Fusing Personal and Environmental Cues for Identification and Segmentation of First-Person Camera Wearers in Third-Person Views

Supplementary Material

7. Details of the TF2023 dataset

7.1. Collection Setup

Recruitment. We recruited a total of 21 participants as actors for the TF2023 dataset. All of the recruited participants were university students and over 18 years old at the time of recording. The study protocol was reviewed and approved by our Institutional Review Board (IRB).

Hardware. For data collection, we utilized Yi Action Camera as first-person camera and the built-in camera of a MacBook for third-person camera. The recording resolutions were set at 1080×1920 and 720×1080 for first-person and third-person views, respectively, capturing video at 60 and 30 frames per second (FPS).

Activities. The participants were given instructions to perform various common social interaction activities. Examples of these activities include playing puzzle games, giving presentations, discussing questions on a whiteboard, and taking snack breaks. Typically, one individual wearing the first-person camera took a central role in the interaction, while the other camera wearer did not. For instance, in a presentation scenario, one camera wearer would play the role of the presenter, and the other would act as one of the observers. Recording locations encompassed both indoor and outdoor settings, including labs, classrooms, houses, and walkways. To enhance diversity, participants were also instructed to wear different outfits if they appeared in multiple scenes.

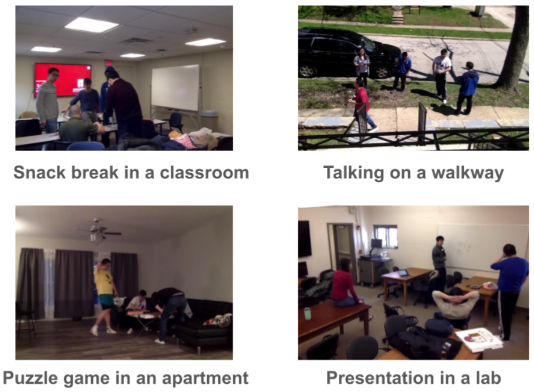
7.2. Post processing

We collected a total of 35 videos, with durations ranging from 5 to 9.5 minutes each. Frame extraction was performed at a rate of 5 frames per second, the same as IUShareView [52]. Each frame consisted of one third-person view, synchronized with two corresponding first-person views. Additionally, the third-person view contained 3-6 segmentation masks, each associated with labels denoting person IDs. An illustrative example is presented in Figure 9 and Figure 10.

All of the frames in TF2023 were hand-labeled by our annotators. We used a script that allowed the annotators to propagate masks from the preceding frame and then make adjustments. Subsequently, two members of our group conducted a comprehensive quality control check on all annotated frames. We accepted the annotations only when both members confirmed the results.

Dataset	IUShareView	TF2023
Number of frames	552	49860
Egoview-mask pairs	2404	296243
Total number of actors	6	21
Avg. actors per scene	2.18	4.29

Table 4. **Quantitative Comparison: IUShareView vs. TF2023.** Egoview-mask pair is the basic unit we used during training, previously referred to as ”combination” in section 3.


Snack break in a classroom

Talking on a walkway

Puzzle game in an apartment

Presentation in a lab

Figure 8. **Sample scenes.**

7.3. Comparison with IUShareView

A quantitative comparison between TF2023 and IUShareView is shown in Tab. 4. In addition to the increase in dataset size, TF2023 also introduces notable enhancements.

Firstly, each third-person view in TF2023 is paired with two synchronized first-person views, an increase from one in IUShareView. This modification aims to mitigate model bias towards camera wearer behaviors (For instance, camera wearers' views usually feature more movement). By having two camera wearers in each scene, the model is forced to focus on the task of relating the first-person view to the third-person view rather than learning binary classification based on camera wearer patterns.

Secondly, TF2023 features more complicated actor interactions compared to the predominantly eating and chatting scenes in IUShareView. In addition, we allowed all actors to move around the environment instead of being stationary.

Furthermore, in TF2023, we carefully partitioned the training and testing sets such that the same scene does not appear in both sets, and an actor does not appear in both sets as a camera wearer.

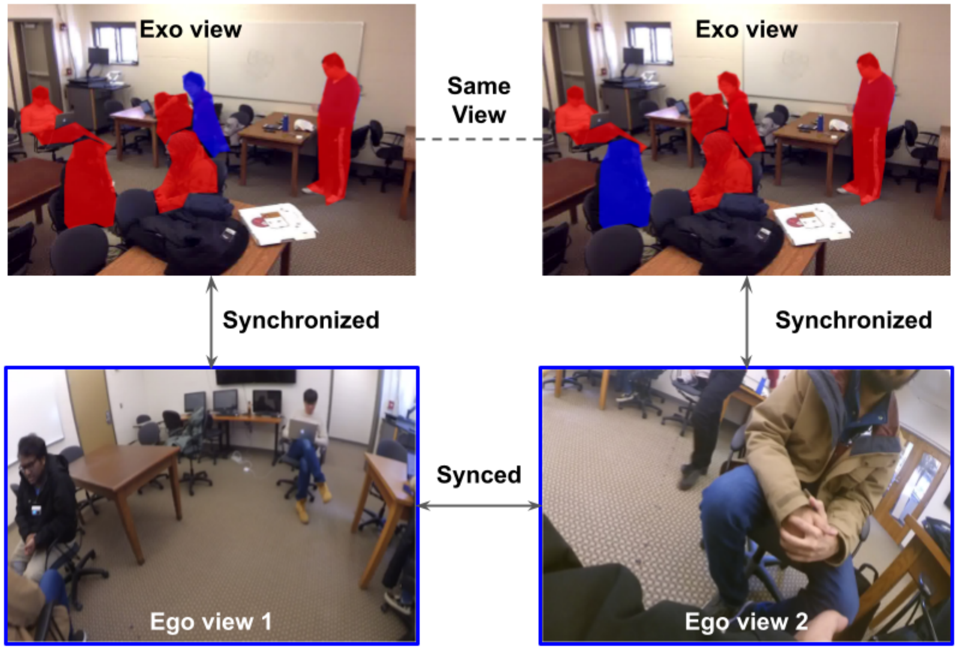


Figure 9. **Annotation samples.** For each frame, our annotators created segmentation masks for all actors in the third-person view. Each segmentation mask is labeled with a personal ID number to denote its alignment with the first-person views. In this illustration, the masks associated with the first-person views are highlighted in blue.

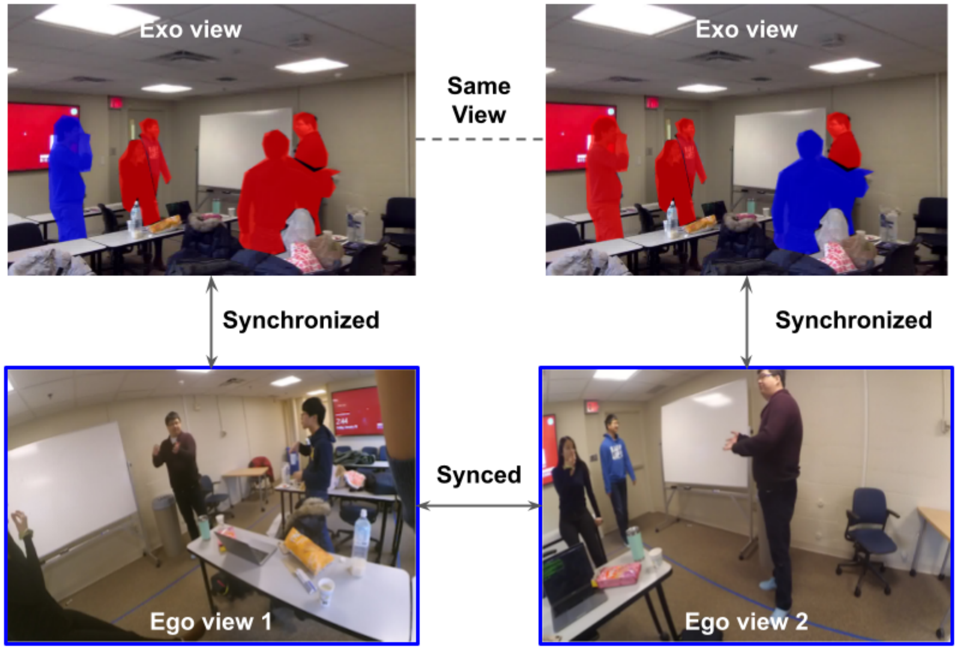


Figure 10. **Annotation samples.** Annotation samples of another scene, the annotation logic is the same as Fig. 9

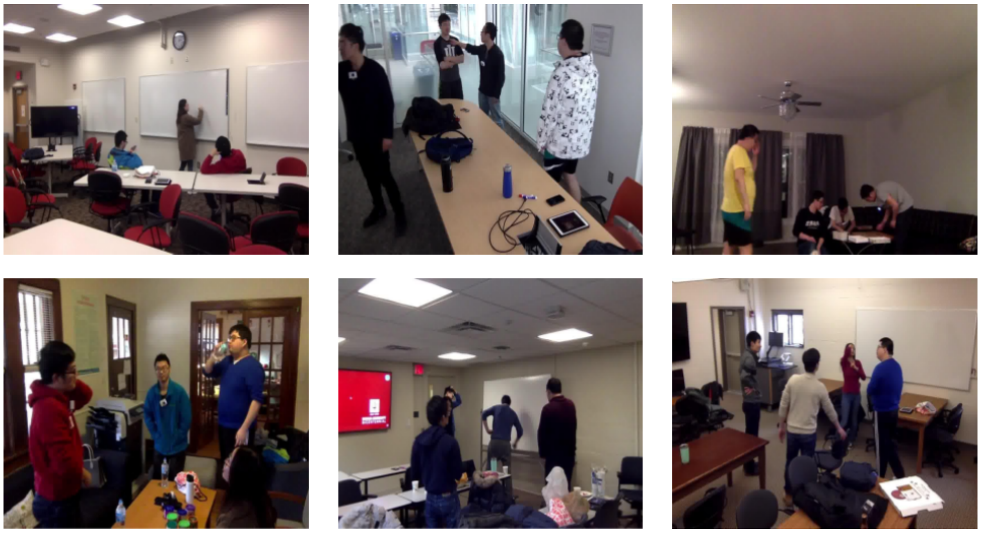


Figure 11. **Dataset samples** (Third-person views)

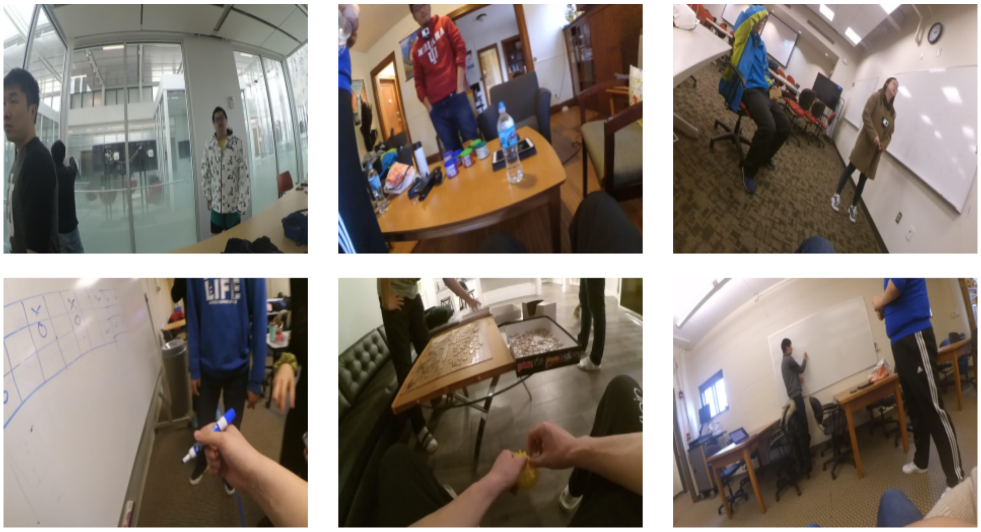


Figure 12. **Dataset samples** (First-person views)