2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Vehicle Intention Classification Using Visual Clues

Marvin Klemp[1], Royden Wagner[1], Kevin Rösch[1,2], Martin Lauer[1], and Christoph Stiller[1]
Code: https://github.com/KIT-MRT/vif

***Abstract*— Classifying intentions of other traffic agents is an essential task for intelligent transportation systems. To simplify this task, vehicles are equipped with various illumination systems, including turn indicators, emergency lights, rear lights, and brake lights. We extend the Waymo open perception dataset with ground truth annotations for different visual intentions to develop methods designed to classify the state of such systems. Furthermore, we propose the VISUAL INTENTION FORMER, a two-step transformer-based architecture to classify visual intentions in image sequences of tracked traffic participants. We use a vision transformer to extract image features, which are passed into a transformer encoder that reasons about temporal dependencies among them. We evaluate against different baseline architectures where our proposed method achieves state-of-the-art results. Additionally, we conduct an in-depth performance analysis of our method regarding different input sequence lengths, vehicle headings, and daytime conditions.**

I. INTRODUCTION

For safe driving, an intelligent transportation system (ITS) must perceive its environment, predict the actions of other traffic agents, and plan its own actions accordingly. Humans heavily depend on visual cues to perceive and predict the behavior of other traffic participants. These cues come from diverse illumination systems, such as headlights, daytime running lights, turn indicators, rear lights, brake lights, parking lights, and emergency lights. For example, an active turn indicator signals to surrounding traffic participants the intention to change lanes. We refer to such information as visual intentions.

Most, state-of-the-art prediction algorithms operate through a two-stage process. Initially, a perception system detects and tracks traffic agents. Subsequently, each tracked agent is passed with its relevant features to a prediction algorithm. The majority of prediction methods require both an environmental description and the current [1] or past [2], [3] physical attributes (such as location and velocity) of other traffic agents. A comprehensive examination of motion prediction methods is available in [4].

Only a small number of prediction methods incorporate visual features. For instance, [5] employs the front camera image as an input in addition to other features. Unfortunately, this trend is also shown in recent datasets. None of the large scale motion prediction datasets such as the Waymo Open Dataset (motion subset) [6], the Lyft level 5 dataset [7], or the Argoverse 2 dataset (prediction subset) [8], contain camera sensor data.

[1]Authors are with, KIT Karlsruhe Institute of Technology, Institute of Measurement and Control Systems, Engler-Bunte-Ring 21, 76131 Karlsruhe, Germany {firstname.lastname}@kit.edu
[2]Authors are with, FZI Research Center for Information Technology

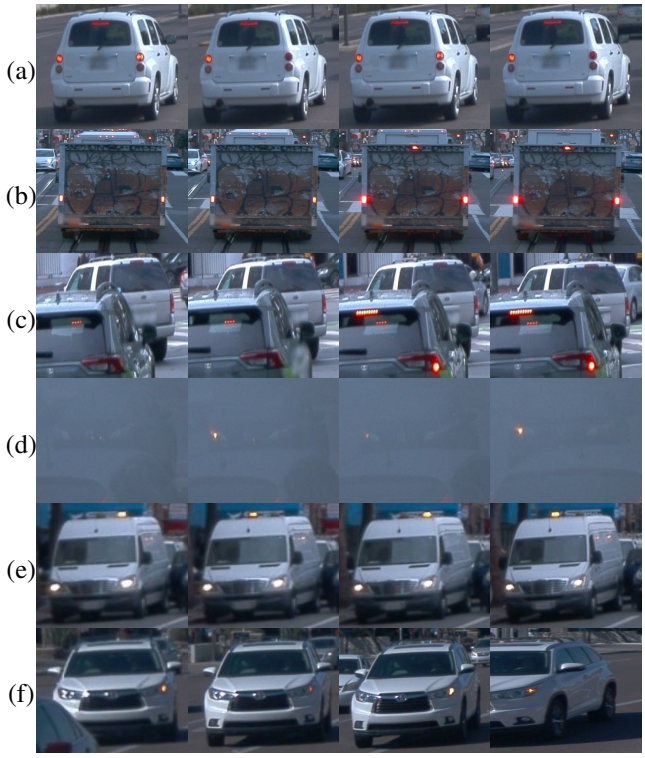


Fig. 1: Visualization of diverse vehicles, conditions, and illumination setups. (a) A car braking while simultaneously indicating a right turn. The brake, rear, and turn indicators share the same light source. (b) A truck with its rear lights on in the initial two frames. In the third and fourth frames, the truck applies the brakes, resulting in a more intense light. (c) The black vehicle with active brake lights significantly occludes the tracked silver truck. (d) Adverse weather conditions. (e) External active emergency lights on top of the vehicle. (f) A vehicle front facing the ego vehicle. The turn indicator on the right hand side is active, indicating a left hand turn.

While not all visual intentions hold equal significance for prediction, we hypothesize that incorporating visual intentions based on turn indicators, emergency lights, and brake lights could enhance the accuracy and confidence of motion prediction algorithms. Hence, as a first step towards validating this assumption, we extend the perception subset of the Waymo Open Dataset with precise ground truth annotations for such visual intentions. Furthermore, leveraging these annotations, we propose, train, and evaluate a transformer-based technique to classify these visual intentions.

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

Our main contributions are summarized as follows:

- We extend the Waymo Open Dataset (perception subset) with ground truth annotations regarding the visual intention of vehicles.
- We propose a transformer-based deep learning architecture to classify the visual intention of a vehicle.
- In our evaluations, we demonstrate state-of-the-art performance compared to different baseline methods across multiple metrics.

II. RELATED WORK

There is little public research available on the classification of visual intentions. However, there are manufacturers that offer products capable of such tasks.

In [9], a dataset and a method are presented for classifying the state of a vehicle's brake lights, and the left and right turn indicator. The dataset contains 91,068 frames in different road and daytime conditions. However, the dataset solely offers annotations for the rear of vehicles, omitting annotations for vehicles approaching from the opposite direction. Their approach operates by extracting two key features from each frame: a brightness feature on a scale of 0 to 255 and an action feature (none, brakes on, brakes off, left, right). Using these features, a probabilistic graph model infers the continuous illumination state of the vehicle.

In [10], a dataset and a convolutional LSTM-based method are presented for the classification of the turn, and emergency lights. Their dataset is extensive, containing 1,257,591 frames captured under diverse conditions. Moreover, the dataset includes annotations for both front and rear facing vehicles. However, the dataset has not been made publicly available.

Furthermore [11] addresses the task of recognizing turn signals of other vehicles in an European highway scenario. Their approach involves light spot detection, followed by an FFT-based feature extraction, and ultimately an AdaBoost classification using the obtained features.

Unfortunately, none of these publications provides code.

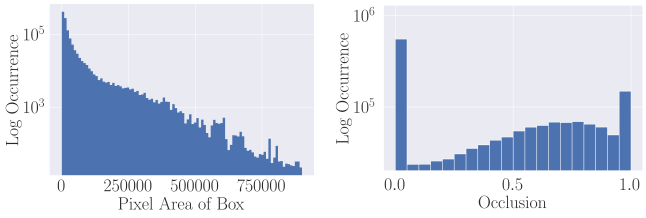
III. DATASET

A. Overview

The Waymo Open Dataset is structured into 20 seconds long sequences. Each sequence contains annotations for traffic participants and sensor data from various modalities, including LiDAR and camera. The dataset contains annotations for 798 training and 202 validation sequences.

We extend the ground truth annotations for the front-facing camera with the intended turn direction, and the state of the rear, break, and emergency lights. The annotation of small vehicles is challenging on its own. Determining if a visual intention is present in a small image can be uncertain and could lead to inaccuracies in the ground truth. In order to prioritize annotation accuracy, we choose to favor annotation quality and exclude objects with an area smaller than 1000 pixels.

The dataset includes a variety of different conditions as it was captured across multiple cities, various times of day, and diverse weather scenarios. Although regulations exist for turn indicators to have a frequency between 1 Hz and 2 Hz [12], they are not always followed. Additionally, modern cars may have exotic turn indicators, such as LED strips indicating the turn direction. Moreover, regulations are not worldwide. While in the US red and yellow colored turn indicators are allowed, the EU enforces turn indicators to be yellow. Hence, as vehicle illumination setups are not uniform, classifying their visual intentions extends beyond challenging environmental conditions to understanding the shape, color, and characteristics of the diverse illumination systems. This results in the need of a diverse large scale dataset. Fig. 1 illustrates examples extracted from the Waymo Open Dataset, showcasing a variety of vehicles, weather conditions, times of day, and illumination setups.



(a) (b)

Fig. 2: Histograms showing a) the area of the projected 2d bounding boxes and b) the occlusion parameter.

B. Annotations

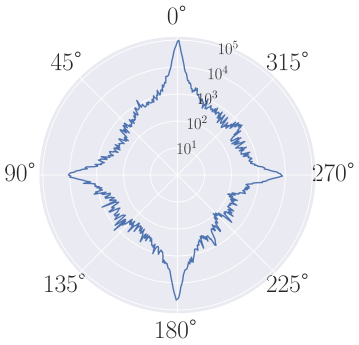


Fig. 3: The distribution of vehicle headings on a logarithmic scale.

To be leveraged for prediction tasks, the turn intention annotations are established from the perspective of the annotated vehicle. If a vehicle shares the same heading as the ego vehicle and its right turn indicators are active, a right turn intention is annotated. Conversely, if a vehicle is front-facing the ego vehicle and its right turn indicators are active, a left turn intention is annotated. This results in the need that methods for turn intention classification require to understand the heading of a vehicle. Fig. 3 shows the distribution of the vehicle heading within the dataset.

The brake and rear lights are typically colored red and can exist as either separate light sources or share the same source.

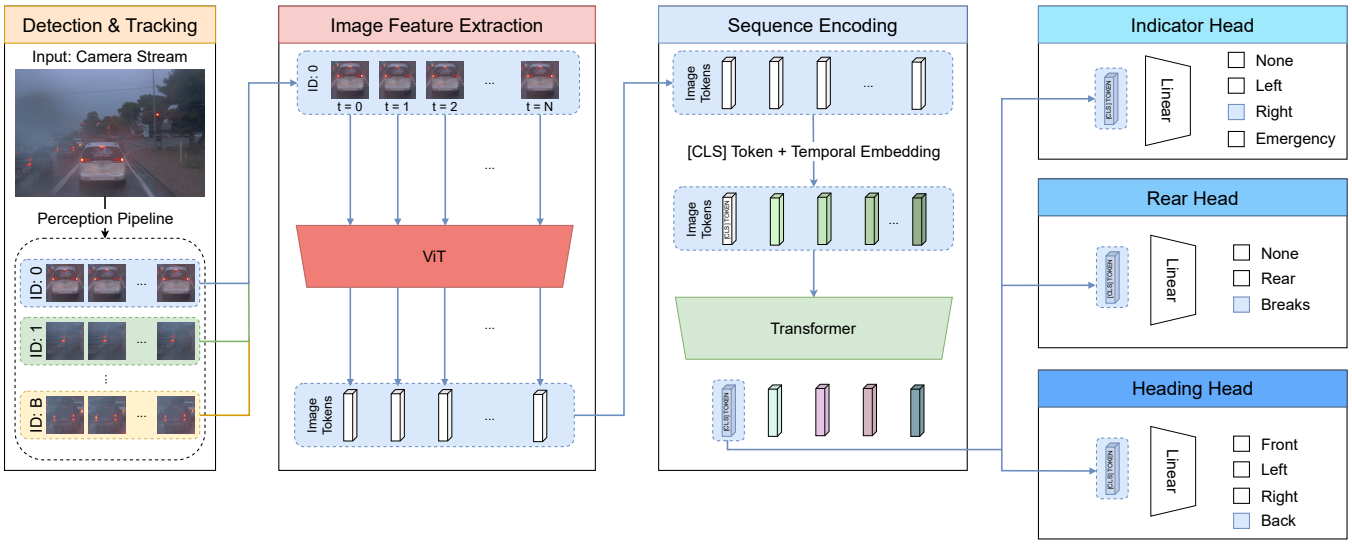


Fig. 4: Visualization of the VISUAL INTENTION FORMER architecture. The detection and tracking of other traffic participants is assumed to be given. Initially in the image feature extraction stage, an image feature vector for each image within a vehicle track is extracted by a vision transformer. Subsequently, these feature embeddings are passed into the sequence encoding stage. In this, a transformer reasons about temporal dependencies among the image feature tokens and encodes information about the sequence in a [CLS] token. Finally, different classification heads predict the visual intentions using the [CLS] token.

In both configurations, the brake light emits a higher intensity compared to the rear light (Fig. 1, b). To distinguish between the brake light and rear light, both classes are annotated.

The classification of emergency lights is a hard task, as emergency lights come in various forms, to name a few, external flashing lights, alternating flashing lights, and the same light sources as the turn indicators.

Furthermore, a challenging aspect is, how strongly a vehicle is occluded (Fig 1, c). We calculate an occlusion parameter between 0 to 1 by calculating the overlap between bounding boxes within a frame and provide this information as ground truth. Fig. 2, b shows the distribution of the occlusion parameter.

C. Data Overview

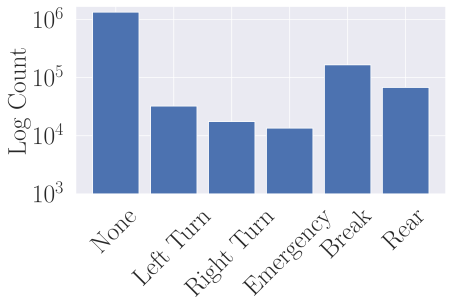


Fig. 5: Class distribution of the visual intention annotations.

In total we annotated 1.552.397 frames based on 25327 vehicle tracks with the following visual intentions NONE, LEFT, RIGHT, EMERGENCY, REAR, and BREAK. Figure 5 provides insight into the class distribution. Mostly, no visual intention is observable. Brake and rear lights rank as the second most intention. Notably, the distribution of left and right turns is nearly balanced.

IV. METHOD

We introduce the VISUAL INTENTION FORMER (VIF), a transformer-based approach designed to classify the visual intention of vehicles. Our method requires a batch of vehicle tracks, where each track consists of the past N image crops of the corresponding vehicle, and classifies the visual intention of each vehicle within the batch at the last frame. We assume the presence of a fully functional perception pipeline, such as [13] and [14] for vehicle detection and tracking, respectively. A full visualization of our method is shown in Fig. 4.

A. Image Feature Extraction

Transformer models [15] learn dependencies among a fixed number of input tokens. To employ a transformer architecture for visual intention classification, we begin by extracting relevant image features and representing them as token vectors. For this purpose, we use a vision transformer [16] architecture (ViT), which was pretrained on the ImageNet 1k challenge [17]. We first resize all image within the batch to a shape of 224x224 pixels and subsequently normalize the pixel values. Following this, we reshape the batch of vehicle tracks [B, N, C, H, W], where B is the number of vehicle tracks, N the number of images per track, and C, H, W are the channel, height, and width dimension, into a batch of images [B*N, C, H, W] and generate image feature embeddings using the ViT model. Afterwards, we reshape the batch of embeddings back into the original batch of

TABLE I: Number of parameters for both image feature extraction and sequence encoding transformer models.

Parameter	Image Feature Extraction	Sequence Encoding
Token dimension	768	768
Number of heads	16	16
Head dimension	64	64
Depth	6	2
MLP dimension	1536	1536

vehicle tracks. Where each track now consists of a sequence of image embeddings. For the image feature extraction stage, we evaluated different CNN-based approaches. However, as the output of CNN-based backbones, such as ResNet [18] are three dimensional, a projection to a one dimensional feature vector has to be learned. The ViT architecture has the advantage that generated feature representations are already a one dimensional feature vector. Hence, the output can be directly passed into the sequence encoding transformer.

B. Sequence Encoding

Once each image in the sequence is encoded as a token vector, we prepend a learnable [CLS] token [19]. In natual language processing tasks, a [CLS] token primarily learns key informations about the entire sentence required for specific downstream tasks. Additionally, we incorporate a learned position embedding for each token. Since this embedding encapsulates the temporal order of images within the sequence, we reference to it as the temporal embedding. Afterwards, we pass this tokens set into a transformer encoder and use the [CLS] token as output. Parameters of the image feature extraction and sequence encoding transformer are outlined in Table I.

C. Visual Intention Classification

Since various vehicle lights can be concurrently active, the classification of the visual intention is a multi-label problem. To address this, we use multiple classification heads.

The REAR head is responsible for classifying the state of the rear and brake lights. Consequently, the output classes include NONE, REAR, or BREAKS. This head is implemented through a linear projection of the [CLS] token into three neurons, each corresponding to one class.

For the classification of the LEFT, RIGHT, and EMERGENCY states, the INDICATOR head is introduced. In a similar manner, a linear projection is employed based on the [CLS] token.

Given that the accuracy of turn intention heavily depends on the vehicles heading, we additionally enhance our method with a HEADING head. The head is responsible for learning the vehicle orientation, enabling the INDICATOR head to use this information for the classification indicator-based visual intentions. We model this head as classification task with the classes BACK, RIGHT, FRONT, LEFT and calculate the target class based on the vehicle heading. To further improve the prediction results, we apply a median filter on the five past predictions per vehicle track.

D. Training

Due to the absence of annotations for the test set, we split the validation set evenly into two sets of 101 sequences each. This configuration yields a total of 798 training sequences, along with 101 sequences for both validation and testing purposes.

During the training process, for each vehicle track, we use a sliding window approach to segment the complete track into multiple sub-tracks, each being as long as the number of images fed into our method. The target labels are extracted from the last image in each sliding window. Afterwards, we apply multiple augmentation methods to each sub-track. First, with probability of 0.5, we horizontally flip all images within the track. It's important to note that in the event of flipping, the left and right turn intention targets need to be adjusted accordingly. Furthermore, with a probability of 0.5, we augment a batch of images using the augmentation method described in [20]. If this augmentation method is applied, it is applied with consistent parameters across each vehicle track. We train for a maximum of 200 epochs and employ early stopping. All classification heads are trained using focal loss [21]. The total loss is calculated by summing the individual losses. Experiments have shown that keeping the ViT for the image feature extraction frozen lead to sub-optimal results. Hence, we also fine-tune the ViT during training. We use a batch size of 16, the AdamW optimizer [22] with an initial learning rate of $5 * 10^{-6}$, and cosine annealing [23] without warm restarts with a final learning rate of $1 * 10^{-5}$.

V. EVALUATION

We perform our experiments on the 101 test sequences containing visual intentions. Before evaluating on these sequences, we train all methods on the merged training and validation splits using a sequence length of 10 input frames.

A. Baselines

Regarding the baseline methods, we focus on deep learning based methods. As related work does not provide code, we implemented multiple baseline methods. Our adaptation of the DEEPSIGNALS approach presented in [10] required modifications due to dataset variations. In detail, our dataset does not provide per frame illumination annotations (UNKNOWN, OFF, ON). Hence, we remove the head responsible for this classification task. Moreover, since our dataset includes annotations for rear and brake lights, we expand the model by incorporating an additional head to classify the intention based on visual clues from these lights.

Following the recommendation of [9], we apply general time sequence analysis techniques for the purpose of classifying the visual intention of a vehicle. Consequently, we adapt both the 3DRESNET [24] architecture, and the VIVIT [25] architecture to classify visual intentions.

B. Primary Test Dataset Evaluation

Our evaluation primarily relies on the F1-Score for the indicator, and rear visual intention classification task. We

TABLE II: Evaluation results on the full test dataset. Best results are indicated in bold.

Model	F1 Indicator	F1 Rear	F1 Heading
3DRESNET [24]	0.55 ± 0.01	0.75 ± 0.03	0.68 ± 0.02
VIVIT [25]	0.59 ± 0.02	0.77 ± 0.02	0.80 ± 0.01
DEEPSIGNALS [10]	0.64 ± 0.02	0.76 ± 0.04	0.83 ± 0.03
VIF (ours)	$\mathbf{0.70} \pm 0.01$	$\mathbf{0.78} \pm 0.02$	$\mathbf{0.93} \pm 0.02$

conduct three separate training runs and report both the mean and standard deviation of the results. The heading classification task is generally considered as task to help the model understand indicator-based visual intentions. The results of the evaluation are shown in Table II. With a mean F1-Score of 0.55 and 0.75 the 3DRESNET performs worst in the indicator and rear test evaluation, respectively. The second worst performing model is the VIVIT. It is significantly stronger in the heading F1-Score and provides a minor improvement in indicator-based visual intention F1-Score. In comparison, the DEEPSIGNALS method was able to learn all tasks. However, in [10] the method achieves an F1-Score for indicator intention classification around 0.70. Whereas in our evaluation, on our test dataset, a mean F1-Score of 0.64 is achieved. As both datasets are large scale, we suspect that their architecture head to classify the state of the light source (UNKNOWN, OFF, ON) for each frame significantly boosts the models performance. The VIF method achieves state-of-the-art performance, resulting in an mean F1-Score of 0.70 for indicator and 0.78 for rear intention classification. Across all methods, the F1-Score for rear classification does not change significantly. We suspect that this is primarily the case because the rear/brake signals are visually easier to learn compared to turn indicator signals.

For subsequent evaluations, we employ the VIF model, which achieves the highest performance in the indicator-based intention classification task. Fig. 6 shows the confusion matrix for the indicator-based intention classification. It is noteworthy that the method achieves a slightly lower performance for emergency intentions compared to turn intentions. We suspect this to be caused due to an in-balance in the dataset.

TABLE III: Evaluation results of the VIF method on the full test dataset regarding different input sequence lengths. Best results are indicated in bold.

Sequence Length	F1 Indicator	F1 Rear	F1 Heading
5	0.68	**0.81**	0.92
10	**0.72**	0.79	**0.94**
15	0.66	**0.81**	0.93
20	0.64	0.77	0.92
25	0.60	0.78	0.93

C. Sequence Length

As the sequence length of the tracked traffic agents is an essential parameter of our method, we conduct an evaluation with different sequence lengths. Table III shows the evaluation results. We evaluate sequence lengths of 5, 10, 15, 20, and 25 input frames. Considering the indicator-based intention classification F1-Score, shorter sequence lengths return more promising results, where a sequence length of 10 achieves the best results. The classification of break, rear lights, and vehicle heading is not affected by the input sequence length and achieves consistent results using different input lengths. We assume that this is the case as visual features necessary for accurately executing these classifications do not undergo significant changes over time.

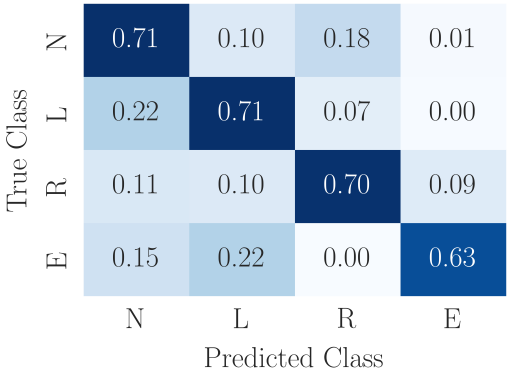


Fig. 6: Confusion matrix for the indicator-based visual intention classification on the test dataset. Labels are N for none, L for left, R for right, E for emergency.

TABLE IV: Evaluation results on the test dataset grouped by the time of the recording.

Task	Day	Night	Dusk/Dawn
F1 Indicator	0.68	0.91	0.50
F1 Rear	0.82	0.70	0.50
F1 Heading	0.93	0.93	0.90

D. Daytime Evaluation

In this experiment, we split the test dataset into different groups based on the daytime. Evaluation results are shown in Table IV. For the indicator and rear intention classification task, the model performs worse during dusk and dawn. This is a typical limitation for perception systems, caused by glares and other vision impairing effects typically occurring during the rising and setting sun. Interestingly, the indicator-based visual intention classification performs significantly stronger in the night. We assume that this is caused by the strong color difference between the indicator and the rest of the car in dark scenes. Hence, indicator-based intentions can be seen more easily and the classification improves. In comparison, the rear and break intention classification performs significantly worse in the night. We assume, that the reduction in performance is caused by the need to further distinguish between the rear and break lights. During day time the rear lights are off, while break lights occur once a vehicle breaks. However, in low-light conditions, such as during the night, the rear lights of a vehicle are always on. When a vehicle starts to brake, they emit a higher intensity.

TABLE V: Evaluation results on the test dataset grouped by the vehicle heading class.

Task	Back	Right	Front	Left
F1 Indicator	0.68	0.80	0.72	0.64
F1 Rear	0.69	0.93	0.91	0.96

E. Vehicle Heading Evaluation

In this experiment, we evaluate the intention classification with respect to the vehicle heading. Therefore, we split the test dataset into separate splits based on the target vehicle heading class. For indicator-based intentions, the results show, that the method is not classifying intentions uniformly regarding the vehicle heading. For back- and front-facing vehicles, the method performs slightly better for front facing vehicles. Meanwhile, vehicles oriented to the left result in decreased classification, while those oriented to the right show improved classification. Regarding the rear-based intention classification, the method performs strongly for vehicles oriented to the RIGHT, FRONT, LEFT. However, this is expected as most of ground truth labels for these cases are NONE because the rear lights are not or barely visible. For back facing vehicles, the method achieves an F1-Score of 0.69. Results are presented in Table V.

F. Architecture Heads

We conduct evaluations to assess the performance of our method with respect to different classification heads. Training the architecture solely with the indicator head led to a performance degradation of 0.12 in F1-Score compared to the architecture with all heads. This validates the assumption that the heading head plays a crucial role in enabling the model to understand the distinction between front- and rear-facing vehicles. Furthermore, we repeat the training of the model containing only the indicator head with inverted class labels for the indicator-based visual intentions. Hence, if a vehicle has the FRONT orientation in the ground truth, the LEFT and RIGHT indicator intention class labels are inverted. If a vehicle has any other orientation, the ground truth label are not inverted. Hence, the model does not have to learn the difference between front- and rear-facing vehicles. However, even with this configuration, the method performs worse in F1-Score compared to the method containing the heading head. Additionally we train the method with all heads and the inverted indicator-based visual intention class labels. Nevertheless this did not lead to further improvements in the indicator intention F1-Score. The results of this evaluation are shown in Table VI.

TABLE VI: Evaluation results for the VIF architecture with varying heads. I, R, H indicate that the indicator, rear, and heading heads are present, respectively. A * describes that the class labels for the indicator-based visual intentions are inverted for vehicles with the FRONT heading.

Model	F1 Indicator	F1 Rear	F1 Heading
I	0.60	-	-
I*	0.68	-	-
IH	0.71	-	0.92
IR	0.61	0.80	-
IRH	0.72	0.79	0.94
IRH*	0.71	0.80	0.93

VI. LIMITATIONS

Transformer-based models require a fixed size of input tokens. Hence, the proposed method also requires a fixed input sequence length. This leads to the limitation, that the visual intention of a vehicle can only be classified after 10 frames, or 1 second (using a camera sampling frequency of 1 Hz). While this is not a critical limitation, we assume that downstream prediction tasks will perform better without this limitation. A straight forward hypothesis to remove this limitation is to zero pad image feature tokens until the maximum sequence length is reached. However, in a first evaluation this approach led to unpromising results.

VII. CONCLUSION

In this paper, we addressed the task of classifying visual intentions. We extended the Waymo Open Dataset with ground truth annotations and proposed a transformer-based method to solve this task. We train and evaluate our approach against different baseline methods and show state-of-the-art results. In future works, we want to remove the limitation of a fixed input sequence length. Furthermore, we will extend this pipeline to include visual intentions of vehicles into a motion prediction method.

ACKNOWLEDGMENT

We acknowledge funding by the Karlsruhe School of Optics and Photonics (KSOP). Additionally, parts of this work were funded by the German Federal Ministry for Economic Affairs and Climate Action under the grant 19A21045G. Furthermore, this work was supported by the Helmholtz Associations Initiative and Networking Fund on the HAICORE@FZJ partition.

REFERENCES

[1] F. J. Wirth, "Conditional behavior prediction of interacting agents on map graphs with neural networks," Ph.D. dissertation, Karlsruher Institut für Technologie (KIT), 2023.

[2] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2980–2987.

[3] R. Wagner, M. Klemp, C. F. Lopez, and O. S. Tas, "Road barlow twins: Redundancy reduction for motion prediction," in *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023. [Online]. Available: https://openreview.net/forum?id=HI12AzV3ZA

[4] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. A. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behaviour prediction for autonomous driving applications: A review," *CoRR*, vol. abs/1912.11676, 2019. [Online]. Available: http://arxiv.org/abs/1912.11676

[5] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun, "End-to-end contextual perception and prediction with interaction transformer," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5784–5791.

[6] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," *CoRR*, vol. abs/1912.04838, 2019. [Online]. Available: http://arxiv.org/abs/1912.04838

[7] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," *CoRR*, vol. abs/2006.14480, 2020. [Online]. Available: https://arxiv.org/abs/2006.14480

[8] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," 2023.

[9] W. Song, S. Liu, T. Zhang, Y. Yang, and M. Fu, "Action-state joint learning-based vehicle taillight recognition in diverse actual traffic scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18 088–18 099, 2022.

[10] D. Frossard, E. Kee, and R. Urtasun, "Deepsignals: Predicting intent of drivers through visual signals," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9697–9703.

[11] B. Fröhlich, M. Enzweiler, and U. Franke, "Will this car change the lane? - turn signal recognition in the frequency domain," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 2014, pp. 37–42.

[12] SAE International, *Side Turn Signal Lamps for Vehicles Less than 12 m in Length*, Std., 08 2014. [Online]. Available: https://www.sae.org/standards/content/j914_201408/

[13] A. Bochkovskiy, C. Wang, and H. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[14] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

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

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[20] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 613–18 624. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[23] ——, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[24] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet?" in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[25] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lui, and C. Schmid, "Vivit: A video vision transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6816–6826.