Pedestrians' Gaze Object Estimation in Traffic Scene Considering Relationships among Multiple Pedestrians

Hiroyuki Tatsumi, Daisuke Deguchi, Seigo Ito and Hiroshi Murase[a]

[a]Graduate School of Informatics, Nagoya University, Nagoya, Japan

ABSTRACT

In this paper, we address the task of estimating the objects that each pedestrian is gazing at in traffic environments with multiple pedestrians. Gaze object estimation is the task of determining "who is gazing at what" such as Pedestrian A looking at Pedestrian B. We propose a gaze object estimation method that considers relationships among multiple pedestrians. In other words, we consider not only the target person but also what surrounding people are looking at. Specifically, the Transformer encoder processes the features of all pedestrians simultaneously to extract features that encode their mutual relationships. These output features are then mapped via Gaze-Role Projection to two role-specific features for each pedestrian: the looker feature and the lookee feature. Finally, the network is trained to maximize the similarity between the looker's feature and the lookee's feature for the correct gaze pair. In our experiments, we created the PEDESTRIAN AWARENESS DATASET. This dataset was created by preparing scenarios that predefined pedestrian actions and gaze objects, and conducting filming based on these scenarios. As a result of the experiment, we confirmed that the proposed method improves estimation accuracy.

Keywords: Pedestrian's Gaze Object Detection, Gaze Estimation, Traffic Scene, Interaction Among Pedestrians

1. INTRODUCTION

In recent years, autonomous mobile robots have been increasingly deployed in environments with multiple pedestrians, such as shopping malls. For smooth movement of both pedestrians and robots, robots need to understand interactions among pedestrians. Since pedestrians act based on their field of vision, information about "who is looking at what" is effective for understanding interaction. As shown in Fig.1, gaze object estimation determines that pedestrians A and B are looking at each other, pedestrian C is looking at pedestrian B, and pedestrian D doesn't have a specific gaze object. This study addresses the gaze object estimation task of estimating "who is looking at what." In this context, "object" includes pedestrians, vehicles, traffic lights, and other entities. Various studies have been conducted on estimating human gaze. Racasens et al.[1] created the Gazefollow dataset and proposed a gaze estimation model that takes head region images as input and outputs a heatmap representing a person's gaze. However, when pedestrians are distant from the viewpoint of an autonomous mobile robot, obtaining clear head images (mainly the eye region) is difficult. Thus, applying the method of Recasens et al.[1] to the traffic scenes targeted in this paper is difficult. Additionally, gaze estimation using heatmaps can face difficulties in discrimination when gaze objects are densely clustered. Tonini et al.[2] proposed a Transformer-based method for end-to-end object detection and gaze estimation. The method detects objects in the gaze direction and outputs their class and bounding box coordinates. However, there are still limitations for distant pedestrians because gaze estimation relies on head features, and gaze object estimation via heatmaps has difficulty distinguishing multiple objects in close proximity.

Therefore, our research group has been working on gaze object estimation using features of the entire pedestrian, not depending solely on head images. In the research by Hata et al.,[3] pose information and positional coordinates of pedestrians are extracted from outdoor traffic scene images to estimate whether pedestrians are looking at the ego-vehicle. Murakami et al.[4] proposed a method that learns the relationship between pedestrians' full-body features and features of objects detected by an object detector using a Transformer to estimate gaze objects. This approach does not depend on clear head images (mainly the eye region) or heatmaps, making it applicable to distant pedestrians and crowded environments. However, independent estimation, which does not consider what other pedestrians are looking at, has the problem that a pedestrian's gaze object becomes ambiguous when multiple gaze object candidates exist in the direction of the body orientation. Specifically, in

Hiroyuki Tatsumi: E-mail: tatsumih@vislab.is.i.nagoya-u.ac.jp

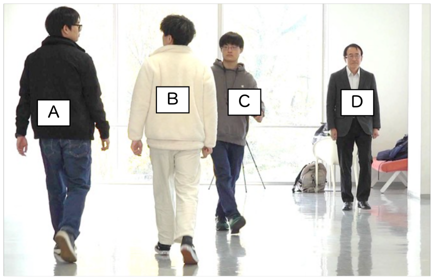


Figure 1: Complex interactions

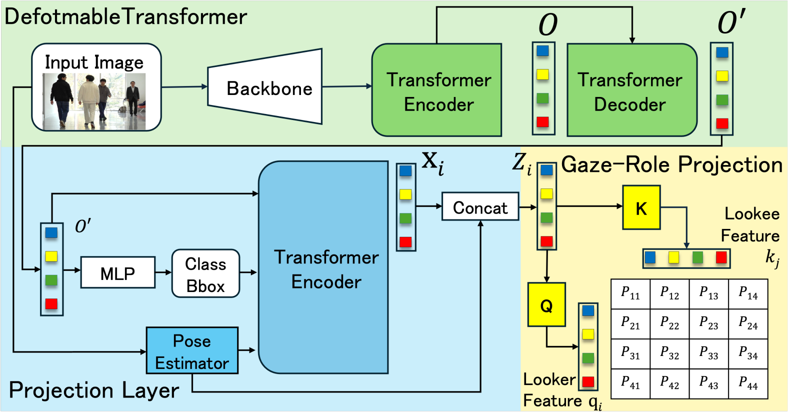


Figure 2: Gaze Object Estimation Network

Fig.1, there are multiple pedestrians in the direction Pedestrian A is facing. In this case, it is unclear which pedestrian A is gazing at if we only focus on Pedestrian A. On the other hand, when we humans estimate the gaze objects of surrounding pedestrians, we consider the gaze objects of other pedestrians together. In the example of Fig.1, A and B are walking side by side in the same direction, and B appears to be gazing at A, so we can estimate that A and B are in the same group, and it is highly likely that A is also gazing at B. This paper performs gaze object estimation that considers the relationships among multiple pedestrians, imitating this human method of gaze object estimation. Additionally, Murakami et al.[4] assumed that pedestrians always have gaze objects. In reality, there are many cases where people don't have specific gaze objects, such as when vaguely looking in the direction of travel. If a person isn't looking at anything specific, but we think they are, we might misinterpret the interaction. To solve these issues, this paper proposes a gaze object estimation method that considers relationships among multiple pedestrians and includes the presence or absence of pedestrian's gaze.

- While conventional methods performed gaze object estimation for each pedestrian independently, this paper improves estimation accuracy by considering relationships among multiple pedestrians in gaze object estimation.
- This paper introduces the PEDESTRIAN AWARENESS DATASET, which assumes environments where autonomous mobile robots operate, for the gaze object estimation task. It is based on scenarios where pedestrians follow predefined gaze targets and walking paths.

2. PROPOSED METHOD

In this paper, for the purpose of improving the accuracy of gaze object estimation, we perform gaze object estimation that considers relationships among multiple pedestrians. Specifically, we optimize every pedestrian's "looker feature $\mathbf{q}_i$" and "lookee feature $\mathbf{k}_j$" using the loss function shown in Eq.(1).

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{N} Y_{ij}\log(P_{ij}) \qquad (1)$$

$$P_{ij} = \frac{\exp\left(\frac{\mathbf{q}_i^T\mathbf{k}_j}{\sqrt{d}}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\mathbf{q}_i^T\mathbf{k}_j}{\sqrt{d}}\right)} \qquad (2)$$

Through optimization, the looker feature $\mathbf{q}_i$ learns to represent who pedestrian i is likely gazing at, and the lookee feature $\mathbf{k}_j$ learns to represent who is likely gazing at object j. This loss function is calculated only for pedestrians with ground truth labels for gaze objects. The model output P_{ij} is the probability that pedestrian i is gazing at object j (include pedestrian). It is obtained by applying the softmax function for normalization to the score calculated by the inner product of $\mathbf{q}_i$ and $\mathbf{k}_j$ (Eq.(2)). Additionally, Y_{ij} is the ground truth that takes the value 1 when pedestrian i is gazing at object j, and 0 otherwise. Both $\mathbf{q}_i$ and $\mathbf{k}_j$ are optimized simultaneously because the loss function L propagates gradients based on backpropagation to each feature. As shown in Eq.(3) and Eq.(4), $\mathbf{q}_i$ and $\mathbf{k}_j$ are updated by weighted sums of $\mathbf{k}_j$ and $\mathbf{q}_i$, respectively. With the weight given by the error term $P_{ij} - Y_{ij}$, they are updated to increase their inner product when $Y_{ij} = 1$ and to decrease it when $Y_{ij} = 0$.

$$\frac{\partial L}{\partial \mathbf{q}_i} = \sum_{j=1}^{N} (P_{ij} - Y_{ij}) \frac{\mathbf{k}_j}{\sqrt{d}} \quad (3) \qquad \frac{\partial L}{\partial \mathbf{k}_j} = \sum_{i=1}^{N} (P_{ij} - Y_{ij}) \frac{\mathbf{q}_i}{\sqrt{d}} \quad (4)$$

As described above, by bringing the looker features $\mathbf{q}_i$ and lookee features $\mathbf{k}_j$ closer together or farther apart through a single loss function, we optimize the gaze objects of multiple pedestrians. The gaze object estimation network that realizes the proposed method consists of three modules: Deformable Transformer which acquires features corresponding to each object; Transformer encoder which simultaneously processes features corresponding to each object and captures relationships among objects; Gaze-Role Projection which divides the features of each pedestrian and object into looker features $\mathbf{q}_i$ and lookee features $\mathbf{k}_j$ and finally outputs the estimation results of gaze objects. Each module is explained below.

2.1 Configuration of Gaze Object Estimation Network

1) Deformable Transformer The input image ($\mathbb{R}^{C \times H \times W}$) is fed into Backbone (ResNet50[5]) to acquire image features. The Deformable Transformer[6] with an Encoder-Decoder architecture takes these features as input and outputs Object Queries $O' = \{\mathbf{oq}_1, \mathbf{oq}_2, \ldots, \mathbf{oq}_N\}$, where each $\mathbf{oq}_i$ is a feature corresponding to a specific object in the image.
2) Projection Layer The Projection Layer consists of a Pose Estimator and a Transformer Encoder. The Pose Estimator estimates pedestrian poses within each BBox and outputs Pose Features P''. BBox coordinates and classes are predicted from Object Queries through an MLP. The Transformer Encoder[7] takes Object Queries O', Pose Features P'', BBox, and class embeddings, and through self-attention, each object query $\mathbf{oq}_i$ interacts with all features of all objects to output $\mathbf{x}_i$.
3) Gaze-Role Projection Gaze-Role Projection takes $\mathbf{z}_i = [\mathbf{x}_i; {p''}_i]$ as input, which concatenates the pedestrian feature vector obtained from the Transformer encoder and the Pose Feature P'' of each pedestrian. We term this module "gaze-role projection" because it linearly projects each object's features into two distinct roles: "looker" and "lookee." From the input $\mathbf{z}_i$, it separates into "looker features $\mathbf{q}_i$" and "lookee features $\mathbf{k}_j$" using two projection matrices $\mathbf{W}_q$ and $\mathbf{W}_k$. Pedestrians are assigned both roles (transformed into both $\mathbf{q}_i$ and $\mathbf{k}_j$), while objects other than pedestrians are assigned only the "lookee" role (transformed only into $\mathbf{k}_j$). Finally, the inner product of $\mathbf{q}_i$ and $\mathbf{k}_j$ is taken and normalized by the softmax function to output the probability P_{ij} that pedestrian i is gazing at object j (Eq.(2)).

In this method, since Gaze-Role Projection simultaneously estimates the gaze objects of multiple pedestrians using the features $\mathbf{x}_i$ of all objects output by the Transformer encoder. Thus, $\mathbf{oq}_i$ corresponding to pedestrians in the Transformer encoder can acquire gaze information that considers the gaze objects of surrounding pedestrians. This realizes gaze object estimation that considers relationships among multiple pedestrians.

3. EXPERIMENT

We perform pedestrian gaze object estimation of pedestrians using the gaze object estimation network. The experimental settings and results are described below.

3.1 Pedestrian Awareness Dataset

We created the PEDESTRIAN AWARENESS DATASET, which annotates the gaze objects of multiple pedestrians. Although existing several datasets assume environments surrounding autonomous mobile robots such as shopping malls, including MOT16,[8] they do not have gaze object annotations. Additionally, the "Gaze Target Dataset in Traffic Scenes" constructed by Murakami et al.[4] includes annotations of pedestrian gaze objects, but since it adds annotations from a third-person view to traffic scene datasets published by Waymo and others,[9] the ground truth of gaze objects is unknown. Furthermore, this dataset targets outdoor traffic scenes, which differs from the environments surrounding autonomous mobile robots assumed in this paper. Therefore, there is no existing dataset that includes both interactions among pedestrians and the ground truth of these pedestrians' gaze objects.

Thus, we constructed the PEDESTRIAN AWARENESS DATASET, which includes ground truth annotations of pedestrians' gaze objects in environments surrounding robots. Ground truth annotation is inherently impossible through estimation from third-person view. Approaches to obtain ground truth include asking actual pedestrians about their gaze objects afterward or specifying them in advance. However, asking afterward may result in uncertain participant memory and complicated interviewing, making it unsuitable for large-scale data collection. Therefore, we created the dataset through scenario-based experiments where gaze objects were specified in advance. Specifically, assuming environments surrounding autonomous mobile robots, we prepared multiple scenes in advance that include interactions among pedestrians (specifying which pedestrians are gaze objects and walking directions), conducted video recording following the scenes, and performed frame extraction and annotation. The gaze objects of pedestrians were kept the same for a certain period to reduce annotation cost. The five scenes are shown below.

- Passing (include Fig.1)
- Overtaking
- Following
- Crossing
- Free movement (with fixed gaze target pedestrian)

Among pedestrians, there are pedestrians gazing at specific pedestrians, pedestrians who don't have specific gaze objects, and pedestrians without gaze object annotations. Gaze object annotations and annotations indicating no specific gaze object were given only to pedestrians whose entire head is visible and pedestrians whose gaze object is partially visible within the frame. Annotations for each pedestrian contain pedestrian ID, gaze object ID, label indicating no specific gaze object, label for gazing at smartphone, and BBox coordinates.

3.2 Experiment

Each value in Table 1 is from the PEDESTRIAN AWARENESS DATASET detected by Deformable DETR.[6] For the experiment, we only evaluate the gaze estimation accuracy for cases where Deformable DETR successfully detects both the pedestrian and ground truth gaze object. The gaze object candidates when estimating with the gaze object estimation network are pedestrians only, and at test time, only images with four pedestrians are used. Additionally, during training, only the parameters of the Projection Layer are updated, and the Backbone and Deformable DETR[6] are initialized with pre-trained weights. Among the proposed methods, the one that utilizes Pose Estimator is denoted as ours1 (oq+pose), and the one that does not utilize Pose Estimator is denoted as ours2 (oq). In the experiment, we compare three methods including Murakami et al.'s method.[4] And we modify the design of Murakami's method to enable training and inference for pedestrians who don't have specific gaze objects. We also utilize RTMPose[10] as the Pose Estimator.

3.3 Results and Discussion

Fig.3 evaluates three methods (ours1 (oq+pose), ours2 (oq), and Murakami) using Top-K Accuracy. Top-K Accuracy indicates the rate at which the ground truth is within the top K results when the gaze object estimation results are sorted by probability in descending order. For Top-1 Accuracy, ours1 (oq+pose) achieved 79.96%, ours2 (oq) achieved 64.15%, showing improvement compared to Murakami[4]'s 51.53%. This result demonstrates that gaze object estimation considering surrounding pedestrians, especially when utilizing pose information, improves estimation accuracy.

Tables 2, 3, and 4 show the confusion matrices for each of the three methods. "Wrong" in the True Positive region refers to cases where the estimated gaze object differs from the ground truth gaze object. Below we discuss observations regarding the confusion matrices. Murakami[4]'s method has more false negatives compared to the two proposed methods. Specifically, out of 10,625 pedestrians who actually have gaze objects, 3,080 (29.0%) are incorrectly estimated as "not having a specific gaze object." This tendency is likely influenced by the structure of the training data. From Table 1, in the training data, the proportion of "pedestrians without specific gaze objects" is approximately 45% (12,779/28,504). On the other hand, pedestrians "with specific gaze objects" account for approximately 55% (15,725/28,504), but since the correct answer must be selected from an average

Table 1: Breakdown of dataset used in the experiment

Item	Training	Test
Number of images	13,170	7,669
Number of pedestrians with specific gaze objects	15,725	10,625
Pedestrians without specific gaze objects	12,779	7,996
Average number of gaze object candidates	2.5	3.0

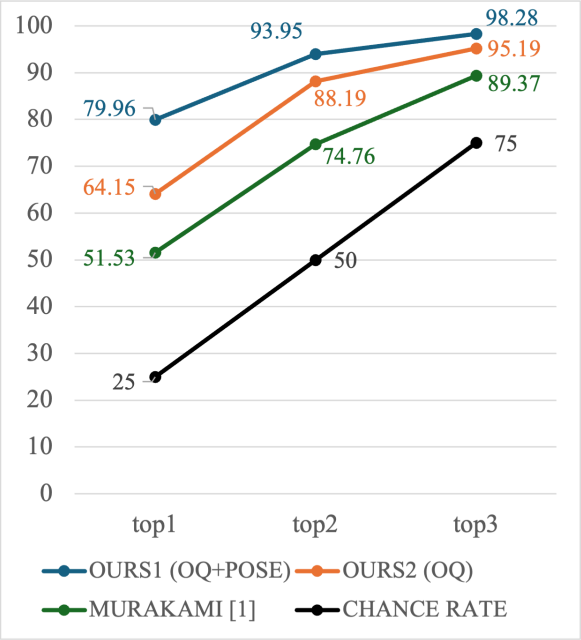


Figure 3: Results

Table 2. OURS1 (OQ+POSE)

	Predicted: Positive	Predicted: Negative	Total
Actual: Positive	correct: 8,259 wrong: 1,070	1,296	10,625
Actual: Negative	1,366	6,630	7,996
Total	10,695	7,926	18,621

Table 3. OURS2 (OQ)

	Predicted: Positive	Predicted: Negative	Total
Actual: Positive	correct: 6,601 wrong: 1,385	2,639	10,625
Actual: Negative	2,652	5,344	7,996
Total	10,638	7,983	18,621

Table 4. MURAKAMI[4]

	Predicted: Positive	Predicted: Negative	Total
Actual: Positive	correct: 3,943 wrong: 3,602	3,080	10,625
Actual: Negative	2,343	5,653	7,996
Total	9,888	8,733	18,621

of 2.5 candidates, the accuracy when randomly selected would be approximately 22% (55% × 1/2.5). In other words, in the training data, estimating “not having a specific gaze object” has twice the probability of being correct compared to estimating “gazing at a specific person.” Therefore, when identifying the gaze object is difficult, the model tends to select “not having a specific gaze object,” which has a higher probability of being correct. In contrast, ours2 (oq) reduced false negatives to 2,639 (24.8%). This is likely because by simultaneously considering the gaze relationships of multiple pedestrians through the Transformer encoder, it became possible to identify gaze objects of pedestrians that were difficult to identify through independent estimation.

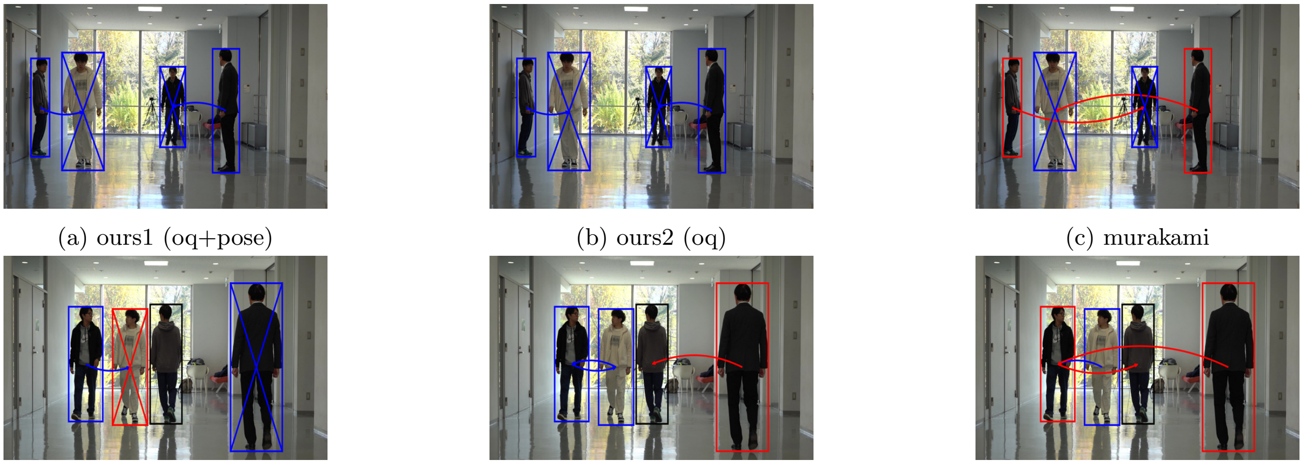
Additionally, Fig.4 visualizes part of the estimation results. Each row in Fig.4 visualizes the estimation results using each model on the same image.

(a)∼(c) show one frame from a crossing scene. Both ours1 (oq+pose) and ours2 (oq) accurately estimate that the pedestrians at both ends are looking at the pedestrian crossing in front of them. However, Murakami[4] makes incorrect estimations for two pedestrians.

(d)∼(f) show one frame from a passing scene where there is a group with mutual gaze. For the 543 pedestrians in total included in this scene, the estimation accuracy was ours2 (oq) (67.3%), ours1 (oq+pose) (58.9%), and Murakami[4] (47.9%). Additionally, among the 128 frames containing pedestrians looking at each other, the rate of correctly estimating both pedestrians with mutual gaze was 50.0% for ours2 (oq), 18.0% for Murakami,[4] and 15.6% for ours1 (oq+pose). It is possible that gaze object estimation considering relationships among multiple pedestrians and gaze object estimation predicted from pose are conflicting.

4. CONCLUSION

In this paper, we proposed an estimation method to estimate what multiple pedestrians surrounding autonomous mobile robots are looking at in environments such as shopping malls. We also constructed the PEDESTRIAN AWARENESS DATASET as a dataset that contains gaze objects considering interactions among pedestrians. With the proposed method, by considering not only the pedestrian being estimated but also what other pedestrians are looking at, we were able to improve the estimation accuracy of gaze objects. And when considering pose information, we were able to significantly improve the accuracy. On the other hand, accuracy improvement is expected in cases where gaze objects inferred from relationships among pedestrians and gaze objects inferred from pose information differ.



(a) ours1 (oq+pose) (b) ours2 (oq) (c) murakami

(d) ours1 (oq+pose) (e) ours2 (oq) (f) murakami

Figure 4: visualization of result: Arrows indicate estimated gaze targets. Cross marks indicate no specific gaze object. Blue: correct estimation. Red: incorrect estimation. Black BBoxes: pedestrians without annotations.

ACKNOWLEDGMENTS

Support for this work was given by the Toyota MotorCorporation (TMC) and JSPS KAKENHI Grant Number 23K28164. However, note that this paper solely reflects the opinions and conclusions of its authors and not TMC or any other Toyota entity.

REFERENCES

[1] Recasens, A., Khosla, A., Vondrick, C., and Torralba, A., "Where are they looking?," in [*Advances in Neural Information Processing Systems (NIPS)*], 199–207 (2015).

[2] Tonini, F., Dall'Asen, N., Beyan, C., and Ricci, E., "Object-aware gaze target detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 21860–21869 (2023).

[3] Hata, R., Deguchi, D., Hirayama, T., Kawanishi, Y., and Murase, H., "Detection of distant eye-contact using spatio-temporal pedestrian skeletons," in [*IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*], 2730–2737 (2022).

[4] Murakami, H., Chen, J., Deguchi, D., Hirayama, T., Kawanishi, Y., and Murase, H., "Pedestrian's gaze object detection in traffic scene," in [*Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*], 333–340 (2024).

[5] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 770–778 (2016).

[6] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J., "Deformable detr: Deformable transformers for end-to-end object detection," in [*Proceedings of the 9th International Conference on Learning Representations (ICLR)*], (2021).

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," in [*Advances in Neural Information Processing Systems (NIPS)*], 6000–6010 (2017).

[8] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K., "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831* (2016).

[9] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al., "Scalability in perception for autonomous driving: Waymo open dataset," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 2443–2451 (2020).

[10] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., and Chen, K., "Rtmpose: Real-time multi-person pose estimation based on mmpose," in [*arXiv preprint*], (2023). arXiv:2303.07399.