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

Ziwei Zhao*, Yuchen Wang*, Chuhua Wang
Indiana University Bloomington
{zz47,wang617,cw234}@iu.edu

Abstract

As wearable cameras become more popular, an important question emerges: how to identify camera wearers within the perspective of conventional static cameras. The drastic difference between first-person (egocentric) and third-person (exocentric) camera views makes this a challenging task. We present PersonEnvironmentNet (PEN), a framework designed to integrate information from both the individuals in the two views and geometric cues inferred from the background environment. To facilitate research in this direction, we also present TF2023, a novel dataset comprising synchronized first-person and third-person views, along with masks of camera wearers and labels associating these masks with the respective first-person views. In addition, we propose a novel quantitative metric designed to measure a model's ability to comprehend the relationship between the two views. Our experiments reveal that PEN outperforms existing methods. The code and dataset are available at https://github.com/ziweizhao1993/PEN.

1. Introduction

Egocentric (first-person) computer vision has received increased attention in the last few years, propelled by new large-scale datasets [16, 41] and expanding capabilities of Augmented Reality (AR) and Virtual Reality (VR) technologies [1, 26, 51], including devices such as head-mounted displays and smart glasses. Egocentric vision research includes a wide variety of problems such as action recognition [43, 48], view synthesis [13, 29, 30], and temporal alignment [43, 53].

In this work, we consider a relatively underexplored task: identifying and segmenting people wearing first-person (egocentric) cameras. Specifically, our objective is to identify and segment camera wearers given a third-person view of a scene and several first-person views indicating their identity, synchronized with the third-person view, as depicted in Fig. 1. This task is important for many applications in which multiple first- and third-person cameras are in the same environment. For example, in immersive teaching environments using virtual reality, a teacher may need to promptly identify a student encountering a question within their view. Similarly, in security scenarios, identifying a police officer in a surveillance camera becomes crucial when their body camera detects anomalies.

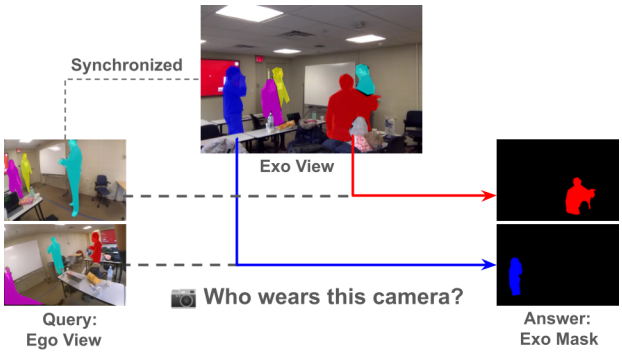

Figure 1. **Problem definition**: Given a third-person (exo) view and one or more first-person (ego) views synchronized with the third-person view, the objective is to predict the segmentation masks of the camera wearers associated with the egocentric views.

This task poses unique challenges compared to conventional identification or re-identification tasks, where a person's appearance is usually the key evidence [7, 54, 57]. First, the camera wearer's appearance is not visible within their own field of view (FOV). Second, the camera wearer's limited, forward-facing viewpoint may not align with the exocentric camera's FOV. In addition, the presence of the camera wearer in the third-person view introduces occlusions, further complicating the extraction of shared visual cues.

While prior studies [14, 52] have explored related topics, our task is different from them in several ways. First, we do not rely on any information from previous frames. Prior research [52] necessitates the use of such information in two forms: the ground truth mask of the camera wearer in

*These authors contributed equally.

the preceding frame (referred to as the "pre-mask") as well as motion information in both the first-person view and the third-person view in the form of optical flow. We find it unrealistic to assume access to such information in many real-world scenarios. This could be due to various reasons, such as the unavailability of this information due to occlusions or privacy concerns, the high cost of storing and accessing this data, or the need for the model to run fully autonomously.

Second, prior work [14, 52] uses body-worn cameras as "third-person" views. However, this does not accurately reflect the characteristics of real-world third-person cameras, which typically have the capability to view an entire scene from a higher vantage point.

Third, our research focuses on scenarios where multiple camera wearers are present in the same scene. This requires that our model not only identify who wears a camera but also associates the camera wearer candidates with their respective first-person views.

Modern computer vision backbones, such as Vision Transformer (ViT) [11] and Swin Transformer [32], have demonstrated the capability to effectively memorize the patterns of camera wearer appearances, scene backgrounds, and other cues in the third-person view. Consequently, as we show, they can predict the identity of the camera wearer in the largest existing dataset for this task (IUShareView [52]) without relying on the first-person view, which should be impossible. This suggests that the model is overfitting to the third-person input. This observation emphasizes the need for both a more complex and challenging dataset and a method for quantifying the extent to which a model uses information from the egocentric view, or conversely, how much it overfits to the third-person input.

Our contributions can be summarized as follows:

- We present a new dataset, comprising synchronized first-person and third-person views, accompanied by masks corresponding to each camera wearer and other actors. Our dataset surpasses the current state-of-the-art by two orders of magnitude in terms of the total number of frames and masks. Additionally, it is much more complex in terms of actor interaction and the average number of actors present in each scene.
- We introduce a novel quantitative metric (**EgoRate**), designed to assess the degree to which a model leverages information from the egocentric view.
- We propose a novel model that not only outperforms prior work and baseline methods but also achieves a higher performance on our introduced quantitative metric.

2. Related Work

Joint first and third video understanding. There has been significant progress in bridging the gap between first-person (egocentric) and third-person (exocentric) video understanding. Several studies have focused on relating visual representations across these two views to enhance various tasks. Interesting work includes considering both first- and third-person views for understanding human actions [5, 15, 43, 46, 48, 49], generating first-person video sequences from third-person views [13, 30, 31], temporal alignment to identify the corresponding frame in different views [43, 53], and retrieving the correct exocentric video given an egocentric video query and vice versa [6, 39]. Work in robotics [27, 35] has investigated integrating egocentric and exocentric views to identify "action possibilities" in objects and to enable precise robotic manipulation, while [10, 50] study pose estimation that uses cross-view information as a guide. Other work includes summarizing first-person videos from third-person views [25], using third-person data to improve egocentric vision models [28], and using the predicted region of attention (ROA) of both first- and third-person videos to guide co-segmentation [56].

Person identification and segmentation has been studied using both egocentric and exocentric views, such as cross-view matching involving a top-view camera [2–4, 18, 20–22]. This camera setting provides less appearance information but offers a better angle to overlook the entire scene. Our research builds upon the task of identifying camera wearers in a side-view third-person camera, as initially introduced in [14, 52]. Fan et al. [14] take into account both spatial and temporal information from both perspectives and create a joint embedding space from first- and third-person matches. Xu et al. [52] introduces a segmentation task and runs person identification and segmentation concurrently. Our methods overcome their limitation of requiring information from past frames. In addition, we enhance the framework to accommodate multiple camera wearers simultaneously in the same third-person view.

Cross-view matching. Cross-view matching involves aligning and identifying the corresponding elements between different perspectives to improve spatial understanding. Recent work has revolutionized feature point matching with Graph Neural Network (GNN) and attention mechanisms. SuperGlue [40] employs GNNs in processing keypoint descriptors from dual images, enhanced with positional encodings to distinguish similar patterns. LoFTR [44] utilizes both self-attention and cross-attention to update cross-view features, with additional improvements introduced by [9]. Meanwhile, [8, 42] focus on improving efficiency, while [12, 45] attempt to bridge the gap between dense and sparse methods. GlueStick [38] jointly establishes correspondences between points and lines. In our work, we choose GlueStick as the matching model between first- and third-person views due to its superior performance when compared to other available methodologies.

Egocentric and exocentric datasets. Ego4D [16] captures a wide range of daily activities around the world, enriched with densely narrated videos and a variety of annota-

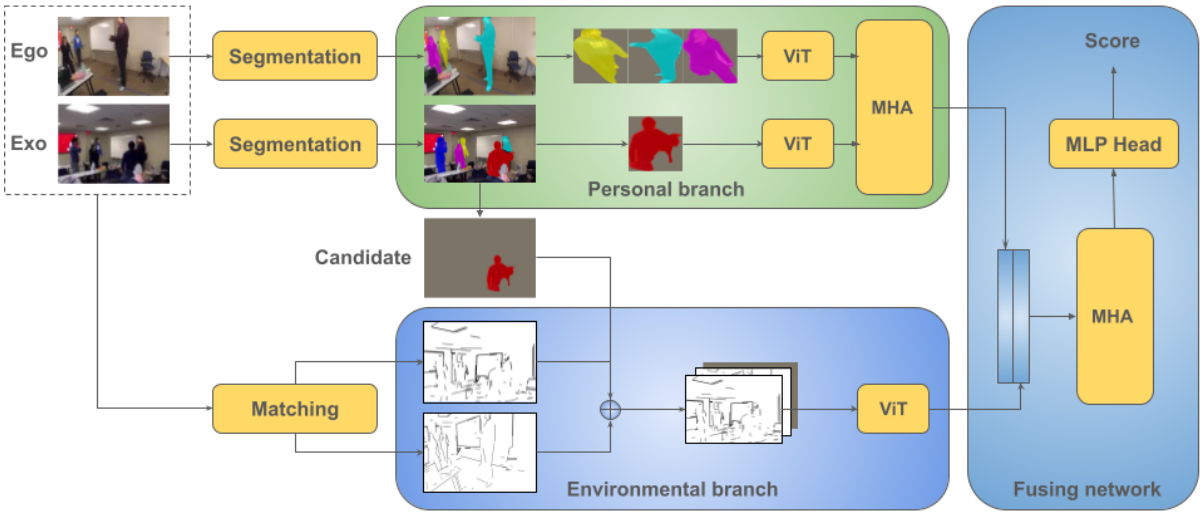

Figure 2. Overview of the PersonEnvironmentNet (PEN). Given a first-person (ego) view and a third-person (exo) view, the personal branch analyzes the relationship of individuals. Concurrently, the environmental branch utilizes a matching model to understand geometric cues. Finally, the outputs from the individual and environmental branches are integrated through feature fusion. The final score is a numerical value ranging from 0 to 1, indicating the confidence score that the candidate mask is associated with the query first-person view.

tions. Ego-Exo4D [17] is a large scale video dataset with synchronized egocentric and exocentric views. UTokyo Ego-Surf [55] involves groups of two or three people wearing cameras in both indoor and outdoor settings while engaging in face-to-face conversations. IUShareView [14] includes first-person videos with 3-4 participants performing everyday activities in various indoor settings. Elfeki et al. [13] present a dataset containing 531 temporally aligned egocentric and exocentric video pairs of an actor performing a range of actions. Charades-Ego [43] encompasses a collection of first- and third-person videos with temporal annotations and action classes, while Han [19] synchronizes top and horizontal views and label subjects with bounding boxes and ID numbers. Assembly101 [41] contains videos of participants assembling and disassembling toy vehicles, notable for its multi-view recordings. CvMHAT [23] consists of synchronous drone-mounted cameras and multiple horizontal-view cameras.

3. Methodology

Given the first-person (egocentric) view of one or more camera wearers, our objective is to identify and segment these individuals, as observed by a third-person camera positioned to capture the entire scene. Our work differs from previous research [52] in multiple ways: Firstly, we have eliminated the dependency of the camera wearer's mask in the previous frame (pre-mask), a condition not always met in real-world scenarios. Pre-mask also gives the model more information than necessary, given that the mask of the same person typically undergoes minimal changes between successive frames. Secondly, prior work [52] used one of the two wearable cameras as the third-person view, limiting each third-person view to only one camera wearer. In contrast, our task can have multiple camera wearers in the same third-person view. This means that we cannot rely on a simple binary classification for each person to determine if they are camera wearers. Instead, our methodology involves predicting a score for each candidate based on each first-person view. Furthermore, our third-person camera configuration more closely resembles typical real-world camera angles, offering a comprehensive overview of the entire scene.

In this section, we first introduce a two-stream baseline for this task (Sec. 3.1). Subsequently, we present a novel model: PersonEnvironmentNet (PEN), which is comprised of three modules: a personal branch (Sec. 3.2) that learns the relationship between individuals in the two views, an environmental branch (Sec. 3.3) designed to capture geometric cues in the background environment, and a fusing model (Sec. 3.4) that integrates information from the two preceding branches.

3.1. Two-stream Baseline

First, we introduce a two-stream baseline for this task. This design shares similarities with [52], but utilizes more recent, transformer-based vision backbones. This baseline allows us to demonstrate a traditional model design for this task and evaluate other methodologies without the confounding variable of backbone performance.

We apply two backbones (e.g., ViT), one dedicated for first-person input, and the other one for third-person input, as depicted in Fig. 3. We modify the first convolution layer of the third-person ViT to accept 4-channel input, as explained below.

During training, the model is given combinations of a third-person view, a first-person view, and a mask for a can-

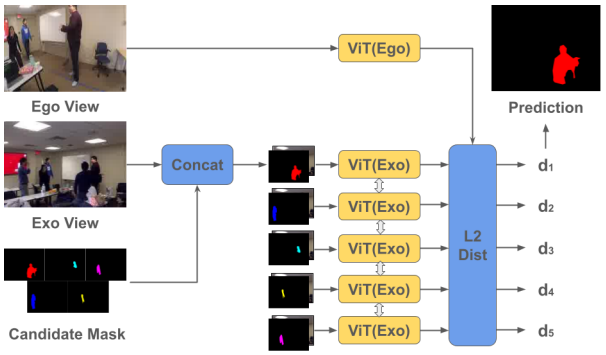

Figure 3. **Structure and inference pipeline of the two-stream baseline.** Given a first-person (ego) view, a third-person (exo) view, and a pool of candidate masks, the model selects the mask with the lowest L2 distance between the output features.

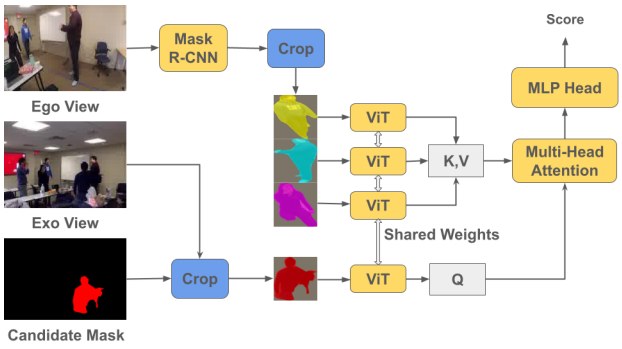

Figure 4. **Structure of the personal branch**. Given a first-person (ego) view, a third-person (exo) view, and a candidate mask, we apply a Mask R-CNN to find individuals in the ego view and establish their connection to the candidate in the exo view.

didate in the third-person view. These combinations can be either positive or negative; a positive combination signifies that the candidate's mask corresponds to the camera wearer whom the first-person view belongs to, while a negative combination indicates otherwise. To process these inputs, we start by concatenating the candidate's mask with the third-person input, creating a 4-channel input. Subsequently, the 3-channel first-person input and the 4-channel third-person input are individually passed through their respective backbones. Finally, we apply a contrastive loss on the outputs,

$$L_{\text{contrastive}} = \frac{1}{2N}\sum_{i=1}^{N} y \cdot d_i^2 + (1-y)\cdot \max(\alpha - d_i, 0)^2. \quad (1)$$

Here N is the total number of elements in the output vectors, d_i indicates the difference of the i_{th} element between the first-person output and the third-person output, and y is a binary indicator variable that is 1 if the combination is positive, and 0 if negative. α is a constant margin.

During inference, each mask candidate, in combination with both the first-person view and the third-person view, is fed into the two-stream model. The mask candidate with the smallest L2 distance between the first-person output and the third-person output is selected as the final prediction. The structure and inference pipeline of the two-stream baseline is shown in Fig. 3.

3.2. Personal Branch

While the two-stream baseline method implicitly learns the connection between first-person and third-person perspectives, it is vulnerable to overfitting on the third-person input. To address this issue, we propose our novel model, Person-EnvironmentNet (PEN). Our approach places greater emphasis on integrating cues from diverse sources between the first-person and third-person viewpoints, thereby enhancing overall performance and reducing the over-dependency on third-person input.

To begin, we introduce a personal branch, as illustrated in Fig. 4. The objective of this branch is to establish a connection between the two views by identifying the individuals present in them. Our input configuration is the same as the two-stream baseline, incorporating a first-person view, a third-person view, and a candidate mask. During training, we first utilize a Mask R-CNN to generate masks and extract all visible individuals from the first-person view. Simultaneously, we use the third-person view and the candidate mask to crop out the candidate in the third-person view. Subsequently, we resize all extracted individuals to the same size and pass them through the same ViT to extract their features. We then apply a multi-head attention layer [47], which learns the relationship between the appearance of every individual in the first-person view of the camera wearer and the appearance of the third-person candidate. This mechanism effectively captures the likelihood of the candidate being able to see these individuals. The result is then fed through a multi-layer perceptron (MLP) head followed by a Sigmoid activation function. The final output is a numerical value between 0 and 1, indicating the confidence score that this candidate is associated with the queried first-person view. Finally, we apply binary cross-entropy loss (BCE loss) to the output, where a target value of 1 denotes a positive combination and 0 for a negative combination,

$$L_{\text{BCE}} = -y \cdot \log(p) + (1-y) \cdot \log(1-p). \quad (2)$$

In the above loss function, y denotes the target value associated with the combination, while p signifies the output score generated by the personal matching branch.

3.3. Environmental Branch

Considering the motion-intensive characteristics of egocentric cameras, we find it important to explore multiple sources of input. For instance, the camera wearer's field of view may encompass only a small subset of all candidates visible in the third-person view. To augment the personal

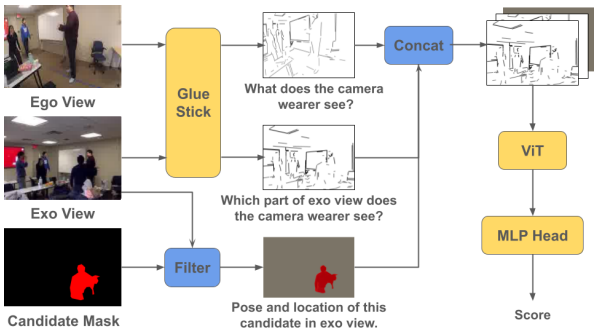

Figure 5. **Structure of the environmental branch.** Given a first-person (ego) view, a third-person (exo) view, and a candidate mask, the environmental branch applies GlueStick to match the two views and infer geometric cues.

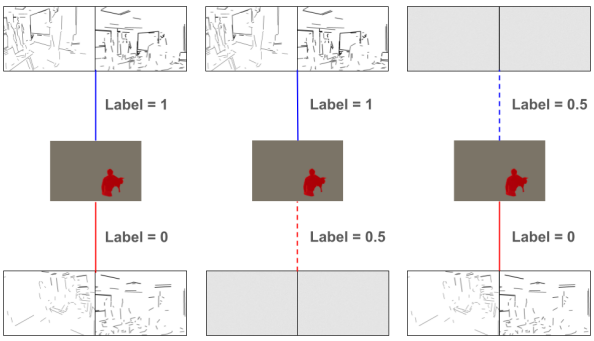

Figure 6. **Adding "ghost samples."** When a candidate does not have an associated positive combination or negative combination, we generate a random mask with label 0.5 to discourage the model from overfitting on the candidate's appearance.

branch outlined in Sec. 3.2, we introduce a novel environmental branch that leverages the information in the background environment from both first- and third-person views to deduce geometric cues between the two perspectives.

To accomplish this, we first apply a matching network on the two views. Specifically, we use GlueStick [38], which applies an attention-based Graph Neural Network (GNN) on a set of points, their associated descriptors, and a set of line segments connecting these points.

Subsequently, we utilize the matched lines generated by Gluestick to produce masks for both the first-person and third-person views, with the line intensity representing confidence. We also use the third-person view and the candidate mask to filter out the background of the third-person view, retaining only the candidate, as shown in Fig. 5. Then, we concatenate the two line masks and the filtered candidate along the channel dimension and apply a Vision Transformer (ViT) and a Multi-Layer Perceptron (MLP) head with a Sigmoid activation function, similar to the structure of the personal branch outlined in Sec. 3.2. The underlying concept of this module is that the two line masks signify which part of the third-person view the camera wearer can perceive, while the filtered candidate denotes its pose and location in the third-person view. The loss function of the environmental branch follows Eq. (2), with p denoting the output score of the environmental branch.

To enhance the training process, we involve the inclusion of "ghost samples" for this task. This measure is particularly important when certain candidates appear exclusively as either negative or positive candidates, where 'always negative' signifies a lack of an associated first-person view, while 'always positive' indicates that they only appear as camera wearers. Consequently, the model may learn to predict these candidates by memorizing their appearances, undermining the objective of understanding the relationship between the first-person and third-person views. To address this issue, we introduce "ghost samples" for these candidates. These ghost samples have randomly generated masks and are assigned a label of 0.5, signifying a state of neither positive nor negative. This action forces the model to learn from the environmental cues (lines) rather than relying solely on the visual appearance of the candidates, as visualized in Fig. 6.

3.4. Fusing Network

The two branches introduced in Sec. 3.2 and Sec. 3.3 both exhibit limitations when operating independently. For instance, the personal branch may struggle to gather sufficient information to deduce relationships between the two views when the Field Of View (FOV) of the first-person camera lacks a sufficient number of individuals. Additionally, disparities between the first-person and third-person views present challenges for the environmental branch, as matching networks such as Gluestick are designed to match views with adequate shared FOVs. However, in the case of first-person and third-person, the geometric disparity between the two views can be too significant, causing the matching network to fail. An extreme instance occurs when the camera wearer is looking in the direction of the third-person camera, resulting in a 180-degree difference and no shared FOV between the two views.

To address these limitations, we introduce a fusing network to capitalize on the information from both sources harnessed by the two branches in the PEN model. This is achieved by removing the MLP heads of the two branches, concatenating their output features along the channel dimension, and applying a multi-head attention layer, followed by an MLP head and Sigmoid activation. The final output of the fusing network is a numerical value ranging between 0 and 1, representing the confidence score that the candidate mask is associated with the query first-person view. We then apply the loss function in Eq. (2) on the output of the fusing model. A visual representation of the full PEN model is shown in Fig. 2.

4. Experiments

First, we evaluated the two-stream baseline on IUShareView [52], the previously largest dataset available for this task. We implemented the two-stream baseline with ViT backbones using the PyTorch framework [37], and we utilized pre-trained weights from TorchVision [36]. We trained this baseline on IUShareView for a total of 100 epochs, with a batch size of 32, and an empirically set learning rate of 6×10^{-5}. During inference, we provided the model with combinations of a first-person view, a third-person view, and several mask candidates provided by the IUShareView dataset. Then we selected the mask candidate with the minimal L2 distance between the outputs of the first-person stream and the third-person stream. We evaluated the performance using the accuracy metric from [52].

The two-stream baseline achieved an astounding 99.3% accuracy. However, further investigation revealed that the model did not truly capture the relationship between first- and third-person as we hoped. Instead, it achieved this high accuracy by overfitting solely to the third-person input.

In the context of this work, we define overfitting as the scenario in which a model excels on familiar data (e.g., IUShareView) but fails to generalize to new observations (e.g., real-world applications). This phenomenon often occurs when a model memorizes specific patterns or biases in the dataset, and may be challenging to detect through conventional testing methods, as the testing set may share the same dataset biases as the training sets.

To illustrate this issue, we introduce an evaluation mechanism that substitutes all first-person inputs with random tensors generated from a normal distribution. The underlying concept is that if the model had truly learned the relationship between first- and third-person perspectives, replacing the first-person input with random tensors would lead to model failure, resulting in random guesses.

Furthermore, we propose a novel quantitative metric, **Egocentric Utilization Rate (EgoRate)** to evaluate how much a model utilizes information from the first-person view,

$$EgoRate = 1 - \frac{Acc_{overfit} - Acc_{guess}}{Acc - Acc_{guess}}. \quad (3)$$

In the equation above, Acc represents the accuracy of the normal test, Acc_{guess} denotes the accuracy of a random guessing model, and $Acc_{overfit}$ signifies the accuracy of the model when first-person inputs are replaced with randomly generated tensors.

Essentially, **EgoRate** quantifies the extent to which performance improvement, when compared to random guessing, can be attributed to the model's understanding of the connection between first-person and third-person views. To clarify, let's consider two extreme scenarios.

- In the case of a model that entirely overfits on the third-person input, meaning that it can identify the camera wearer in the third-person view without utilizing any information from the first-person view, its overfit accuracy matches the normal accuracy, resulting in an EgoRate of 0. This indicates that the model's performance is solely achieved by overfitting on the third-person input.
- Conversely, if a model fully leverages the relationship between first-person and third-person perspectives, its overfit accuracy drops to random guessing, leading to an EgoRate of 1. This suggests that the model's performance can be entirely attributed to its comprehension of the link between these two views.

Therefore, EgoRate serves as a valuable metric for quantifying the model's ability to comprehend the relationship between first-person and third-person viewpoints, instead of overfitting on third-person input only.

4.1. Dataset

The two-stream baseline was able to achieve a 99.3% accuracy on the IUShareView [52] dataset due to both the high performance of the Vision Transformer (ViT) backbone, and the limitation of size and diversity of the IUShareView dataset. To address this, we introduced a novel dataset, TF2023, and primarily conducted the evaluation of our proposed model (PEN) on this new dataset.

IUShareView dataset was released in [52] as the largest dataset for the first-person and third-person cross-view matching and segmentation problem. It contains 1824 pairs of cross-view synchronized images for training and 580 pairs for testing.

TF2023 dataset was collected and annotated by us with the aim of better understanding the cross-view relationship in egocentric videos. The actors and scenes were carefully partitioned into training and testing sets, ensuring that no camera wearer appearing in the training set appears in the test set. Moreover, the scenes in the testing set differed from those in the training set. For TF2023, we collected and annotated 208,794 pairs of cross-view synchronized images for training and 87,449 pairs for testing.

On average, each scene in TF2023 featured 4.29 actors in the third-person view, surpassing the average of 2.18 actors in IUShareView. Furthermore, TF2023 included more complicated participant interactions including puzzle games and presentation scenes, in contrast to the predominantly eating and chatting activities in IUShareView. We believe that the increase in both scale and complexity will make TF2023 particularly valuable for training and evaluating cross-view matching problems in the egocentric vision community, especially when utilizing larger models.

Methodology	Backbone	Loss function	IoU↑	Acc↑(%)	EgoRate↑(IoU)	EgoRate↑(Acc)
Random guess	N/A	N/A	0.21	25.2	N/A	N/A
Third-first [52]	FCN [33]	Contrastive	0.26	31.1	0.52	0.52
Two-Stream baseline	ViT	Contrastive	0.33	39.9	0.31	0.32
Two-Stream baseline	ViT	BCE loss	0.27	31.9	0.24	0.31
Two-Stream baseline	Swin	Contrastive	0.31	36.9	0.11	0.11
Two-Stream baseline	Swin	BCE loss	0.26	31.4	0.11	0.05
Personal branch	ViT	BCE loss	0.40	48.9	0.95	0.94
Environmental branch	ViT	BCE loss	0.39	47.2	0.95	0.95
Fusing (Add)	ViT	BCE loss	0.43	52.5	0.95	0.94
Fusing (Concat)	ViT	BCE loss	0.43	52.4	0.96	0.96
Fusing (Self-attention)(PEN)	ViT	BCE loss	**0.44**	**52.9**	**0.97**	**0.96**

Table 1. **Experiment results on TF2023.** Our method (PEN) achieved the highest performance in IoU, Acc, and EgoRate.

4.2. Experimental Results

Setup We conducted evaluations of our model (PEN) on our proposed dataset, TF2023. We implemented our model on PyTorch, and all ViTs in our model utilized pre-trained weights from the torchvision library. All experiments were conducted on a single Nvidia Tesla V100s 32GB GPU.

Our training process comprised multiple stages. Initially, we fixed the ViT backbone of the personal branch and trained the multi-head attention layer and MLP head for 3 epochs. This phase was conducted with a batch size of 32, and we used the AdamW optimizer [34] with a learning rate of 6×10^{-5}. Subsequently, we unfroze the ViT backbone and continued training with the same hyperparameters for an additional 30 epochs.

Next, we trained the environmental branch for 30 epochs, using a batch size of 64 and the AdamW optimizer with a learning rate of 6×10^{-5}.

Finally, we froze both the personal branch and environmental branch backbones and conducted fine-tuning of the self-attention layer and MLP head for 30 epochs with a batch size of 64 and the AdamW optimizer with a learning rate of 5×10^{-6}.

To evaluate the effectiveness of each methodology, we employed two metrics: Intersection over Union (IoU) and Accuracy (Acc) of the mask. For the purpose of our evaluation, a mask is considered accurate if the IoU value surpasses 0.5. In contrast to prior study [52], we used Mask R-CNN [24] to generate a pool of potential masks rather than relying on a set of ground truth masks provided by the dataset. This choice resulted in an overall reduction in the numbers across all methodologies, given that Mask R-CNN's segmentation results may miss or over-segment candidates. Nevertheless, we adopted this evaluation approach as it better simulates real-world scenarios where perfect segmentation is unattainable.

Specifically, we utilized a pre-trained Mask R-CNN with a score threshold of 0.9 to generate a pool of candidate masks for evaluation. The pool of candidate masks is shared for all evaluated methods to ensure fairness.

Comparison with other methods Given the novelty of our research question, we were unable to identify any existing state-of-the-art method for direct comparison. The most closely related method we found is the "third-first" model introduced in [52]. The third-first model was designed to perform both segmentation and identification given the ground truth mask in the preceding frame (referred to as the pre-mask). For the purpose of our evaluation, we excluded the segmentation branch of the third-first model, as it relies on the availability of the ground truth pre-mask. Furthermore, it is rare for the mask of the same candidate to change significantly between two consecutive frames. Nevertheless, we adopted the third-first model as a baseline by modifying its identification branch to accept the candidate mask from the current frame as input, allowing it to predict whether it corresponds to the first-person view.

We also compared our model against the two-stream baseline introduced in Sec. 3.1, using two backbones: Vision Transformer (ViT) [11] and Swin Transformer [32]. Since our model (PEN) does not contain a two-stream structure, we utilized Binary Cross-Entropy loss (BCE loss) instead of the contrastive loss used in the third-first model [52] and the two-stream baselines. To verify that the observed performance improvement was not solely attributed to this change in the loss function, we also adjusted the two-stream baseline model to utilize BCE loss. This was achieved by concatenating the output from the first-person stream and the third-person stream, followed by attaching an MLP head with Sigmoid activation. We implemented a random guess model by randomly selecting a candidate mask generated by a pre-trained Mask R-CNN as output.

The evaluation results on TF2023 are presented in Tab. 1. To provide a more comprehensive assessment, we also extended our experiments to include the IUShareView dataset. As shown in Tab. 2, the two-stream baseline demonstrated

IUShareView	Acc(%)	EgoRate(Acc)
Random Guess	50	N/A
Third-first [52]	74.5	0.20
Two-stream (ViT)	99.3	0.0
PEN(Ours)	99.7	0.77

Table 2. Evaluation results on IUShareView.

TF2023	IoU	Acc(%)	EgoRate(IoU/Acc)
Random Guess	0.21	25.2	N/A
Two-stream (ViT)	0.33	39.9	0.31/0.32
Without ghost case	0.39	47.1	0.67/0.65
With ghost cases	0.39	47.2	0.95/0.95

Table 3. Effect of adding ghost cases for the environmental branch.

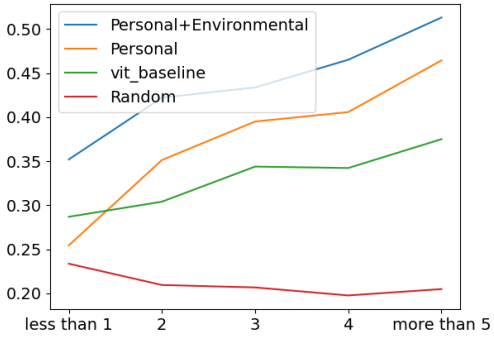

Figure 7. Variation in performance (IoU) with different numbers of people visible in the first-person view.

that a high accuracy can be achieved by overfitting on the third-person input. PEN was able to achieve similar performance while maintaining a high EgoRate.

4.3. Ablation Studies

Tab. 1 illustrates the effectiveness of each module within the PEN model. Both the personal branch and the environmental branch demonstrate a significant improvement in both IoU and Acc performance, compared to the baselines. Furthermore, we observed a further enhancement in performance when these two branches were combined via the fusing network. Additionally, as depicted in Figure 7, adding the environmental branch to the personal branch proves effective, especially when the number of visible people in the first-person view is low.

Our evaluation also includes the evaluation of three fusion methodologies: self-attention fusing, concatenate fusion, and add fusing. The latter two methods were implemented by concatenating or adding the outputs of the two branches before the MLP head. As shown in Tab. 1, the variation in performance across these methods is minor, indicating a consistent enhancement in performance through the fusion of output features from the personal and environmental branches. We selected the self-attention fusion approach as our final design, due to its superior performance during our evaluation.

The impact of incorporating ghost samples during the training of the environmental branch is presented in Tab. 3. Compared to conventional training, the inclusion of ghost samples did not yield substantial improvements in IoU or Acc. However, it led to a significant increase in the EgoRate metric, affirming its effectiveness in discouraging the model from overfitting on third-person input.

5. Conclusion

In this paper, we addressed the challenge of identifying and segmenting camera wearers in a third-person view, given their associated first-person views. We introduced a novel dataset, TF2023, comprising synchronized first-person and third-person frames, segmentation masks for all individuals in third-person frames, and labels associating them with the first-person frames. Compared to previous datasets, TF2023 is larger in terms of the number of frames and masks, while also containing more complex scenes and interactions.

Furthermore, we proposed a new quantitative metric, EgoRate, designed for assessing the model's tendency to overfit on third-person input.

In addition, we presented a novel method that integrates information from two sources: a personal branch dedicated to matching the individuals in the two views, and an environmental branch focusing on geometric cues. Our evaluation showed our proposed method outperformed previous state-of-the-art on both TF2023 and existing dataset, while also exhibiting a higher EgoRate.

6. Acknowledgements

The authors gratefully acknowledge Prof. David Crandall's guidance and feedback on earlier versions of this work. This work was supported in part by the National Science Foundation under award DRL-2112635 to the AI Institute for Engaged Learning. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Asmaa Saeed Alqahtani, Lamya Foaud Daghestani, and Lamiaa Fattouh Ibrahim. Environments and system types of virtual reality technology in stem: A survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(6), 2017. 1

[2] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 253–268. Springer, 2016. 2

[3] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[4] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1353–1366, 2018. 2

[5] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018. 2

[6] Shervin Ardeshir and Krishna Regmi. Egotransfer: Transferring motion across egocentric and exocentric domains using deep neural networks, 2016. 2

[7] Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and vision computing*, 32(4):270–286, 2014. 1

[8] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. *International Conference on Computer Vision (ICCV)*, 2021. 2

[9] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. *European Conference on Computer Vision (ECCV)*, 2022. 2

[10] Ameya Dhamanaskar, Mariella Dimiccoli, Enric Corona, Albert Pumarola, and Francesc Moreno-Noguer. Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition*, 138: 109358, 2023. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 7

[12] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[13] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji. From third person to first person: Dataset and baselines for synthesis and retrieval. *arXiv preprint arXiv:1812.00104*, 2018. 1, 2, 3

[14] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017. 1, 2, 3

[15] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, Chiho Choi, and Behzad Dariush. Weakly-supervised online action segmentation in multi-view instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13780–13790, 2022. 2

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 3

[18] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. Multiple human association between top and horizontal views by matching subjects' spatial distributions. *arXiv preprint arXiv:1907.11458*, 2019. 2

[19] Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yunjun Zhang, Liang Wan, and Song Wang. Complementary-view multiple human tracking. In *AAAI Conference on Artificial Intelligence*, 2020. 3

[20] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2021. 2

[21] Ruize Han, Yiyang Gan, Jiacheng Li, Feifan Wang, Wei Feng, and Song Wang. Connecting the complementary-view videos: joint camera identification and subject association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2416–2425, 2022.

[22] Ruize Han, Yiyang Gan, Likai Wang, Nan Li, Wei Feng, and Song Wang. Relating view directions of complementary-view mobile cameras via the human shadow. *International Journal of Computer Vision*, 131(5):1106–1121, 2023. 2

[23] Ruize Han, Wei Feng, Feifan Wang, Zekun Qian, Haomin Yan, and Song Wang. Benchmarking the complementary-view multi-human association and tracking. *International Journal of Computer Vision*, 132(1):118–136, 2024. 3

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[25] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018. 2

[26] Miao Hu, Xianzhuo Luo, Jiawen Chen, Young Choon Lee, Yipeng Zhou, and Di Wu. Virtual reality: A survey of enabling technologies and its applications in iot. *Journal of Network and Computer Applications*, 178:102970, 2021. 1

[27] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, pages 1–1, 2022. 2

[28] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021. 2

[29] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1843–1847. IEEE, 2020. 1

[30] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021. 1, 2

[31] Gaowen Liu, Hugo Latapie, Ozkan Kilic, and Adam Lawrence. Parallel generative adversarial network for third-person to first-person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1917–1923, 2022. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 7

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 7

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[35] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *CVPR*, 2022. 2

[36] TorchVision maintainers and contributors. TorchVision: PyTorch's Computer Vision library, 2016. 6

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[38] Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. Gluestick: Robust image matching by sticking points and lines together. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9706–9716, 2023. 2, 5

[39] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019. 2

[40] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[41] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 1, 3

[42] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12517–12526, 2022. 2

[43] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person

videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[44] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[45] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[46] Thanh-Dat Truong and Khoa Luu. Cross-view action recognition understanding from exocentric to egocentric perspective. *arXiv preprint arXiv:2305.15699*, 2023. 2

[47] 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. 4

[48] Sagar Verma, Pravin Nagar, Divam Gupta, and Chetan Arora. Making third person techniques recognize first-person actions in egocentric videos. In *ICIP*, 2018. 1, 2

[49] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5250–5261, 2023. 2

[50] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3446–3455, 2021. 2

[51] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. Virtual reality. *Business & Information Systems Engineering*, 62:455–461, 2020. 1

[52] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652, 2018. 1, 2, 3, 6, 7, 8

[53] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *arXiv preprint arXiv:2306.05526*, 2023. 1, 2

[54] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1

[55] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Ego-surfing first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5445–5454, 2015. 3

[56] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First-and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[57] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

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.

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.

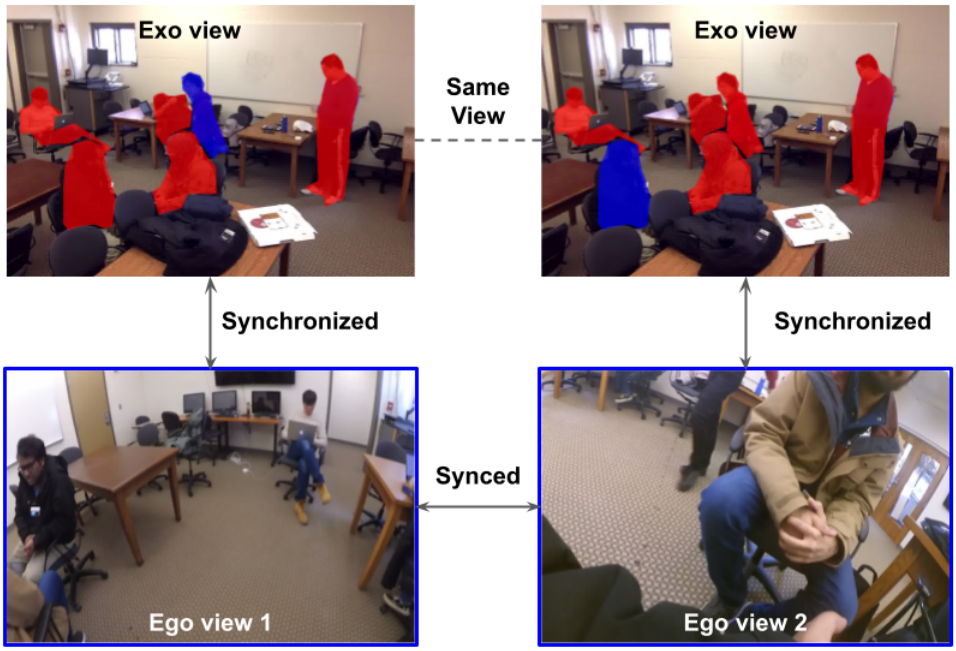

Figure 9. **Annotation samples.** For each frame, our annotators created segmentation masks for all actors in the third-person view. Each segmentation mask is labeled with a personal ID number to denote its alignment with the first-person views. In this illustration, the masks associated with the first-person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

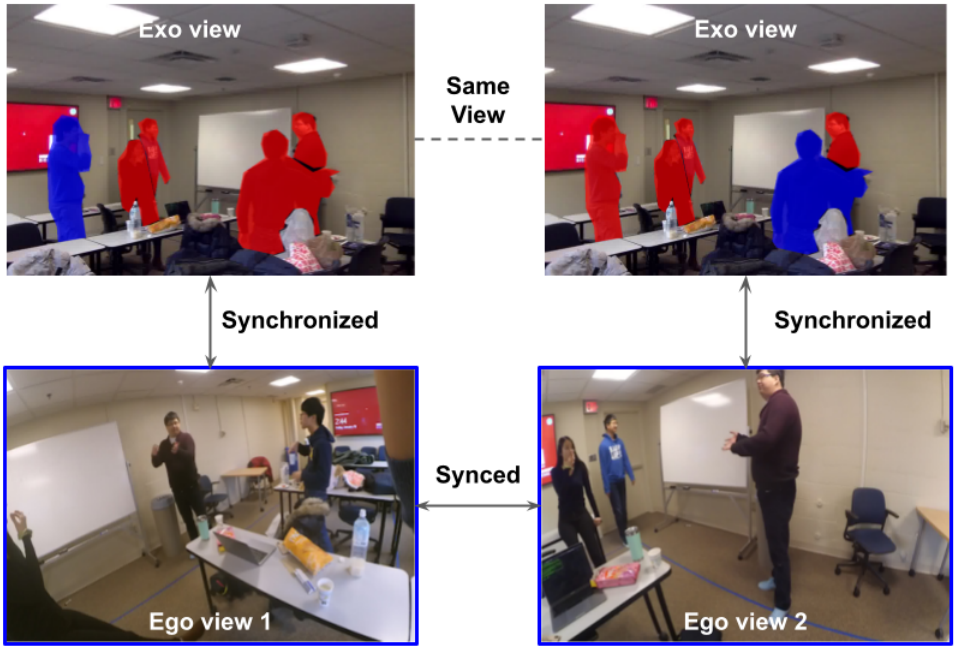

Figure 10. **Annotation samples.** Annotation samples of another scene, the annotation logic is the same as Fig. 9

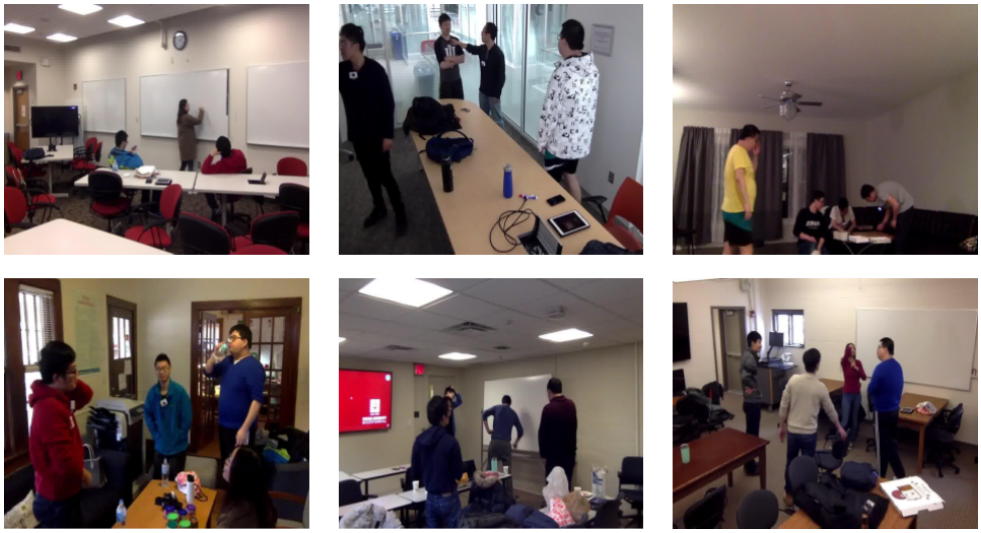

Figure 11. **Dataset samples** (Third-person views)

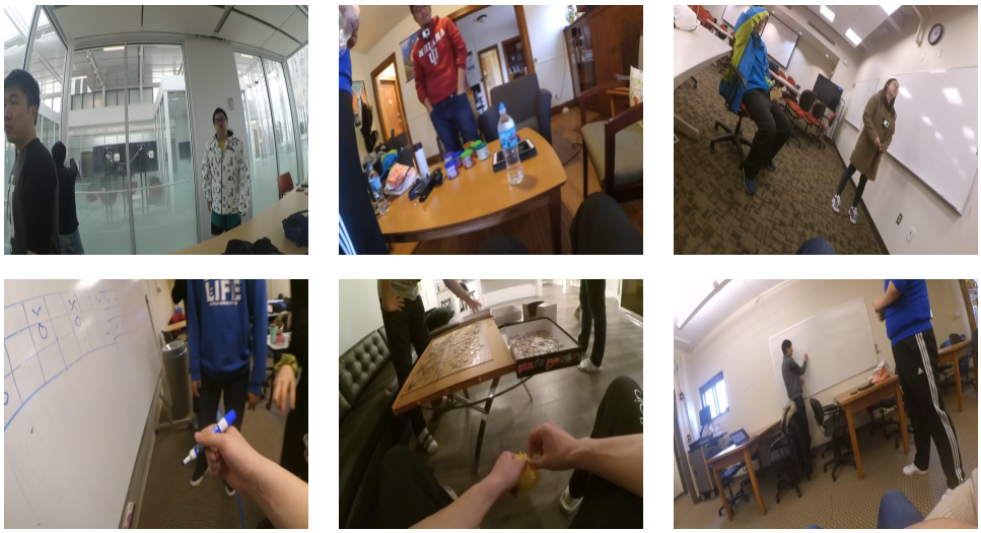

Figure 12. **Dataset samples** (First-person views)