Multimodal Human Mesh Recovery for Stand-off Triage in Mass Casualty Scenarios

Aniket Agarwal

CMU-RI-TR-25-66

July 28, 2025

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Professor László Attila Jeni *(chair)*
Professor Srinivasa Narasimhan
Mosam Dabhi

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

To my family, friends, and mentors.

Abstract

Mass-casualty triage robots must judge a victim's ability to move—even when light is scarce, geometry is ambiguous, or some sensors fail. We tackle this challenge with a multimodal mesh-recovery framework that fuses RGB, LiDAR, and infrared (IR) inputs to reconstruct full-body SMPL meshes from whichever subset of sensors is available. Built on a frozen HMR 2.0 backbone, a lightweight transformer-based Modality Unifier is trained with random modality dropout so the same weights handle any sensor combination offline.

To supervise and benchmark such fusion, we curate two mesh-annotated datasets: STCrowd-Mesh (∼10 k RGB+LiDAR pedestrian frames) and LLVIP-Mesh (∼15 k aligned RGB-IR pairs). On STCrowd-Mesh, adding LiDAR to RGB lowers MPJPE from 86.9 mm to 75.8 mm (–14.3%) and PA-MPJPE from 63.1 mm to 57.8 mm (–5.3 mm), confirming LiDAR's value for absolute spatial accuracy. In low-light LLVIP-Mesh scenes, fusing IR with RGB yields consistent but smaller gains, indicating complementary appearance cues.

We deploy the trained model offline on real-world DARPA Triage Challenge field logs recorded from a mobile Spot robot. The recovered SMPL meshes are used to infer motor alertness by analyzing joint movement patterns across time, classifying casualties into normal, abnormal, or unresponsive categories. These results show that dense, modality-flexible 3D pose estimation can enable remote physiological assessment, offering a promising step toward fully autonomous triage in complex and degraded environments.

Acknowledgments

To begin with, I would like to express my heartfelt gratitude to my advisor Professor László Jeni for his continued support and guidance throughout my masters and during this work. Tackling through many challenges and research ideas through discussions with him have been the highlight of my MSR journey at CMU.

I also thank Professor Srinivasa Narasimhan & Professor Sebastian Scherer for their guidance throughout the DARPA Triage Challenge. I learned a lot on how to think from first principles from Professor Srinivasa and also pretty much all about core robotics from Professor Sebastian.

My teammates in DARPA Triage Challenge, specifically Kabir Kedia, Mayank Mishra, Yaoyu Hu, Ceci Morales & James Emilian, taught me a lot, not just from technical perspective, but also in a personal capacity on how to deal with problems that may arise in such a large project. The constant late night grinds towards the challenge deadline, while dauting, have undoubtedly been a great learning experience. I would also love to thank the PI's Dr Kimberly Elenberg & Professor John Galeotti for their support and constant motivation through some of the toughest times through our project.

I would like to sincerely thank my committee members, Professor Srinivasa and Mosam Dabhi for their invaluable feedback on my thesis project and motivating me to think about potential pitfalls in the initial drafts.

To my parents, Parameshwar and Indu, and my sisters, Shipra and Sanjana, who have always been my side throughout my academic journey: motivating and supporting me through thick and thin. Their support and encouragement have been the paramount reason of whatever I have achieved till now and get to be in this life. I am also grateful to have my friends Vedant, Krish, Karan, Parth, Rohit, Alex and Raymond by my side throughout the two years of my master's program.

Funding

This work has been funded by the DARPA Triage Challenge Contract.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mass casualty incidents, such as those caused by natural disasters, terrorist attacks, or large-scale accidents, present overwhelming challenges to first responders and medical teams. In such high-pressure scenarios, the rapid and accurate assessment of casualties is critical for prioritizing care and improving survival outcomes [30]. Traditionally, triage in these situations relies heavily on manual observation and subjective judgment, which can be error-prone and time-consuming, especially in environments with limited visibility, obstructions, or dynamic threats.

To address this problem, the **DARPA Triage Challenge** [6] was introduced as a benchmark initiative to explore the use of autonomous systems and AI for assessing physiological and behavioral indicators of casualties [4, 18, 28]. One of its key thrusts is the development of machine perception systems capable of evaluating ocular, motor or distress indicators without physical contact, even in degraded or uncertain environments.

This thesis is motivated by the vision set forth by the DARPA Triage Challenge. We propose a multimodal perception framework that leverages RGB cameras, LiDAR, and infrared (IR) sensors to robustly estimate human pose and activity. Unlike traditional monocular pose estimation systems that are prone to failure in poor lighting or occluded conditions [16], our approach fuses complementary modalities to generate dense pose representations.

Beyond pose estimation, we also demonstrate how such a system can be used in downstream triage-relevant tasks, such as detecting movement or alertness levels,

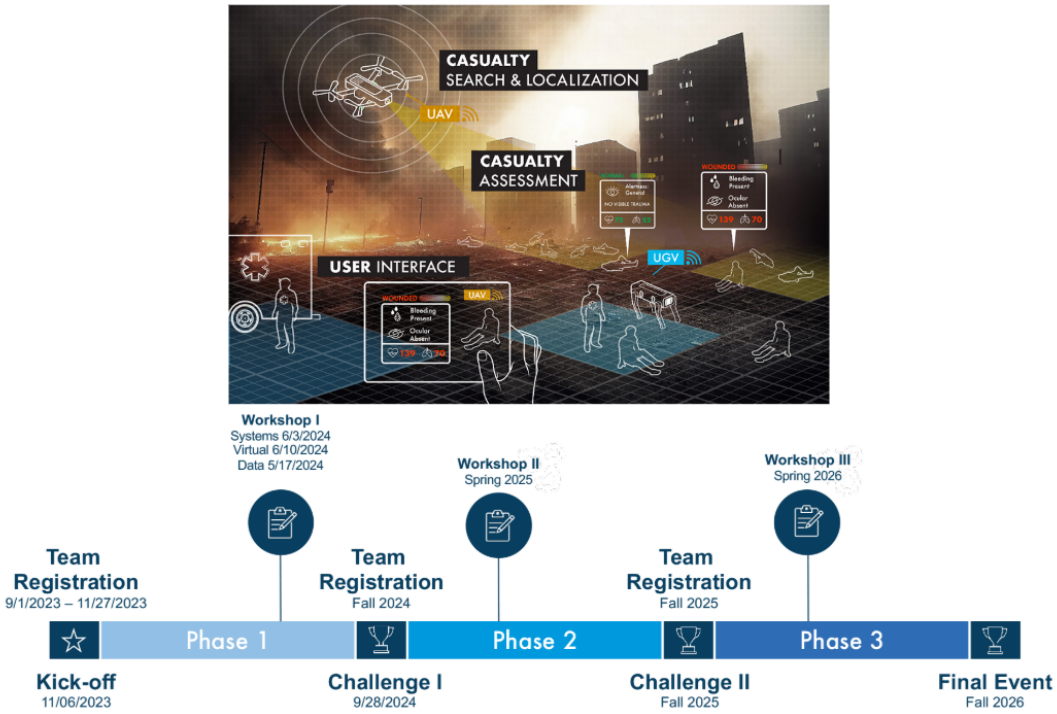

Figure 1.1: The DARPA Triage Challenge is a 3 year long challenge. Each year follows a similar pattern of a workshop (mainly for data collection) and the main challenge where teams compete. The ultimate goal of this challenge is to develop autonomous robots with ability of remote physiological signs assessment.

classifying states like unconsciousness or postural instability, and ultimately aiding autonomous triage decision-making.

1.1 The DARPA Triage Challenge (DTC)

The DARPA Triage Challenge [6] is a research initiative launched by Defense Advanced Research Projects Agency (DARPA) to spur the development of new AI systems specifically targeted towards the problem of triage. It is a 3-year long challenge with the end goal of developing robots that can autonomously assess casualties in a simulated mass-casualty scenario and remotely detecting their physiological signs (Fig 1.1). While there are several tracks in the DTC challenge, CMU team (**Team Chiron**) participated in the *Systems Track* where the primary task is for teams to perform stand-off evaluations of simulated casualties on a certain number of courses.

Figure 1.2: *(Left)* Team Chiron at DTC Challenge Year-1 with our (pet) robots. *(Right)* A closer look at the sensor payload used in Year-1. We used an iphone as our primary rgb sensor attached to a gimbal for necessary movements. We also utilized zed cameras for locomotion, speaker/microphone for specific tasks, with the main computational overhead being handled by an onboard Nvidia Jetson Orin.

DTC Year-1 Challenge

In the first year (2023-2024) of DARPA Triage Challenge Systems Track, the primary objective was to develop robots (UGVs/UAVs) from ground up with the proper sensor payload and algorithm deployment for remote physiological sign assessment. In Fig 1.2 you can see the swarm of robots we used for the challenge (2RCs, 1 wheelchair, 1 spot, 1 UAV). Each robot was selected to leverage its unique mobility capabilities and sensor mounting flexibility, enabling robust performance across different terrains and operational conditions. For instance, tracked UGVs (2RCs) provided excellent stability and traction on uneven or loose surfaces, making them ideal for traversing rubble or outdoor environments. The wheelchair platform offered a stable, low-speed option for controlled environments like pavement or flat indoor testbeds, with a high payload capacity for heavier sensors. The Boston Dynamics Spot robot, with its quadruped design, excelled in navigating tight or cluttered spaces and stepping over obstacles, allowing it to inspect hard-to-reach victims. Finally, the UAV (drone)

added aerial coverage, enabling rapid area exploration and top-down inspection, especially in scenarios where ground access was restricted.

For the challenge itself, there were three tracks, namely Urban, Crash and Battlefield (as seen in Fig 1.3) that we were evaluated on. Each track came with its own sets of difficulties in terms of uneven terrain, heavy occlusions, wide casualty dispersion, etc. Each course demanded a different operational strategy, and the heterogeneous fleet allowed us to play to each robot's strengths; whether it was the terrain adaptability of the tracked UGVs, the agility of Spot in confined areas, or the aerial reach of the UAV. For the actual physiological signs estimation and a brief of the methods/algorithms we utilized for the same, we outline as below (Since this is not the main focus of this thesis we outline the methods used for each of them briefly):

1. **Severe Hemorrhage:** We employed segmentation models [21] in conjunction with IoU-based spatial heuristics to identify regions of excessive blood pooling, indicative of severe hemorrhage.
2. **Respiratory Distress:** We used an RGB-based human pose estimation model [13] to detect postures commonly associated with respiratory distress, complemented by analysis of arrhythmic chest movement patterns derived from temporal variations in upper-body pixel intensity.
3. **Vitals:** This involved estimating Heart Rate (HR) and Respiratory Rate (RR). HR was computed using Independent Component Analysis (ICA) [10] over facial regions in RGB imagery. RR was estimated by analyzing chest motion using dense optical flow [23] to capture periodic breathing patterns.
4. **Trauma:** Trauma detection encompassed both localized injury identification and amputation assessment. For injury detection, we fine-tuned the multimodal vision-language model LLaVA [20] using a custom dataset of prompt-image pairs focused on visible injuries. For amputation detection, we used 2D human pose estimators (YOLOv11 [13]) to obtain full-body keypoints from multiple views around the casualty. These were then triangulated and lifted into 3D, allowing us to apply limb-length heuristics to infer missing limbs.
5. **Alertness:** The goal was to assess a casualty's responsiveness by evaluating ocular, verbal, and motor cues using multimodal sensor data. For *ocular*

Figure 1.3: The three evaluation tracks from the DARPA Triage Challenge Year-1 Systems Track—*Urban*, *Crash*, and *Battlefield*. Each track simulated a different disaster scenario with unique environmental and triage challenges. Urban involved tight alleyways and buildings, Crash featured burnt vehicles and uneven terrain, and Battlefield replicated wide, grassy areas with dispersed casualties. These settings tested our system's ability to perform physiological sign assessment across varying terrains, lighting conditions, and casualty layouts.

alertness, we employed Apple Vision's facial landmark detection system [2] to extract eye keypoints, and determined eye state (open/closed) based on a threshold over the Euclidean distance between eyelid landmarks. For *verbal alertness*, we used OpenAI's Whisper model [27] to transcribe ambient audio captured by onboard microphones, followed by a bag-of-words analysis to detect the presence of speech. Lastly, *motor alertness* was evaluated by tracking 3D joint trajectories over time using the lifting model used for amputation detection; movement magnitude thresholds were applied to identify spontaneous limb or body motion.

To summarize, Year 1 of the DARPA Triage Challenge Systems Track laid the groundwork for building autonomous, sensor-rich robotic systems capable of remote

physiological sign assessment in complex, outdoor environments. By combining a heterogeneous robot fleet with task-specific algorithms across multiple sensing modalities, we demonstrated a functional triage system that operated across diverse terrain and casualty scenarios. Our efforts culminated in a successful performance during the final evaluation, where we secured an overall **4th place** among participating teams. This first year served as a crucial foundation, technically and operationally, paving the way for the more ambitious goals and autonomy-focused challenges of Year 2.

DTC Year-2 Challenge

The second year of the DARPA Triage Challenge Systems Track brought a significant shift in focus and complexity compared to Year 1. While the first year primarily evaluated perception pipelines under supervised or teleoperated robot control, Year 2 emphasized complete autonomy—requiring robots to independently navigate through challenging environments, localize casualties, and assess their physiological states without any human intervention. The operating conditions were also substantially harsher: many scenarios took place in low-light or fully dark settings, pushing the need for robust multimodal sensing using infrared and LiDAR data when RGB information became unreliable. One of the major additions in Year 2 was the requirement to precisely geolocate each casualty and report their position as part of the evaluation. The test courses themselves became more dynamic and realistic, often involving cluttered terrain, occluded viewpoints, and even walking or talking actors simulating live casualties. These changes collectively elevated the difficulty of the challenge and placed a strong emphasis on autonomy, perception under degraded sensing, and accurate spatial understanding. This set the stage for more advanced integration of navigation, localization, and physiological sign estimation in our robotic systems.

Fig 1.4 shows our robotic platform and sensor payload designed to meet these challenges. The multimodal sensor suite enabled robust perception in degraded conditions, while the autonomy payload and arm-mounted sensors supported navigation and fine-grained victim assessment.

These increasing demands for autonomous reasoning in degraded environments, coupled with the need for robust human state estimation, highlighted a critical

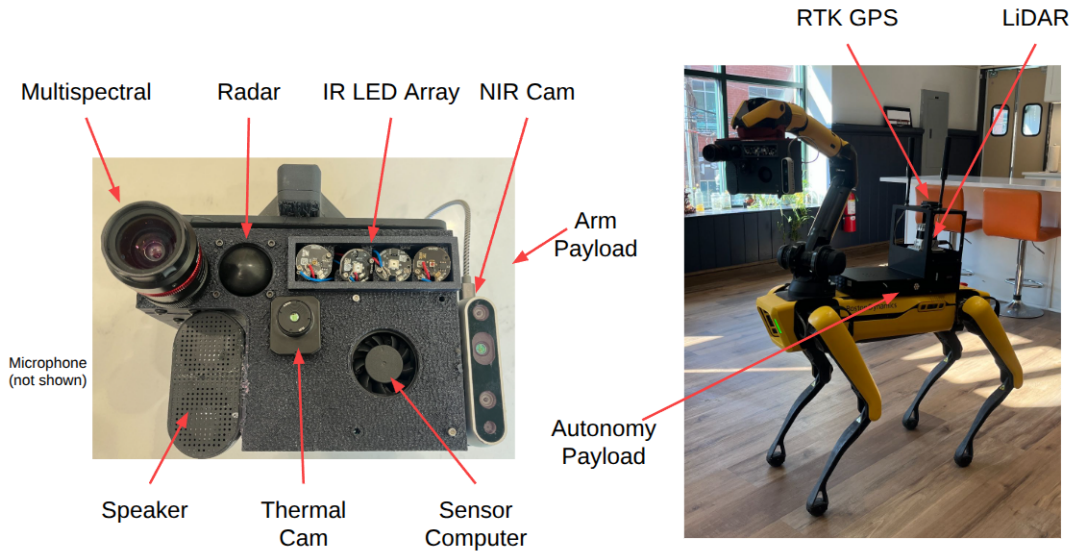

Figure 1.4: Sensor payload and robotic platform used in Year-2 of the DARPA Triage Challenge Systems Track. *Left:* A custom perception payload equipped with multispectral, radar, IR, NIR, and thermal sensors, along with a speaker, microphone (not shown), and onboard computation for real-time inference. *Right:* The Boston Dynamics Spot robot outfitted with an autonomy payload, LiDAR for mapping and navigation, RTK GPS for precise localization, and an arm-mounted perception suite for multimodal data capture. This configuration enabled fully autonomous operation and multimodal triage assessment in challenging environments.

bottleneck in the perception pipeline: the accurate and consistent understanding of human pose from noisy, incomplete, or modality-limited inputs. As traditional monocular pose estimation methods often failed under occlusions, poor lighting, or low-resolution inputs, it became evident that a more resilient solution was required, one that could effectively fuse complementary cues from multiple sensing modalities. This realization served as the key motivation behind the design and development of our *multimodal pose detection framework*, which forms the core contribution of this thesis.

1.2 Mutimodal Pose Detection: Motivation

In the course of building autonomous systems for the DARPA Triage Challenge Year-2, one of the most critical perception bottlenecks we identified was the need

for robust and spatially detailed human pose understanding in complex, real-world environments. Keypoint-based methods, while efficient, often failed under occlusions, poor lighting, or limited sensor conditions, leading to hinderance in downstream tasks such as trauma localization, motion tracking, and physiological sign assessment.

To overcome these challenges, we turned to dense 3D mesh representations of the human body, which provide richer structural information, including surface normals and body region segmentation. This was especially important in our setup, where the perception payload was mounted on the **arm of the Spot robot**, requiring viewpoint-aware autonomy to optimally position sensors for casualty assessment. Additionally, given the variability and complementary strengths of the sensing modalities(RGB, LiDAR, and infrared), we were motivated to develop a unified *multimodal pose detection framework* capable of adapting to sensor availability while producing consistent, high-fidelity mesh outputs. This thesis introduces such a system and details its design, training, and real-world evaluation.

Despite considerable advances in RGB-only mesh estimation, such as HMR [12], SPIN [17], and HMR 2.0 [8], these methods often struggle under occlusion, poor lighting, or complex scene clutter. Occlusion-aware variants like OCHMR [14] improve robustness in crowded environments, yet remain fundamentally limited by their reliance on RGB input alone. To overcome these single-modality bottlenecks, recent research has explored alternative sensing pathways: LiDAR-HMR [7] reconstructs human meshes from sparse point clouds, RF-Avatar [37] leverages radio-frequency reflections to handle occlusions, and M4esh [34] uses mmWave radar for mesh estimation in visually adverse settings.

Notably, MMPedestron [35] introduces a unified modular encoder for multimodal pedestrian detection, supporting inputs such as RGB, infrared, LiDAR, depth, and event streams. While highly effective for classification and detection tasks, MMPedestron does not address dense pose recovery. In this work, we take inspiration from its modality-unification strategy and extend it to the mesh regression domain. By integrating a similar fusion design into a frozen HMR 2.0 backbone, we develop the first framework capable of estimating full-body human meshes from any subset of RGB, LiDAR, and infrared data—enabling resilient pose recovery in diverse triage conditions.

To address these limitations, we propose a unified architecture for dense 3D human

pose estimation that can natively operate with RGB, LiDAR, and infrared inputs, used individually or in any combination. Built atop the HMR 2.0 [8] mesh regression backbone, our model embeds each modality into patch tokens, processes them with a lightweight transformer [31], and uses a learnable modality unifier to seamlessly handle missing or noisy input streams. To our knowledge, this is the first mesh regression system with such flexibility across all three complementary modalities.

A further contribution lies in dataset creation. We extend two existing datasets to enable mesh-supervised multimodal learning: *STCrowd-Mesh* augments RGB+LiDAR scenes with SMPL annotations, and *LLVIP-Mesh* provides RGB+infrared pairs similarly annotated for full-body mesh regression. These resources allow rigorous evaluation and training in modality-variable scenarios, laying essential groundwork for robust multimodal perception.

This thesis makes the following key contributions:

1. A **unified multimodal mesh recovery architecture**—built on an HMR 2.0 backbone with a lightweight transformer fusion module and a learnable modality unifier—that is the first to natively ingest and regress dense human pose meshes from any combination of *RGB*, *LiDAR*, and *infrared* inputs.
2. The creation and release of two **mesh-supervised multimodal datasets**:
   - *STCrowd-Mesh*: extending synchronized RGB+LiDAR streams with SMPL mesh annotations.
   - *LLVIP-Mesh*: extending aligned RGB+infrared pairs with full-body SMPL labels.
3. Demonstrated **practical applications** of our framework, including motor alertness detection and real-world deployment on Spot robots, showcasing dense human mesh pose estimation for viewpoint-aware autonomy and casualty assessment.

Chapter 2

Background and Related Works

2.1 Background

SMPL Model

The SMPL (Skinned Multi-Person Linear) model [22] is a parametric model of the human body that enables realistic 3D mesh reconstruction by regressing body shape and pose parameters. It outputs a triangulated mesh of 6890 vertices that adapts to different identities and articulations using learned blend shapes and skinning weights.

Formally, the SMPL model is defined as:

$$M(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{T}, \mathcal{S}, \mathcal{P}, \mathcal{W}, \mathcal{J}), \tag{2.1}$$

where:

- $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ represents the **pose parameters** — relative joint rotations of K body parts.
- $\boldsymbol{\beta} \in \mathbb{R}^{|\beta|}$ denotes the **shape parameters** — controlling body proportions.
- $\mathbf{T} \in \mathbb{R}^{6890 \times 3}$ is the **template mesh** in canonical T-pose.
- $\mathcal{S}$ and $\mathcal{P}$ are the learned **shape** and **pose blend shapes**.
- $\mathcal{W}$ is the set of **skinning weights** for linear blend skinning (LBS).
- $\mathcal{J}$ is the **joint regressor**, mapping vertices to joint locations.

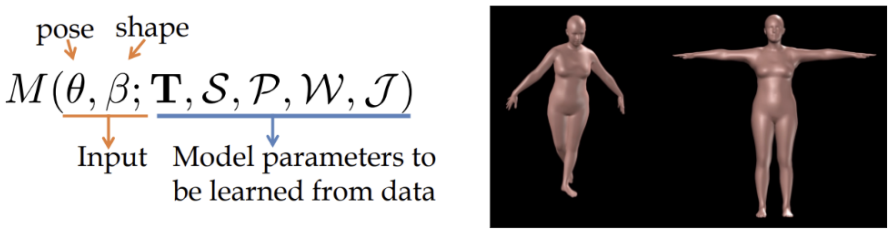

Figure 2.1: The SMPL model takes shape ($\boldsymbol{\beta}$) and pose ($\boldsymbol{\theta}$) as input, along with learned model parameters $\mathbf{T}, \mathcal{S}, \mathcal{P}, \mathcal{W}, \mathcal{J}$, to produce a realistic human mesh.

The shape parameters $\boldsymbol{\beta}$ deform the template mesh to represent different body types, while the pose parameters $\boldsymbol{\theta}$ articulate the mesh via LBS. The result is a realistic, fully skinned mesh ready for animation or 3D reasoning. Fig 2.1 illustrates the full SMPL pipeline and its components.

This formulation allows SMPL to generalize well across different human body shapes and poses, making it an effective representation for dense pose estimation tasks.

Sensor Modalities

RGB: RGB images capture dense texture and color information from the visible spectrum. They are typically used in computer vision tasks such as detection, segmentation, and pose estimation. However, RGB data is sensitive to lighting conditions and occlusion, making it unreliable in low-light or cluttered environments.

LiDAR: Light Detection and Ranging (LiDAR) sensors emit laser pulses and measure the time-of-flight to estimate depth. LiDAR provides accurate 3D structure, especially useful for spatial geometry, scene understanding, and surface estimation. In human-centric tasks, LiDAR is effective for estimating coarse body shapes but may suffer from sparsity on soft or absorptive surfaces like clothing or skin.

Infrared (IR): Infrared imaging captures thermal radiation or near-infrared reflectance depending on the sensor type. IR modalities are particularly useful in low-light or nighttime conditions where RGB fails. They can reveal body outlines, limb temperature gradients, or heat-based cues, but lack fine texture and structural

detail.

Each of these modalities provides complementary information: RGB offers high-resolution appearance cues, LiDAR encodes geometric depth, and infrared supports visibility in adverse lighting. Combining them enables more robust and context-aware perception systems for real-world applications.

2.2 Related Works

Human Mesh Recovery and 3D Pose Estimation

Human mesh recovery (HMR) refers to the task of estimating a full 3D mesh of the human body from sensor data such as images or point clouds. Classical 3D pose estimation methods focused on predicting a sparse set of keypoints [25, 26, 29], often using intermediate 2D joint predictions followed by lifting to 3D. While effective for coarse motion understanding, keypoints alone lack the surface-level detail needed for applications like contact reasoning or fine-grained geometry analysis. To address this, parametric models like SMPL [22] have become central to mesh recovery pipelines. HMR [12] pioneered end-to-end mesh regression from a single RGB image by learning to predict SMPL pose and shape parameters. Subsequent works such as SPIN [17] improved robustness via iterative refinement and self-supervised losses. HMR 2.0 [8] introduces temporal priors and depth cues to enhance realism and stability. More recent methods like KBody [39] and PARE [15] address occlusion via part-based or attention-guided strategies, while others [36] extend mesh regression to hands and face. Despite strong results in controlled settings, these RGB-only methods degrade under occlusion, low-light, or complex backgrounds. Non-SMPL mesh recovery methods [33, 38] attempt to directly predict vertex positions, trading semantic structure for flexibility. However, SMPL-based techniques remain dominant due to their compactness, physical realism, and compatibility with downstream pose analysis tools.

Multimodal Datasets for Human Perception

The demand for robust human understanding across diverse scenarios has led to datasets that combine multiple sensing modalities. The CMU Panoptic Studio [11]

provides synchronized multi-view RGB and depth for social interaction analysis. PROX [9] captures human-environment interactions with RGB-D and IMU sensors. The EPFL DensePose dataset [1] annotates dense correspondences but relies only on RGB. For outdoor scenarios, STCrowd [5] offers RGB and LiDAR in crowded environments, but lacks mesh annotations. LLVIP [32] introduces paired RGB and infrared images for low-light pedestrian detection. The JRDB dataset [24] combines RGB, LiDAR, and audio for social navigation, yet focuses on bounding boxes and trajectories. While these datasets highlight the value of multimodal sensing, few provide ground-truth mesh supervision, and even fewer align modalities at per-frame or per-person granularity. This thesis addresses this gap by augmenting STCrowd and LLVIP with SMPL mesh annotations, creating two new multimodal benchmarks tailored for mesh recovery.

Multimodal Pose and Mesh Recovery

To overcome the limitations of single-modality systems, several works have explored multimodal fusion for pose or mesh estimation. LiDAR-HMR [7] adapts HMR to sparse point clouds by encoding depth distributions. RF-Avatar [37] uses RF signals to infer meshes through walls and occlusions. M4esh [34] incorporates mmWave radar for robust mesh reconstruction in adverse environments. In the domain of pedestrian detection, MMPedestron [35] proposes a modular transformer-based encoder that dynamically fuses RGB, LiDAR, infrared, and depth for robust perception. While focused on bounding boxes, its design inspires our modality unifier for mesh regression. AdaptiveFusion [3] shows benefits of token-level fusion across RGB and depth in pose estimation but lacks infrared support and mesh output. Our approach builds on these insights by fusing RGB, LiDAR, and infrared within a transformer-based encoder and regressing SMPL meshes using a frozen HMR 2.0 backbone. Unlike prior work, we support any subset of modalities, enabling deployment under sensor degradation. This multimodal and flexible design is critical for real-world triage settings, motivating the unified architecture presented in this thesis.

Chapter 3

Dataset Creation

Here we outline the process of creating the pseudo-SMPL annotations for the STCrowd and LLVIP dataset and creating our STCrowd-Mesh & LLVIP-Mesh datasets.

3.1 STCrowd-Mesh

To enable mesh-supervised training in outdoor multimodal settings, we extend the STCrowd [5] dataset by generating SMPL mesh annotations for each individual. STCrowd comprises approximately 10,891 synchronized image-LiDAR frames of crowded street scenes captured using a 128-beam LiDAR and RGB cameras. Each frame includes high-quality 3D bounding boxes and instance-level annotations for pedestrians, making it an ideal foundation for multimodal mesh generation. We outline the process for generating the ground-truth SMPL parameters from this dataset below and a brief of it can also be seen in Fig 3.1.

Initial Mesh Estimation. We begin by applying the HMR 2.0 [8] model to each RGB image crop around a person's 2D bounding box to obtain coarse SMPL parameters: pose $\boldsymbol{\theta}_0 \in \mathbb{R}^{72}$ and shape $\boldsymbol{\beta}_0 \in \mathbb{R}^{10}$. These parameters define a human mesh $\mathbf{V}$ through the SMPL function:

$$\mathbf{V} = \text{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta}) \tag{3.1}$$

Here, $\mathbf{V} \in \mathbb{R}^{6890\times 3}$ is the mesh vertex matrix, and SMPL is a parametric body model.

Figure 3.1: STCrowd mesh fitting pipeline. Top left: synchronized RGB image and 2D detection; bottom left: associated LiDAR points; center: coarse mesh prediction via HMR 2.0; right: refinement via LiDAR-guided optimization. Red shows initial mesh, green shows final mesh.

LiDAR Point Cloud Filtering. Given raw LiDAR point cloud data $\mathbf{P}_{\text{raw}} \in \mathbb{R}^{M\times 3}$, we filter it using two stages:

- **Camera View Filtering:** We keep points that project within the image bounds and a predefined depth range.
- **3D Bounding Box Filtering:** Using the provided 3D bounding box, we extract person-centric regions around the LiDAR centroid $\mathbf{c}_{\text{bbox}}$ and spatial extent $\mathbf{s}_{\text{bbox}}$:

$$\mathbf{P}_{\text{filtered}} = \{\mathbf{p} \in \mathbf{P}_{\text{raw}} \mid |\mathbf{p} - \mathbf{c}_{\text{bbox}}| \leq \alpha \cdot \mathbf{s}_{\text{bbox}}/2\} \tag{3.2}$$

SMPL-LiDAR Optimization. To better align the initial mesh with 3D geometry, we optimize a composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{chamfer}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} \tag{3.3}$$

where $\mathcal{L}_{\text{chamfer}}$ measures point-wise distance between mesh vertices $\mathbf{v}_i$ and LiDAR

points $\mathbf{p}_j$:

$$\mathcal{L}_{\text{chamfer}} = \frac{1}{N}\sum_{i=1}^{N}\min_j \|\mathbf{v}_i - \mathbf{p}_j\|^2 + \frac{1}{M}\sum_{j=1}^{M}\min_i \|\mathbf{p}_j - \mathbf{v}_i\|^2 \tag{3.4}$$

$$\mathcal{L}_{\text{shape}} = \|\boldsymbol{\beta}\|_2^2 \quad ; \quad \mathcal{L}_{\text{pose}} = \|\boldsymbol{\theta}_b\|_2^2 \tag{3.5}$$

Coordinate Alignment and Scaling. We compute a robust scale factor s to match the extent of the predicted mesh and LiDAR region:

$$s = \text{median}\left(\frac{\mathbf{e}_{\text{lidar}}}{\mathbf{e}_{\text{mesh}} + \epsilon}\right) \tag{3.6}$$

where $\mathbf{e}_{\text{lidar}}$ and $\mathbf{e}_{\text{mesh}}$ denote the bounding box sizes of the LiDAR and mesh, respectively. The final transformed mesh is obtained as:

$$\mathbf{V}_{\text{transformed}} = \text{Transform}(\text{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta}), s, \phi, \mathbf{c}_{\text{lidar}}) \tag{3.7}$$

where ϕ is the rotation from the camera to LiDAR coordinate frame.

Algorithm 1 outlines this smpl lidar fitting optimization process.

Algorithm 1 SMPL-LiDAR Fitting Optimization

Input: Initial SMPL parameters $\boldsymbol{\theta}_0, \boldsymbol{\beta}_0$, filtered LiDAR points $\mathbf{P}_{\text{filtered}}$
Initialize: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$, $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_0$
Compute scale factor s between LiDAR and mesh extents
Initialize Adam optimizer and loss weights $\lambda_{\text{shape}}, \lambda_{\text{pose}}$
for $t \leftarrow 1$ **to** T **do**
    $\mathbf{V} \leftarrow \text{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta})$
    $\mathbf{V}_{\text{trans}} \leftarrow \text{Transform}(\mathbf{V}, s, \phi, \mathbf{c}_{\text{lidar}})$
    Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{chamfer}} + \lambda_{\text{shape}} \cdot \mathcal{L}_{\text{shape}} + \lambda_{\text{pose}} \cdot \mathcal{L}_{\text{pose}}$
    Update $\boldsymbol{\theta}, \boldsymbol{\beta}$ via Adam optimizer
end for
Output: Refined SMPL parameters $\boldsymbol{\theta}^*, \boldsymbol{\beta}^*$

Filtering Criteria and Dataset Statistics. To ensure high-quality supervision, we filter the STCrowd dataset to retain only individuals with low to moderate

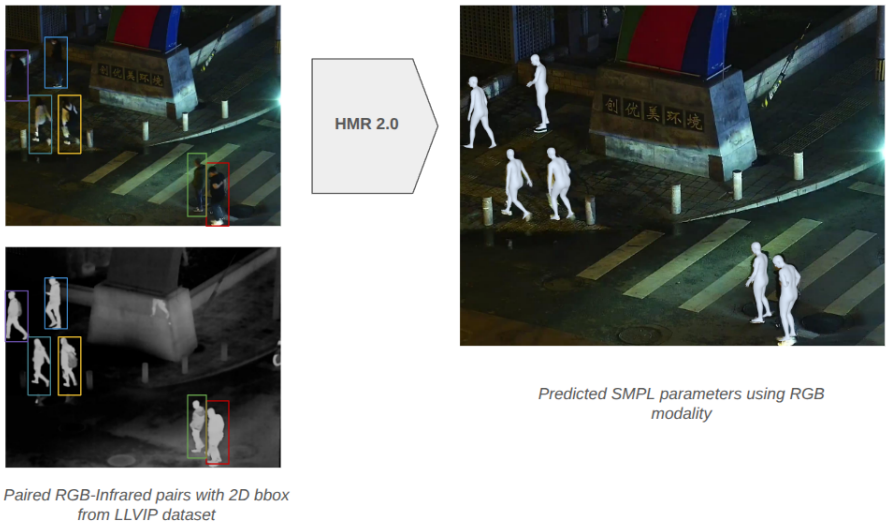

Figure 3.2: LLVIP mesh generation pipeline. Top left: RGB image with 2D person bounding boxes; bottom left: corresponding infrared image; center: HMR 2.0 model applied to RGB crop; right: predicted SMPL mesh overlaid on original RGB image. Mesh is reused for both RGB and infrared supervision.

occlusion (occlusion level 0 or 1) and sufficient 3D evidence (LiDAR point count > 70). After filtering, we obtain **10,342** valid RGB-LiDAR frames with average of approximately 4 persons per frame with refined mesh annotations. These are split into 75% for training and 25% for evaluation.

3.2 LLVIP-Mesh

The LLVIP dataset [32] consists of over 15488 aligned RGB-infrared image pairs captured in low-light conditions for pedestrian detection. Each frame includes 2D bounding boxes for all visible persons, with pixel-level alignment between RGB and infrared views. However, it lacks dense mesh or 3D pose supervision.

Pseudo Mesh Generation Using HMR 2.0. To extend LLVIP with mesh annotations, we adopt a pseudo-labeling strategy using a pre-trained RGB-based mesh recovery model. We utilize the 2D bounding boxes provided in LLVIP to crop individual person regions from the RGB image. These crops are passed through the HMR 2.0 [8] model to estimate SMPL parameters:

- Pose $\boldsymbol{\theta} \in \mathbb{R}^{72}$,
- Shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$,
- Mesh $\mathbf{V} = \mathrm{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$.

The predicted SMPL mesh acts as a pseudo-ground-truth annotation for the cropped RGB region.

Annotation Transfer to Infrared. Since LLVIP provides tightly aligned RGB-IR pairs, we reuse the same SMPL mesh annotations for both the RGB and IR modalities of a person. This enables modality-agnostic training using shared mesh targets, without requiring 3D annotations in the infrared domain.

Final LLVIP-Mesh Statistics. The full LLVIP dataset contains approximately 15488 RGB-IR image pairs with valid 2D bounding boxes. Since no additional filtering is needed, we directly annotate all subjects using HMR 2.0. The resulting **LLVIP-Mesh** dataset consists of 15488 RGB-infrared subject pairs, each associated with a shared SMPL mesh label. These are split into 75% for training and 25% for evaluation.

Chapter 4

Multimodal People Pose Detection: Approach

Our mesh recovery system follows a transformer-based, modality-agnostic pipeline inspired by HMR 2.0 [8] and MMPedestron [35]. We unify RGB, infrared, and LiDAR modalities for robust full-body mesh recovery in diverse outdoor conditions. An overview is shown in Figure 4.1.

4.1 Multimodal Input

2D Representation of Modalities All modalities are represented as 2D images. LiDAR point clouds are converted to pseudo-depth images by projecting points into the image plane using known calibration parameters and encoding depth via intensity or disparity-like encodings. The infrared modality is directly used as grayscale images. RGB frames are used as standard 3-channel inputs.

Person Detection We employ a person detector to extract bounding boxes from each modality. During training, we use ground-truth bounding boxes. For RGB modality, we utilize ViTDet [19] as our base detector, while for other modalities like LiDAR and IR, we use the MMPedestron detector [35].

Modality Unification To enable joint reasoning over multiple modalities, we adopt the modality unifier architecture proposed in MMPedestron [35]. Each modality stream is passed through a shared patch embedding module, converting modality-

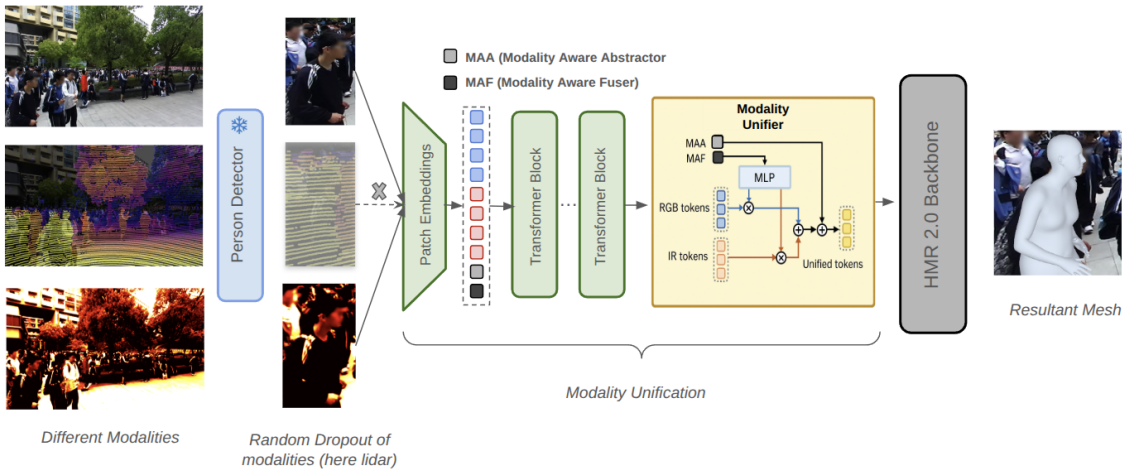

Figure 4.1: Overview of our multimodal mesh recovery pipeline. Given RGB, LiDAR, and infrared inputs, person detection is first applied independently to each modality. Cropped person-centric regions are passed through shared patch embedding and transformer blocks, with learnable MAA (Modality-Aware Abstractor) and MAF (Modality-Aware Fuser) tokens facilitating cross-modal attention. The Modality Unifier aggregates features using modality-specific confidence weights predicted by an MLP. During training, random dropout of modalities (e.g., LiDAR here) encourages robustness to missing or degraded inputs. The fused tokens are fed into the HMR 2.0 backbone to regress the SMPL mesh parameters, yielding the final 3D body mesh.

specific images into a token sequence. These are followed by stacked transformer blocks that refine the visual representation.

To enable efficient fusion of these heterogeneous tokens, we prepend two learnable tokens per sample: the *Modality-Aware Abstractor (MAA)* and *Modality-Aware Fuser (MAF)*. These tokens guide the transformer to identify relevant modality information and to assess the importance of each modality. The **MAA** token aims to assess the importance of each modality in the fusion process. The **MAF** token aims to collect the domain knowledge related to input modalities.

A multilayer perceptron (MLP) uses the MAF token to predict confidence weights $\mathbf{c}$ for each modality. These weights, combined with modality validity indicators w_i, allow for robust aggregation:

$$\mathbf{X}_{\text{uni}} = \frac{\sum_{i=1}^{m} w_i c_i \mathbf{X}_i}{\sum_{i=1}^{m} w_i} + \mathbf{x}_{\text{MAA}} \tag{4.1}$$

where m is the number of modalities, c_i and $\mathbf{X}_i$ denote the confidence and token sequence respectively, $\mathbf{x}_{\text{MAA}}$ is the feature of MAA token, and w_i indicates whether the modality is valid (1) or missing (0). The result is a unified token representation $\mathbf{X}_{\text{uni}}$ that fuses complementary features across modalities. This aggregation allows our model to flexibly reason over any subset of available modalities, enabling robustness under sensor failure or occlusion.

4.2 SMPL Regression with HMR 2.0

The unified tokens $\mathbf{X}_{\text{uni}}$ are passed into the frozen HMR 2.0 [8] transformer head to obtain SMPL parameters $\Theta = [\boldsymbol{\theta}, \boldsymbol{\beta}, \pi]$, where $\boldsymbol{\theta} \in \mathbb{R}^{72}$ represents body pose, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ denotes body shape, and π is the weak-perspective camera translation.

Loss Functions

We supervise the model using a combination of pose, keypoint and adversarial losses similar to the setting in HMR 2.0 [8]:

- **SMPL Parameter Loss:** Applied to all samples for supervison over SMPL parameter predictions:

$$\mathcal{L}_{\text{smpl}} = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \tag{4.2}$$

- **3D Keypoint Loss:** Supervised with L1 loss between predicted and GT 3D keypoints:

$$\mathcal{L}_{\text{kp3D}} = \|\mathbf{X} - \mathbf{X}^*\|_1 \tag{4.3}$$

- **2D Keypoint Reprojection Loss:** Supervised using L1 loss between reprojected and GT 2D joints:

$$\mathcal{L}_{\text{kp2D}} = \|\pi(\mathbf{X}) - \mathbf{x}^*\|_1 \tag{4.4}$$

- **Adversarial Loss:** A discriminator D_k is applied over SMPL joints to regularize plausible human pose predictions:

$$\mathcal{L}_{\text{adv}} = \sum_k (D_k(\boldsymbol{\theta}_k, \boldsymbol{\beta}) - 1)^2 \tag{4.5}$$

4.3 Training Strategy

We design our own two-phase training strategy aimed at stabilizing multimodal learning and improving generalization. In the first phase, we freeze the HMR 2.0 backbone and train only the modality unifier and transformer layers. This allows the earlier layers to specialize in modality alignment without prematurely disrupting the pretrained SMPL regressor. Freezing the regression head also prevents noisy gradients from early fusion layers from degrading learned priors.

We apply a random modality dropout during training with a probability p for each input modality. This encourages the model to robustly handle missing or corrupted sensors. When a modality is dropped, we replace it with empty padded tokens and mask it within attention layers.

After a fixed number of epochs, we unfreeze the HMR 2.0 layers and continue training the entire pipeline end-to-end. This second phase allows the entire network to jointly optimize modality fusion and mesh prediction. This fine-tuning stage helps the regressor adapt better to fused inputs and ultimately boosts performance across unseen combinations of modalities.

Chapter 5

Experiments & Results

We conduct a thorough evaluation of our proposed multimodal mesh recovery framework across multiple challenging datasets. Our experiments aim to validate the efficacy of our modality unification strategy and assess performance improvements brought by leveraging complementary modalities (RGB, LiDAR, IR). We also explore model robustness under modality dropout and highlight generalization across diverse real-world conditions.

5.1 Implementation Details

Our training is performed entirely in the multimodal setting, leveraging the **STCrowd-mesh** and **LLVIP-mesh** datasets. We adopt a two-phase training strategy: we first train only the initial layers of our architecture (including the transformer blocks and modality unifier) for 60 epochs. Subsequently, we unfreeze the deeper layers of the HMR 2.0 backbone and continue training for an additional 30 epochs. During training, we employ *modality dropout* with a probability $p = 0.4$ to encourage robustness under missing or corrupted modalities.

5.2 Quantitative Results

We evaluate our multimodal mesh-recovery framework using two standard metrics:

- MPJPE (*Mean Per-Joint Position Error*): the average Euclidean distance (in millimetres) between predicted and ground-truth 3-D joint positions without alignment; it reflects absolute 3-D localisation accuracy.
- PA-MPJPE (*Procrustes-Aligned MPJPE*): the same distance computed after rigidly aligning predictions to ground-truth; it highlights pose quality while factoring out global translation, rotation, and scale.

Baseline methods. We compare against three image-based, unimodal mesh-recovery models:

- HMR [12]: a weakly-supervised ResNet-based regressor trained with adversarial priors on SMPL parameters.
- SPIN [16]: improves HMR via a self-optimisation loop and iterative refinement.
- HMR 2.0 [8]: a ViT-based successor that offers stronger generalisation and higher resolution.

Each baseline was fine-tuned on the corresponding modality-specific training subset (e.g., RGB-only or IR-only) of the STCrowd-Mesh and LLVIP-Mesh datasets for fair comparison. We omit LiDAR-only results because the pretrained person detector [35] (used to extract person crops for inference) fails to reliably detect people using LiDAR-projected images alone, making standalone LiDAR settings infeasible.

Table 5.1: MPJPE (mm) on STCrowd-Mesh and LLVIP-Mesh test set. Lower is better.

Method	**STCrowd-Mesh**		**LLVIP-Mesh**		
	RGB only	RGB+LiDAR	RGB only	IR only	RGB+IR
HMR	116.7	–	121.3	153.6	–
SPIN	102.3	–	113.1	145.3	–
HMR 2.0	87.4	–	93.4	113.2	–
Multimodal HMR (Ours)	**86.9**	**75.8**	**93.2**	**107.5**	**91.5**

Analysis. Several insights emerge from Tables 5.1 and 5.2.

(i) LiDAR improves absolute 3D positioning. On STCrowd-Mesh, integrating LiDAR with RGB leads to a sharp MPJPE drop from 86.9mm to 75.8mm (–14.3%). The corresponding PA-MPJPE gain is smaller (–5.3mm), implying LiDAR

Table 5.2: PA-MPJPE (mm) on STCrowd-Mesh and LLVIP-Mesh test set. Lower is better.

Method	**STCrowd-Mesh**		**LLVIP-Mesh**		
	RGB only	RGB+LiDAR	RGB only	IR only	RGB+IR
HMR	97.3	–	102.4	141.5	–
SPIN	71.2	–	86.4	124.3	–
HMR 2.0	63.5	–	68.4	98.7	–
Multimodal HMR (Ours)	**63.1**	**57.8**	**68.3**	**94.5**	**66.9**

mainly corrects for global misalignment (scale, distance) rather than fine-grained pose structure. This validates LiDAR's value in improving absolute 3D mesh placement under diverse crowd densities.

(ii) Infrared aids low-light perception but has weaker 3D signal. LLVIP-Mesh results show RGB-only MPJPE at 93.2mm, which drops modestly to 91.5mm when IR is added. Interestingly, IR-only performance is substantially worse for all methods (e.g., 107.5mm MPJPE for our method), indicating that while IR contributes contrast in low-light conditions, it lacks explicit geometric depth. The fusion of RGB and IR, however, still benefits both metrics consistently, e.g., PA-MPJPE improves from 68.3mm (RGB) to 66.9mm (RGB+IR).

(iii) Classic baselines fail to close the domain gap. Across both datasets, HMR and SPIN trail HMR 2.0 by wide margins, even when fine-tuned. This is evident especially in IR-only settings, where their MPJPEs exceed 140mm on LLVIP. These results indicate that transformer-based architectures like HMR 2.0 handle domain shifts better, especially thermal and nighttime scenes, but still plateau under unimodal input.

(iv) Fusion offers complementary robustness. HMR 2.0 Multimodal consistently improves upon the RGB-only HMR 2.0 baseline without needing architectural modifications. For instance, MPJPE improves by 11.6mm (STCrowd) and 1.7mm (LLVIP). The gain is larger when adding LiDAR than IR, highlighting the superior spatial grounding LiDAR offers. Yet the improvement from RGB+IR remains valuable, especially when LiDAR is unavailable.

(v) Dropout-trained fusion models preserve unimodal performance. Despite being trained with dropout (p=0.4), our multimodal HMR retains strong unimodal results: e.g., 86.9/63.1mm on STCrowd-Mesh with RGB only, outperforming

or matching the HMR 2.0 RGB-only baseline. This validates our unifier's ability to fuse while remaining robust to missing modalities at inference time.

In summary, modality fusion leads to significant benefits in both absolute accuracy (MPJPE) and pose structure (PA-MPJPE), especially in scenes requiring spatial grounding (via LiDAR) or low-light robustness (via infrared). The unified model not only improves performance but also retains flexibility in degraded-sensor conditions—crucial for real-world field deployments.

5.3 Qualitative Results

We visualize mesh predictions from our multimodal HMR model on both STCrowd-Mesh and LLVIP-Mesh datasets to assess qualitative improvements over unimodal baselines.

STCrowd-Mesh. In Fig. 5.1, we present paired RGB and LiDAR frames (left), RGB-only predictions (center), and RGB+LiDAR predictions (right). The multimodal predictions show significantly improved alignment with body contours and reduced artifacts in crowded and depth-challenging regions, especially for occluded or distant subjects. The additional LiDAR signal helps disambiguate overlapping poses and improves overall spatial grounding.

LLVIP-Mesh. Fig. 5.2–5.4 show predictions for scenes captured under low-light conditions using paired RGB and infrared inputs. RGB-only models often fail in such scenarios due to illumination loss or occlusions. Our RGB+IR model recovers more accurate and complete meshes, particularly for pedestrians under shadows (Fig. 5.2), overlapping subjects (Fig. 5.3), and cluttered scenes with multiple interactions (Fig. 5.4). These results validate the effectiveness of infrared sensing in supplementing RGB cues under adverse visual conditions.

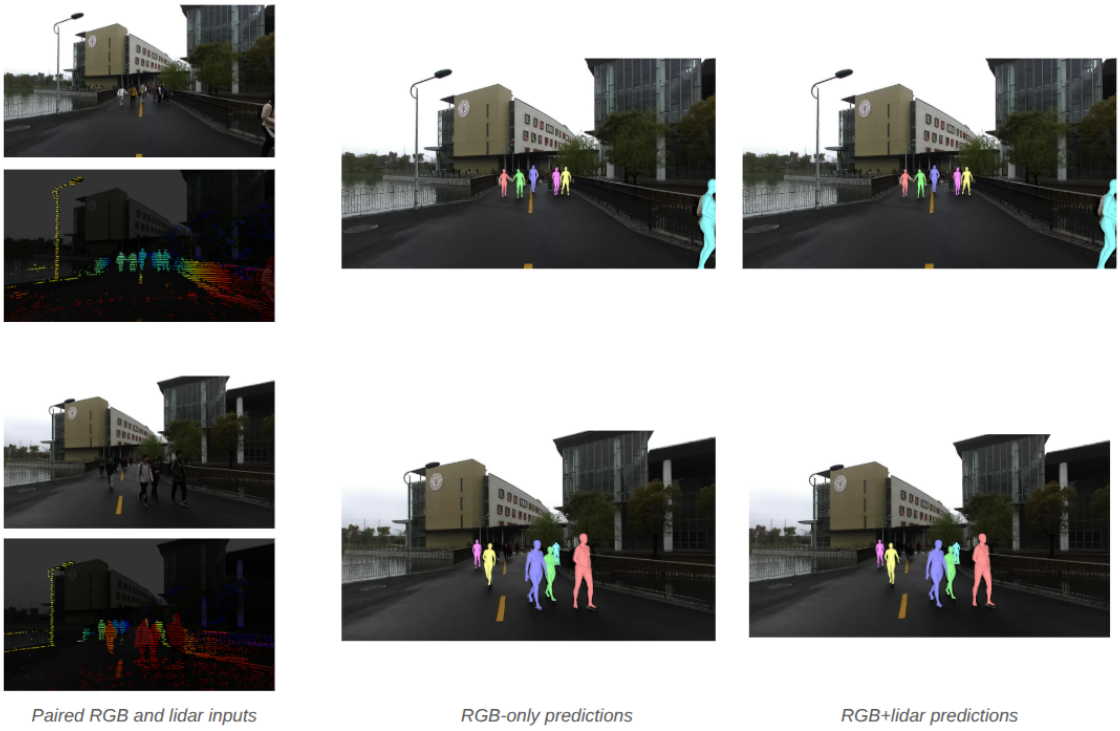

Figure 5.1: Qualitative Results on STCrowd-Mesh dataset

Figure 5.2: RGB+IR fusion improves pose recovery for occluded and poorly lit pedestrians.

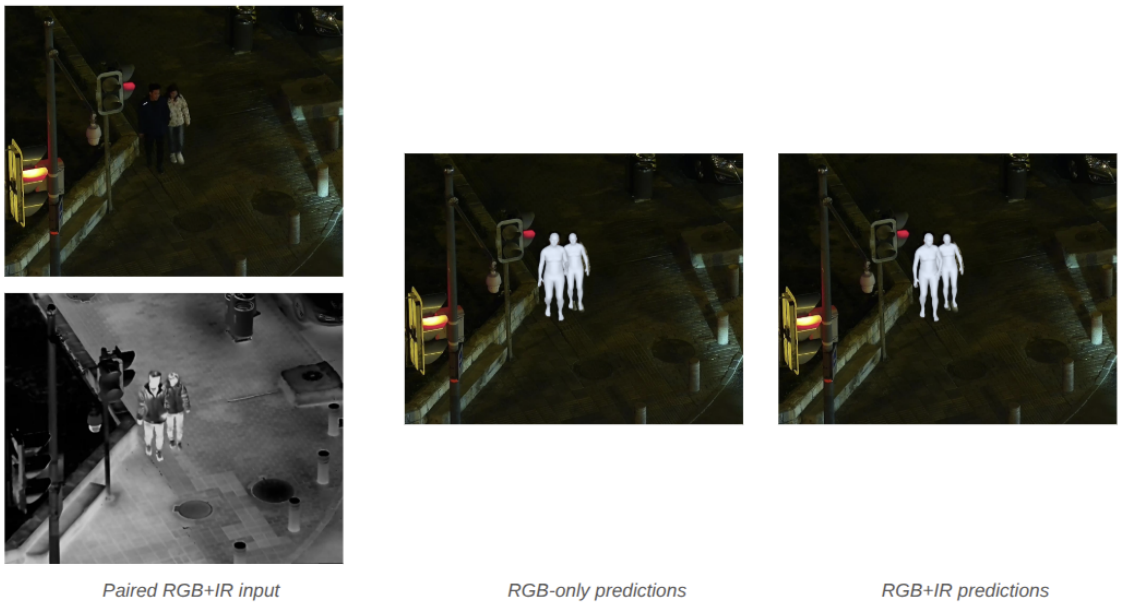

Figure 5.3: RGB+IR model separates overlapping subjects better than RGB-only predictions.

Figure 5.4: RGB+IR predictions yield cleaner and more complete multi-person meshes.

Chapter 6

Results on Real-World data & Alertness Detection

In this section, we evaluate our method on real-world multimodal data collected during the Year-2 workshop of the DARPA Triage Challenge, held in March 2025. Unlike curated benchmark datasets, this field data features outdoor conditions, uneven illumination, occlusions, and practical deployment noise, posing significant generalization challenges. We demonstrate the qualitative performance of our mesh recovery model and further use the predicted meshes to infer motor alertness of casualties.

6.1 Qualitative Mesh Recovery on DARPA Data

We apply our multimodal mesh recovery pipeline to the RGB and LiDAR streams recorded during the field tests. While our full pipeline supports RGB, LiDAR, and Infrared modalities, the IR camera failed during the workshop runs, so only RGB and LiDAR were used in this evaluation. A visualization of the workshop data can be seen in Fig 6.1.

To ensure robustness and downstream application, we limit the model to predict at most one mesh per frame. Specifically, we select the most confident person bounding box from the upstream detector module and discard other candidates. This avoids failure cases in crowded scenes and aligns with our downstream focus on assessing

Figure 6.1: Sample snapshot from DARPA workshop Year-2 data. Top: RGB frame captured by the Spot arm camera. Bottom: LiDAR visualization in the global frame with detected robot links. The person sitting in front is selected via bounding box confidence and serves as the primary target for mesh recovery.

physiological signs.

For LiDAR–RGB fusion, we first bring the LiDAR point cloud into the RGB camera's coordinate frame. Since the LiDAR was mounted on the rear of the Spot robot and the RGB camera on the arm, we apply both a static transform (from calibration) and dynamic TFs published by the robot during runtime. Once the LiDAR points are transformed into the RGB frame, we project them using the camera intrinsics to generate a sparse depth image aligned with the RGB view. This fused pair is then fed into our modality-unified HMR network for inference. We see some qualitative results in Fig 6.2.

Figure 6.2: Multimodal mesh overlay results across four scenes. Each row shows RGB+LiDAR projections and mesh reconstruction.

6.2 Motor Alertness Assessment

Given that our real-world deployment setting lacks labeled ground truth for pose or mesh annotations, directly validating the quality of mesh recovery becomes challenging. To indirectly evaluate the utility and reliability of our recovered meshes, we perform a downstream task: assessing the motor alertness of casualties.

According to the DARPA Triage Challenge (DTC) guidelines, motor alertness is categorized into three levels—**normal**, **abnormal**, and **absent**—based on the presence and intensity of voluntary limb movements.

To perform this analysis, we leverage the recovered SMPL mesh from our multi-modal HMR pipeline. Specifically, we focus on the `body_pose` parameter from the SMPL model, which encodes the relative 3D rotations of 23 joints (excluding the root pelvis) in the form of 3×3 rotation matrices. These matrices represent the rotation of each joint relative to its parent in the kinematic chain and are inherently invariant to global translation or rotation of the camera — making them ideal for motion estimation in mobile settings such as a moving robot platform.

We subsample the sensor stream to 1 frame per second (1 FPS) to capture meaningful temporal differences while filtering out high-frequency jitter. For each joint of interest, specifically the ankles, knees, elbows, and wrists, we compute the angular difference in degrees between consecutive frames using the geodesic distance on $SO(3)$.

We then define the following classification logic:

- **Normal:** At least 3 limb movements with angular change $> 15°$
- **Abnormal:** At least 1 movement $> 15°$, or 3+ movements between $10°$ and $15°$
- **Absent:** All movement below $10°$, or fewer than 3 mild movements

This scheme ensures high sensitivity to limb movements while avoiding spurious noise-induced detections. Algorithm 2 outlines the full procedure.

We present several qualitative examples in Fig 6.3–6.6, showcasing frame-wise mesh predictions and corresponding joint rotations used for motor alertness assessment. *Note that although only the RGB frames are shown for visualization clarity, the underlying mesh predictions are generated using our RGB+LiDAR pipeline.*

Algorithm 2 Motor Alertness Detection from SMPL Pose

Input: Sequence of frames $\{F_t\}$ at 1 FPS with SMPL `body_pose` parameters
Initialize: counters $C_{>15} \leftarrow 0$, $C_{10-15} \leftarrow 0$
for each joint j in {ankle, knee, elbow, wrist} **do**
    for each consecutive pair (F_t, F_{t+1}) **do**
        Compute relative rotation: $R_{\text{rel}} = R^j_{t+1} \cdot (R^j_t)^{-1}$
        Compute angular change θ (in degrees) from R_{rel}
        if $\theta > 15$ **then**
            $C_{>15} \leftarrow C_{>15} + 1$
        else if $10 \leq \theta \leq 15$ **then**
            $C_{10-15} \leftarrow C_{10-15} + 1$
        end if
    end for
end for
if $C_{>15} \geq 3$ **then**
    Classify as: Normal
else if $C_{>15} \geq 1$ **or** $C_{10-15} \geq 3$ **then**
    Classify as: Abnormal
else
    Classify as: Absent
end if

Figure 6.3: Example 1 (Abnormal): Casualty standing with limited movement. A single joint exceeds the 15° threshold while others remain low. This is correctly classified as **abnormal**.

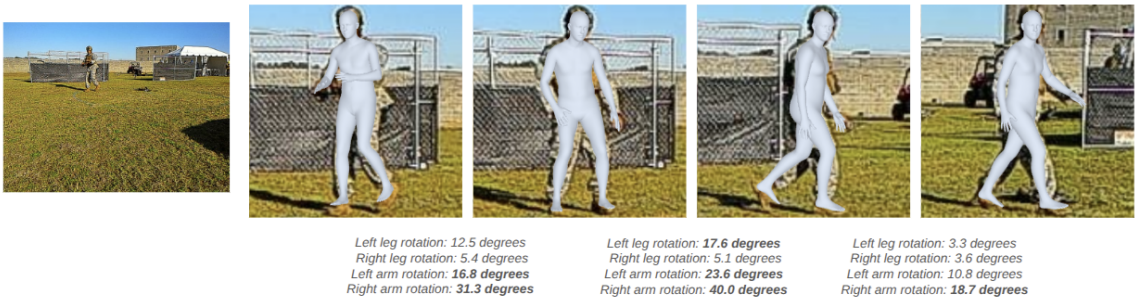

Figure 6.4: Example 2 (Normal): Casualty walking across the field. Strong, repeated limb motions across multiple joints with values well beyond 15° result in a confident classification of **normal** motor alertness.

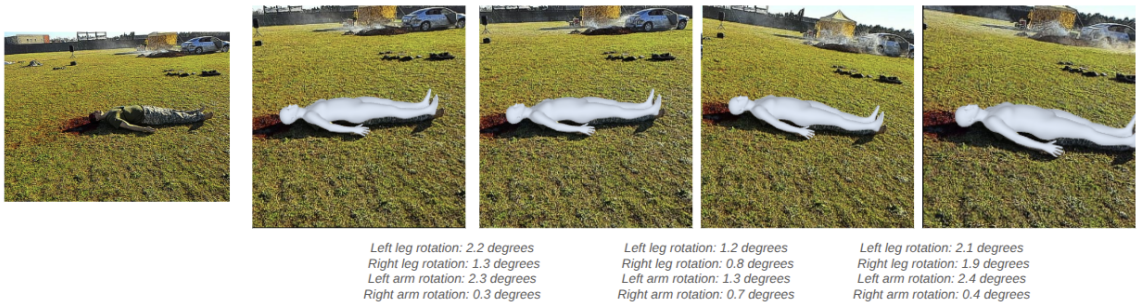

Figure 6.5: Example 3 (Absent): Casualty lying on the ground with negligible joint rotation. All movements stay below 3°, leading to an accurate classification of **absent** alertness.

How good was our heuristic for predicting motor alertness on DTC Workshop data?

During the DARPA Triage Challenge Year-2 Workshop, we conducted two real-world runs: one during daylight and another at night. Due to hardware complications in the night-time run (including IR and LiDAR stream interruptions), we restrict our evaluation to the day-time run, which consisted of 15 casualties.

Each casualty was manually annotated for ground-truth motor alertness category using visual inspection and team-provided metadata. The ground-truth distribution included: 9 cases labeled as Normal, 1 as Abnormal, and 5 as Absent.

Our heuristic-based prediction method successfully identified all 9 **Normal** and the single **Abnormal** case. However, it misclassified 3 out of the 5 **Absent** cases as **Normal**. These false positives are primarily attributed to mesh prediction errors

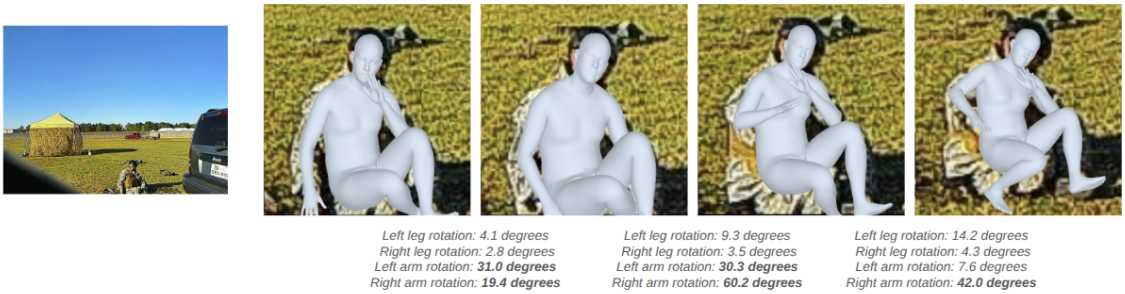

Figure 6.6: Example 4 (Normal): Seated casualty exhibiting significant voluntary motion in both arms and legs, with several joint rotations exceeding 30°. This case is correctly classified as **normal** motor alertness.

such as incorrect joint alignment, unusual body orientation, or occlusion-induced distortions. In some instances, even minimal arm or leg movement, misestimated due to poor mesh fitting or oblique camera angles, triggered angular spikes above the defined threshold, leading to an erroneous alert classification. Some representative failure cases are shown in Fig 6.7

We summarize the results using the confusion matrix shown in Table 6.1.

Table 6.1: Confusion matrix for motor alertness classification on 15 casualties from DARPA Workshop (Day run).

Ground Truth →	**Normal**	**Abnormal**	**Absent**
Predicted: Normal	9	0	3
Predicted: Abnormal	0	1	0
Predicted: Absent	0	0	2

Despite a few false positives in the **Absent** category, the overall performance indicates that our SMPL-based heuristic is well-calibrated for detecting meaningful motor activity, especially under constrained and noisy field conditions.

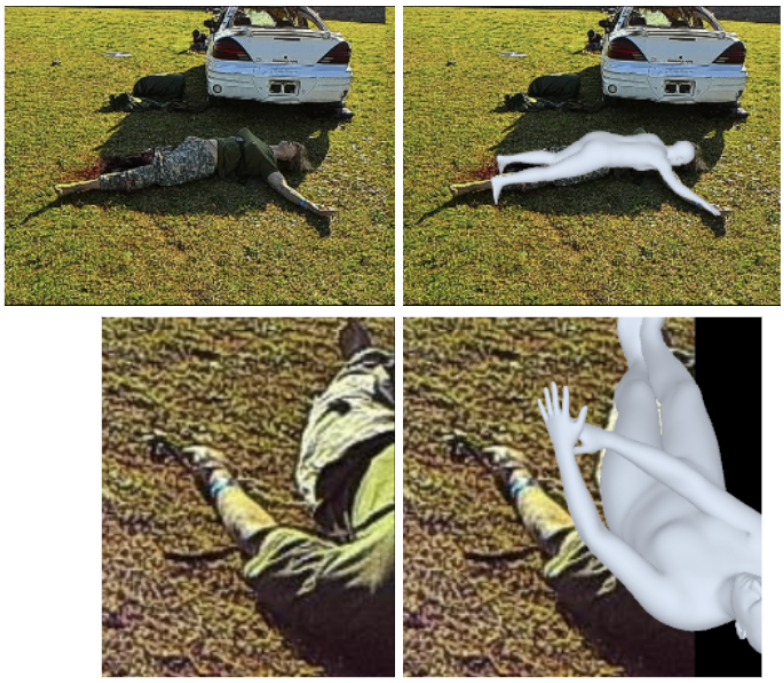

Figure 6.7: Representative failure case where our algorithm incorrectly classified a casualty as **Normal** despite the ground truth being **Absent**. The mesh overlay exhibits incorrect hand orientation and inflated limb rotation, likely due to poor camera angle and self-occlusion/occlusion from nearby objects. These artifacts resulted in large estimated angular changes across frames, falsely triggering our motor alertness heuristic.

Chapter 7

Future Work & Conclusion

7.1 Future Work

1. **Model optimization for real-time deployment.** The current pipeline runs at around 1 FPS on an NVIDIA Jetson Orin. We plan to reduce model size, use lighter transformer variants, and convert components to TensorRT with INT8 quantization to target 5–10 FPS for deployment.
2. **Improving robustness under occlusion.** Some failure cases are due to occlusions and poor viewpoints. Adding temporal information, occlusion-aware losses, and using LiDAR to fill missing parts can improve mesh stability.
3. **Training with field-collected data.** To handle domain shifts, we will label a subset of DARPA field recordings to create an RGB+LiDAR+IR dataset. Training with this data should reduce prediction errors and help support downstream tasks like alertness detection.

7.2 Conclusion

This thesis set out to answer a simple question: *Can dense 3-D human meshes be recovered reliably in the wild when a robot has access to multiple but intermittently available sensors?* Our answer is an emphatic "yes." We introduced the first unified framework that ingests any subset of RGB, LiDAR, and infrared inputs and

regresses a full-body SMPL mesh through a lightweight transformer equipped with a modality-aware unifier. To supervise such fusion we produced two new benchmarks, STCrowd-Mesh and LLVIP-Mesh, each containing per-person SMPL annotations aligned with multimodal imagery. Extensive experiments showed that fusing LiDAR resolves absolute-depth ambiguity while infrared restores pose fidelity in low light, yielding consistent improvements over strong RGB-only baselines on both benchmarks. Crucially, the same model generalises to DARPA field data, where its meshes enable downstream motor-alertness assessment, an essential physiological sign in mass-casualty triage. By demonstrating that multimodal fusion, lightweight architecture, and task-oriented evaluation can coexist in a single deployable system, this work lays a practical foundation for next-generation robots capable of stand-off, autonomous casualty assessment.

Bibliography

[1] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 729–746, 2018. 2.2

[2] Apple. Apple vision documentation. https://developer.apple.com/documentation/vision. 5

[3] Anjun Chen, Xiangyu Wang, Zhi Xu, Kun Shi, Yan Qin, Yuchi Huo, Jiming Chen, and Qi Ye. Adaptivefusion: Adaptive multi-modal multi-view fusion for 3d human body reconstruction. *arXiv preprint arXiv:2409.04851*, 2024. 2.2

[4] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 1

[5] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19608–19617, 2022. 2.2, 3.1

[6] DARPA. Darpa triage challenge. https://triagechallenge.darpa.mil/, 2023. 1, 1.1

[7] Bohao Fan, Wenzhao Zheng, Jianjiang Feng, and Jie Zhou. Lidar-hmr: 3d human mesh recovery from lidar. *arXiv preprint arXiv:2311.11971*, 2023. 1.2, 2.2

[8] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 1.2, 2.2, 3.1, 3.2, 4, 4.2, 5.2

[9] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2.2

[10] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. 3

[11] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2.2

[12] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1.2, 2.2, 5.2

[13] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 2, 4

[14] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1715–1725, 2022. 1.2

[15] Bugra Kocabas, Chun-Hao P Huang, and Otmar Hilliges. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 2.2

[16] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 1, 5.2

[17] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 1.2, 2.2

[18] Changzhi Li, Julie Cummings, Jeffrey Lam, Eric Graves, and Wenhsing Wu. Radar remote monitoring of vital signs. *IEEE Microwave magazine*, 10(1):47–56, 2009. 1

[19] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 4.1

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 4

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying

dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 1

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2.1, 2.2

[23] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981. 3

[24] Christian Martin, Gaurav Pandey, Oliver Saurer, and Sebastian Scherer. Jrdb: A socially-aware robot dataset for navigation in human environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 2.2

[25] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *ACM Transactions on Graphics (TOG)*, volume 36, pages 1–14. ACM, 2017. 2.2

[26] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7025–7034, 2017. 2.2

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 5

[28] Revanth Krishna Senthilkumaran, Mridu Prashanth, Hrishikesh Viswanath, Sathvika Kotha, Kshitij Tiwari, and Aniket Bera. Artemis: Ai-driven robotic triage labeling and emergency medical information system. *arXiv preprint arXiv:2309.08865*, 2023. 1

[29] Xiao Sun, Bin Xiao, Fangyin Liang, Yichen Wei, and Xiaoou Tang. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 2.2

[30] SALT Mass Casualty Triage. Concept endorsed by the american college of emergency physicians, american college of surgeons committee on trauma, american trauma society, national association of ems physicians, national disaster life support education consortium, and state and territorial injury prevention directors association. *Disaster med public health prep*, 2(4):245–246, 2008. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1.2

[32] Xiaofeng Wan, Xi Li, Zhuo Zhou, Panfei Wang, and Lei Zhao. Llvip: A visible-infrared paired dataset for low-light pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3496–3504, 2021. 2.2, 3.2

[33] Feng Xu, Yuxiao Zhang, and Li Cheng. Monocular neural performance capture with real-time hand and face tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3363, 2022. 2.2

[34] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 391–406, 2022. 1.2, 2.2

[35] Yi Zhang, Wang Zeng, Sheng Jin, Chen Qian, Ping Luo, and Wentao Liu. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In *European Conference on Computer Vision*, pages 430–448. Springer, 2024. 1.2, 2.2, 4, 4.1, 5.2

[36] Yuxiao Zhang, Yifan Wang, Feng Xu, and Li Cheng. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11659–11668, 2021. 2.2

[37] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10113–10122, 2019. 1.2, 2.2

[38] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12234–12243, 2021. 2.2

[39] Hao Zhu, Xingyu Wang, Zerong Zheng, Ziwei Liu, and Wei Liu. Kbody: Towards accurate and realistic 3d human body recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2925–2935, 2023. 2.2