Dynamic Group Detection using VLM-augmented Temporal Groupness Graph (Supplementary Material)

Kaname Yokoyama Chihiro Nakatani Norimichi Ukita
Toyota Technological Institute
{sd25444,sd23501,ukita}@toyota-ti.ac.jp

7. Implementation Details

7.1. Architecture

Our network mainly consists of the CLIP image encoder (ψ), the framewise trajectory encoder (ϕ), the temporal trajectory encoder (χ), and the FC network (ρ), as shown in Fig. 3 of the main paper. Regarding the framewise trajectory encoder ϕ, we employ a three-layered 1D convolution network followed by batch normalization and ReLU activation layers as with [46]. The transformer encoder [36] is employed as the temporal trajectory encoder χ. This encoder has 16 heads and 24 layers. The FC network ρ consists of two full-connection layers followed by ReLU activation layers. Layer normalization is also located in front of the two full-connection layers.

7.2. Hard Negative Training and Easy Negative Pruning at Inference for Stability and Efficiency

In our joint learning, we have an issue of imbalance between in-group and non-group pairs because most people observed in an image are in different groups. That is, most pairs are labeled to be non-group. Such imbalance induces unstable and inefficient training.

Distance-based hard negative training To avoid this imbalance issue, this paper proposes to focus on hard-negative non-group samples. This hard negative is defined based on the distance between people. In nature, in-group people are close to each other. This assumption suggests training only non-group pairs, in each of which two people are close, as hard negative samples. In our implementation, for each person of interest, k-nearest neighbor non-group people are trained as hard negatives at each frame.

Easy negative pruning Since in-group people are not largely far apart in nature, bruteforce pairing at inference is useless and inefficient. Based on this assumption, we propose the following two edge pruning schemes, distance-based pruning and feature-similarity-based pruning schemes.

Distance-based pruning. For each person, only k-nearest neighbor people are connected. $k = 2$ for experiments on JRDB and PANDA, and $k = 3$ for those on Café.

Feature-similarity-based pruning. Edges are pruned if their weights (i.e., P_g in the temporal groupness graph) are less than a pre-defined threshold, Th_e.

7.3. Training Time

The total training times for JRDB, CafeV, CafeP, and PANDA are 7h (4h for CLIP finetuning + 3h for joint learning), 17h (3h for CLIP finetuning + 14h for joint learning), 17h (3h for CLIP finetuning + 14h for joint learning), and 22h (16h for CLIP finetuning + 6h for joint learning), respectively.

8. Experiments on PANDA

As mentioned in the main paper, the PANDA dataset is not used for exeprimentas shown in the main paper because PANDA has only static group annotations. Instead, all experiments for static group detection shown in the main paper are conducted on PANDA and shown in Sec. 8.

8.1. Dataset

PANDA dataset [5] consists of 10 high-resolution quasi bird's-eye-view videos of outdoor scenes, including many people. The image resolution ranges between around 15k $\times$ 25k and 25k $\times$ 35k pixels. Each observed person in each frame is annotated with various labels such as a bounding box, a location, and occluded or not. Following [3, 5], nine and one videos are used for training and test, respectively, and 2fps videos are sampled from the original videos. As a result, 1,799 and 234 frames are sampled and used for training and test, respectively. Each sampled video observes 145 and 75 groups on average.

8.2. Details

The details different from those shown in Sec. 4.2 in the main paper are as follows.

Table 8. Quantitative results of static group detection on PANDA. The best score in each metric on each dataset is colored **red** (also in all the following Tables).

Method	Precision	Recall	F1
G2L [5]	0.293	0.160	0.207
Dis.Mat+ [7]	0.429	0.120	0.188
GNN w/ GRU [3]	0.419	0.173	0.245
ARG [6]	0.349	0.200	0.254
S3R2 [3]	0.559	0.507	0.532
GroupTrans [8]	0.750	0.545	0.632
Ours	**0.813**	**0.693**	**0.748**

Table 9. Different visual prompts for CLIP on PANDA.

Dataset (Task)	Model	I-Enc	Group detection		
		F1	Precision	Recall	F1
PANDA (static)	No prompt	0.599	0.683	0.573	0.623
	Mask	0.642	0.710	0.587	0.642
	A circle	0.643	0.696	0.640	0.667
	Two circles	**0.662**	**0.813**	**0.693**	**0.748**

Table 10. Different CLIP finetuning labels, w/ and w/o occlusion on PANDA.

Dataset (Task)	Model	Precision	Recall	F1
PANDA (static)	Ours w/o Occ	0.697	**0.707**	0.702
	Ours	**0.813**	0.693	**0.748**

Table 11. GA-CLIP image features vs visual features on PANDA.

Dataset (Task)	Model	I-Enc	Group detection		
		F1	Precision	Recall	F1
PANDA (static)	ResNet50	0.589	0.467	0.653	0.544
	ViT-L/16	0.552	0.630	0.680	0.654
	DINOv2	0.630	0.607	**0.720**	0.659
	Ours	**0.662**	**0.813**	0.693	**0.748**

Table 12. Feature ablation study on PANDA.

Dataset (Task)	Model	Precision	Recall	F1
PANDA (static)	Ours w/o App.	0.809	0.507	0.623
	Ours w/o Traj.	0.610	0.667	0.637
	Ours	**0.813**	**0.693**	**0.748**

Architecture: Louvain needs 505 MiB on PANDA.

Hyper Parameters: While $D_x = 8$ on JRDB and Café as mentioned in the main paper, $D_x = 10$ on PANDA because the face direction is added as an additional visual attribute. The number of frames used for trajectory features is $T = 51$ on PANDA. The effect of T on group detection is validated in Sec. 10.6.

8.3. Comparison with SoTA methods

The results of static group detection on PANDA are shown in Table 8. The results of all previous methods come from [3, 8]. Our method outperforms the previous methods in all metrics. For example, our method improves F1 by 0.116 on PANDA, compared with GroupTransformer [8], the best previous method.

8.4. Detailed Analyses

8.4.1. Different Visual Prompts for CLIP

Table 9 shows the contributions of visual prompts. In addition to group detection, the results of three-class classification (i.e., "individual people," "a group of people," and "occlusion") using the image encoder are also shown in the "I-Enc" row of Table 9. We can see that our proposed visual prompt using (d) two circles is the best in all metrics.

8.4.2. Occlusion Handling with CLIP Finetuning

The three classes (i.e., "individual people," "a group of people," and "occlusion") are used in our proposed finetuning of the CLIP image encoder. The effectiveness of the "occlusion" class is verified by ablating it in our CLIP finetuning scheme.

The results are shown in Table 10. We can see that "occlusion" improves precision and F1 significantly (i.e., 0.116 and 0.046 improvements in precision and F1, respectively), while recall is slightly degraded (i.e., 0.014).

8.4.3. GA-CLIP image features vs Visual Features

Large-scale pretrained vision-based encoders, ResNet [2], ViT [1], and DINOv2 [4], replace the CLIP image encoder to verify the effectiveness of VLM for group detection. In addition to group detection, three-class classification with the image encoder is also evaluated.

Table 11 shows that in terms of precision and F1, our method using GA-CLIP outperforms all the others by a large margin. The gaps from the second-best scores are $0.183 = 0.813 - 0.630$ and $0.089 = 0.748 - 0.659$ in precision and F1, respectively. While our method is the second-best in recall, the gap from the best is not large (i.e, $0.027 = 0.720 - 0.693$). This proves that higher-level spatial contexts useful for group detection are represented in GA-CLIP compared with image encoders trained only with images.

8.4.4. Ablation of Trajectory and Image Features

The contributions of trajectory and image features are validated by ablating either of them, as shown in Table 12. In this feature ablation, we can see that our method with both features is better than that uses one of them in all metrics.

In precision on PANDA, ours w/o Traj. is worse than ours w/o App. This may be because the relative locations of a pair are well-observed in bird's-eye-view images of

Figure 11. Visual results of dynamic group detection on PANDA. People enclosed by the same color are in the same group.

Figure 12. Failure cases of our method on PANDA. Detected in-group members are marked by rectangles with the same color.

PANDA but their walking directions are not easy to understand. This results in overdetections by ours w/o Traj. In F1, however, our method outperforms both ours w/o App. and Traj.

8.4.5. Inference Time

On a NVIDIA RTX 6000 Ada, the Louvain algorithm needs 0.048 seconds/frame for the temporal groupness graph on PANDA. The cost on PANDA is larger than those of JRDB and Café, which are shown in the main paper, because more people are observed in each frame on PANDA. The total inference time is 0.60 secs/frame on PANDA. Since each video consists of 234 frames on PANDA, the inference times for each video are 139 secs.

8.5. Success and Failure Cases

Success cases on PANDA are shown in Fig. 11. In the upper example, two people enclosed by the red boxes are correctly detected as in-group people at t_1 and t_2. Furthermore, after they split at t_3, our method can find that they are not in the same group. The lower example is a more difficult case. Our method can correctly track the dynamic group change so that two people merge and split at t_2 and t_3, respectively.

Figure 12 shows two failure cases of our method on PANDA. In the upper case, a sales staff interacts with a customer. During this interaction, the trajectory features are not reliable because they do not walk. In addition, since their interaction is short in time and not visually clear. In the lower case, since two individuals walk side by side, they are detected as in-group members. Furthermore, even after their path splits at the intersection at t_3, they are still erroneously detected as a group. This error may be caused due to trajectory features computed from a long-term observation (i.e., $T = 51$ in PANDA), while this long-term observation increases the robustness to instantaneous framewise observation error.

9. Experiments on CaféV and CaféP

While the Café dataset consists of CaféV and CaféP, their mean results are shown in the main paper due to the page limitation. Section 9 shows the separate results of CaféV and CaféP.

9.1. Dataset

The detail of the Café dataset is described in the main paper. All clips in this dataset are split into training, validation, and test splits in two different ways, split-by-view and split-by-place, which are called CaféV and CaféP, respectively. In CaféV and CaféP, the test split has only unseen viewpoints and places, respectively.

9.2. Comparison with SoTA methods

The results of static and dynamic group detections on CaféV and CaféP are shown in Table 13 and Table 14. Our method outperforms all others in terms of all metrics by a large gap.

9.3. Detailed Analyses

Experimental results shown in Tables 15, 16, and 17 correspond to Tables 5, 7, and 8 in the main paper, respectively. Overall, Tables 15, 16, and 17 show the superiority of our proposed component on CaféV and CaféP as with on Café whose results are shown in the main paper.

10. Additional Experiments

Several additional experiments, not included in the main paper for the page limitation, are presented in this section.

10.1. Zero-shot Binary Classification using CLIP

While only several typical results of our preliminary experiments are shown in Fig. 4 of the main paper, more detailed quantitative results are shown in Table 18. Remember that, in our preliminary experiments, binary classification accuracy (i.e., "individual people" or "a group of people") is verified using a softmax layer connected to the pretrained CLIP image and text encoders, which are not finetuned with the "occlusion" class[1]. In addition to a raw bounding box (i.e., (a) in Fig. 8 of the main paper), three visual prompts (b), (c), and (d) of Fig. 8 of the main paper are verified quantitatively in Table 18. Two red circles for specifying a pair of people of interest are (d), which is proposed in our work.

[1]The results of three-class classification (i.e., "individual people," "a group of people," or "occlusion") are shown in Table 4 of the main paper.

Table 13. Quantitative results of static group detection on CaféV and CaféP.

Method	Precision	Recall	F1
CaféV			
S3R2 [3]	0.598	0.704	0.647
GroupTrans [8]	0.278	0.420	0.335
Ours	**0.771**	**0.887**	**0.825**
CaféP			
S3R2 [3]	0.591	0.714	0.647
GroupTrans [8]	0.296	0.430	0.351
Ours	**0.739**	**0.900**	**0.812**

Table 14. Quantitative results of dynamic group detection on CaféV and CaféP.

Method	Cluster	Precision	Recall	F1
CaféV				
S3R2[3]	LP	0.614	0.524	0.565
	CNM	0.603	0.681	0.639
	Louvain	0.615	0.630	0.622
Group Trans[8]	LP	0.234	0.212	0.222
	CNM	0.093	0.067	0.078
	Louvain	0.293	0.331	0.311
Ours	LP	**0.782**	0.776	0.779
	CNM	0.685	0.905	0.780
	Louvain	0.721	**0.907**	**0.803**
CaféP				
S3R2[3]	LP	0.534	0.533	0.533
	CNM	0.545	0.723	0.621
	Louvain	0.521	0.637	0.574
Group Trans[8]	LP	0.182	0.177	0.179
	CNM	0.080	0.071	0.075
	Louvain	0.227	0.275	0.249
Ours	LP	**0.689**	0.759	0.722
	CNM	0.603	**0.939**	0.735
	Louvain	0.634	0.900	**0.744**

As alternatives, the pair is enclosed by one circle in (c), and other people are masked in (b).

As shown in Table 18, binary classification accuracy is better on JRDB than Café and PANDA because a person's appearance is better observed in first-person-view images of JRDB than in bird's-eye-view images of Café and PANDA. The best accuracy is obtained in (b) mask on JRDB. This is because, in (b) where other people except for a target pair are masked, it is easy for CLIP to focus on the target pair to evaluate whether or not this pair is in the same group.

However, (1) such unnatural masked images may cause a domain gap between these masked images and their original raw images used in CLIP pretraining. In addition, (2) other people around the target pair are also informative in deeply evaluating the groupness of this pair. Due to these two reasons, even on JRDB, the accuracy is just a little higher than

Table 15. Different visual prompts for CLIP on CaféV and CaféP.

Dataset (Task)	Model	I-Enc	Group detection		
		F1	Precision	Recall	F1
CaféV (static)	No prompt	0.630	0.629	0.827	0.715
	Mask	0.839	0.736	0.852	0.790
	A circle	0.798	0.701	0.822	0.757
	Two circles	**0.905**	**0.771**	**0.887**	**0.825**
CaféV (dynamic)	No prompt	0.630	0.550	0.830	0.662
	Mask	0.839	0.706	0.850	0.772
	A circle	0.798	0.673	0.816	0.738
	Two circles	**0.905**	**0.721**	**0.907**	**0.803**
CaféP (static)	No prompt	0.521	0.545	0.711	0.617
	Mask	0.806	0.731	0.893	0.804
	A circle	0.718	0.603	0.778	0.680
	Two circles	**0.861**	**0.739**	**0.900**	**0.812**
CaféP (dynamic)	No prompt	0.521	0.459	0.715	0.559
	Mask	0.806	**0.643**	0.893	**0.748**
	A circle	0.718	0.541	0.775	0.637
	Two circles	**0.861**	0.634	**0.900**	0.744

Table 16. GA-CLIP image features vs visual features on CaféV and CaféP.

Dataset (Task)	Model	I-Enc	Group detection		
		F1	Precision	Recall	F1
CaféV (static)	ResNet50	0.657	0.591	0.781	0.673
	ViT-L/16	0.544	0.611	0.760	0.677
	DINOv2	0.740	0.604	0.765	0.675
	Ours	**0.905**	**0.771**	**0.887**	**0.825**
CaféV (dynamic)	ResNet50	0.657	0.551	0.774	0.644
	ViT-L/16	0.544	0.558	0.760	0.643
	DINOv2	0.740	0.582	0.774	0.664
	Ours	**0.905**	**0.721**	**0.907**	**0.803**
CaféP (static)	ResNet50	0.669	0.608	0.868	0.715
	ViT-L/16	0.546	0.451	0.372	0.408
	DINOv2	0.731	0.574	0.751	0.651
	Ours	**0.861**	**0.739**	**0.900**	**0.812**
CaféP (dynamic)	ResNet50	0.669	0.494	0.845	0.623
	ViT-L/16	0.546	0.452	0.365	0.404
	DINOv2	0.731	0.523	0.752	0.617
	Ours	**0.861**	**0.634**	**0.900**	**0.744**

Table 17. Feature ablation study on CaféV and CaféP.

Dataset (Task)	Model	Precision	Recall	F1
CaféV (static)	Ours w/o App.	0.657	0.807	0.724
	Ours w/o Traj.	0.646	0.838	0.729
	Ours	**0.771**	**0.887**	**0.825**
CaféV (dynamic)	Ours w/o App.	0.595	0.831	0.694
	Ours w/o Traj.	0.637	0.839	0.724
	Ours	**0.721**	**0.907**	**0.803**
CaféP (static)	Ours w/o App.	0.620	0.873	0.725
	Ours w/o Traj.	0.691	0.873	0.771
	Ours	**0.739**	**0.900**	**0.812**
CaféP (dynamic)	Ours w/o App.	0.508	0.868	0.641
	Ours w/o Traj.	0.583	0.864	0.696
	Ours	**0.634**	**0.900**	**0.744**

Table 18. Visual prompts for zero-shot binary classification using CLIP.

Dataset	Visual prompt	Accuracy
JRDB	(a) No prompt	0.518
	(b) Mask	**0.561**
	(c) A circle	0.535
	(d) Two circles	0.516
Café	(a) No prompt	0.501
	(b) Mask	0.513
	(c) A circle	**0.516**
	(d) Two circles	0.509
CaféV	(a) No prompt	0.502
	(b) Mask	0.520
	(c) A circle	**0.527**
	(d) Two circles	0.514
CaféP	(a) No prompt	0.500
	(b) Mask	**0.505**
	(c) A circle	0.502
	(d) Two circles	0.502
PANDA	(a) No prompt	0.480
	(b) Mask	0.478
	(c) A circle	**0.509**
	(d) Two circles	0.479

Table 19. CLIP trained with multi-label learning.

Dataset (Task)	Model	Precision	Recall	F1
JRDB (static)	Ours w/ Multi	0.735	0.805	0.768
	Ours	**0.742**	**0.844**	**0.790**
JRDB (dynamic)	Ours w/ Multi	0.682	0.744	0.712
	Ours	**0.724**	**0.820**	**0.769**
PANDA (static)	Ours w/ Multi	0.667	0.640	0.653
	Ours	**0.813**	**0.693**	**0.748**

50% (i.e., chance level).

Why are (d) two circles worse than (b) mask in the preliminary experiments? Furthermore, (c) circling is also better than (d) two circles in the preliminary experiments. Our interpretation of these results is that circling for specifying people is potentially useful even in the pretrained CLIP, as validated by the accuracy of (c) circling. This may be because this circling is also trained in the pretrained CLIP. However, we need the CLIP finetuning for more explicit training to specify a pair without interference from the background and other objects. As a result of this CLIP finetuning, our proposed method (i.e., (d) two circles) is better than the other three visual prompts, (a), (b), and (c), as shown in Table 4 of the main paper.

10.2. CLIP trained with Multi-label Learning

While the probabilities of the three classes ("individual people," "a group of people," and "occlusion") are estimated so that the sum of these three probabilities is one in our proposed CLIP finetuning, these three classes are not necessarily exclusive. That is, while "individual people" and "a group of people" are exclusive, either of these two classes can be observed with "occlusion." This fact motivates us to verify the effectiveness of multi-label learning for CLIP finetuning. For this verification, the group label, $C_g \in \{\text{individual people, a group of people}\}$, and the visibility label, $C_v \in \{\text{occlusion, visible}\}$, are independently classified with each softmax layer in the finetuning process. Note that Café is not used for this experiment because "occlusion" is not annotated on Café.

Table 19 shows that the aforementioned multi-label classification is not beneficial. This observation can be interpreted as follows. If two people in a pair are both visible, the group label classification is not difficult, resulting in successful multi-label classification. However, if at least one of two people in the pair is occluded, it is difficult to classify the group label correctly. In such a case, the unreliable results of the group label classification disturb the finetuning process. In our proposed three-class classification, on the other hand, if at least one of two people in the pair is occluded, the probability of the occlusion class gets higher, and those of the other two classes (i.e., "individual people" and "a group of people") are lower. That is, our proposed three-class classification does not have to classify the group label correctly. For the above reasons, our proposed three-class classification works better than the multi-label classification task.

10.3. Model Capacity of CLIP

Table 20 shows the group detection results obtained with CLIPs of different model capacities. The model capacity of ViT-L is larger than that of ViT-B. The cropped bounding box is resized to 224×224 pixels in all models except for ViT-L/14@336px, where it is resized to 336×336 pixels. A value following the slash denotes the size of each patch fed into ViT: e.g., a patch with 32×32 pixels is used in ViT-B/32. The performance increases as the model capacity grows and is not saturated yet, even with the largest model. Thus, we expect further improvement with larger image encoders.

10.4. Joint Learning Strategies

Table 21 shows the group detection results of different learning strategies of CLIP during joint learning: (1) jointly trained, (2) jointly trained using a smaller learning rate (i.e., 2e-7) compared with the one for other networks (i.e., as described in Sec. 4.3.1 of the main paper), and (3) fixed (i.e., our strategy). The strategy (2) is verified because the

Table 20. CLIPs with different model capacities.

Dataset (Task)	CLIP	I-Enc	Group detection		
		F1	Precision	Recall	F1
JRDB (static)	ViT-B/32	0.573	0.679	0.757	0.716
	ViT-B/16	0.599	0.739	0.760	0.749
	ViT-L/14	0.631	0.725	0.757	0.740
	ViT-L/14@336px	**0.666**	**0.742**	**0.844**	**0.790**
JRDB (dynamic)	ViT-B/32	0.573	0.597	0.651	0.623
	ViT-B/16	0.599	0.680	0.658	0.669
	ViT-L/14	0.631	0.700	0.720	0.710
	ViT-L/14@336px	**0.666**	**0.724**	**0.820**	**0.769**
Café (static)	ViT-B/32	0.712	0.581	0.740	0.649
	ViT-B/16	0.767	0.631	0.837	0.719
	ViT-L/14	0.791	0.668	0.854	0.750
	ViT-L/14@336px	**0.885**	**0.756**	**0.893**	**0.819**
Café (dynamic)	ViT-B/32	0.712	0.543	0.734	0.620
	ViT-B/16	0.767	0.576	0.830	0.680
	ViT-L/14	0.791	0.608	0.839	0.703
	ViT-L/14@336px	**0.885**	**0.681**	**0.904**	**0.776**
CaféV (static)	ViT-B/32	0.710	0.598	0.695	0.643
	ViT-B/16	0.772	0.645	0.864	0.738
	ViT-L/14	0.786	0.678	0.826	0.745
	ViT-L/14@336px	**0.905**	**0.771**	**0.887**	**0.825**
CaféV (dynamic)	ViT-B/32	0.710	0.587	0.683	0.631
	ViT-B/16	0.772	0.613	0.855	0.714
	ViT-L/14	0.786	0.654	0.806	0.722
	ViT-L/14@336px	**0.905**	**0.721**	**0.907**	**0.803**
CaféP (static)	ViT-B/32	0.715	0.560	0.794	0.657
	ViT-B/16	0.761	0.615	0.805	0.697
	ViT-L/14	0.797	0.657	0.887	0.755
	ViT-L/14@336px	**0.861**	**0.739**	**0.900**	**0.812**
CaféP (dynamic)	ViT-B/32	0.715	0.490	0.794	0.606
	ViT-B/16	0.761	0.531	0.801	0.639
	ViT-L/14	0.797	0.554	0.879	0.680
	ViT-L/14@336px	**0.861**	**0.634**	**0.900**	**0.744**
PANDA (static)	ViT-B/32	0.618	0.686	0.640	0.662
	ViT-B/16	0.640	0.685	0.665	0.676
	ViT-L/14	0.649	0.758	0.667	0.709
	ViT-L/14@336px	**0.662**	**0.813**	**0.693**	**0.748**

CLIP image encoder is already finetuned before joint learning, while the other networks (i.e., ϕ, χ, and ρ) are trained from scratch.

Table 21 shows that our method in which CLIP is fixed during joint learning is the best regardless of the learning rate for CLIP. This may be because the joint learning causes CLIP to overfit to the training data of the group detection dataset, which is much smaller than the huge amount of training data used for CLIP pretraining.

10.5. Hard Negative Training and Easy Negative Pruning at Inference for Stability and Efficiency

Feature-similarity-based easy negative pruning at inference:

Figure 13 shows the effect of our easy negative pruning in training by varying the threshold of pairwise groupness probabilities, Th_e. Th_e ranges from 0 to 0.95 at an interval of 0.05. With $Th_e = 0$, all edges remain, and the F1 score is low. F1 increases as Th_e gets greater. In all the datasets, after F1 reaches the max, it drops. Th_e with the max F1 is different among JRDB, PANDA, and Café. While the max F1 scores are observed around at $Th_e = 0.9$ on JRDB and PANDA, F1 reaches around at $Th_e = 0.6$ on Café. This difference is caused probably due to difficulty in image-based

Table 21. Different joint learning strategies. In "JL," CLIP is jointly trained.

Dataset (Task)	Model	I-Enc	Group detection		
		F1	Precision	Recall	F1
JRDB (static)	JL	0.121	0.654	0.785	0.713
	JL w/ small lr	0.643	0.660	0.824	0.733
	Ours	**0.666**	**0.742**	**0.844**	**0.790**
JRDB (dynamic)	JL	0.121	0.595	0.684	0.636
	JL w/ small lr	0.643	0.589	0.771	0.668
	Ours	**0.666**	**0.724**	**0.820**	**0.769**
Café (static)	JL	0.265	0.639	0.841	0.728
	JL w/ small lr	0.646	0.647	0.855	0.737
	Ours	**0.885**	**0.756**	**0.893**	**0.819**
Café (dynamic)	JL	0.265	0.557	0.856	0.674
	JL w/ small lr	0.646	0.554	0.846	0.680
	Ours	**0.885**	**0.681**	**0.904**	**0.776**
CaféV (static)	JL	0.224	0.660	0.836	0.742
	JL w/ small lr	0.513	0.611	0.838	0.707
	Ours	**0.905**	**0.771**	**0.887**	**0.825**
CaféV (dynamic)	JL	0.224	0.608	0.861	0.713
	JL w/ small lr	0.513	0.551	0.825	0.661
	Ours	**0.905**	**0.721**	**0.907**	**0.803**
CaféP (static)	JL	0.313	0.615	0.846	0.712
	JL w/ small lr	0.805	0.690	0.876	0.772
	Ours	**0.861**	**0.739**	**0.900**	**0.812**
CaféP (dynamic)	JL	0.313	0.497	0.849	0.627
	JL w/ small lr	0.805	0.558	0.872	0.702
	Ours	**0.861**	**0.634**	**0.900**	**0.744**
PANDA (static)	JL	0.175	0.797	0.627	0.702
	JL w/ small lr	0.175	0.687	0.613	0.648
	Ours	**0.662**	**0.813**	**0.693**	**0.748**

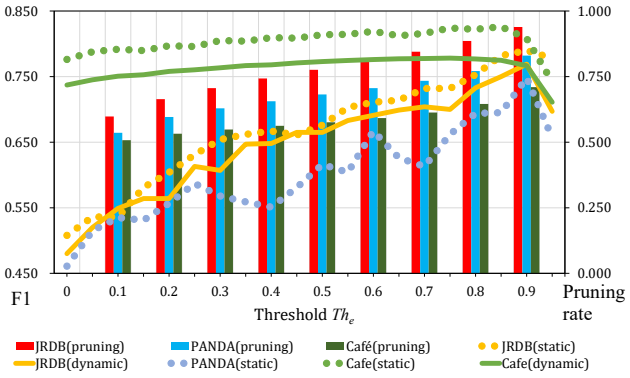

Figure 13. Results of group detection with a varying threshold, Th_e, for temporal groupness graph construction. The bars indicate the pruning rates of the pairs of people depending on Th_e. Solid and dotted lines indicate the F1 scores of the results of dynamic and static detections, respectively.

group probability estimation on Café in which people do not walk and close to each other in most frames. This difficulty results in many lower probability scores, which require our method to evaluate many lower probability scores. As a result, the threshold of pairwise groupness probabilities, Th_e, gets lower.

Distance-based hard negative training and Distance-based easy negative pruning at inference: By varying k of k-nearest neighbor sampling, the effects of the hard negative training process and the easy negative test-time pruning process are verified, as shown in Fig. 14

Figure 14 shows that hard negative training using $k = 1$ works better than $k = 2, 3, 4$ in many cases. For test-time pruning, when we focus on hard negative training using $k = 1$, $k = 2$ is the best at inference on JRDB (static), JRDB (dynamic), and PANDA (static). These results can be naturally interpreted as follows. As mentioned at the beginning of Sec. 10.5, the number of negative pairs (i.e., people in different groups) is significantly greater than that of positive pairs (i.e., in-group members). In addition, most people are alone, and groups consisting of two in-group members are greater in number than other groups. Due to this imbalance and this group statistics, 1-nearest neighbor is sufficient at training. At test time, on the other hand, the 1-nearest neighbor person may or may not be an in-group member. However, $k = 1$ at inference leads to a significantly sparsely-connected graph, resulting in unstable group detection. This may be why $k = 2$ works better than $k = 1$ at inference. Note that graph clustering using the temporal groupness graph allows us to connect $(k + 1)$-nearest neighbor people via temporal edges connecting each person's nodes in subsequent frames. That is, three or more people can be detected as in-group members even with $k = 2$ (as shown in Fig. 15, for example).

On Café also, hard negative training works better with $k = 1$ in many cases. This is because the number of negative pairs is much greater than that of positive pairs in Caf'e as well as in JRDB and PANDA. On the other hand, $k = 3$ works better at inference in Café, while $k = 2$ is the best in JRDB and PANDA. This may be because, in Café, many groups consist of more people (e.g., four or more people) than in-group people in JRDB and PANDA. That is, k at inference should be greater to detect more in-group people in each gruop in Café.

10.6. Analysis of Input Frame Changes

Table 22 shows the results in which the number of input frames for computing the trajectory features is changed.

We can see that the best result is obtained with $T = 51$ (i.e., F1 = 0.748) on PANDA. The F1 score is largely improved from $T = 11$ (i.e., F1 = 0.671) to $T = 51$. These results reveal that long-term trajectory features allow us to improve the performance of group detection, e.g., group detection robust against framewise estimation error of groupness probabilities and group detection using long-term relationships between different trajectories. The F1 score decreases with more frames (i.e., $T = 61$). This performance drop (i.e., $0.063 = 0.748 - 0.685$) implies that too long-term temporal information prevents graph clustering from detecting short-term group formations (e.g., two people greet each other for two seconds).

On the other hand, the number of frames in each video (denoted by N_f) is much smaller than in JRDB and Café than in PANDA. Therefore, all frames in each video are used to compute trajectory features on JRDB and Café, as mentioned in the main paper. That is, $T = N_f$ in all experiments except for those shown in Table 22. N_f is between 28 and 115 frames on JRDB and around 60 frames on Café.

Table 22. Analysis of Input Frame Changes.

Dataset (Task)	Input frame	Precision	Recall	F1
JRDB (static)	T=11	0.677	0.715	0.696
	T=31	0.686	0.805	0.740
	T=51	0.707	0.721	0.714
	T=71	0.711	0.832	0.767
	T=91	0.719	0.729	0.724
	T=All (Ours)	**0.742**	**0.844**	**0.790**
JRDB (dynamic)	T=11	0.644	0.750	0.693
	T=31	0.619	0.757	0.681
	T=51	0.700	0.704	0.702
	T=71	0.649	0.761	0.701
	T=91	0.691	0.743	0.716
	T=All (Ours)	**0.724**	**0.820**	**0.769**
Café (static)	T=11	0.669	0.855	0.750
	T=31	0.672	0.857	0.753
	T=All (Ours)	**0.756**	**0.893**	**0.819**
Café (dynamic)	T=11	0.618	0.857	0.717
	T=31	0.616	0.862	0.719
	T=All (Ours)	**0.681**	**0.904**	**0.776**
CaféV (static)	T=11	0.654	0.835	0.733
	T=31	0.649	0.838	0.731
	T=All (Ours)	**0.771**	**0.887**	**0.825**
CaféV (dynamic)	T=11	0.647	0.838	0.730
	T=31	0.642	0.848	0.731
	T=All (Ours)	**0.721**	**0.907**	**0.803**
CaféP (static)	T=11	0.687	0.878	0.771
	T=31	0.700	0.879	0.780
	T=All (Ours)	**0.739**	**0.900**	**0.812**
CaféP (dynamic)	T=11	0.583	0.879	0.701
	T=31	0.586	0.879	0.704
	T=All (Ours)	**0.634**	**0.900**	**0.744**
PANDA (static)	T=11	0.676	0.667	0.671
	T=21	0.725	0.667	0.694
	T=31	0.700	0.653	0.676
	T=41	0.794	0.667	0.725
	T=51 (Ours)	**0.813**	**0.693**	**0.748**
	T=61	0.721	0.653	0.685

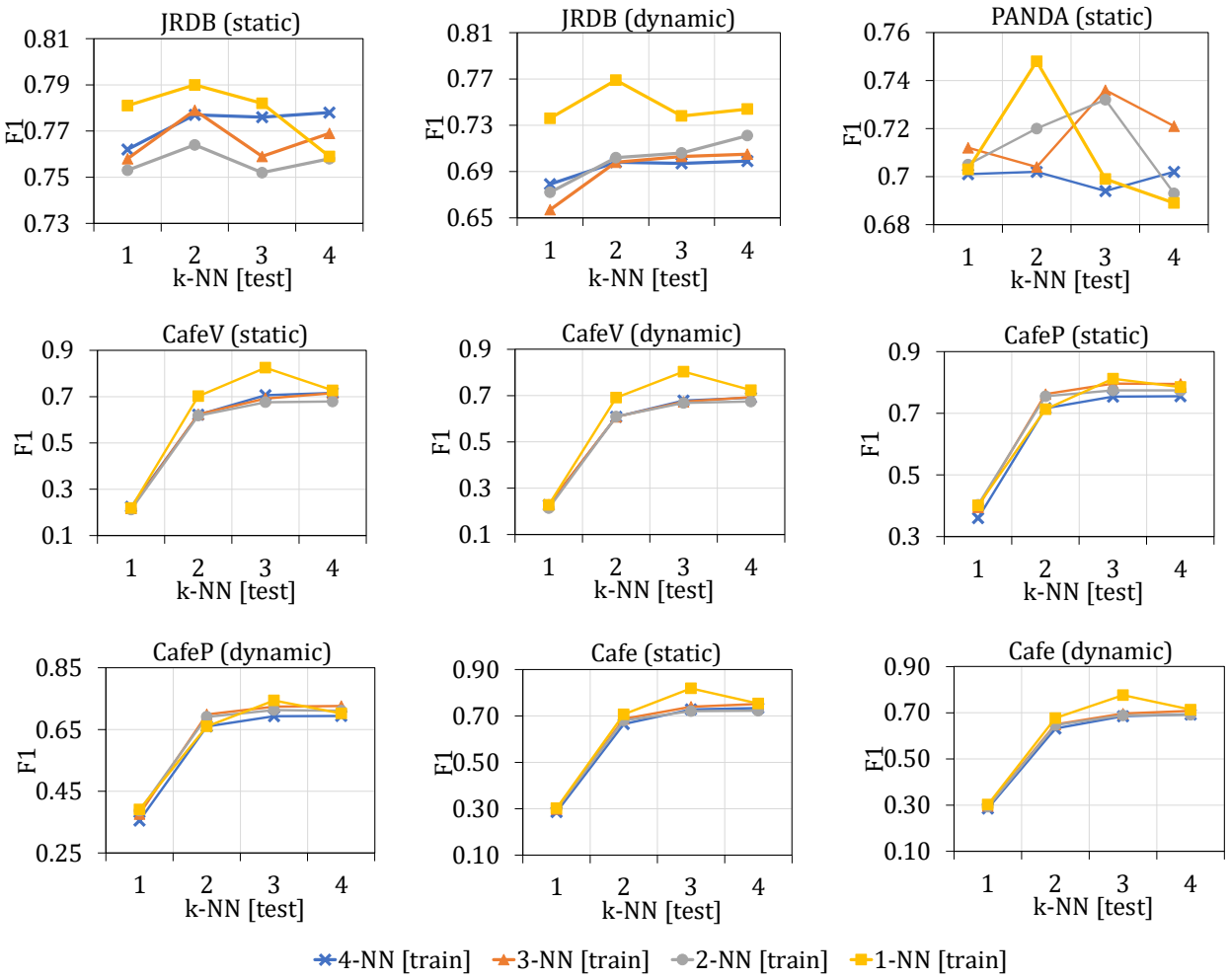

Figure 14. Effects of hard-negative sampling on group detection.

In Table 22, if $N_f > T$, T frames are used for trajectory features. If $N_f \leq T$, N_f frames are used.

The best F1 score is obtained with T=All on both JRDB (static) and JRDB (dynamic). The best F1 score is obtained with a larger T on JRDB than on PANDA. This may be because even non-group people are close to each other in first-person view images of JRDB. That is, many frames are required to observe that such non-group people move away from in-group people.

As well as JRDB, Café needs all frames, T=All, for achieving the best F1 score. This may be because birs'-eye-view images of Café are not captured directly above the scene but diagonally above the scene. That is, many people are mutually occluded and close to each other in images. Therefore, many frames are required to observe that non-group people move away from in-group people in Café.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and D. Weissenborn et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[3] J. Li, R. Han, H. Yan, and Z. Qian et al. Self-supervised social relation representation for human group detection. In *ECCV*, 2022. 1, 2, 4

[4] M. Oquab, T. Darcet, T. Moutakanni, and H. V. Vo et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 2

[5] X. Wang, X. Zhang, Y. Zhu, and Y. Guo et al. PANDA: A gigapixel-level human-centric video dataset. In *CVPR*, 2020. 1, 2

[6] J. Wu, L. Wang, L. Wang, and J. Guo et al. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019. 2

[7] X. Zhan, Z. Liu, J. Yan, D. Lin, and et al. Consensus-driven

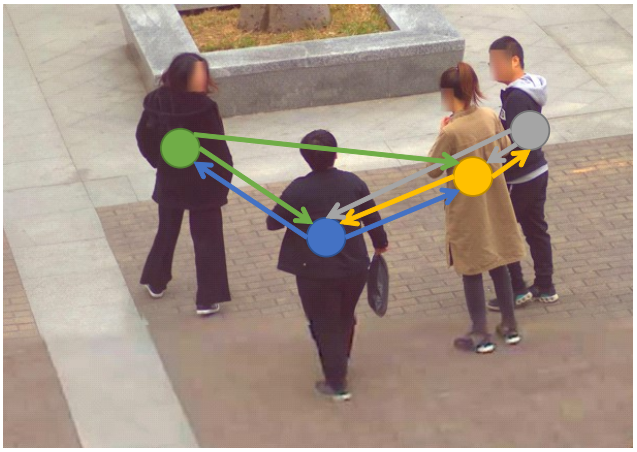

Figure 15. An example of detecting a group of four people using 2-NN. Each node represents a person, and the nodes connected by arrows indicate the individuals selected as 2-NN. The leftmost person and the rightmost person are also connected through the two individuals in between.

propagation in massive unlabeled data for face recognition. In *ECCV*, 2018. 2

[8] J. Zhang, L. Gu, Y. Lai, and X. Wang et al. Toward grouping in large scenes with occlusion-aware spatio-temporal transformers. *IEEE Trans. Circuits Syst. Video Technol.*, 34(5): 3919–3929, 2024. 2, 4