

# 1 Supplementary material - The MIPT dataset

This document contains additional information about the MIPT dataset. Example images for some sequences are shown in Figure 1, while Table 1 contains a brief description of each sequences and suggests a split into training, validation and test data. The subset assignment is chosen so that all subsets have approximately the same level of difficulty. Restriction to the multimodal depth+thermal sequences results in a 5/2/2 (Training/Validation/Test) split, while using the entire dataset results in a 12/4/4 distribution (see Table 1).

The duration of the sequences ranges from about 100 to 300 seconds. We use a custom multimodal sensor that include an Orbbec Astra<sup>1</sup> depth sensor and a FLIR Lepton 3.5<sup>2</sup> thermal camera module combined in a single housing. The combined module was mounted at height between 2.2 and 3 meters in a downward facing monitoring pose.

All sequences show an indoor scene of a single room or hallway between  $12m^2$  and  $32m^2$  in size. The largest rooms approach the limits of the depth sensor used, which has a maximum detection depth of 8 meters. The locations are a mix of office spaces, apartments, and workshop areas.

The number of people in a scene ranges from 2 up to more than 10 people, with 2 being the most frequent number. A higher number of people in a scene was not deemed to be sensible because many of the spaces tend to be small, with areas as little as  $12m^2$ . Thus, even with only two actors, our recordings feature frequent occlusions between people, or people and objects. The dataset is almost exactly balanced in terms of gender, while the age of the actors range roughly from 20 to 60 years, with a bias towards younger participants. Activities are mostly scripted to guarantee a high degree of movement, poses and behavioral patterns in a short time frame.

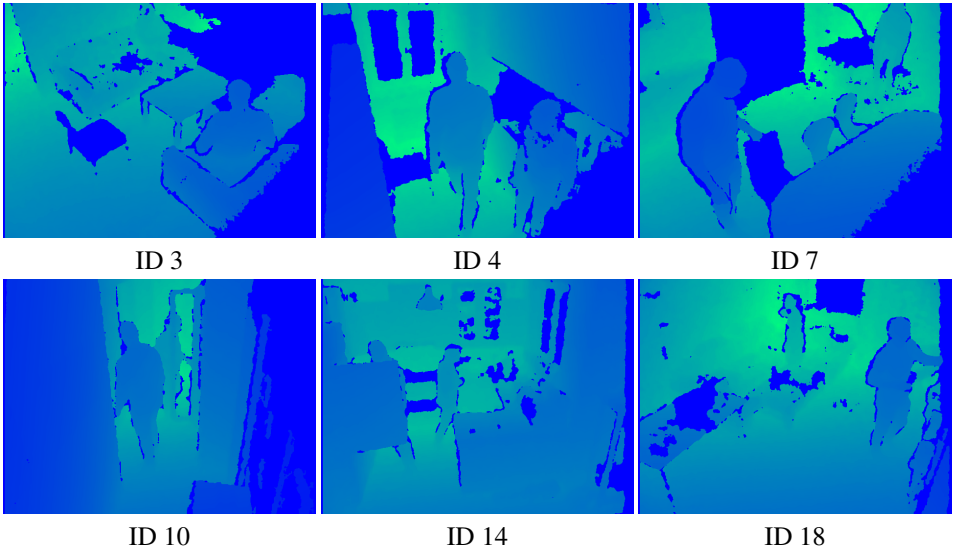


Figure 1: Sample depth frames for six of the sequences in the MIPT dataset.

<sup>1</sup><https://orbbec3d.com/product-astra-pro/> (last accessed on September 3rd 2021)

<sup>2</sup><https://www.flir.com/products/lepton/> (last accessed on September 3rd 2021)

**Table 1: Sequence overview. Available modalities (Depth, Thermal), room sizes with special characteristics and suggested subset.**

ID	Modalities	Comment	Room size (estimate)	Suggested subset
0	D+T	A crowded office space featuring three desks in a room of medium size. Many person-object and person-person occlusions.	20m <sup>2</sup>	Training
1	D+T	A hallway scene recorded at a steep downward angle. People entering and leaving the scene in multiple directions.	15m <sup>2</sup> (visible area)	Validation
2	D+T	A small breakroom area. The scene features people around a coffee table and sleeping on a couch.	15m <sup>2</sup>	Training
3	D+T	A larger breakroom area. Similar to sequence ID 2 but different environment.	20m <sup>2</sup>	Training
4	D+T	A narrow kitchen area. People standing or sitting at the window while talking.	15m <sup>2</sup>	Test
5	D+T	A larger laboratory environment populated with desks and other instruments.	30m <sup>2</sup>	Validation
6	D+T	An open corridor and stairway environment.	20m <sup>2</sup>	Training
7	D+T	A conference room featuring a large table and many positions for people to sit in.	20m <sup>2</sup>	Test
8	D+T	An office space with sloped roofs. Narrow with many person-person occlusions.	20m <sup>2</sup>	Training
9	D	A large smokers' room. Many people entering and leaving the room. Up to 6 people at the same time.	25m <sup>2</sup>	Training
10	D	A narrow, long corridor with many person-person occlusions.	12m <sup>2</sup>	Training
11	D	A small entrance area, partly out of the sensors field of view.	15m <sup>2</sup>	Validation
12	D	A short hallway area. People entering and leaving on two sides.	19m <sup>2</sup>	Training
13	D	Two people performing housework in a kitchen.	12m <sup>2</sup>	Test
14	D	A kitchen and living room area. Content similar to sequence ID 13.	20m <sup>2</sup>	Training
15	D	A large joint living room and kitchen area. Two people cooking and setting a table.	32m <sup>2</sup>	Training
16	D	A large dining room. Many large objects in the scene causing person-object occlusions.	23m <sup>2</sup>	Training
17	D	A living room and TV area.	16m <sup>2</sup>	Validation
18	D	A large living room area similar to sequence ID 17.	24m <sup>2</sup>	Test
19	D	A large living room and TV area similar to sequence ID 17 and 18.	30m <sup>2</sup>	Training