, 58] (see Table 2). As these methods do not utilize any temporal information they are only compared against our best configuration without a background model (Res50 + BTA + BEV). We outperform all methods in inference time and achieve second best results on the AHS metric.

Imaging Modalities We evaluate detection performance under variations of the imaging modality (see Table 3). These tests are only trained, validated and tested on the multimodal subset of MIPT using a Mob2 + BTA + BG configuration. Use of thermal and RGB information (RGB not included in the public dataset) in addition to depth leads to a distinct performance increase compared to the depth-only baseline. Overall we achieve best results using a depth + thermal combination, suggesting that these modalities are complementary and form a better combination than the more traditionally used depth + RGB. Tests using only RGB or thermal modalities did not converge during training.

Table 1: Ablation study comparing combinations of backbones, background model (BG), broadcasted target assignment (BTA) and the birds-eye-view module (BEV). With (top section) and without (bottom section) temporal information.

Configuration	AHS	AP	fps (TITAN X)	fps (Jetson)
Res50	0.284	0.386	25.8	1.6
Res50$_{IoU}$	0.280	0.390	25.6	1.6
Res50 + BTA	0.406	0.475	25.6	1.6
Res50 + BEV	0.410	0.470	10.3	0.2
Res50 + BTA + BEV	**0.544**	**0.639**	10.3	0.2
Mob2	0.271	0.330	71.6	11.3
Mob2 + BTA	0.372	0.437	71.5	11.3
Res50 + BG	0.459	0.575	24.8	1.6
Res50 + BG$_{Stack}$	0.419	0.532	24.6	1.6
Res50 + BG + BTA	0.578	0.650	24.8	1.6
Res50 + BG + BEV	0.503	0.579	10.0	0.2
Res50 + BG + BTA + BEV	**0.592**	**0.664**	10.0	0.2
Mob2 + BG + BTA	0.452	0.530	68.5	11.1

Table 2: Comparison with state-of-the-art methods. Best method is marked in bold, second best in italic.

Configuration	AHS	AP	fps (TITAN X)
VoteNet [51]	0.508	0.634	*6.3*
PointPillars [31]	0.452	0.558	6.0
H3DNet [68]	0.532	*0.657*	4.2
Group-Free 3D [36]	**0.558**	**0.673**	3.9
Ours (Res50 + BTA + BEV)	*0.544*	0.639	**10.3**

Table 3: Evaluation of detection models using different imaging modalities.

Modalities	AHS	AP
Depth	0.336	0.408
Depth + Thermal	**0.403**	**0.486**
Depth + RGB	0.377	0.461

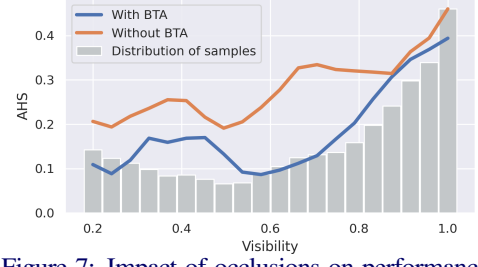


Figure 7: Impact of occlusions on performance in detection

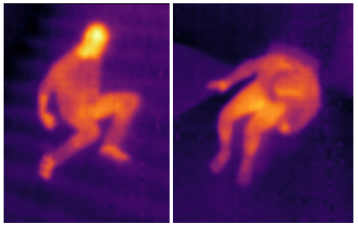


Figure 8: Unusual human poses causing inconsistent labeling.

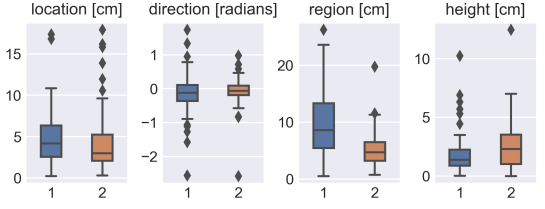
Occlusion performance We examine the impact of broadcasted target assignment on detection performance with respect to occlusions (see Figure 7). The visibility of a person in the scene is estimated using a heuristic based on the 3D bounding box and its surrounding depth readings. For a visibility of less than 90% and especially in the $60-80\%$ region we can see a tremendous improvement in AHS when using BTA compared to regular target assignment. The BTA approach is especially effective on datasets featuring tall and narrow objects, and smaller environments, where the subject intersects with a large number of cells. This makes our system less suitable for evaluation on publicly available datasets that either do not predominantly feature people, or have scenes of different scales, such as pedestrian detection in the KITTI benchmark [19] (scene scales are up to $10x$ in each dimension compared to our dataset).

Baseline Tracking We provide baseline tracking results for our best Res50 and Mob2 configurations (see Table 4). The detection system is supplement with the DeepSort [65] tracker, modified for tracking in 3D space instead of image space. In line with the MOT20 challenge [16] and the KITTI

Table 4: Baseline tracking evaluation using HOTA, CLEARMOT and Mostly-Tracked/ Partly-Tracked/Mostly-Lost metrics.

Method	HOTA	DetA	AssA	MOTA	MOTP	MT	PT	ML
Res50 (BTA + BEV + BG)	41.4%	50.4%	34.3%	51.1%	69.0%	7	13	1
Mob2 (BTA + BG)	34.2%	46.5%	25.5%	43.8%	69.4%	6	13	2

Figure 9: Statistical evaluation of labeling deviations by human annotators (left) and evaluation of human annotation (right).



Metric	Human 1	Human 2
AHS @ 0.25	0.931	0.979
AHS @ 0.5	0.669	0.583
AP @ 0.25	1.000	1.000
AP @ 0.5	0.861	0.810

benchmark [18] we report HOTA [33], CLEAR MOT [6] and mostly-tracked (MT)/partly-tracked (PT)/mostly lost (ML) [34] metrics. Required ground truth tracking files in the MOT Challenge format are included with the MIPT dataset.

Future work – Error of human annotation We demonstrate an effect observed in 3D human detection. Bounding boxes for human detection tend to be more ambiguous than for rigid objects (examples in Figure 8). Consequently, even under strict annotation guidelines we observe large deviations between labels by two human annotators on the same data sequence compared to the ground truth provided in the MIPT dataset. We analyse variances in location, direction, region (width and length) and height labels compared to ground truth (Figure and Table 9). This variance within dataset labels introduces an upper bound to the performance achievable on a dataset such as MIPT. Thus we propose the development of a detection metric derived from AP and AHS that place a stronger emphasis on localization accuracy over bounding box widths and heights as this is the more accurately labeled parameter.

6 Conclusion

We presented a foundation for identity-preserving human detection, both in terms of a specialized detection system and suitable data. Our system is highly configurable; evaluation showed that it can perform the current state-of-the-art, while also being configurable for inference on edge devices. We demonstrated the value of our innovative broadcasted target assignment (BTA) scheme, the birds-eye-view module and background model, and demonstrated their impact (or lack of impact) on inference speed. To further demonstrate strengths, our BTA scheme was more closely scrutinized in terms of dependence on target occlusions. The efficacy of a multimodal depth and thermal approach was demonstrated and it was shown that these modalities complement each other better than the common RGB + depth combination. Baseline tracking results were provided using metrics in line with the MOT20 challenge and the KITTI benchmark. Last we propose future work on specialized evaluation metrics for 3D human detection.

Acknowledgment

This work has been partly funded by the Austrian security research program KIRAS of the Federal Ministry of Agriculture, Regions and Tourism (BMLRT) under Grant 873495.

References

[1] Hamid Aghajan, Juan Carlos Augusto, Chen Wu, Paul McCullagh, and Julie-Ann Walkden. Distributed vision-based accident management for assisted living. In *International Conference on Smart Homes and Health Telematics*, pages 196–205. Springer, 2007.

[2] Troy J Allard, Richard K Wortley, and Anna L Stewart. The effect of cctv on prisoner misbehavior. *The Prison Journal*, 88(3):404–422, 2008.

[3] Umar Asif, Benjamin Mashford, Stefan Von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, and Stefan Harrer. Privacy preserving human fall detection using video data. In *Machine Learning for Health Workshop*, pages 39–51. PMLR, 2020.

[4] Flavia Benetazzo, Alessandro Freddi, Andrea Monteriù, and Sauro Longhi. Respiratory rate detection algorithm based on rgb-d camera: theoretical background and experimental results. *Healthcare Technology Letters*, 1(3):81–86, 2014.

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[6] Pascal Birnstill, Daoyuan Ren, and Jürgen Beyerer. A user study on anonymization techniques for smart video surveillance. In *12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2015.

[7] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5561–5569, 2017.

[8] Wassim Bouachir, Rafik Gouiaa, Bo Li, and Rita Noumeir. Intelligent video surveillance for real-time detection of suicide attempts. *Pattern Recognition Letters*, 110:1–7, 2018.

[9] T. Callemein, Kristof Van Beeck, and T. Goedemé. How low can you go? privacy-preserving people detection with an omni-directional camera. *ArXiv*, abs/2007.04678, 2019.

[10] Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[11] Chien-Yen Chang, Belinda Lange, Mi Zhang, Sebastian Koenig, Phil Requejo, Noom Somboon, Alexander A Sawchuk, and Albert A Rizzo. Towards pervasive physical rehabilitation using microsoft kinect. In *6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 159–162. IEEE, 2012.

[12] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017.

[13] Albert Clapés, Julio CS Jacques Junior, Carla Morral, and Sergio Escalera. Chalearn lap 2020 challenge on identity-preserved human detection: Dataset and results. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 801–808. IEEE, 2020.

[14] Rita Cucchiara, Andrea Prati, and Roberto Vezzani. A multi-camera vision system for fall detection and alarm generation. *Expert Systems*, 24(5):334–345, 2007.

[15] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman Sadeh, and Mahadev Satyanarayanan. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1387–1396. IEEE, 2017.

[16] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[17] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88 (2):303–338, June 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012.

[20] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[22] Thomas Heitzinger and Martin Kampel. Ipt: A dataset for identity preserved tracking in closed domains. In *25th International Conference on Pattern Recognition (ICPR)*, pages 8228–8234. IEEE, 2021.

[23] Gabriel Hermosilla, Javier Ruiz-del Solar, Rodrigo Verschae, and Mauricio Correa. A comparative study of thermal face recognition methods in unconstrained environments. *Pattern Recognition*, 45(7):2445–2459, 2012.

[24] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019.

[25] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504–516, 2017.

[26] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Temporally consistent 3d pose estimation in the interventional room using discrete mrf optimization over rgbd sequences. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 168–177. Springer, 2014.

[27] Michal Kepski and Bogdan Kwolek. Fall detection using ceiling-mounted 3d depth camera. In *2014 International conference on computer vision theory and applications (VISAPP)*, volume 2, pages 640–647. IEEE, 2014.

[28] Mate Krišto and Marina Ivasic-Kos. An overview of thermal face recognition methods. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1098–1103. IEEE, 2018.

[29] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[30] Soon-kak Kwon. Face recognition using depth and infrared pictures. *Nonlinear Theory and Its Applications, IEICE*, 10(1):2–15, 2019.

[31] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[32] Colin Lea, James Facker, Gregory Hager, Russell Taylor, and Suchi Saria. 3d sensing algorithms towards building an intelligent intensive care unit. *AMIA Summits on Translational Science*, 2013:136, 2013.

[33] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2953–2960. IEEE, 2009.

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[35] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li. Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on Cybernetics*, 43(5):1442–1452, 2013.

[36] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021.

[37] Matthias Luber, Luciano Spinello, and Kai O Arras. People tracking in rgb-d data with on-line boosted target models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3844–3849. IEEE, 2011.

[38] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

[39] Thomas B Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[40] Andreas Mogelmose, Chris Bahnsen, Thomas Moeslund, Albert Clapés, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 301–307, 2013.

[41] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy expectations and preferences in an iot world. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 399–412, 2017.

[42] Shota Nakashima, Yuhki Kitazono, Lifeng Zhang, and Seiichi Serikawa. Development of privacy-preserving sensor for person detection. *Procedia-Social and Behavioral Sciences*, 2(1):213–217, 2010.

[43] Mark S. Nixon, Tieniu N. Tan, and Rama Chellappa. *Human Identification Based on Gait*. Int. Series on Biometrics. Springer-Verlag, Dec. 2005.

[44] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose, Thomas B Moeslund, and Sergio Escalera. Multi-modal rgb–depth–thermal human body segmentation. *International Journal of Computer Vision*, 118(2):217–239, 2016.

[45] Francesco Pittaluga and Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2215–2226, 2016.

[46] Francesco Pittaluga, Aleksandar Zivkovic, and Sanjeev J Koppal. Sensor-level privacy for thermal cameras. In *2016 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2016.

[47] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 791–799. IEEE, 2019.

[48] Christopher Pramerdorfer, J Strohmayer, and Martin Kampel. Sdt: A synthetic multi-modal dataset for person detection and pose classification. In *IEEE International Conference on Image Processing (ICIP)*, pages 1611–1615. IEEE, 2020.

[49] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.

[50] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, 2018.

[51] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9277–9286, 2019.

[52] Re3data.Org. Cornell activity datasets: Cad-60 & cad-120, 2016. URL http://service.re3data.org/repository/r3d100012216.

[53] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, 2017.

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[55] Manola Ricciuti, Susanna Spinsante, and Ennio Gambi. Accurate fall detection in a top view privacy preserving configuration. *Sensors*, 18(6):1754, 2018.

[56] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Fall detection from human shape and motion history using video surveillance. In *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW)*, volume 2, pages 875–880. IEEE, 2007.

[57] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[58] JoonOh Seo, SangUk Han, SangHyun Lee, and Hyoungkwan Kim. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29(2): 239–251, 2015.

[59] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304. Ieee, 2011.

[60] Martin Simon, Stefan Milz, Karl Amende, and H. Groß. Complex-yolo: Real-time 3d object detection on point clouds. *arXiv preprint arXiv:1803.06199*, abs/1803.06199, 2018.

[61] Luciano Spinello and Kai O Arras. People detection in rgb-d data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3838–3843. IEEE, 2011.

[62] Luis Enrique Sucar, Felipe Orihuela-Espina, Roger Luis Velazquez, David J Reinkensmeyer, Ronald Leder, and Jorge Hernández-Franco. Gesture therapy: An upper limb virtual reality-based motor rehabilitation platform. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(3):634–643, 2013.

[63] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement*, 35(5):807, 2014.

[64] Di Wang, Fei Dai, and Xiaopeng Ning. Risk assessment of work-related musculoskeletal disorders in construction: State-of-the-art review. *Journal of Construction Engineering and Management*, 141(6):04015008, 2015.

[65] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[66] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):780–785, 1997.

[67] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012.

[68] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *Proceedings of the European Conference on Computer Vision*, 2020.

[69] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.