Cooperative Probabilistic Trajectory Forecasting under Occlusion

Anshul Nayak
Virginia Tech
anshulnayak@vt.edu

Azim Eskandarian
Virginia Tech.
eskandarian@vt.edu

***Abstract*— Perception and planning under occlusion is necessary for safety-critical tasks. Modern autonomy has been increasingly relying on communication between agents to overcome this. However, communicating big data under adverse conditions and limited bandwidth may not be always feasible. Relative pose estimation between connecting vehicles sharing a common field of view can be a computationally cheap and effective way of communicating location of other traffic agents under such adverse conditions. Therefore, in the current study, we use cooperative perception between communicating agents to reliably estimate and then predict the trajectory of a pedestrian which is occluded in the ego agent reference frame. We show that if we know the relative orientation between connecting vehicles as well as the pedestrian coordinate in the other agent's frame of reference, we can reliably estimate the future trajectory of the pedestrian. Moreover, to ensure safety guarantees downstream, we make probabilistic prediction of the future states.**

I. Introduction

Modern day autonomy relies on accurate detection and forecasting of other agents for navigation. Recently, end-to-end forecasting pipelines were developed which take raw sensor data and forecast the future intention of other agents [1][2]. Typically, sensors continuously perceive the object during forecasting. Yet, there are situations where the object may be partially or fully occluded, rendering detection and forecasting of such objects quite challenging. Recent advances in communication between multiple traffic agents have been used to address detection and forecasting under occlusion[3][4]. In such a scenario, an object occluded from ego agent's perspective is detected by other agents such as vehicles and infrastructure and the information is shared through vehicle-to-vehicle (v2v) or vehicle-to-everything (v2x) communication respectively [5][6]. However, effectively communicating rich sensor information from lidar and camera across multiple agents is expensive and may result in high latency. To overcome this, only necessary information such as position, orientation and velocity of occluded object can be shared by establishing cooperative perception between agents sharing a common field of view. In cooperative perception, each agent recovers its own pose based on shared visual features establishing relative orientation between communicating agents. Relative orientation is established by a rigid body transformation with known rotation and translation between a pair of communicating agents. Once relative orientation is established, critical information about occluded object as observed by other agents can be obtained by the ego agent in real-time [7]. Although past research have focused on both cooperative perception [8] and relative orientation, to the best of our knowledge, cooperative perception under occlusion for prediction and planning has been unexplored.

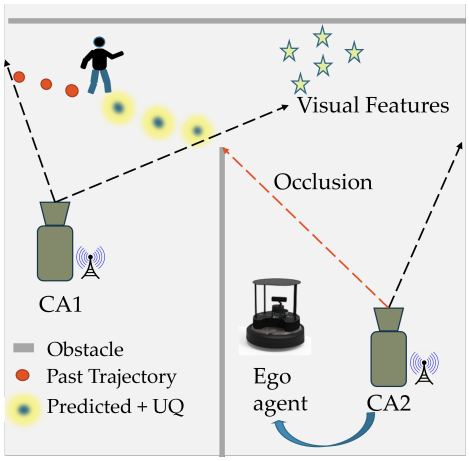
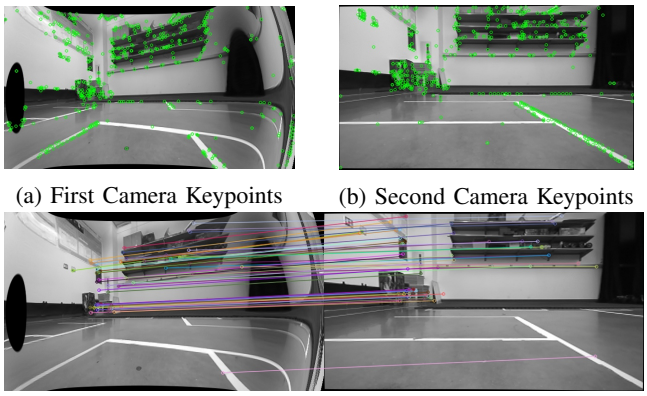


Fig. 1: A Schematic of cooperative trajectory prediction under occlusion where the pedestrian is visible to Connected Agent 1 (CA1) while occluded from CA2.

In this paper, we focus on cooperatively forecasting the trajectory of an occluded object in an uncertain scenario based on relative orientation (see Figure 1). In the schematic, both the connected agents CA1 and CA2 can be infrastructures with sensors mounted in a traffic scenario or a multi-camera setting [9]. Both the agents share common visual features such that each agent can recover its own pose and eventually obtain the relative orientation of the other communicating agents. However, the pedestrian is only visible to the CA1 and occluded from the view of the CA2. This makes it difficult for the ego agent to obtain the pedestrian's current state for predicting the future states and ensuring safe motion planning. Meanwhile, with the established cooperative perception with CA2, CA1 can send the occluded pedestrian's observed trajectory in real-time and the pedestrian's future states can then be predicted by CA2 through an end-to-end prediction network [11]. Note that CA2 will receive the pedestrian's location in its own frame though pose recovery and rigid body transformation and share that information with ego agent. Our contributions are outlined below: (1) we demonstrate that cooperative perception can be reliably used to recover the past trajectory of the occluded object, (2) the proposed method can accurately predict the future trajectory of the occluded object obtained through cooperative perception.



(a) First Camera Keypoints (b) Second Camera Keypoints

(c) Feature Matching

Fig. 2: Feature Description and Matching: (a-b) Detection of keypoints (corners, blobs, edges) (c) Feature matching between the image pair

II. Proposed Method

A. Relative Pose Estimation

For connected agents, relative pose can be established between multiple agents sharing common visual features. This process consists of two fundamental steps; feature detection and matching followed by pose recovery.

Feature Detection and Matching: The feature detection algorithm finds salient points such as corners, edges or flat surfaces in an image (Figure 2a,2b). Each feature is then described using pixel information of a small patch around it. Descriptors can be gradient-based (like SIFT [16], KAZE) which rely on orientation of gradients in the patch or binary (like ORB [17], AKAZE) which generate an unique binary key for every feature. Once the features and their descriptors have been located for a pair of images, feature matching is applied based on the vectorial distance between descriptors (see Figure 2c). Popular matching methods include the Brute Force (BF) and FLANN-based matcher. However, outliers may be still present with good matches which corrupt the overall pose recovery process. Therefore, RANSAC algorithm with 1000 iterations and 99% confidence has been applied to the matches to reject any outliers.

Pose Recovery: Any point in 3D world gets registered in the image plane through a simple transform $\mathrm{x} = \mathrm{P}\,\mathbf{X}$, where x is homogeneous image coordinate and $\mathbf{X}$ represents 3D coordinates. The matrix $\mathrm{P} = \mathrm{K}[\mathrm{R}|\mathrm{t}]$ stores camera parameters, including the intrinsic camera calibration matrix (K) and extrinsic parameters such as rotation (R) and translation (t), establishing the pose between image pairs with a common field of view. When two cameras capture identical features, the 3D feature coordinates, recorded as homogeneous coordinates in images x and x' for each camera, must satisfy $\mathrm{x}'^{\mathrm{T}}\,\mathrm{F}\,\mathrm{x} = 0$. Here, F denotes the fundamental matrix, derived via Direct Linear Transform (DLT) based on a set of 'n' matches between image pair. Given the knowledge of camera intrinsics K, the essential matrix E can be computed from F as $\mathrm{E} = \mathrm{K}'^{\mathrm{T}}\mathrm{F}\mathrm{K}$.

Relative pose estimation between two connected agents looking at the same visual features can be achieved through the singular value decomposition of the Essential Matrix E into rotation matrix $\mathrm{R} \in \mathbb{R}^{3\times3}$ and translation vector $\mathrm{t} \in \mathbb{R}^{3\times1}$. $\mathrm{E} = [\mathrm{t}]_{\mathrm{x}}\,\mathrm{R}$. The rotation matrix can be further transformed into corresponding Euler angles $[\psi, \theta, \phi]$, unambiguously. However, the essential matrix E is scale-invariant and the absolute distance can not be recovered. Therefore, true distance between cameras, d_{true} is provided to recover the scale factor for translation vector t. Rotation matrix and translation vector can facilitate coordinate transformation of any object in the ego agent's frame of reference.

B. Probabilistic Trajectory Prediction

Cooperative perception enables ego agent to obtain the occluded object's states and then predict the future trajectory of the occluded object. Since, deterministic trajectory prediction of occluded objects can be error-prone, we probabilistically predict the future trajectory with uncertainty bounds for more robustness. This uncertainty-inclusive trajectory prediction can enable safer and more reliable motion planning [14] [15]. For deep learning based trajectory prediction models, some of the popular techniques to approximate the uncertainty include Monte Carlo (MC) dropout [18] or deep ensembles (DE) [19]. In this paper, we use MC dropout as it offers uncertainty estimates without significant changes to the neural network (NN) architectures. Specifically, MC Dropout is applied during inference and it introduces stochastic dropout of weights at each layer with some probability, Bernoulli(p_i). The inference process is repeated for N times for the same input, x^* to generate a distribution of outputs $\{y_1^*, y_2^*, ..., y_N^*\}$. The mean and the variances of the distribution can then be computed from the obtained samples.

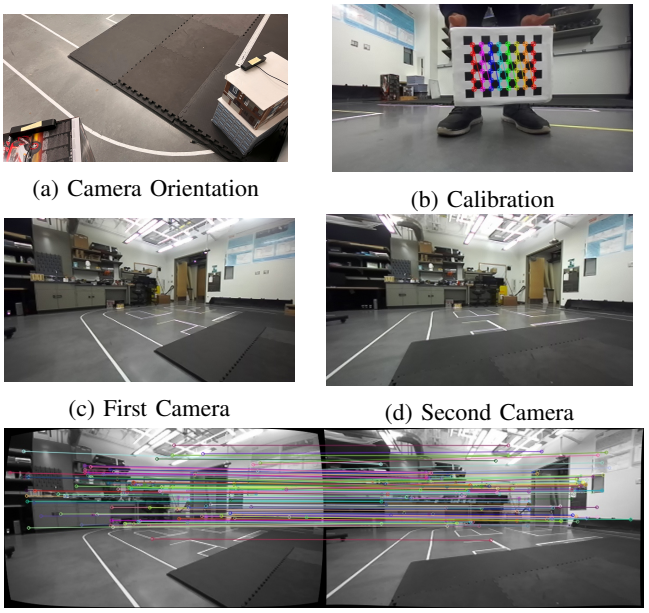
Encoder Decoder Model In this paper, we design an LSTM-based encoder-decoder architecture [**?**] to forecast pedestrian trajectories over varying time horizons. The encoder transforms the input trajectory sequence $\{\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_T\}$ for T time steps, into an encoded space vector 'e' using a nonlinear function, i.e. e = g(x). An ablation study revealed the advantage of encoding both the position and velocity, $\mathrm{X} = \{x, y, u, v\}$, over just encoding the position. This encoded information is subsequently used by the decoder to predict future states $\{\mathrm{X}_{T+1}, \mathrm{X}_{T+2}, ..., \mathrm{X}_{T+F}\}$. The NN predicts the future position of pedestrians $\hat{x}$ and $\hat{y}$. During inference, MC dropout with probability p is applied to infer the distribution for future trajectory estimates.

III. EXPERIMENTS

A. Cooperative Perception

For establishing cooperative perception and obtaining the relative orientation in real-time, we placed two depth cameras with some known orientation as shown (Figure 3a). Multi-camera grab with software synchronisation was performed to ensure that both the cameras capture images simultaneously. Each camera was calibrated using a standard 9" x

7" checkerboard pattern to obtain the camera intrinsics, K (Figure 3b). Both the cameras were exposed to almost similar visual features, albeit from different perspective owing to the location and orientation of each camera. The different perspective view as seen from each camera has been represented in figures 3c and 3d respectively. Images are transformed into gray scale and relative pose estimation is obtained based on the steps mentioned in Sec.II-A. First, feature description and matching are carried out based on common matched features from both images (Figure 3e). Further, with camera intrinsics, the fundamental matrix, F and essential matrix, E are obtained.



(a) Camera Orientation (b) Calibration

(c) First Camera (d) Second Camera

(e) Feature Matching

Fig. 3: Relative pose estimation between two cameras with common visual features

For the current experiment, the ground truth relative orientation represented using Euler angles was obtained directly from the internal gyroscope and accelerometer of the camera (Figure 3a). Since, the cameras are on a flat surface, the roll and pitch angles were negligible while the ground truth yaw orientation was $19.12°$. Further, the true distance between cameras, d_{true} was also measured to estimate the exact scale = $\frac{d_{true}}{\|t\|}$ of the translation vector t. We have tabulated the results for relative pose estimation (Tab. I). The average feature descriptors on each image along with the total good matches for the image pair were considerably high with a low matching time of 0.54 seconds to obtain the fundamental matrix F. The experimentally evaluated rotation matrix $\mathrm{R}^{3\mathrm{x}3}$ and translation vector $\mathrm{t}^{3\mathrm{x}1}$ between the cameras are below.

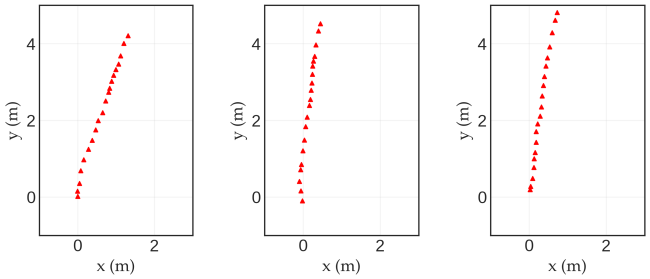
$$\mathrm{R} = \begin{bmatrix} 0.927 & -0.0447 & 0.370 \\ 0.0233 & 0.997 & 0.062 \\ -0.372 & -0.048 & 0.926 \end{bmatrix} \quad \mathrm{t} = \begin{bmatrix} 1.163 \\ 0.066 \\ 0.040 \end{bmatrix} \tag{1}$$

The rotation matrix R can be converted to Euler angles, rpy = [1.44, -3.018, 21.878] closely matches the ground truth orientation obtained from imu pose data within the camera (Tab. I). The rotation matrix is non-degenerate as the diagonal elements of the matrix R are close to identity.

Ground Truth	rpy = [1.31, -1.767, 19.12]
Average Estimate	rpy = [1.44, -3.018, 21.878]
Average Feature points	1290
Good Matches	128
Matching Time	0.54 secs

TABLE I: Results for Relative Orientation

Figure 4a represents the pedestrian trajectory in CA1's reference frame. We compute the transformation of pedestrian states to CA2's reference frame using the estimated relative orientation between cameras $[\mathrm{R}|\mathrm{t}]$. Since, the essential matrix $\mathrm{E} = \mathrm{K}_1^{\mathrm{T}}\mathrm{F}\mathrm{K}_2$ computes the relative orientation of the second camera with respect to first camera, we use inverse rigid body transformation, $[\mathrm{X}', \mathrm{Y}', \mathrm{Z}']^{\mathrm{T}} = \mathrm{R}^{\mathrm{T}} \times ([\mathrm{X}, \mathrm{Y}, \mathrm{Z}]^{\mathrm{T}} - \mathrm{t})$ to transform the pedestrian trajectory from first camera $[\mathrm{X}, \mathrm{Y}, \mathrm{Z}]$ to second camera's $[\mathrm{X}', \mathrm{Y}', \mathrm{Z}']$ reference frame (Figure 4b). Figure 4b, 4c represent transformed trajectory and ground truth trajectory in the second camera's reference frame. The plots show that average Euclidean between the trajectories is minimum. Thus, using cooperative perception, one can reliably estimate the relative orientation between two CAs and use that information to obtain the location of any occluded object in its own frame of reference.



(a) First Camera (b) Transformed (c) Second Camera

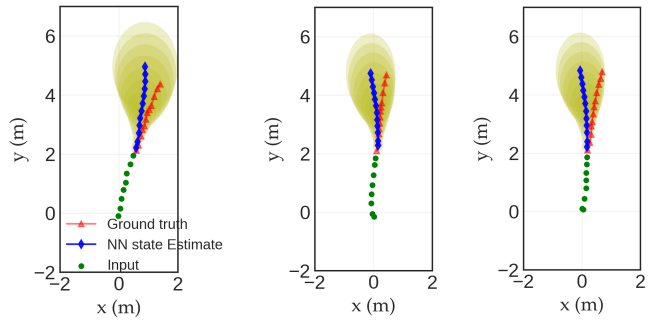
Fig. 4: Pedestrian trajectory in the frame of reference of (a) CA1 (b) Transformed trajectory of first camera using relative pose (c) CA2 assuming no occlusion.

B. Uncertainty-Inclusive Trajectory Forecasting

From the same orientation of cameras (Figure 3b), simultaneous object detection and tracking of the pedestrian was carried using a simple Mask R-CNN [21]. The object detection module accurately classifies and tracks the pedestrian providing 3D world coordinates for position and velocity in real-time. Figure 5a, 5b represent the tracking of pedestrian from two different perspective as seen by individual camera. The trajectories were collected with the camera recording at 30 frames per second for a duration of 8 seconds. The sampling time is set at 12 frames such that the camera obtains the object's position and velocity every 0.4 seconds. Every single trajectory with the duration of 8 seconds results in 20 $\{x, y, u, v\}$ samples, out of which 8 samples (3.2 secs) rep-

(a) First Camera (b) Second Camera

Fig. 5: Simultaneous Object detection and tracking of the pedestrian from two cameras having some relative orientation



(a) 1^{st} camera coordinates (b) 1^{st} camera transformed coords (c) 2^{nd} camera coordinates

Fig. 6: Uncertainty-inclusive trajectory prediction

resent past trajectory while 12 samples (4.8 secs) represent the ground truth which will be used to validate against the NN prediction. The weights of the NN model are trained on publicly available datasets namely ETH[22] and UCY[23]. End-to-end training was carried out minimising the Gaussian NLL loss with Adam optimizer and a learning rate of $1e^{-3}$. NN model was trained for 150 epochs with a batch size of 32. The model was compiled and fit using train and test data. During real-time inference, only model parameters such as trained weights and biases were considered which makes the inference process computationally cheap.

In figure 6, we show the uncertainty-inclusive prediction for the pedestrian trajectory as observed by the first camera. The model takes 8 input states (●) to predict 12 states into future. ▲ represents the actual ground truth trajectory of the pedestrian. Further, the plot shows the mean predicted path (◆) alongwith the 1Σ covariance ellipse to quantify uncertainty during prediction. The plot shows that the ADE between predicted NN state estimate and ground truth is small with the ground truth lying within the 1Σ covariance.

In order to test the reliability of the cooperative perception during trajectory prediction, we transform the coordinates of the original trajectory in CA1's reference which includes both the input and ground truth states to the CA2's frame using rigid-body transform. The transformation was applied to the position coordinates, $\{x, y\}$ and the transformed trajectory was used for prediction. Figure 6b represents the trajectory prediction for transformed coordinates of the CA1 in the CA2's reference after applying relative pose transformation. Meanwhile, figure 6c represents the trajectory prediction for the original trajectory of the pedestrian as seen by CA2 assuming there is no occlusion. The transformed coordinates as well as prediction of the future states for the pedestrian in figure 6b closely matches the original trajectory in figure 6c. To quantify the dissimilarity between probabilistic predicted states, we computed the Kulback-Leibler (KL) divergence at each future time, $\{X_{T+1}, X_{T+2}, ..., X_{T+F}\}$. Assuming, q $\sim \mathcal{N}_k(\mu_q, \Sigma_q)$ and p $\sim \mathcal{N}_k(\mu_p, \Sigma_p)$ represent the bivariate distribution of future predicted states between figure 6b and 6c respectively at any future time $T+F$, then KL-divergence is:

$$KL(p||q) = \frac{1}{2}[log\frac{|\Sigma_q|}{|\Sigma_p|} - d + tr(\Sigma_q^{-1}\Sigma_p) + (\mu_q - \mu_p)^T\Sigma_q^{-1}(\mu_q - \mu_p)] \quad (2)$$

where d=2 for a bivariate distribution. Similarly, the Shannon entropy which signifies the information of a true distribution is represented as $H(p) = \frac{1}{2}log((2\pi e)^d det(\Sigma_p))$. In order to describe the distribution q at each state, we need $KL(p||q)$ bits of more information over the true distribution $H(p)$. In Fig.7, results indicate that the KL-divergence reduces with time which shows that the predictive distributions become more and more similar between the trajectories. Further, $\frac{KL(p||q)}{H(p)}$ which shows how much extra bits of information would be required if we know the true distribution, p increases. Overall, our results show that cooperative perception and trajectory prediction can be reliably combined in adverse scenarios like occlusion producing almost similar results to the ground truth predictions assuming there is no occlusion.

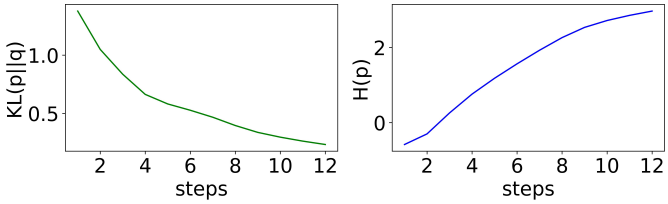


Fig. 7: KL divergence and shanon entropy, H(p) for predicted states.

IV. Conclusion

In this paper, we show the importance of combining cooperative perception for obtaining relative pose estimation and trajectory prediction under occlusion. First, we showed how cooperative perception can be utilised by two cameras sharing common visual features to obtain the accurate relative orientation between them. We also performed pedestrian trajectory tracking in one camera's frame of reference and transformed the coordinates using relative pose to another camera's frame. Our results show that trajectory transformation followed by prediction from another agent's frame of reference was almost similar to the prediction results in the original camera's reference assuming no occlusion. This shows cooperative perception and trajectory forecasting can be combined for prediction and planning under occlusion. In future, the research can be extended to dynamic pose estimation where the ego agents can establish relative pose through visual odometry and simultaneously forecast trajectory of occluded object.

REFERENCES

[1] Casas, Sergio, Wenjie Luo, and Raquel Urtasun. "Intentnet: Learning to predict intention from raw sensor data." In Conference on Robot Learning, pp. 947-956. PMLR, 2018.

[2] Luo, Wenjie, Bin Yang, and Raquel Urtasun. "Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 3569-3577. 2018.

[3] Wang, Tsun-Hsuan, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction." In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 605-621. Springer International Publishing, 2020.

[4] Ruan, Jiageng, Hanghang Cui, Yuhan Huang, Tongyang Li, Changcheng Wu, and Kaixuan Zhang. "A Review of Occluded Objects Detection in Real Complex Scenarios for Autonomous Driving." Green Energy and Intelligent Transportation (2023): 100092.

[5] Sridhar, S., and A. Eskandarian. Cooperative Perception in Autonomous Ground Vehicles Using a Mobilerobot Testbed. IET Intelligent Transport Systems, Vol. 13, No. 10, 2019, pp. 1545–1556.

[6] A. Eskandarian, C. Wu, and C. Sun. Research Advances and Challenges of Autonomous and Connected Ground Vehicles. IEEE Transactions on Intelligent Transportation Systems, Vol. 22, No. 2, Feb. 2021, pp. 683-711. https://doi.org/10.1109/TITS.2019.2958352

[7] Nayak, A., A. Eskandarian, P. Ghorai, and Z. Doerzaph. A Comparative Study on Feature Descriptors for Relative Pose Estimation in Connected Vehicles. In ASME International

[8] Ghorai, Prasenjit, Azim Eskandarian, Young-Keun Kim, and Goodarz Mehr. "State estimation and motion prediction of vehicles and vulnerable road users for cooperative autonomous driving: A survey." IEEE transactions on intelligent transportation systems 23, no. 10 (2022): 16983-17002.

[9] Goldhammer, Michael, Elias Strigel, Daniel Meissner, Ulrich Brunsmann, Konrad Doll, and Klaus Dietmayer. "Cooperative multi sensor network for traffic safety applications at intersections." In 2012 15th International IEEE Conference on Intelligent Transportation Systems, pp. 1178-1183. IEEE, 2012.

[10] Nayak, Anshul, Azim Eskandarian, and Zachary Doerzaph. "Uncertainty estimation of pedestrian future trajectory using Bayesian approximation." IEEE Open Journal of Intelligent Transportation Systems 3 (2022): 617-630.

[11] Nayak, Anshul, Azim Eskandarian, Zachary Doerzaph, and Prasenjit Ghorai. "Pedestrian Trajectory Forecasting Using Deep Ensembles Under Sensing Uncertainty." arXiv preprint arXiv:2305.16620 (2023).

[12] Kim, Seong-Woo, Wei Liu, Marcelo H. Ang, Emilio Frazzoli, and Daniela Rus. "The impact of cooperative perception on decision making and planning of autonomous vehicles." IEEE Intelligent Transportation Systems Magazine 7, no. 3 (2015): 39-50.

[13] Yu, Ming-Yuan, Ram Vasudevan, and Matthew Johnson-Roberson. "Occlusion-aware risk assessment for autonomous driving in urban environments." IEEE Robotics and Automation Letters 4, no. 2 (2019): 2235-2241.

[14] Kahn, G., A. Villaflor, V. Pong, P. Abbeel, and S. Levine. Uncertainty-aware Reinforcement Learning for Collision Avoidance. arXiv preprint arXiv:1702.01182, 2017

[15] Wu, Xihui, Anshul Nayak, and Azim Eskandarian. "Motion planning of autonomous vehicles under dynamic traffic environment in intersections using probabilistic rapidly exploring random tree." SAE International Journal of Connected and Automated Vehicles 4, no. 12-04-04-0029 (2021): 383-399.

[16] Lowe, D. G.. Distinctive Image Features from Scaleinvariant Keypoints. International Journal of Computer Vision, Vol. 60, No. 2, 2004, pp. 91–110.

[17] Rublee, E., V. Rabaud, K. Konolige, and G. Bradski. ORB: An Efficient Alternative to SIFT or SURF. In International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.

[18] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In international conference on machine learning, pp. 1050-1059. PMLR, 2016.

[19] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems 30 (2017).

[20] Hartley, Richard, and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003.

[21] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

[22] Pellegrini, Stefano, Andreas Ess, and Luc Van Gool. "Improving data association by joint modeling of pedestrian trajectories and groupings." In European conference on computer vision, pp. 452-465. Springer, Berlin, Heidelberg, 2010.

[23] L. Leal-Taix e, M. Fenzi, A. Kuznetsova, B. Rosenhahn and S. Savarese, "Learning an Image-Based Motion Context for Multiple People Track- ing," 2014 IEEE Conference on Computer Vision and Pattern Recogni- tion, 2014, pp. 3542-3549, doi: 10.1109/CVPR.2014.453.