

# What Happened 3 Seconds Ago? Inferring the Past with Thermal Imaging

Zitian Tang<sup>1</sup> Wenjie Ye<sup>1</sup> Wei-Chiu Ma<sup>2</sup> Hang Zhao<sup>1,3</sup>

<sup>1</sup>IIS, Tsinghua University <sup>2</sup>CSAIL, MIT <sup>3</sup>Shanghai Qi Zhi Institute

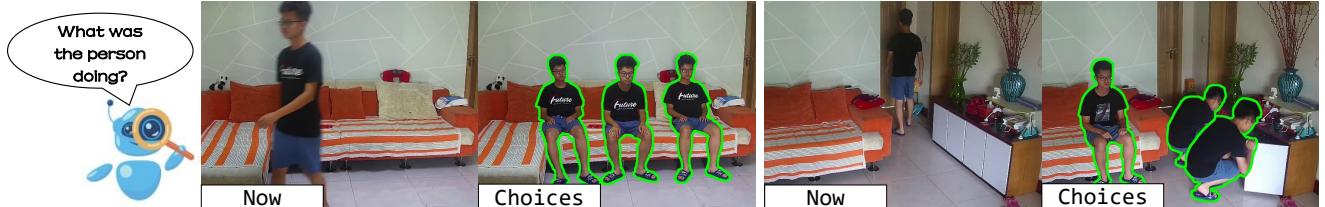


Figure 1. Can you tell what the person was doing 3 seconds ago? Inferring past human motions suffers from huge uncertainty. See the answers in Fig. 2.

## Abstract

Inferring past human motion from RGB images is challenging due to the inherent uncertainty of the prediction problem. Thermal images, on the other hand, encode traces of past human-object interactions left in the environment via thermal radiation measurement. Based on this observation, we collect the first RGB-Thermal dataset for human motion analysis, dubbed Thermal-IM. Then we develop a three-stage neural network model for accurate past human pose estimation. Comprehensive experiments show that thermal cues significantly reduce the ambiguities of this task, and the proposed model achieves remarkable performance. The dataset is available at <https://github.com/ZitianTang/Thermal-IM>.

## 1. Introduction

Imagine we have a robot assistant at home. When it comes to offering help, it may wonder what we did in the past. For instance, it wonders which cups were used, then cleans them. Or it can better predict our future actions once the past is known. But how can it know this? Consider the images in Fig. 1. Can you tell what happened 3 seconds ago? An image contains a wealth of information. The robot may extract geometric and semantic cues, infer the affordance of the scene, and imagine how humans would interact and fit in the environment. Therefore, in the left image, it can confidently deduce that the person was sitting on the couch; however, it is not sure where. Similarly, it can imagine many possibilities in the right image but cannot be

certain. Indeed, given a single RGB image, the problem is inherently ill-posed.

In this paper, we investigate the use of a novel sensor modality, thermal data, for past human behavior analysis. Thermal images are typically captured by infrared cameras, with their pixel values representing the temperature at the corresponding locations in the scene. As heat transfer occurs whenever there is contact or interaction between human bodies and their environment, thermal images serve as strong indicators of where and what has happened. Consider the thermal images in Fig. 2. With thermal images, we can instantly determine where the person was sitting. This is because the objects they contacted were heated, leaving behind bright marks. If a robot assistant is equipped with a thermal camera, it can more effectively infer the past and provide better assistance. Otherwise, we may need a camera installed in every room and keep them operational throughout the day.

With these motivations in mind, we propose to develop a system that, given an indoor thermal image with a human in it, generates several possible poses of the person  $N$  ( $N = 3$ ) seconds ago. To achieve this goal, we first collect a Thermal Indoor Motion dataset (Thermal-IM) composed of RGB, thermal, and depth videos of indoor human motion with estimated human poses. In each video, the actor performs various indoor movements (*e.g.*, walking, sitting, kneeling) and interacts with different objects (*e.g.*, couch, chair, cabinet, table) in a room. Then we design a novel, interpretable model for past human pose estimation. The model consists of three stages: the first stage proposes where the human might have been 3 seconds ago, leveraging the most discernible information in thermal images. The second stage infers what action the human was performing. Finally, the

Corresponding to: hangzhao@mail.tsinghua.edu.cn.



Figure 2. **Thermal images to the rescue:** Thermal images encode traces of past human-object interactions, which can help us infer past human behavior and understand objects’ affordance. In this work, we focus on estimating human body poses a few seconds ago.

third stage synthesizes an exact pose.

Experiments show that our method managed to generate plausible past poses based on the locations and shapes of thermal cues. These results are more accurate than the RGB-only counterparts, thanks to the reduced uncertainty of past human movements. Furthermore, our model automatically and implicitly discovers the correlation between thermal mark intensity and time.

The contributions of this work are the following:

- We make the first attempt at a novel past human motion estimation task by exploiting thermal footprints.
- We construct the Thermal-IM dataset, which contains synchronized RGB-Thermal and RGB-Depth videos of indoor human motion.
- We propose an effective three-stage model to infer past human motion from thermal images.

## 2. Related Works

**Thermal imaging in machine learning:** A thermal camera captures the far-infrared radiation emitted by any object (known as black body radiation), which is robust in varied illumination conditions. This property helps improve the performances of semantic segmentation and tracking systems significantly in urban scenes. Ha *et al.* [7] releases the first RGB-Thermal image segmentation dataset and verifies the benefit of incorporating thermal images, especially in night-time scenes. Subsequently, plenty of datasets and models about semantic segmentation [5, 15, 26–29, 39, 43] and tracking [13, 14, 16–18, 20, 22, 31, 36–38, 40, 41] are proposed. These works propose various model structures to investigate the best way to fuse RGB and thermal features.

There are a few works making use of other characteristics of thermal imaging. Based on the property that most glass is opaque to infrared light, Huo *et al.* [10] recognizes glass based on RGB-Thermal image pairs. Their method significantly outperforms the RGB counterpart. As hand-object contact can leave apparent marks on objects, Brahmbhatt *et al.* [1] proposes a dataset recording contact maps for human grasps. They use a generative adversarial network model to predict how humans grasp a given object.

Their results reveal various aspects affecting human grasping behavior.

Our work is the first study on the relationship between thermal imaging and indoor human motion. Solving this task requires a deep understanding of how a thermal mark’s location, shape, and intensity relate to human behavior.

**Human motion prediction:** Human motion prediction aims to predict the 2D or 3D future poses, given one’s pose history. A wide range of techniques are used to tackle this task regardless of the scene context, such as graphical models [2], recurrent neural networks [6, 12, 24, 34], graph convolutional networks [23, 42], and temporal convolutional networks [9, 19]. Moreover, [4, 30, 33, 35] consider pose history together with image context to predict future poses. However, these methods only concern a local patch around the human rather than the whole background scene.

To predict 3D future human motion, Cao *et al.* [3] proposes a method composed of three modules, GoalNet, PathNet, and PoseNet. Given an image and pose history, a VAE-based GoalNet predicts a few possible human torso positions in the future. Afterward, PathNet, an Hourglass model [25], generates a route from the current human position to each predicted future one. Finally, Transformer-based PoseNet synthesizes a pose at each point along a route. It is worth noticing that the last two modules are deterministic, and the PoseNet is not provided with scene context. Wang [32] develops a GAN-based model to generate plausible future human motion in a given image. Their method comprises two stages. The first stage generates motion trajectories conditioned on the scene, while the second stage approximates the pose distribution given the scene and the trajectory.

Our work focuses on inferring human motion in the past rather than the future. These two tasks are similar regardless of the direction of the time flow. Hence, the works above inspire our model design. While these works take a historical pose sequence into account, our work infers the past only according to a single frame and a static human pose in it.

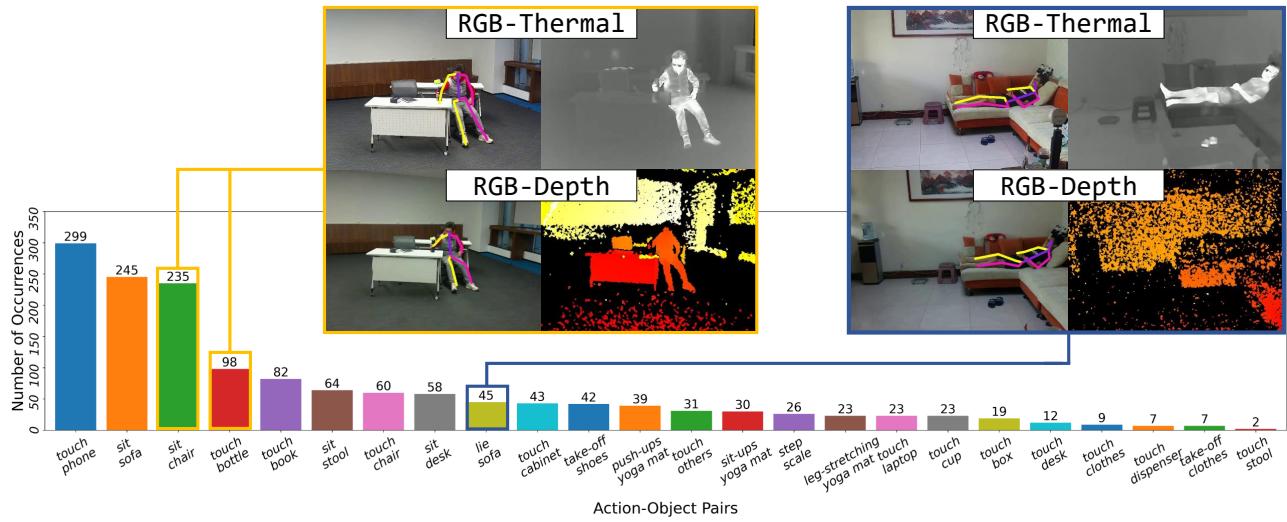


Figure 3. **Statistics of our Thermal Indoor Motion (Thermal-IM) dataset.** Our dataset comprises 783 synchronized RGB-Thermal and RGB- Depth video clips, with 24 types of human-object interactions. We also provide estimated 2D and 3D human poses in each frame.

### 3. Thermal Indoor Motion dataset

Previous RGB-Thermal image datasets are mostly about urban scenes rather than indoor scenes. Moreover, none of them focus on human-scene interactions. This motivates us to collect the Thermal Indoor Motion (Thermal-IM) dataset. It contains synchronized RGB-Thermal and RGB- Depth videos with estimated human poses about a person moving and interacting with objects in indoor scenes.

We collect the Thermal-IM dataset using an RGB- Thermal camera (Hikvision DS-2TD4237T-10) and an RGB-Depth camera (Intel RealSense L515). The resolution of the RGB-Thermal camera is  $1080 \times 1920$  for the RGB channel and  $288 \times 384$  for the thermal channel. That of the RGB-Depth camera is  $480 \times 640$ . These two cameras record videos simultaneously, and their extrinsic parameters are estimated.

During data collection, an actor performs several preset actions, leaving cues in thermal images. In total, we collect 783 video clips,  $\sim 560k$  frames in 15 FPS ( $\sim 10.4$  hours). 74% of the videos involve one actor and two different rooms, which is the main part we use to develop our method. The rest is a held-out part for the generalization test in Sec. 5.5, engaging one another actor or room. We record the videos from various viewing angles and rearrange the objects in the rooms to ensure scene diversity.

We implement a pose estimation pipeline to derive smooth and accurate 3D pose sequences in the videos. Details of the pose estimation process are in Appendix. We manually annotate the start and end time of each human- object interaction in the videos. There are 24 different types of action-object pairs present in the dataset. Statistical details and examples are in Fig. 3.

Although 3D poses and depth point clouds are available in the dataset, we concentrate on a 2D version of the proposed task - inferring 2D poses in the image space. Therefore, we obtain 2D poses by projecting the 3D ones to the image plane of the RGB-Thermal camera to conduct our work. Note that the actor is usually stationary, in which case motion inference is trivial. We filter out the clips where the average displacement per joint is less than 45 pixels in 3 seconds. The remaining 110k frames serve as the data for our proposed task.

### 4. Method

This work aims to infer what a person in a thermal image was doing  $N$  seconds ago. We set  $N = 3$  since we empirically find that it takes at most 3 seconds for a person to complete an action. If  $N$  is too small, one can infer past poses directly from current poses without any context. On the other hand, if  $N$  gets larger, the thermal cues may disappear, and the uncertainty of the past increases.

Due to the inherent uncertainty of human motion, our model makes stochastic predictions, *i.e.*,  $M$  possible 3s-ago poses of the person. Thermal images provide plenty of cues telling where the people were and how they interacted with the environment. It is challenging for a model to understand this information and make plausible inferences.

Formally, given a thermal image  $I \in \mathbb{R}^{H \times W}$ , the goal is to generate  $M$  2D poses of the person 3 seconds ago, denoted by  $q_{1:M} \in \mathbb{R}^{M \times J \times 2}$ . Here  $(H, W)$  is the size of the images, and  $J$  is the number of joints in a human pose. We also provide the current pose  $p \in \mathbb{R}^{J \times 2}$  of the person in the image so that a model can focus on inferring the past instead of struggling to recognize the person first.

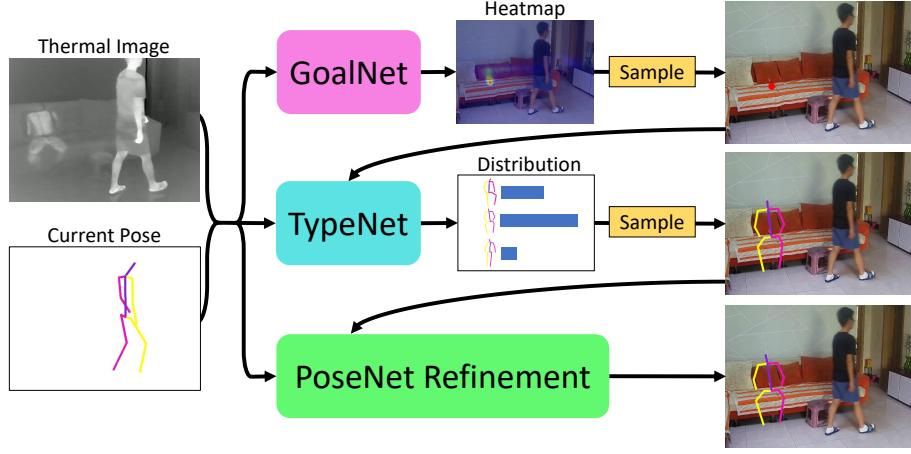


Figure 4. **Overview of our pipeline.** Given a thermal image and an estimated human pose as input, GoalNet first predicts the distribution of the person’s location 3 seconds ago and samples one from it. Then, TypeNet predicts pose type distribution and samples one pose from them. Finally, PoseNet refines the pose to match the input observation better. The RGB images are for visualization purposes only.

We propose a three-stage framework illustrated in Fig. 4 to tackle this task. It has three components, GoalNet, TypeNet, and PoseNet. In the first stage, GoalNet proposes possible positions where the human was 3 seconds ago. Next, TypeNet assigns a possible pose type (sitting, standing, walking, *etc.*) at each proposed position. Finally, PoseNet synthesizes a pose of the assigned type at each proposed position. The rest of this section will discuss the motivation and details of our method.

#### 4.1. GoalNet

When a thermal image like Fig. 2 is shown to humans, one can intuitively figure out where the actor used to be by recognizing the bright marks on the objects. Consequently, one can confidently tell that the actor was around the bright mark or on a path connecting that position with the current position 3 seconds ago. Therefore, our first stage GoalNet  $\mathcal{G}$  is designed to capture the distribution of human position.

Let  $r \in \mathbb{R}^2$  denote the torso joint position of a pose  $q \in \mathbb{R}^{J \times 2}$  that we want to generate. We sample  $r$  from a distribution  $P(r) \in \mathbb{R}^{H \times W}$ , where

$$P(r) = \mathcal{G}(I, H_p) \quad (1)$$

is predicted by GoalNet based on the image and current human pose. Here we use  $H_x \in \mathbb{R}^{L \times H \times W}$  to denote the heatmap representation of a series of 2D positions  $x \in \mathbb{R}^{L \times 2}$ . Thus  $H_p$  here is in shape  $J \times H \times W$ .

We use an Hourglass model [25] as the architecture of GoalNet. This model can not only capture local thermal cues but also consider the context information to generate a plausible position distribution in the image space.

#### 4.2. TypeNet

After the torso position is specified, the next step is to generate a pose at that location. Instead of drawing a human pose directly, one may first speculate what action the character was doing there, such as if the one was sitting or standing and if the one was facing to the left or right. Moreover, the possible human poses are diverse even at a particular position, *e.g.*, it is plausible to stand to both the left and right in some circumstances. Therefore, we need first specify a *pose type* (action) at each position before synthesizing an explicit pose.

To derive pose labels, we cluster all the poses in the training set into several groups, and each group corresponds to a pose type. In practice, we align the torso joints of all poses, represent a  $J$ -joint pose as a  $2(J - 1)$ -dimensional vector, and apply the K-Means algorithm in Euclidian space to form 200 clusters.

The second stage TypeNet  $\mathcal{T}$  then gives a distribution  $P(z) \in \mathbb{R}^{200}$  over all pose types at the proposed position  $r$  according to the inputs, where  $z$  denotes a pose type index. Formally,

$$P(z) = \mathcal{T}(I, H_p, H_r). \quad (2)$$

As a typical image classification task, ResNet18 [8] serves as the backbone of TypeNet. We can sample a pose type from the distribution  $P(z)$  as the actor’s action that we infer.

#### 4.3. PoseNet

The final step is synthesizing a human pose of type  $z$  at location  $r$ . At this step, the detailed information in the image determines the pose’s size and the joints’ accurate positions. We develop PoseNet  $\mathcal{P}$  to infer the pose while being aware of this information.

PoseNet is also an Hourglass model like GoalNet. It gives a heatmap  $P(q) \in \mathbb{R}^{(J-1) \times H \times W}$  for all joints of  $q$  except the torso joint. Instead of feeding the pose type index  $z$  into PoseNet, we paint the  $z$ -th cluster's center pose at position  $r$  as input. Hence, PoseNet is refining a given pose rather than generating a pose from scratch. Formally, we have

$$P(\tilde{q}) = \mathcal{P}(I, H_p, H_r, H_{C_z+r}), \quad (3)$$

$$\forall 1 \leq j \leq J-1, \tilde{q}_j = \arg \max_{x,y} P(\tilde{q}_j = (x,y)), \quad (4)$$

$$q = [\tilde{q}, r], \quad (5)$$

where  $\tilde{q}$  denotes a human pose without the torso joint and  $C_z$  is the  $z$ -th pose cluster center.

#### 4.4. Learning

We split the main part of the dataset mentioned in Sec. 3 into training, validation, and test sets in terms of video clips and train our model with the training set.

The three modules are trained separately, using the ground truth in the last step as input and supervised by the labels in the current step. As the predictions of all modules are probability distributions, we utilize Cross Entropy Loss ( $\mathcal{L}_{CE}$ ) as their training objectives. Let  $\hat{r}$  and  $\hat{q}$  be the ground truth torso position and human pose 3 seconds ago, and let  $\hat{z}$  be the pose type of  $\hat{q}$ . The training losses for GoalNet  $\mathcal{G}$ , TypeNet  $\mathcal{T}$ , and PoseNet  $\mathcal{P}$  are

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{CE}(\mathcal{G}(I, H_p), \hat{r}), \quad (6)$$

$$\mathcal{L}_{\mathcal{T}} = \mathcal{L}_{CE}(\mathcal{T}(I, H_p, H_{\hat{r}}), \hat{z}), \quad (7)$$

$$\mathcal{L}_{\mathcal{P}} = \sum_{j=1}^{J-1} \mathcal{L}_{CE}(\mathcal{P}_j(I, H_p, H_{\hat{r}}, H_{C_z+\hat{r}}), \hat{q}_j). \quad (8)$$

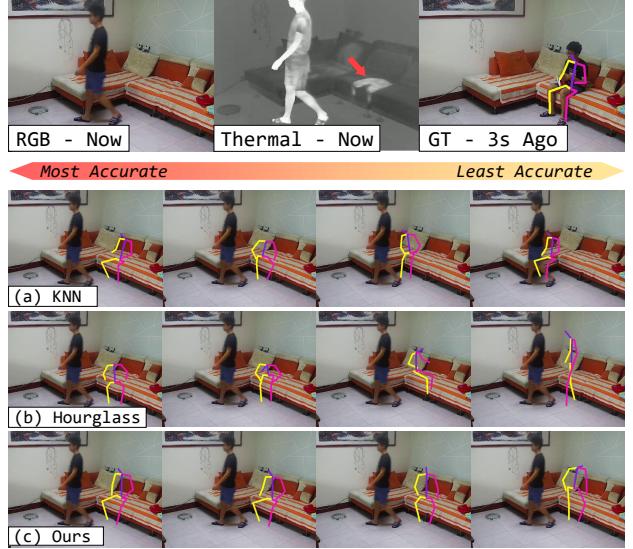
#### 4.5. Inference

This task requires a model to give  $M$  possible answers for each test sample. To do this, we sample  $M$  torso positions  $r$  at the GoalNet stage and then run TypeNet and PoseNet once for each sampled position.

In TypeNet, rather than sampling among all pose types, we find that top- $k$  sampling with  $k = 5$  leads to the best performance. That is, we sample the pose type from the five types with the highest probabilities given by TypeNet. At the PoseNet stage, the position with the highest weight in each joint's heatmap is picked as the final prediction.

### 5. Experiments

In this section, we first evaluate the effectiveness of our approach on the Thermal-IM dataset. Then we investigate the importance of different modalities for inferring past human behavior. Finally, we comprehensively study the characteristic of our model.



**Figure 5. Comparison against baselines.** We sort the predictions of each approach based on MPJPE and show the 1st, 5th, 10th, and 20th poses from left to right. KNN fails to infer where the person was. Hourglass is able to locate where the person was, but their predictions do not comply with the affordance of the scene. Our method, in contrast, produces reasonable and accurate estimates.

#### 5.1. Evaluation metrics

**MPJPE:** We calculate the Mean Per Joint Position Error (MPJPE) [11] between the generated poses and the ground truth to evaluate their similarity. Specifically, MPJPE is the average Euclidian distance in the number of pixels from each joint to its corresponding answer. Due to the uncertainty of the past, 30 poses are generated for each test sample by each model. We report the average MPJPE of the top-1/3/5 ones closest to the ground truth.

**Negative log-likelihood (NLL):** As our modules yield probability distributions, we can determine the probability of each pose. To quantify the accuracy of our model, we utilize the NLL of the actual poses as a metric. We report this metric for all methods that support likelihood estimation.

**Semantic score:** Note that one can randomly synthesize diverse poses ignoring the scene context to achieve low top- $k$  MPJPE. However, such poses may be implausible in the scene. We use semantic score [21] to measure how many generated poses are plausible in the given contexts. Specifically, we construct a dataset containing RGB images with plausible and implausible poses based on Thermal-IM (including the held-out part) and train a binary classifier to distinguish them. Plausible poses are the 3s-ago poses, and implausible poses are derived by randomly replacing, shifting, and perturbing the plausible ones. The classifier achieves a

Method	MPJPE			NLL	Semantic Score(%)
	Top 1	Top 3	Top 5		
KNN	19.26	24.53	28.44	N/A	61.94
Hourglass	23.80	27.99	31.03	136.23	67.05
Ours	<b>18.33</b>	<b>22.25</b>	<b>25.25</b>	<b>103.75</b>	<b>82.11</b>

Table 1. **Evaluation results of our model and baselines.** Our model outperforms all the baselines in all metrics.

Input	MPJPE			NLL	Semantic Score(%)
	Top 1	Top 3	Top 5		
RGB	22.06	27.21	31.12	105.03	<b>87.56</b>
Thermal	<b>18.33</b>	<b>22.25</b>	<b>25.25</b>	<b>103.75</b>	82.11
RGB-T	19.23	23.52	26.76	<b>103.75</b>	85.46
T w/o pose	19.62	24.00	27.27	104.38	80.55

Table 2. **Ablation study of model input.** The thermal model achieves the best MPJPE and NLL, while the RGB model has the highest semantic score. The RGB-T model access both modalities but does not provide a better performance in any metric. Once the current pose is not provided, the thermal model can still achieve competitive results.

test accuracy of 85.77%. The semantic score for a method is defined as the ratio of generated poses recognized as plausible by the classifier. Examples of training data and implementation details are in Appendix.

## 5.2. Baselines

**KNN:** We first construct a pool of current-past pose pairs from the training set. Since the adjacent video frames are alike, we sample one frame every 15 frames. Next, given a test human pose, we leverage K-Nearest Neighbor (KNN) to retrieve 30 closet samples from the pool. Finally, we treat the corresponding past pose as the results.

**Hourglass:** We adapt the state-of-the-art 2D pose estimation model [25] as our second baseline. Given input observation(s), we first predict a distribution map for each joint of the *past* pose. Since independently sampling each joint may result in unrealistic poses, we then exploit the human poses from the training set and evaluate their likelihood with the predicted distribution. This ensures that the estimated poses are always realistic. Finally, we select 30 poses with the highest likelihood. In practice, we consider 1/200 poses from the training set.

## 5.3. Evaluation results

As shown in Tab. 1, our method outperforms the baselines significantly across all metrics. It is able to recover

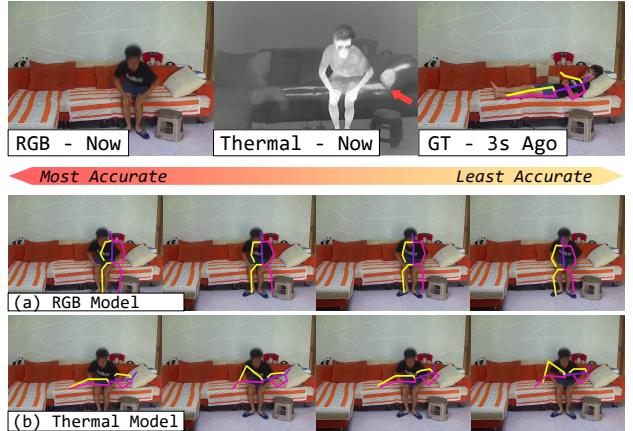


Figure 6. **Ablation study of input modality.** In this example, it is hard to infer the person’s previous action through the RGB image. With the thermal image, however, one can easily deduce that the person was lying on the couch. Consequently, the thermal model provides much more accurate predictions.

plausible human poses 3 seconds ago accurately. Some qualitative results are shown in Fig. 5. In the thermal image, the bright mark implies that the person was sitting on the sofa. Our method observes this and synthesizes several poses sitting or getting up from there. In contrast, KNN either retrieves poses sitting in other places or gives implausible answers - a pose sitting on nothing. As for Hourglass, although it succeeds in locating the place where the person was sitting, the estimated poses do not comply with the affordance of the sofa.

## 5.4. Ablation studies

We first investigate the importance of different modalities for inferring the past. Then we study whether the availability of the current human pose will affect the model performance. We refer the readers to Appendix for ablation on the modules.

**Importance of different modalities:** As shown in Tab. 2, our model performs best on MPJPE and NLL when taking thermal images as input. However, the semantic score is higher when RGB images are included. We conjecture this is because the details in the scene are more apparent in the RGB domain.

We show two qualitative comparisons in Fig. 6 and Fig. 7. The horizontal thermal mark on the sofa (see Fig. 6) implies that the person was lying there. With thermal images, the model can infer various lying poses. Its RGB counterpart, however, cannot notice this and only synthesizes sitting poses. As for Fig. 7, there is no thermal cue on any object, indicating that the person did not touch anything in the short past. The thermal model thus only generates

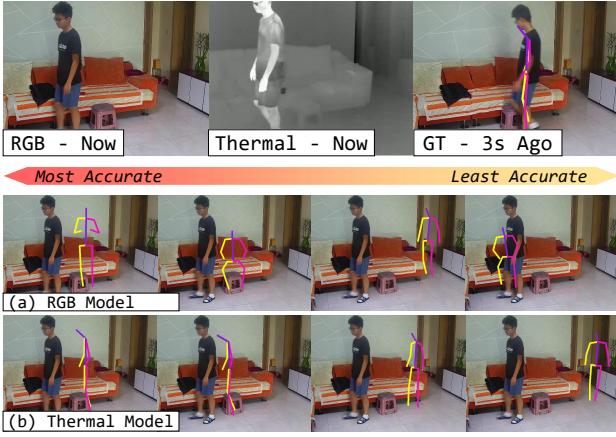


Figure 7. **Ablation study of input modality.** In this example, the person did not touch any object. While the thermal model learns this, the RGB model gives poses sitting on the sofa.

walking poses. In contrast, the RGB model fails to capture the difference and predicts multiple sitting poses.

Additionally, although the RGB-Thermal model benefits from richer information, its performance falls between that of the RGB and thermal models. This observation suggests that early fusion methods, such as concatenation at the input level, fail to capture cross-modal interactions effectively. Further research is needed to develop more effective fusion methods that can capitalize on the complementary nature of RGB and thermal modalities.

**Current pose as input:** To infer the past, it is crucial for a model to know the human pose in the current frame. Once the model is provided with the current pose, it does not need to implicitly learn to recognize the human. However, one may have trouble estimating the current pose in practice. To tackle this issue, we train a model that does not require current poses. As shown in Tab. 2, the performance degrades a bit but is still competitive. If the human pose is unavailable in practice, our method without input body pose can serve as an effective alternative.

## 5.5. Analysis

**Effect of thermal intensity:** After a human-object interaction, the thermal mark left on the object gradually gets dimmer and finally vanishes. Thus the intensity of the mark reveals information about when the interaction happened. To see whether our model learns about this knowledge, we manually modify the mark brightness in a given image and then examine how our model’s prediction changes.

Fig. 8 shows the varied heatmap predictions of GoalNet, representing the distributions of the person’s position 3 seconds ago. When the mark on the sofa is bright, the heatmap density tends to converge at the mark’s position. On the

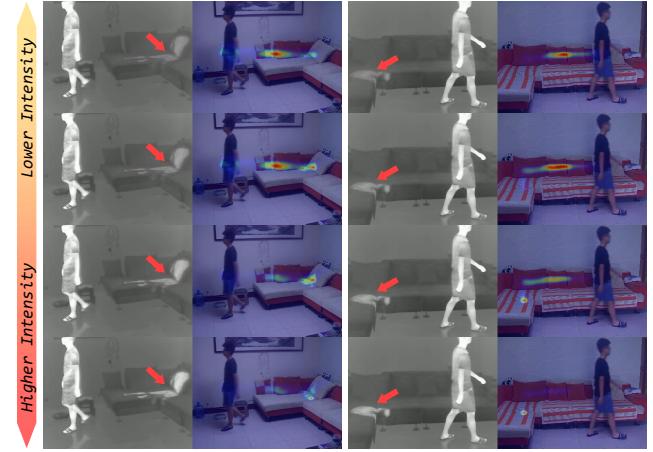


Figure 8. **Varied GoalNet predictions when modifying the thermal intensity.** The intensity of the bright marks in the images increases from top to bottom. As it increases, the inferred distribution of the character’s 3s-ago position gets closer to the thermal mark.

contrary, it is close to the person when the mark is dim. This result coincides with our intuition. A mark is bright only if the interaction was just over in the past few seconds; therefore, it is more likely that a person was there 3 seconds ago. Conversely, if a mark is dim, the person probably had already left there 3 seconds ago. This experiment suggests that our model understands the time information contained in thermal mark intensity.

**Generalization:** The videos in the Thermal-IM training set only involve one actor and two different rooms. It is critical to investigate whether a model trained on it generalizes well when the actor and environment are changed. To this end, we use the held-out part of the Thermal-IM dataset mentioned in Sec. 3 to conduct generalization experiments. This part of data involves several factors changed from the training set, including the arrangement of objects, the background, the actor, and the room. These changes are illustrated in Fig. 9.

We test our RGB and thermal models over these four cases to examine how these changes influence them. Tab. 3 shows the results. Although the model performances deteriorate compared to Tab. 2, the thermal model is still the best in MPJPE and NLL when changing the arrangement, background, or room. Particularly, when a new background or room is involved, the performance decrease of the thermal model is much smaller than that of its RGB counterpart. We hypothesize that reason is that when seeing new objects and backgrounds, the RGB model tries to identify things humans can interact with; instead, the thermal model infers the past by finding the thermal marks in certain shapes without considering object identification. The latter mechanism is

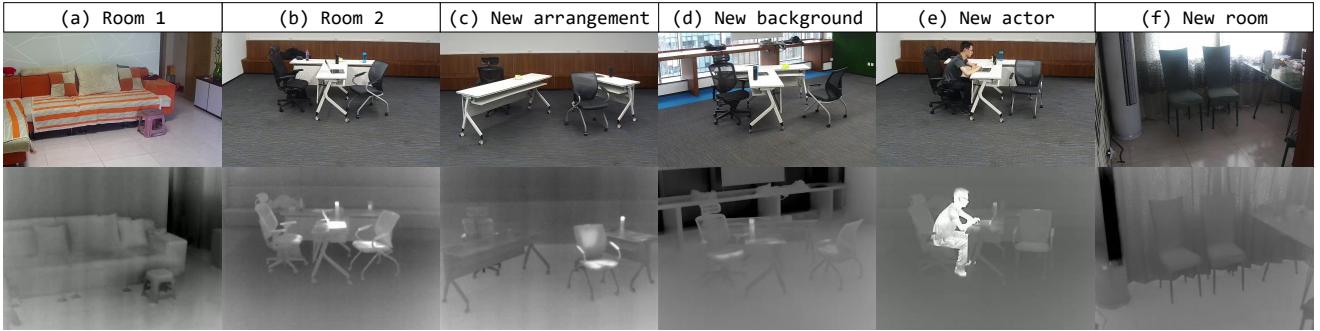


Figure 9. **A demonstration of our generalization test.** Our training data is collected from two rooms, (a) Room 1 and (b) Room 2. The latter four are samples of extra data for the generalization test; (c) Room 2 with a new object arrangement; (d) Room 2 with a new background; (e) A new actor in Room 2; (f) A new room different from Room 1 and 2.

Changed Factor	Modality	MPJPE			NLL	Semantic Score(%)
		Top 1	Top 3	Top 5		
Arrangement	RGB	21.27	26.38	30.42	107.10	<b>93.69</b>
	Thermal	<b>20.41</b>	<b>25.10</b>	<b>28.36</b>	<b>105.37</b>	89.56
Background	RGB	25.07	30.02	33.47	111.67	<b>83.80</b>
	Thermal	<b>19.85</b>	<b>24.24</b>	<b>27.83</b>	<b>107.82</b>	81.49
Actor	RGB	<b>24.37</b>	29.20	32.77	114.87	<b>91.21</b>
	Thermal	24.60	<b>28.92</b>	<b>31.98</b>	<b>114.26</b>	81.33
Room	RGB	35.05	42.00	47.11	121.14	19.55
	Thermal	<b>23.05</b>	<b>27.59</b>	<b>31.16</b>	<b>112.84</b>	<b>36.88</b>

Table 3. **Generalization test results.** In most cases, the thermal model provides more accurate predictions, while the RGB model achieves higher semantic scores when the room is not changed. Moreover, the thermal model greatly outperforms the RGB model when introducing a new background or room. This certifies that our thermal model is more robust to environmental appearance.

more robust to the appearance changes of the environment.

**Limitations:** In Tab. 3, we observe performance degradation in both RGB and thermal models when a new actor is involved. We ascribe this to the new actor’s different stature and behavioral habits from the one in the dataset. The personal habits introduce new actions the model has never seen, and the different statures indicate different sizes of human poses. However, our model architecture, most notably TypeNet, limits the prediction of poses that appear in the training set. We expect future work on the model design to perform better on actor generalization.

## 6. Conclusions

In this work, we propose to infer past human pose by leveraging thermal imaging. We collect the Thermal-IM dataset containing RGB-Thermal and RGB-Depth videos about indoor human motion with estimated poses. Based on this dataset, a three-stage method is developed to tackle the proposed task. We show that inference of the past becomes

an easier task with thermal images compared to RGB ones. The experiments demonstrate not only our model’s capability of understanding past human location and action but also its awareness of the correlation between thermal mark intensity and time. Some aspects of this task remain to be explored, such as how to effectively fuse RGB and thermal modalities to use their information jointly.

**Acknowledgement** WCM is partially funded by a Siebel scholarship and the MIT-IBM Watson AI Lab.

## References

- [1] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 2
- [2] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM Press, 2000. 2

- [3] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*. 2020. 2
- [4] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. 2
- [5] Fufin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, sep 2021. 2
- [6] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015. 2
- [7] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, sep 2017. 2
- [8] 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)*. IEEE, jun 2016. 4
- [9] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno. Human motion prediction via spatio-temporal inpainting. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2
- [10] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. Apr. 2022. 2
- [11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, jul 2014. 5
- [12] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. 2
- [13] Peng Jingchao, Zhao Haitao, Hu Zhengwei, Zhuang Yi, and Wang Bofan. Siamese infrared and visible light fusion network for rgb-t tracking. Mar. 2021. 2
- [14] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Cebovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, Abdelrahman Eldesokey, Jani Kapyla, Gustavo Fernandez, Abel Gonzalez-Garcia, Alireza Memarmoghadam, Andong Lu, Anfeng He, Anton Varfolomieiev, Antoni Chan, Ardhendu Shekhar Tripathi, Arnold Smeulders, Bala Suraj Pedasingu, Bao Xin Chen, Baopeng Zhang, Baoyuan Wu, Bi Li, Bin He, Bin Yan, Bing Bai, Bing Li, Bo Li, Byeong Hak Kim, Chao Ma, Chen Fang, Chen Qian, Cheng Chen, Chenglong Li, Chengquan Zhang, Chi-Yi Tsai, Chong Luo, Christian Micheloni, Chunhui Zhang, Dacheng Tao, Deepak Gupta, Dejia Song, Dong Wang, Efstratios Gavves, Eunu Yi, Fahad Shahbaz Khan, Fangyi Zhang, Fei Wang, Fei Zhao, George De Ath, Goutam Bhat, Guangqi Chen, Guangting Wang, Guoxuan Li, Hakan Cevikalp, Hao Du, Haojie Zhao, Hasan Saribas, Ho Min Jung, Hongliang Bai, Hongyuan Yu, Hongyuan Yu, Houwen Peng, Huchuan Lu, Hui Li, Jiakun Li, Jianhua Li, Jianlong Fu, Jie Chen, Jie Gao, Jie Zhao, Jin Tang, Jing Li, Jingjing Wu, Jingtuo Liu, Jingqiao Wang, Jinqing Qi, Jinyue Zhang, John K. Tsotsos, Jong Hyuk Lee, Joost van de Weijer, Josef Kittler, Jun Ha Lee, Junfei Zhuang, Kangkai Zhang, Kangkang Wang, Kenan Dai, Lei Chen, Lei Liu, Leida Guo, Li Zhang, Liang Wang, Lian-gliang Wang, Lichao Zhang, Lijun Wang, Lijun Zhou, Linyu Zheng, Litu Rout, Luc Van Gool, Luca Bertinetto, Martin Danelljan, Matteo Dunnhofer, Meng Ni, Min Young Kim, Ming Tang, Ming-Hsuan Yang, Naveen Paluru, Niki Martinel, Pengfei Xu, Pengfei Zhang, Pengkun Zheng, Pengyu Zhang, Philip H.S. Torr, Qi Zhang Qiang Wang, Qing Guo, Radu Timofte, Rama Krishna Gorthi, Richard Everson, Ruize Han, Ruohan Zhang, Shan You, Shao-Chuan Zhao, Shengwei Zhao, Shihu Li, Shikun Li, Shiming Ge, Shuai Bai, Shuosen Guan, Tengfei Xing, Tianyang Xu, Tianyu Yang, Ting Zhang, Tomas Vojir, Wei Feng, Weiming Hu, Weizhao Wang, Wenjie Tang, Wenjun Zeng, Wenyu Liu, Xi Chen, Xi Qiu, Xiang Bai, Xiao-Jun Wu, Xiaoyun Yang, Xier Chen, Xin Li, Xing Sun, Xingyu Chen, Xinmei Tian, Xu Tang, Xue-Feng Zhu, Yan Huang, Yanan Chen, Yanchao Lian, Yang Gu, Yang Liu, Yanjie Chen, Yi Zhang, Yinda Xu, Yingming Wang, Yingping Li, Yu Zhou, Yuan Dong, Yufei Xu, Yunhua Zhang, Yunkun Li, Zeyu Wang Zhao Luo, Zhao-liang Zhang, Zhen-Hua Feng, Zhenyu He, Zhichao Song, Zhihao Chen, Zhipeng Zhang, Zhirong Wu, Zhiwei Xiong, Zhongjian Huang, Zhu Teng, and Zihan Ni. The seventh visual object tracking VOT2019 challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, oct 2019. 2
- [15] Xin Lan, Xiaojing Gu, and Xingsheng Gu. MMNet: Multi-modal multi-stage network for RGB-t image semantic segmentation. *Applied Intelligence*, 52(5):5817–5829, aug 2021. 2
- [16] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, dec 2016. 2
- [17] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. RGB-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, dec 2019. 2
- [18] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2022. 2
- [19] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018. 2
- [20] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph

- learning for RGB-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, oct 2017. 2
- [21] Xuetong Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 5
- [22] Chengwei Luo, Bin Sun, Ke Yang, Taoran Lu, and Wei-Chang Yeh. Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme. *Infrared Physics and Technology*, 99:265–276, jun 2019. 2
- [23] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2
- [24] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. 2
- [25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer International Publishing, 2016. 2, 4, 6
- [26] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. PST900: RGB-thermal calibration, dataset and segmentation network. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2020. 2
- [27] Yuxiang Sun, Weixun Zuo, and Ming Liu. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, jul 2019. 2
- [28] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 18(3):1000–1011, jul 2021. 2
- [29] Johan Vertens, Jannik Zurn, and Wolfram Burgard. HeatNet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, oct 2020. 2
- [30] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017. 2
- [31] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for RGB-t tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 2
- [32] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2
- [33] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 2
- [34] Shuangjiu Xiao C. He Zeng Huang Hao Li Yi Zhou, Zimo Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *International Conference on Learning Representation*, 2018. 2
- [35] Jason Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2
- [36] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end RGB-t tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, oct 2019. 2
- [37] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust RGB-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021. 2
- [38] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. Apr. 2022. 2
- [39] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. ABMDR-Net: Adaptive-weighted bi-directional modality difference reduction network for RGB-t semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2
- [40] Tianlu Zhang, Xueru Liu, Qiang Zhang, and Jungong Han. SiamCDA: Complementarity- and distractor-aware RGB-t tracking based on siamese network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1403–1417, mar 2022. 2
- [41] Xingming Zhang, Xuehan Zhang, Xuedan Du, Xiangming Zhou, and Jun Yin. Learning multi-domain convolutional network for RGB-t visual tracking. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, oct 2018. 2
- [42] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [43] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for RGB-thermal scene parsing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3571–3579, jun 2022. 2

# Appendix

We first introduce the pose estimation pipeline in dataset construction in Appendix A. In Appendix B, we specify the details of the method implementation. We individually evaluate each module in our method, and the results are in Appendix C. We also show some qualitative results in Appendix D to demonstrate our method’s capability of inferring the past. Appendix E is an ablation study on our method pipeline. Finally, we state the approval for our Thermal-IM dataset in Appendix F.

## A. Pose estimation in dataset construction

We implement a two-stage pose extraction pipeline to acquire smooth and accurate 3D poses from RGB video pairs.

In the first stage, we use [4] to estimate a coarse 3D pose for every RGB frame. The resulting poses are in OpenPose [2] *body\_25* skeleton with 25 joints. According to the pose estimation results, we divide the videos into continuous segments, ensuring that the character is always in the view of both cameras in each segment.

In the second stage, we synthesize the monocular pose estimation results of the two cameras to improve the pose quality. We first implement a triangulation step for more accurate depth estimation. Specifically, for each timestamp  $t$ , let  $p_t, q_t$  be the coarse 3D poses in the camera coordinates detected from the two cameras. We find a pair of scales  $a, b$  that minimizes the  $\ell_2$  distance between  $p_t \cdot a$  and  $q_t \cdot b$  in the world coordinate. After that, we use EasyMocap [1] to refine the pose sequence further. It smooths a sequence of 3D poses by optimizing the SMPL body parameters [6].

We further eliminate the failure cases of pose estimation. Specifically, all the poses are clustered into 2000 groups, and we manually filter out the clusters representing contorted poses.

## B. Implementation details

**Data processing:** In our task, we use the first 15 joints out of the 25 joints in the OpenPose skeleton to represent a human pose. The first 15 joints are enough to depict human actions while ignoring the details such as ears and toes.

The input image size for our model is  $288 \times 384$ . And the evaluation metric MPJPE is also computed at this scale. The RGB and thermal lenses of our RGB-Thermal camera have different fields of view, and that of the thermal lens is

Module	Learning rate	Batch size
GoalNet	$5 \times 10^{-5}$	32
TypeNet	$5 \times 10^{-5}$	128
PoseNet	$1 \times 10^{-4}$	32
Semantic score model	$3 \times 10^{-5}$	128

Table 1. Learning rates and batch sizes.



Figure 1. Samples used to develop the semantic score classifier. Plausible ones are samples in the dataset, while implausible ones are derived by random pose replacement, shift, and perturbation.

smaller. We resize the thermal images in a preset way to align with the human poses, which are estimated in RGB image space.

**Model implementation:** The backbones of GoalNet and PoseNet are both an Hourglass model [7] with three blocks, while that of TypeNet is ResNet18 [3]. The sizes of heatmap outputs of GoalNet and PoseNet are  $72 \times 96$ , and they are resized to be  $288 \times 384$  by interpolation.

All modules are trained using the Adam optimizer [5] for 6k batch iterations. The learning rates and batch sizes are in Tab. 1. We use random crop and flip as data augmentation for all of them.

**Semantic score:** The data we use to train the semantic score model contains RGB images with plausible and implausible poses. Plausible poses are the 3s-ago poses, and implausible poses are derived by randomly replacing, shifting, and perturbing the plausible ones. Some samples are shown in Fig. 1.

Given an RGB image and a pose, we want a binary classifier to estimate how likely the pose is plausible. We use

Module	Average $\ell_2$ Distance		
	Top 1	Top 3	Top 5
GoalNet	10.50	15.02	31.12

Table 2. **Evaluation of GoalNet.** We calculate the  $\ell_2$  distances from the top-1/3/5 predicted positions to the ground truth in the number of pixels.

Module	Accuracy		
	Top 1	Top 3	Top 5
TypeNet	10.50	15.02	31.12

Table 3. **Evaluation of TypeNet.** The task of TypeNet is indeed classification, so we evaluate the top-1/3/5 accuracy of its prediction.

ResNet18 as the model and train it with Binary Cross Entropy Loss. It is trained using the Adam optimizer with a weight decay of  $1 \times 10^{-3}$  for 6k batch iterations. The learning rate and batch size are in Tab. 1. Random crop and flip are used as data augmentation.

## C. Individual evaluation of modules

As the three modules in our method are trained separately, we evaluate their performances in their own tasks in the following.

**GoalNet:** For each test instance, GoalNet samples 30 torso joint positions according to the predicted heatmap, and we evaluate how close they are to the ground truth 3s-ago position. We sort the 30 positions by order of their distances to the ground truth and compute the average  $\ell_2$  distance of the top-1/3/5 ones. We show the results in Tab. 2.

**TypeNet:** We evaluate TypeNet as a classifier and report its top-1/3/5 accuracy. The results are in Tab. 3.

**PoseNet:** We examine how the refinement of PoseNet makes an inputted pose type center closer to the ground truth 3s-ago pose. We report the MPJPE of poses before and after refinement in Tab. 4.

## D. More qualitative results

In Fig. 3, we show samples of our method’s synthesized poses in the test set. The involved indoor actions include sitting on a sofa/chair/table, lying on a sofa, touching a cabinet/bottle, and several actions on a yoga mat (sit-ups, push-ups, and leg stretching).

Module	MPJPE	
	Before	After
PoseNet	8.87	8.59

Table 4. **Evaluation of PoseNet.** Given the cluster center pose as input, we evaluate how much our PoseNet can refine it. The table shows the MPJPE from the poses to the ground truth poses before and after PoseNet refinement.

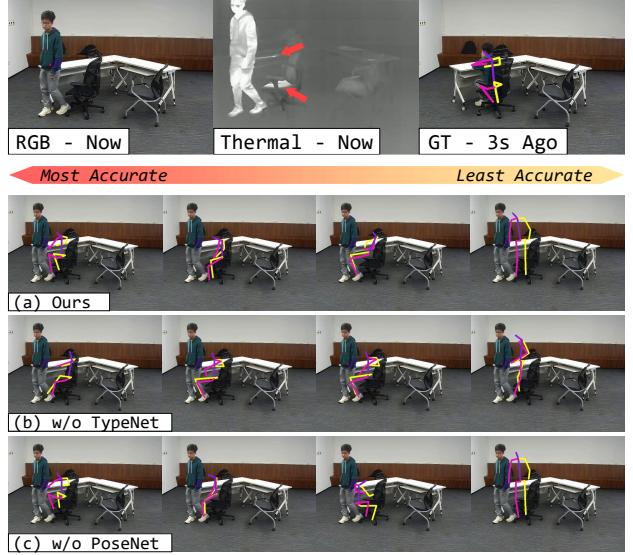
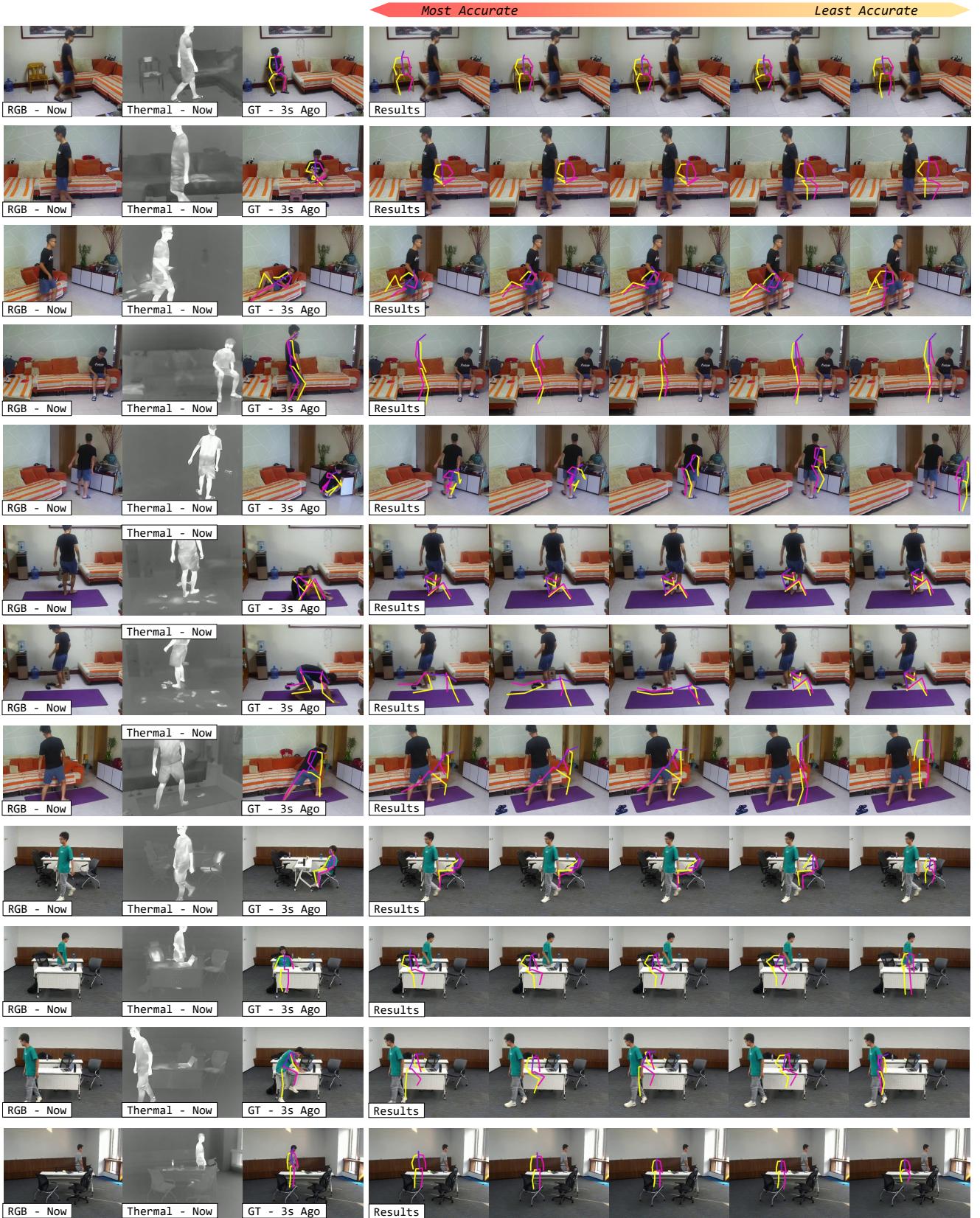


Figure 2. **Ablation study on model architecture.** From the thermal image, we can deduce that the person was sitting on the chair with arms on the table. In the model w/o TypeNet, predicted poses are often out-of-shape. The model w/o PoseNet can hardly provide the pose we desire because the number of pose types is limited. Our full model can refine the center pose of a type to fit with the details in the image, so it successfully generates sitting poses with an arm on the table (the 1st and 3rd column).

## E. Ablation studies on pipeline modules

We implement two versions of our model without TypeNet or PoseNet to see how these modules contribute to our method.

**w/o TypeNet:** In a model without TypeNet, PoseNet generates a pose based on the input image and a root position given by GoalNet. The type of the synthesized pose is not specified here. In some cases, however, various poses are possible at a specific position. The skeleton joints generated by this model cannot be guaranteed to belong to the same pose, which leads to out-of-shape results as Fig. 2(b) shows and low semantic score in Tab. 5. Besides, because the generated poses are far from reality, the top-1 MPJPE is much higher than our complete model, though the top-5



**Figure 3. Visualization results of our model.** For each sample, we sort the 30 predictions in order of MPJPE and show the 1st, 3rd, 5th, 10th, and 20th poses from left to right. Please pay attention to the bright marks pointed out by the arrows in the thermal images.

Modules			MPJPE			NLL	Semantic Score(%)
GoalNet	TypeNet	PoseNet	Top 1	Top 3	Top 5		
✓		✓	19.04	22.45	<b>25.19</b>	112.28	73.12
✓	✓		18.64	22.61	25.65	N/A	76.68
✓	✓	✓	<b>18.33</b>	<b>22.25</b>	25.25	<b>103.75</b>	<b>82.11</b>

Table 5. **Ablation study of removing different components.** Our model (the last row) outperforms the incomplete ones in most metrics, though removing TypeNet provides a slightly lower Top-5 MPJPE. We do not report the NLL for the one without PoseNet since it cannot be calculated in this setting.

MPJPE is competitive.

*Conference on Computer Vision (ECCV)*, pages 483–499. Springer International Publishing, 2016. 1

**w/o PoseNet:** In a model without PoseNet, TypeNet provides a pose type, and the center pose of this type is moved to the GoalNet’s predicted position to serve as an answer. Since the number of pose types is limited, the duplicated pose cannot always fit with the details in the image. In the first column of Fig. 2(a) *vs.* (c), the model without PoseNet simply draws a sitting pose, while our complete model refines it so that the right arm is put on the table. As Tab. 5 illustrates, the refinement served by PoseNet improves both the synthesized poses’ similarity to the answer and the plausibility in the context.

## F. Approval

We have obtained approval for collecting and using the Thermal-IM dataset from the Institutional Review Board of our university department.

## References

- [1] Easymocap - make human motion capture easier. Github, 2021. 1
- [2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1
- [3] 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)*. IEEE, jun 2016. 1
- [4] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*. IEEE, dec 2021. 1
- [5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. Dec. 2014. 1
- [6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1
- [7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European*