Conversational Group Detection with Graph Neural Networks

Sydney Thompson*
Yale University
New Haven, CT, United States
sydney.thompson@yale.edu

Abhijit Gupta*
Anjali W. Gupta
Yale University
New Haven, CT, United States

Austin Chen
Marynel Vázquez
Yale University
New Haven, CT, United States

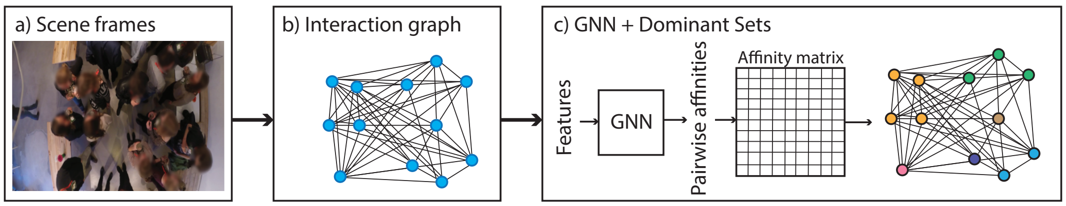

Figure 1: Proposed approach for conversational group detection ((a) includes an example frame from MatchNMingle [5]).

ABSTRACT

We study conversational group detection in varied social scenes using a message-passing Graph Neural Network (GNN) in combination with the Dominant Sets clustering algorithm. Our approach first describes a scene as an interaction graph, where nodes encode individual features and edges encode pairwise relationship data. Then, it uses a GNN to predict pairwise affinity values that represent the likelihood of two people interacting together, and computes non-overlapping group assignments based on these affinities. We evaluate the proposed approach on the Cocktail Party and MatchNMingle datasets. Our results suggest that using GNNs to leverage both individual and relationship features when computing groups is beneficial, especially when more features are available for each individual.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

KEYWORDS

F-formation; clustering; graph neural network

ACM Reference Format:

Sydney Thompson, Abhijit Gupta, Anjali W. Gupta, Austin Chen, and Marynel Vázquez. 2021. Conversational Group Detection with Graph Neural Networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462244.3479963

*Both authors contributed equally to this research.

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479963

1 INTRODUCTION

Conversational group detection has a wide range of applications, including video surveillance [8, 15, 29], displays and exhibits [10, 16], co-located collaboration [21], and interactive playgrounds [17, 22]. Group detection can also enable better spoken language interaction with situated agents [4], non-verbal robot behavior generation [34], and socially aware robot navigation in human environments [26].

Similar to prior work, we approach the problem of conversational group detection by reasoning about human proxemics [12] and conversational formations. During free-standing conversations, people tend to form certain spatial patterns with each other, known as Face Formations or F-Formations in short [18]. F-Formations are varied, adapting to factors such as density and physical environmental constraints. They characterize conversational groups.

We describe a social scene as an interaction graph and explore using a Graph Neural Network (GNN) [2] for conversational group detection. Inspired by Swofford and colleagues [30], we use the GNN to predict pairwise affinities for the graph, which encode the likelihood that two people are part of an F-Formation. Then, we use the affinities to cluster people into conversational groups, as shown in Figure 1. While Swofford and colleagues [30] used a Deep Set [25, 40] architecture to aggregate context from graph nodes when predicting an affinity value, this work advocates in favor of a more general message-passing architecture for reasoning about information in both the nodes and edges. This allows us to reduce feature engineering and more explicitly leverage relational features.

In summary, our main contributions are threefold. First, we propose a novel approach for group detection which relies on a GNN. Second, we conduct experiments on two datasets with varied input features such as position, orientation, and top-down images of participants to demonstrate the efficacy of the proposed model. Third, we open-source our code to facilitate future reproducibility.[1]

2 RELATED WORK

The problem of conversational group detection has traditionally been approached by hand-crafted heuristics and mathematical models [15, 29, 32]. However, advancements in machine learning have

[1] http://gitlab.com/interactive-machines/perception/group_gnn

enabled improved social awareness with greater generalization [11, 14, 30]. In particular, the approach by Swofford et al. [30], called DANTE, outperformed several traditional approaches. DANTE receives spatial features for people in a scene and constructs a fully-connected interaction graph, using the input data as node features. It then computes pairwise affinities by combining the dyad node features with context aggregated using a Deep Set architecture [25, 40]. These affinities are used to partition the graph with the Dominant Sets algorithm [15, 23]. Because DANTE computes context in tandem with a dyad, it relies heavily on hand-crafted feature transformations to preserve rotation and translation invariance.

While DANTE [30] mainly reasons about information encoded in the nodes of a graph, we propose to use a more general message-passing GNN architecture [2] for affinity prediction. The GNN consists of a collection of update and aggregate functions that allow for node and edge information consolidation in a graph. This GNN architecture is a superset of Deep Sets, as discussed in [2].

While GNNs have previously been used for node clustering [3, 19, 31, 38], our problem differs in several key ways. Methods such as [19, 31] require an input affinity matrix, while our GNN must calculate the affinities itself. Also, several prior models for clustering with deep learning require information about the number of clusters [3, 38]; however, we do not know the number of conversational groups in a scene in advance. Lastly, many models (e.g. [6, 13]) for social interaction analysis are designed and evaluated on large graphs (see [36] for five such datasets with an average number of nodes ranging from 13 to 500). In our case, we make predictions over smaller graphs (2-16 nodes) as there is a physical limit to how many people can interact simultaneously in a given place [18].

Three reasons motivate us to predict an affinity matrix with a message-passing GNN. First, the values in an affinity matrix can be thought of as unidimensional edge features and, by design, GNNs are well suited to predict this type of data. Second, a single GNNs can work on graphs with numbers of nodes, which is important when reasoning about varied environments. Third, strategic choices about what features are encoded in the nodes and edges of a graph can make GNNs invariant to spatial rotations and translations. This reduces the amount of pre-processing transformations needed to analyze a scene in comparison to DANTE.

Prior benchmarks in conversational group detection from still images [28–30] commonly consider datasets with a limited number of people, e.g., the Cocktail Party dataset [41] considers six people. Given the relative simplicity of these datasets, we study group detection performance using the recent MatchNMingle dataset [5] made available by the Delft University of Technology. MatchNMingle is a multi-sensor dataset of in-the-wild conversations for the analysis of social interactions. It contains 4446 images of a scene with up to 15 people per frame, as shown in Figure 1.

3 METHOD

This paper studies conversational group detection: partitioning a set of people in a scene into non-overlapping clusters representing interacting groups. Formally, assume that there are n people in the scene and let P be a set of individual feature vectors, $P = \{\boldsymbol{f}_k \mid 1 \leq k \leq n\}$. Then, the groups can be expressed via a clustering schema $C : P \rightarrow \{1, \ldots, n_c\}$, with n_c the number of clusters.

Given a dataset $\mathcal{D}$ of N examples, $\mathcal{D} = \{(P_1, C_1), \ldots, (P_N, C_N)\}$, we frame the group detection problem from a supervised learning perspective as computing a function $h(P_i) = \hat{C}_i$ that estimates cluster assignments. Each predicted $\hat{C}_i$ should be as close as possible to the true C_i for all the examples i in $\mathcal{D}$. Note that in this problem the number of clusters and people may differ across examples.

3.1 Clustering Conversational Interactants

We propose to construct the function $h(P_i) = \hat{C}_i$ using a Graph Neural Network (GNN), followed by the application of the Dominant Sets (DS) algorithm [23]. To this end, we first create a fully-connected *interaction graph* that describes the scene, $G_i^0 = (N_i^0, E_i^0)$, as illustrated in Figure 1. We assign each feature vector $\boldsymbol{f}_k$ to node features $\boldsymbol{n}_k^0$ and edge features $\boldsymbol{e}_{jk}^0$ in the graph, such that:

$$N_i^0 = \{\boldsymbol{n}_k^0 \mid 1 \leq k \leq |P_i|\},\ E_i^0 = \{\boldsymbol{e}_{jk}^0 \mid 1 \leq j, k \leq |P_i|, j \neq k\}$$

The proposed GNN is composed of two graph computation layers, $g(\cdot) = g^2(g^1(\cdot))$. Each layer transforms an input graph G_i^{l-1} into another graph G_i^l, with l indicating the l-th layer without loss of generality. At the last layer of the GNN, $\boldsymbol{e}_{jk}^2 \in E_i^2$ represents the pairwise affinity from node j to node k in the graph.

Based on the pairwise affinities output by the GNN, we construct an affinity matrix A_i for the graph corresponding to the set P_i. A_i is then passed through the DS algorithm [23], which iteratively groups graph nodes into clusters by maximizing the quadratic program $\max_{\boldsymbol{x} \in S_{|P_i|}} \boldsymbol{x}^T A_i \boldsymbol{x}$, where $S_{|P_i|}$ is the standard simplex in $\mathbb{R}^{|P_i|}$. Here, solutions to the quadratic program represent a group of people, the dominant set in the input A_i. Note that every iteration of DS reduces the size of the affinity matrix by removing the data corresponding to the last group that was predicted by the algorithm.

Oftentimes, there will be individuals in a scene that are not in a group conversation. However, the peeling-off strategy employed by DS tends to group together these individuals. To combat this problem, we use the DS stopping criteria from [15] to consider the global context of the complete graph when grouping people.

3.1.1 Graph Neural Network. Each computation layer of the proposed GNN is a *graph network block* comprised of two updates: one for the edges and one for the nodes, following the message-passing architecture described in [2]. If we define the node and edge features for layer l as $\boldsymbol{n}_k^l$ and $\boldsymbol{e}_{jk}^l$, respectively, then the graph network block operates as follows:

$$\boldsymbol{e}_{jk}^{l+1} = \text{edge\_update}\left(\boldsymbol{e}_{jk}^l, \boldsymbol{n}_j^l, \boldsymbol{n}_k^l\right) \tag{1}$$

$$\boldsymbol{n}_k^{l+1} = \text{node\_update}\left(\boldsymbol{n}_k^l, \text{agg}\left(\{\boldsymbol{e}_{jk}^{l+1} | j \neq k\}\right)\right) \tag{2}$$

The edge_update$(\cdot)$ and node_update$(\cdot)$ functions are neural networks that reason about edge or node features in relation to the information in their neighborhood in the graph, as shown in Figure 2. The agg$(\cdot)$ function is a symmetric function that summarizes information in the edge features connected to a given node.

Our motivation for designing our GNN with two graph network blocks stems from the fact that we consider fully-connected interaction graphs with no self loops in this work. Thus, two graph network blocks suffice to make the output affinity values dependent on the information encoded in all the nodes and edges in the graph.

Figure 2: Edge update (left) and node update (right).

3.1.2 Implementation Details. When creating an interaction graph G^0, each person in the scene is associated to a node and edges are created between all individuals. The edge features $\boldsymbol{e}^0$ are derived from a subset of the person features $\boldsymbol{f}$ in order to describe pairwise relationships. For example, if the person features include position information, then we compute an edge feature that corresponds to the distance between people. The remaining individual features are used as node features $\boldsymbol{n}^0$ in the interaction graph.

Our model uses averaging for $\text{agg}(\cdot)$ in eq. (2), and multi-layer perceptrons (MLPs) for the $\text{edge\_update}(\cdot)$ and $\text{node\_update}(\cdot)$ functions in eq. (1) and (2), respectively. We also apply MLPs to both the initial node and edge features in G^0, before the update functions, in order to balance their relative number of features. For example, if the nodes include image-derived features and the edges contain only distance, these MLPs can embed the image into a smaller feature and the distance into a bigger feature to increase their relative importance. The final edge embeddings in the GNN have a size of 1, corresponding to pairwise affinities. We train the GNN using binary cross-entropy on these affinities, as in [30].

Finally, we aggregate the pairwise affinities output by the GNN into a matrix, $\hat{A}_i$, and then compute a symmetric affinity matrix $A_i = \frac{1}{2}(A_i + A_i^T)$. The latter matrix is used by DS to compute groups.

4 EVALUATION

We compare the proposed approach for group detection against baselines on two datasets with different person-level features.

4.1 Datasets

Cocktail Party Dataset [41]. The dataset contains 30 minutes of interactions among six people in a lab environment. The dataset provides position and head orientation information for each individual, and we also consider their body orientation from [35]. Conversational groups are labeled for 320 frames. We use the first 64 frames for testing, the next 64 for validation, and the rest for training. We chose this dataset for our evaluation because of its relative simplicity, high quality features, and popularity [29, 30, 32].

MatchNMingle Dataset [5]. The dataset was recorded over 3 days with a total of 92 participants. We used the "mingle" data, a subset of MatchNMingle where participants engaged in a cocktail party. For each day of recording, this subset includes 10/30 minutes of video from 3 cameras with annotated bounding boxes and "social actions" in 9 categories for each participant. Triaxial acceleration and binary proximity data is provided for 71/92 participants, all labeled at 20Hz. Also, manually-annotated conversational groups are given at 1 Hz. We aggregate these annotations into 600 frames per camera per day. Because there was high variability in group sizes, spacing, and environment obstacles between recordings, we partition each recording individually. The first and last 10% of frames from each recording were used for test, the next 10% of frames from beginning and end were used for validation, and the middle 60% were used for training.

We consider 4 types of features for individuals in MatchNMingle:

- *Position features* (pos) include x, y coordinates for the corresponding person on a video recording.
- *Acceleration features* (accel) are the last 10 accelerometer readings for a person, covering a time window of 0.5 seconds.
- *Image features* (img) are visual embeddings for the person. We compute the embedding by passing a 32×32 cropped section of the recorded image around the person to ResNet [20] and extracting the 512 features in the penultimate layer of the network.
- *Semantic features* (label) encode person actions. The features are computed by aggregating the actions into a 9-dimensional vector that indicates their occurrence per type over the last 0.5 seconds.

4.2 Group Detection Methods

We consider three methods in our evaluation:

(1) **Dist.** Hand-crafted baseline inspired by [15, 39]. The method computes an affinity matrix as $A_{ij} = \exp\left(-d_{ij}/2\sigma^2\right)$, where d_{ij} is the distance between two participants and $\sigma = 2$ meters, following [15]. DS is then applied to obtain groupings, as in [15].

(2) **DANTE.** We implement the DANTE neural network [30] in PyTorch and use DS for clustering, as in [15]. The dyad and context MLPs of DANTE had two layers with 32 and 64 units. The final MLP had 64 and 32 units. All but the last layer used a ReLU activation followed by batch normalization.

For Cocktail Party, DANTE uses position and the orientations as node features, transforming each feature into a coordinate frame centered between each dyad, as in [30]. For MatchNMingle, it uses the position, transformed by the dyadic coordinate frame, and applies the rest of the features without additional transformations.

(3) **GNN.** Our proposed combination of a GNN with DS for group detection. We implement the GNN using PyTorch Geometric to leverage sparse tensor computations. For the edge updates, we use two MLPs of dimensions 128, 64 and 32, 16 for each graph network block. The node updates use MLPs of dimensions 32, 16 and 16, 16. As in DANTE, we use ReLU activations and batch norm. These dimensions were chosen to produce a similar number of parameters to the DANTE models for all input feature combinations.

For Cocktail Party, the GNN uses distance and both angles, transformed into point pair features [9], as edge features. For MatchNMingle, it uses distance for edge features and all other features as node features. When considering position-only features, however, we do not use any node features.

Both DANTE and the GNN were trained using a learning rate of 1e-4 that decays to 1e-6 over 1000 epochs, a batch size of 512, and the Adam optimizer. Early stopping halted training if there was no decrease in the cross-entropy loss after 50 epochs.

4.3 Results

Our main evaluation metric is the Group F1 metric [32]. For a threshold T, the Group F1 metric considers a ground truth cluster with n_g people to be correctly identified if at least $\lceil T \cdot n_g \rceil$ members are grouped together by the algorithm and no more than $\lceil (1-T) \cdot n_g \rceil$

Table 1: Results on Cocktail Party, including average results and std. deviation ($\mu \pm \sigma$) over the test examples. Results in bold are significantly better than those with regular font.

Metric	Dist	DANTE	GNN
F1T1	0.24 ± 0.34	$\mathbf{0.58 \pm 0.43}$	$\mathbf{0.62 \pm 0.41}$
F1T2/3	0.53 ± 0.32	$\mathbf{0.71 \pm 0.35}$	$\mathbf{0.70 \pm 0.37}$

Table 2: Results on MatchNMingle, including average results and std. deviation ($\mu \pm \sigma$) over the test examples. Results in bold are significantly better than those with regular font.

Features	Metric	Dist	DANTE	GNN
pos	F1T1	0.28 ± 0.26	0.32 ± 0.28	0.30 ± 0.26
	F1T2/3	0.38 ± 0.29	$\mathbf{0.43 \pm 0.30}$	0.40 ± 0.28
pos+accel	F1T1	-	0.24 ± 0.25	$\mathbf{0.34 \pm 0.27}$
	F1T2/3	-	0.30 ± 0.28	$\mathbf{0.43 \pm 0.28}$
pos+img	F1T1	-	0.28 ± 0.29	$\mathbf{0.31 \pm 0.28}$
	F1T2/3	-	0.34 ± 0.29	$\mathbf{0.40 \pm 0.29}$
pos+accel	F1T1	-	0.23 ± 0.24	$\mathbf{0.32 \pm 0.28}$
+img	F1T2/3	-	0.28 ± 0.26	$\mathbf{0.42 \pm 0.29}$
pos+accel	F1T1	-	0.27 ± 0.26	$\mathbf{0.36 \pm 0.29}$
+img+label	F1T2/3	-	0.35 ± 0.28	$\mathbf{0.46 \pm 0.29}$

false subjects are identified. We consider two values for T: 2/3 and 1. For example, let the true group be $g = \{1, 2, 4\}$ and the predicted group $\hat{g} = \{1, 2, 5\}$. Here, $n_g = 3$. For T=1, $\lceil T \cdot n_g \rceil = \lceil 1 \cdot 3 \rceil = 3 > |g \cap \hat{g}|$, so the group is not correctly identified. For $T = \frac{2}{3}$, $\lceil T \cdot n_g \rceil = \lceil \frac{2}{3} \cdot 3 \rceil = 2 \leq |g \cap \hat{g}|$ and $\lceil (1-T) \cdot n_g \rceil = \lceil \frac{1}{3} \cdot 3 \rceil = 1 \leq |\hat{g} \setminus g|$, so the group is correctly identified.

Cocktail Party. Table 1 displays the F1T1 and F1T2/3 scores on the Cocktail Party dataset. For both metrics, a Kruskall-Wallis non-parametric test indicated that there was a significant difference between scores by method, with $p < 0.0001$ for F1T1 and $p = 0.0003$ for F1T2/3. Further, Steel-Dwass post-hoc tests showed that in both cases DANTE and the proposed GNN method led to significantly higher performance than the Dist baseline, but the scores for DANTE and the GNN were not significantly different. Given the high F1T2/3 scores for the data-driven methods, we proceeded to evaluate performance on the more complex MatchNMingle dataset, which contains almost 14 times as many frames as Cocktail Party, 2.5 times the maximum number of people per frame, and, unlike Cocktail Party, a variable number of people per frame.

MatchNMingle. Table 2 shows the results based on the features available for group detection. When only position was available, a Kruskal-Wallis test resulted in significant differences for F1T1 ($p = 0.02$) and a Steel-Dwass post-hoc test indicated that DANTE had significantly higher results than the Dist baseline. No other significant pairwise differences were found. For T1T2/3, the Kruskall-Wallis test also showed significant differences ($p = 0.0003$). In this case, DANTE was significantly better than the other two methods.

Wilcoxon tests showed significant differences by method for the combinations of pos, accel, img, and label features in Table 2. The proposed GNN outperformed DANTE in all these cases. Interestingly, the GNN shows increased performance the more features were provided to it, especially including accel and label. However, DANTE is not as effective at incorporating additional features. For example, the F1T1 score for DANTE using all MatchNMingle features is about 5% lower than the score for using only position data, where the proposed GNN results in a 6% increase in performance.

5 LIMITATIONS & FUTURE WORK

We demonstrated the successful application of GNNs to group detection. In principle, the inductive nature of the proposed approach allows our method to run in an online fashion, processing streams of data. However, more tests are needed to verify this in practice. Future work could also evaluate the GNN on other group detection datasets, like CoffeeBreak [7] or Salsa [1].

Unexpectedly, DANTE and the proposed GNN did not benefit from the added image features in the MatchNMingle dataset. Further, in the case of DANTE, performance tended to decrease with more features. There are several possible explanations for this phenomenon. First, the image features from the ResNet [20] model could have been too deep in the network. Low-level features from earlier in the network could be used to fix this issue. Second, the MatchNMingle cameras have visible radial distortion, which we did not correct for because the intrinsic camera parameters are not public. Lastly, the performance drop could be due to challenges combining feature modalities. In this respect, future work could explore using attention mechanisms to fuse data, e.g., as in [24, 33, 37].

We processed the data from different cameras in the MatchNMingle dataset as independent samples, although some of them contained information captured at the same time from different views. Likewise, we did not consider the temporal correlation of data across dataset samples, but this information could improve model prediction [32, 35]. Thus, future work could explore detecting groups across multiple camera views to understand more holistically the environment, and combining GNNs with recurrent neural networks to take advantage of temporal correlations, e.g., as in [27].

6 CONCLUSION

We presented an approach to predict conversational clusters in social scenes, where the number of clusters is unknown a priori. Our results indicate that GNNs can better take advantage of multi-modal data for group detection in comparison to baselines. In particular, the proposed GNN-based model outperformed the previous state-of-the-art approach [30] on the complex MatchNMingle dataset with all types of data except position-only, while requiring less data pre-processing. This suggests that leveraging relational inductive biases in data-driven methods for group detection is beneficial.

ACKNOWLEDGMENTS

Portions of the research in this paper used the MatchNMingle Dataset made available by the Delft University of Technology, Delft, The Netherlands. This work was supported by the National Science Foundation (NSF), Grant No. (IIS-1924802). The findings and conclusions in this paper are those of the authors and do not necessarily reflect the views of the NSF. The authors are also thankful to the Yale Hahn Scholars program for supporting A. W. Gupta.

REFERENCES

[1] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. 2016. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1707–1720.

[2] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 [cs.LG]

[3] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. 2020. Spectral Clustering with Graph Neural Networks for Graph Pooling. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 874–883.

[4] Dan Bohus, Chit W. Saw, and Eric Horvitz. 2014. Directions Robot: In-the-Wild Experiences and Lessons Learned. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris, France) *(AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 637–644.

[5] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 2021. The MatchNMingle Dataset: A Novel Multi-Sensor Resource for the Analysis of Social Interactions and Group Dynamics In-the-Wild During Free-Standing Conversations and Speed Dates. *IEEE Transactions on Affective Computing* 12, 1 (2021), 113–130. https://doi.org/10.1109/TAFFC.2018.2848914

[6] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. 2019. Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 257–266.

[7] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. 2011. Social Interaction Discovery by Statistical Analysis of F-formations. In *Proceedings of the British Machine Vision Conference (BMVC)*, Vol. 2. Citeseer, BMVA Press, 4.

[8] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino. 2013. Human behavior analysis in video surveillance: A Social Signal Processing perspective. *Neurocomputing* 100 (2013), 86–97.

[9] Haowen Deng, Tolga Birdal, and Slobodan Ilic. 2018. PPF-FoldNet: Unsupervised Learning of Rotation Invariant 3D Local Descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 602–618.

[10] Eyal Dim and Tsvi Kuflik. 2014. Automatic Detection of Social Behavior of Museum Visitor Pairs. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2014), 1–30.

[11] Ekin Gedik and Hayley Hung. 2018. Detecting Conversing Groups Using Social Dynamics From Wearable Acceleration: Group Size Awareness. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–24.

[12] Edward Twitchell Hall. 1966. *The Hidden Dimension*. Vol. 609. Garden City, NY: Doubleday.

[13] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[14] Hooman Hedayati, Daniel Szafir, and Sean Andrist. 2019. Recognizing F-Formations in the Open World. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 558–559.

[15] Hayley Hung and Ben Kröse. 2011. Detecting F-Formations as Dominant Sets. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (Alicante, Spain) *(ICMI '11)*. Association for Computing Machinery, New York, NY, USA, 231–238. https://doi.org/10.1145/2070481.2070525

[16] Junko Ichino, Kazuo Isoda, Tetsuya Ueda, and Reimi Satoh. 2016. Effects of the Display Angle on Social Behaviors of the People around the Display: A Field Study at a Museum. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 26–37.

[17] Manuela Jungmann, Richard Cox, and Geraldine Fitzpatrick. 2014. Spatial Play Effects in a Tangible Game with an F-Formation of Multiple Players. In *Proceedings of the Fifteenth Australasian User Interface Conference - Volume 150* (Auckland, New Zealand) *(AUIC '14)*. Australian Computer Society, Inc., AUS, 57–66.

[18] Adam Kendon. 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive.

[19] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs.LG]

[20] Xuelei Li, Liangkui Ding, Li Wang, and Fang Cao. 2017. FPGA accelerates deep residual learning for image recognition. In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. 837–840.

[21] Nicolai Marquardt, Ken Hinckley, and Saul Greenberg. 2012. Cross-Device Interaction via Micro-Mobility and f-Formations. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, 13–22.

[22] Alejandro Moreno, Robby van Delden, Ronald Poppe, and Dennis Reidsma. 2013. Socially Aware Interactive Playgrounds. *IEEE Pervasive Computing* 12, 3 (2013), 40–47.

[23] Massimiliano Pavan and Marcello Pelillo. 2006. Dominant sets and pairwise clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1 (2006), 167–172.

[24] Ashwini Pokle, Roberto Martín-Martín, Patrick Goebel, Vincent Chow, Hans M Ewald, Junwei Yang, Zhenkai Wang, Amir Sadeghian, Dorsa Sadigh, Silvio Savarese, et al. 2019. Deep Local Trajectory Replanning and Control for Robot Navigation. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 5815–5822.

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.

[26] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From Proxemics Theory to Socially-Aware Navigation: A Survey. *International Journal of Social Robotics* 7, 2 (2015), 137–153.

[27] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. 2018. Graph Networks as Learnable Physics Engines for Inference and Control. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 4470–4479.

[28] Francesco Setti, Hayley Hung, and Marco Cristani. 2013. Group detection in still images by F-formation modeling: A comparative study. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 1–4.

[29] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. 2015. F-formation detection: Individuating free-standing conversational groups in images. *PloS one* 10, 5 (2015), e0123783.

[30] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. 2020. Improving Social Awareness Through DANTE: Deep Affinity Network for Clustering Conversational Interactants. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 020 (May 2020), 23 pages.

[31] Anton Tsitsulin, John Palowitch, Bryan Perozzi, and Emmanuel Müller. 2020. Graph Clustering with Graph Neural Networks. arXiv:2006.16904 [cs.LG]

[32] Sebastiano Vascon, Eyasu Z. Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. 2016. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding* 143 (2016), 11–24. https://doi.org/10.1016/j.cviu.2015.09.012

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kai ser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[34] Marynel Vázquez, Elizabeth J. Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E. Hudson. 2017. Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 42–52.

[35] Marynel Vázquez, Aaron Steinfeld, and Scott E. Hudson. 2015. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3010–3017. https://doi.org/10.1109/IROS.2015.7353792

[36] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks? arXiv:1810.00826 [cs.LG]

[37] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-Modal Attention With Semantic Consistence for Image–Text Matching. *IEEE transactions on neural networks and learning systems* 31, 12 (2020), 5412–5425.

[38] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Hierarchical Graph Representation Learning with Differentiable Pooling. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 4805–4815.

[39] Ting Yu, Ser-Nam Lim, Kedar Patwardhan, and Nils Krahnstoever. 2009. Monitoring, recognizing and discovering social networks. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1462–1469.

[40] Manzil Zaheer, Satwik Kottur, Siamak Ravanbhakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J Smola. 2017. Deep Sets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3394–3404.

[41] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. 2010. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*. 37–42.