Intelligent and Converged Networks ISSN 2708-6240
2023, 4(1): 41−49 DOI: 10.23919/ICN.2023.0004

Emma: An accurate, efficient, and multi-modality strategy for autonomous vehicle angle prediction

Keqi Song, Tao Ni, Linqi Song, and Weitao Xu*

Abstract: Autonomous driving and self-driving vehicles have become the most popular selection for customers for their convenience. Vehicle angle prediction is one of the most prevalent topics in the autonomous driving industry, that is, realizing real-time vehicle angle prediction. However, existing methods of vehicle angle prediction utilize only single-modal data to achieve model prediction, such as images captured by the camera, which limits the performance and efficiency of the prediction system. In this paper, we present Emma, a novel vehicle angle prediction strategy that achieves multi-modal prediction and is more efficient. Specifically, Emma exploits both images and inertial measurement unit (IMU) signals with a fusion network for multi-modal data fusion and vehicle angle prediction. Moreover, we design and implement a few-shot learning module in Emma for fast domain adaptation to varied scenarios (e.g., different vehicle models). Evaluation results demonstrate that Emma achieves overall 97.5% accuracy in predicting three vehicle angle parameters (yaw, pitch, and roll), which outperforms traditional single-modalities by approximately 16.7%–36.8%. Additionally, the few-shot learning module presents promising adaptive ability and shows overall 79.8% and 88.3% accuracy in 5-shot and 10-shot settings, respectively. Finally, empirical results show that Emma reduces energy consumption by 39.7% when running on the Arduino UNO board.

Key words: multi-modality; autonomous driving; vehicle angle prediction; few-shot learning

1 Introduction

The past decades have witnessed the explosive development of autonomous driving and electric vehicles such as the Tesla Model series. According to a recent investigation[1], the global electric vehicle market has reached 411 billion US dollar by the end of 2021. Such innovative autonomous driving depends on various sensors (e.g., high-speed camera, radar, and LiDAR) and state-of-the-art artificial intelligence (AI) models. Specifically, these sensors first capture signals that contain driving information, such as road conditions, congestion situations, and vehicle status. Then, the captured signals will be processed and fed into pre-installed deep learning models to generate the best strategy for the current driving condition. Finally, the central control system will conduct the strategy by adjusting the direction and speed of the driving vehicle.

Although existing computer vision based methods[2−5] have achieved great performance in autonomous driving, two main challenges remain unsolved in terms of practicality and efficiency. **(1) Single sensor modality:** Modern self-driving cars are equipped with many different types of sensors, which work independently of each other. Hence, the control system needs to provide several single-modal services by each uncorrelated sensor, which limits the model performance[2−4]. **(2) Low adaptability:** It is known that deep learning requires a significant amount of data to obtain good performance. In reality, however, a challenge is that there exist many unseen scenarios, such as unseen vehicles and unseen road

• Keqi Song, Tao Ni, Linqi Song, and Weitao Xu are with the Shenzhen Research Institute, City University of Hong Kong, Hong Kong, China, and also with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. E-mail: kqsong2-c@my.cityu.edu.hk; taoni2-c@my.cityu.edu.hk; linqi.song@cityu.edu.hk; weitaoxu@cityu.edu.hk.

∗ To whom correspondence should be addressed.

Manuscript received: 2023-02-13; revised: 2023-03-11; accepted: 2023-03-28

conditions. Traditional strategies need to train different models for different scenarios, which consumes enormous computational resources and requires extensive data collection[5].

To address the aforementioned two challenges, we propose Emma, an accurate, efficient, and multi-sensor modality based strategy for vehicle angle prediction in autonomous driving. Vehicle angle prediction is an important component in modern self-driving cars because the car needs the angle parameters as an essential factor to generate the best driving strategy. Figure 1 shows an image captured by the camera when driving the car and it also shows the three angle parameters: yaw, pitch, and roll, which represent the current status of the vehicle. Specifically, Emma leverages both the images captured by the camera and the signals from inertial measurement unit (IMU) sensor with a fusion network to achieve feature fusion and multi-modal prediction. Moreover, we design and implement a few-shot learning module based on the most advanced meta-learning concept[6] to equip Emma with the ability of fast domain adaptation. That is, Emma can quickly adapt to various scenarios, such as new vehicle models and new road conditions. Combining the fusion network with the few-shot learning method, Emma provides a high-accuracy solution to predict vehicle angle while consuming lower energy consumption.

We implement and evaluate Emma in a public autonomous driving dataset that contains over 5000 real-world images as well as 3D accelerometer data. Evaluation results show that Emma achieves an overall 97.5% accuracy in predicting yaw, pitch, and roll and is approximately 16.7% and 36.8% higher than single modality based methods (i.e., image only or IMU only). Additionally, the proposed few-shot learning module shows promising adaptation ability in shifting models from one vehicle model to another. Specifically, Emma achieves an overall 79.8% accuracy in 5-shot learning while realizing an overall 88.3% accuracy in 10-shot learning. Furthermore, we also deploy Emma into mobile systems such as micro-controller and measure the energy consumption compared with traditional single-modal settings. Results show Emma can reduce power consumption by 39.7% compared with traditional strategies in the Arduino UNO platform. We believe Emma can provide an accurate and efficient strategy for predicting the running vehicle’s angle for the growing autonomous driving industry.

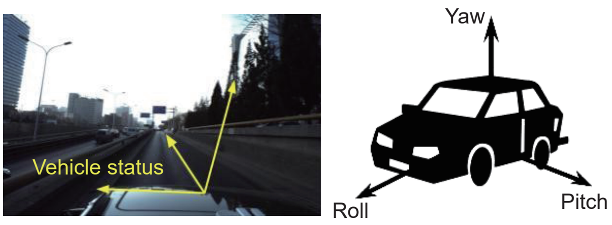


Fig. 1 Example of vehicle angle prediction.

We summarize the contributions as follows.

- We propose a multi-modal sensor fusion strategy for accurate vehicle angle prediction, which is named Emma. Emma achieves much higher accuracy compared to the single-sensor modality based method by fusing the information from both the camera and IMU sensor.
- We design and implement a few-shot learning module based on the concept of meta-learning to enable the ability of domain adaptation of the models and increase the system efficiency. As such, Emma achieves fast adaptation to various scenarios such as different vehicles.
- We conduct extensive experiments to evaluate the performance of Emma. The evaluation results demonstrate that Emma achieves overall 97.5% accuracy in vehicle angle prediction while reducing power consumption by approximately 39.7% compared with traditional methods. In addition, the proposed few-shot learning module also realizes high accuracy in adapting models across different scenarios.

The rest of this paper is organized as follows. Section 2 presents the design and implementation of Emma, including the fusion network and the few-shot learning module. Then, Section 3 shows the evaluation results of Emma which contain the prediction performance and energy consumption. Next, Section 4 introduces several related works including multi-modality and few-shot learning. Finally, Section 5 concludes the paper.

2 Related work

2.1 Vehicle angle prediction

Vehicle angle prediction is a new topic in the autonomous driving industry and recent studies have investigated different methods to detect the real-time vehicle angle. For instance, Huang et al.[7] implemented a deep learning framework for predicting vehicle angles from 3D visual models. Furthermore, Khan et al.[8] proposed a self-supervised vehicle angle prediction method based on the geometric analysis of the captured images. Compared to these single-modal approaches, Emma achieves multi-modality vehicle angle prediction that has demonstrated high accuracy, good domain adaptation ability, and low energy consumption.

2.2 Multi-modality in autonomous driving

Multi-modality sensing improves the model performance by combining not only visual modality captured by the camera but also non-image modalities from other sensors (e.g., IMU, mmWave, and acoustic). As such, it addresses the limitations of traditional vision-based sensing approaches in non-line-of-sight (NLoS) scenarios. For example, Roy et al.[9] proposed a multi-vehicle detection method based on the fusion of modalities, that is, images with seismic, acoustic, and radar data. Chen et al.[10] presented MV3D, a sensory-fusion framework that exploits signals from both LiDAR sensors and cameras for 3D object detection in autonomous driving. Moreover, Pan et al.[11] introduced an acoustic-seismic modality fusion approach to monitor moving vehicles. Emma follows a similar line of research by introducing the first multi-modality strategy for vehicle angle prediction.

2.3 Few-shot learning

Few-shot learning becomes popular recently as it enables deep neural networks (e.g., CNN) to achieve fast adaptation to unseen conditions with only a few samples (e.g., five or ten shots). For instance, Refs. [12−14] presented the advantages of few-shot learning compared with traditional domain shift methods such as transfer learning. Furthermore, OneFi[15] pushes the limits by using only a one-shot sample to achieve quick model adaptation in Wi-Fi sensing. GazeGraph[16] exploits few-shot learning to track eye movements while acquiring few sensitive gaze information. In addition, Refs. [17−19] adopted few-shot learning to realize more practical website fingerprinting attacks. In this paper, we present Emma, the first work that leverages few-shot learning on vehicle angle prediction in autonomous driving.

3 System design

3.1 System overview

Figure 2 presents the system overview of Emma. First, the autonomous driving vehicle takes images from the camera and signals from the IMU sensors (e.g., accelerometer) followed by a signal pre-processing module that removes noise from the raw data. Then, we separately train the image model and the IMU model. Next, we convert the single modality models to feature extractors and feed them into the fusion network to achieve feature fusion and build a multi-modality model. In addition, a few-shot learning module is added for quickly adapting the multi-modality models to different scenarios (e.g., different vehicle models). Finally, the multi-modality model with high adaptability can be exploited for vehicle angle prediction.

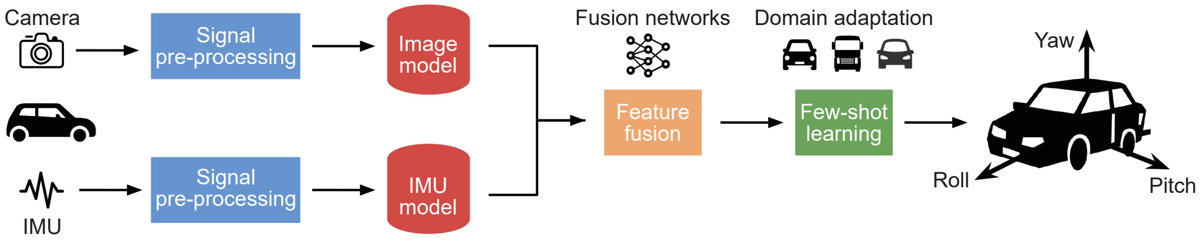


Fig. 2 Overview of Emma.

3.2 Data pre-processing

In a real autonomous driving scenario, data collected by cameras and IMU sensors contain not only vehicle information but also extra interference. To remove the noise from the raw data, we design and implement the signal processing modules for both the image channel and IMU channel. Specifically, we utilize the NAFNet model[20] to denoise images captured by the camera because it achieves a high signal-to-noise rate (SNR) while keeping the structural similarity[21]. For the time-series data captured by IMU sensors, we first apply a Savitzky-Golay (S-G) filter to remove noise in the collected signals without distorting their shapes[22]. Next, the denoised images and the filtered IMU signals are used to train single modality independently. In practice, we apply the default settings of the NAFNet model and set the frame length as 100 with three polynomial orders in the S-G filter.

3.3 Single modality model construction

To build the single modality, we utilize the well-known convolutional neural network (CNN) that has demonstrated promising performance in image classification. Furthermore, CNN-based methods are utilized in one-dimensional time-series signals classification[23−25] because they can capture temporal and spatial features from time series and achieve a high classification accuracy[26−29]. In Emma, we utilize two convolutional layers to extract temporal and spatial features from input data (image or time-series IMU data) and two batch normalization layers to standardize the data and stabilize the learning process. Then, two max-pooling layers reduce the dimension by half and a dropout layer has been added for preventing overfitting. Finally, the flatten layer converts feature maps to one-dimensional and the last fully-connected layers output the predicted class with the highest probability. We set the window size of the convolutional layer as 5, following with max-pooling layers that have window size of 2 and stride of 2. There are 128 filters in the first convolutional layer, 192 in the second, and 300 in the third. In addition, we train the CNN-based models by setting initial learning rate as 0.01 and train every single modality with 300 epochs.

3.4 Fusion network

Due to the difference between image signals and IMU signals, the traditional CNN-based single modality model cannot directly combine and transform them into semantic information[30−32]. To address this inconsistency and achieve feature fusion for multi-modality, we implement a fusion network based on a similar approach proposed by Pandey and Wang[33], which ignores the dynamic distribution of the weight across features from multiple modalities and its architecture is presented in Fig. 3. Specifically, we combine features from the image model and the IMU model by utilizing the joint feature space[34] that employs feature maps of each channel as a feature detector and filter. Then, the fused multi-modality features will be processed and distilled as the extracted feature vectors that contain identical knowledge and informative context while ignoring interference features. Finally, a fully-connected layer and a softmax layer take the multi-modality features to produce the

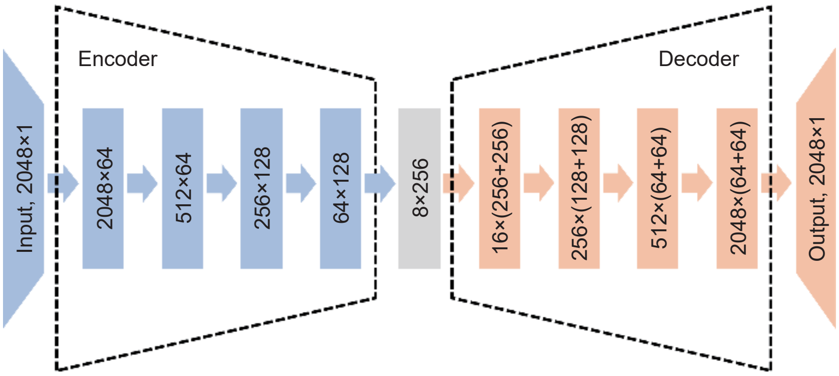


Fig. 3 Fusion network architecture.

output of predicted vehicle angles with the highest probability.

3.5 Few-shot learning module

Although the CNN-based vehicle angle predictor achieves high accuracy with the fused data from multiple modalities, its performance can be impacted by the shifting conditions. One solution is to train multiple predictors for different scenarios (e.g., different vehicle models). However, such a method not only requires a large-scale dataset to ensure good performance but also limits the practicality in varied scenarios. Therefore, considering the varying scenarios in practice, we design a few-shot learning module in Emma based on the concept of meta-learning. Below, we illustrate our proposed algorithm in two stages: the meta-training stage and the deployment stage.

We present the meta-training algorithm for vehicle angle prediction in Algorithm 1. In the meta-training step, we denote the vehicle angle predictor as v and network parameters as θ. After obtaining the optimized initialization parameters θ^*, the vehicle angle predictor can realize fast adaptation to various scenarios (e.g., new vehicle models and new road conditions) with only $K \times N$ samples in autonomous driving. For example, when a new target dataset D_{new} is collected from a different condition ($D_{\text{new}} \cap D_S = \varnothing$), the optimized predictor can quickly adapt to this new scenario T_{new} and obtain the new parameters θ_{new}. We obtain the adapted vehicle angle predictor with fine-tuned parameters θ_{new} for the new scenario T_{new}. In practice, we set α as 0.01 and β as 0.001. In Section 3.3, we comprehensively evaluate the proposed performance of few-shot learning module.

Algorithm 1 Meta-training for vehicle angle prediction

```
Input: $D_S$: source dataset. $v$: vehicle angle predictor.
    $\alpha$ and $\beta$: learning rate hyperparameters.
Output: $v_{\theta^*}$: adapted vehicle angle predictor with optimized
    parameters $\theta^*$.
$\theta \leftarrow \theta_0$, $v_\theta \leftarrow v_{\theta_0}$   ▷ Initialize $v_\theta$ with random parameters $\theta_0$
while not finished do
    $T_S \leftarrow$ generate a batch of scenarios from $D_S$
    for each task $T_i \in T_S$ do
        $S_{T_i} \leftarrow K \times N$ support samples from $T_i$
        $S_{Q_i} \leftarrow K \times N$ query samples from $T_i$ ($S_{T_i} \cap S_{Q_i} = \varnothing$)
        Evaluate $\nabla_\theta L_{T_i}(v_\theta)$ with $S_{T_i}$ and loss $L_{T_i}(v_\theta, S_{T_i})$
        $\theta'_{T_i} \leftarrow \theta_0 - \alpha \nabla_\theta L_{T_i}(v_\theta, S_{T_i})$   ▷ obtain scenario-specific parameters $\theta'_{T_i}$ of $T_i$ using gradient descent.
        Evaluate $L_{T_i}(v_{\theta'_{T_i}})$ with query samples from $S_{Q_i}$.
    $\theta^* \leftarrow \theta_0 - \beta \nabla_\theta \Sigma_{T_i \in T} \mathcal{L}_{T_i}(f_{\theta'_{T_i}}, S_{Q_i})$   ▷ obtain optimized parameters $\theta^*$ that minimizes all scenario losses.
```

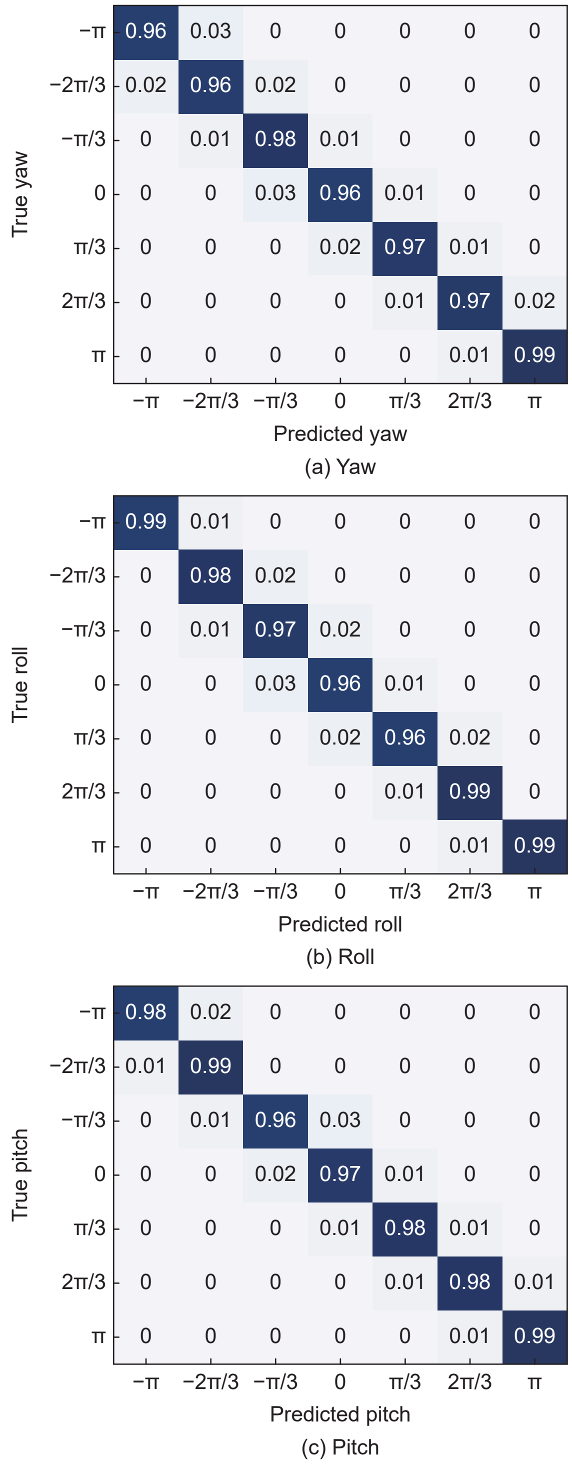
4 Evaluation

4.1 Experimental setup

To evaluate the performance of Emma, we use the ApolloCar3D autonomous driving dataset provided by Peking University and Baidu, the search engine giant of China. Specifically, the dataset was collected from 30 different vehicle models that contain a total of 60 000 labeled 3D car instances as well as 3D accelerometer data (x, y, z) from 5277 real-world images. We leverage the signal processing toolbox from MATLAB for pre-processing the accelerometer data. We implement the CNN-based vehicle angle predictor and the proposed few-shot learning module in Keras 2.3 based on the Tensorflow 2.0 framework. In addition, all data processing is conducted on a desktop running Windows 10 with 32 GB memory, an Intel i7-9700K CPU, and an NVIDIA GeForce RTX 2080Ti GPU.

4.2 Overall performance

To evaluate the effectiveness of Emma, we use the classification accuracy and the corresponding confusion matrices as the performance metrics. We divide the angle ranging from [−π, π] to the approximate seven angles [−π, −2π/3, −π/3, 0, π/3, 2π/3, π] and evaluate the accuracy of the predicted yaw, pitch, and roll. Moreover, we also train models by using single-modal data, i.e., only the images or only the accelerometer data, and further, compare their performance to the multi-modality model. The three parts of Fig. 4 show the confusion matrices results of angle prediction for yaw, pitch, and roll, where Emma achieves accuracy rates of 97.0%, 97.9%, and 97.7% in the mentioned three angle parameters, respectively. Furthermore, to demonstrate the effectiveness of using



(a) Yaw

(b) Roll

(c) Pitch

Fig. 4 Confusion matrices results of angle prediction for yaw, roll, and pitch.

multi-modality, we compare the two results of single modality (i.e., only images or only accelerometer), and the results of Emma, as shown in Fig. 5. It can be seen by fusing the multi-modality data, Emma achieves over 97.5% accuracy, whereas two single modality based methods achieve 80.8% and 60.7% accuracy only. Therefore, exploiting multi-modality can increase the accuracy of vehicle angle prediction by approximately 16.7% and 36.8%.

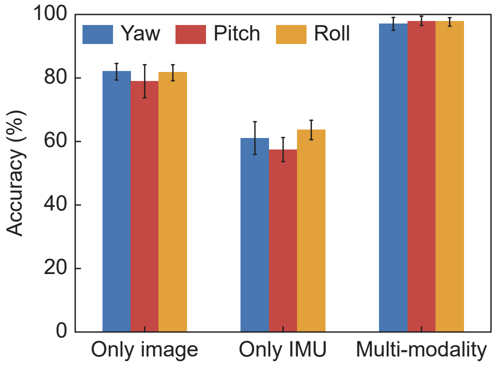


Fig. 5 Performance comparison of single- and multi-modality.

4.3 Few-shot learning evaluation

To evaluate the performance of the few-shot learning module, we apply the model trained from the Audi Q7 to the dataset collected from two cars: Skoda Fabia and MG GT. Specifically, we select the baseline by directly applying the trained model of Audi Q7 to predict the angles of the other two vehicles without adaptation and compare the results with Emma. Figures 6 and 7 present the results of domain adaptation using the few-shot learning module. We can see that the overall accuracy decreases to 18.3% and 6.0% when the trained models are directly used in different vehicles. Nevertheless, Emma achieves 79.8% and 73.1% accuracy rates in 5-shot learning and accuracy rates of 88.3% and 84.0% in 10-shot learning. Therefore, Emma achieves fast domain adaptation to different

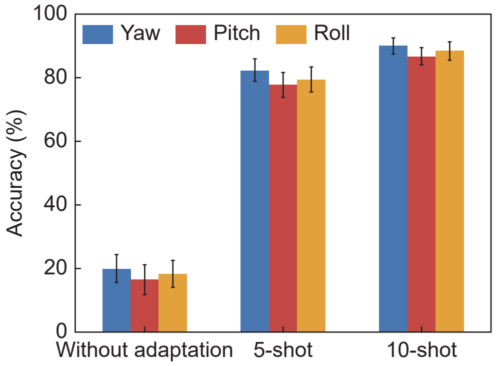


Fig. 6 Domain adaptation result: Audi Q7 → Skoda Fabia.

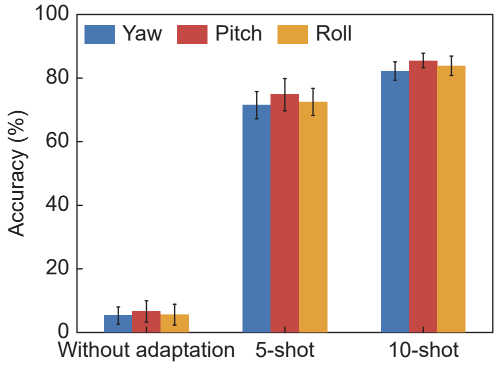


Fig. 7 Domain adaptation result: Audi Q7 → MG GT.

vehicle models while maintaining a high accuracy by exploiting the few-shot learning approach.

4.4 Energy consumption evaluation

Having demonstrated the high performance of Emma in vehicle angle prediction, we further evaluate the energy efficiency of Emma in the wild. Specifically, we deploy the trained DNN models in an Arduino UNO micro-controller and use a Monsoon Power Monitor※ to measure the energy consumption before and after applying Emma. Figure 8 shows the energy consumption of three running statuses: (1) No loads: Arduino UNO is running at default setting without DNN models; (2) Multiple DNN models: Arduino UNO is running pre-installed multiple DNN models for vehicle angle prediction; (3) Emma: Arduino is running pre-installed DNN model with Emma. Hence, we know that the energy consumption of the three strategies is 205 mW, 268 mW, and 243 mW. As such, the energy consumption of running multiple DNN models is approximately 63 mW while leveraging Emma with DNN consuming only 38 mW. Therefore, Emma reduced energy consumption by around 39.7%.

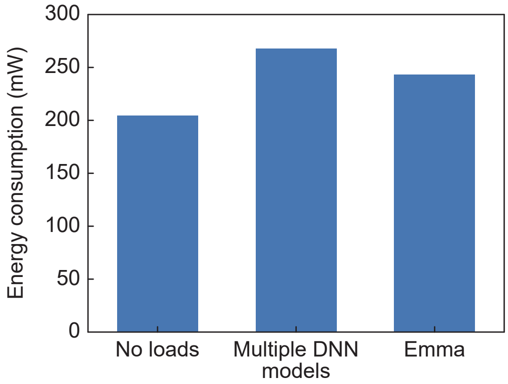


Fig. 8 Energy consumption results.

5 Conclusion

In this paper, we present Emma, an efficient and accurate multi-modality strategy for vehicle angle prediction in autonomous driving. Compared with traditional single-modal methods that adopt only images captured by the camera, Emma leverages signals captured from both camera and IMU sensors and utilizes a fusion network to achieve high accuracy in vehicle angle prediction. In addition, a few-shot learning module enables Emma to become domain adaptive to different scenarios, and the results indicate Emma can maintain a high accuracy across various vehicle models while reducing power consumption by 39.7%. We believe Emma can provide a step-forward solution for vehicle angle prediction in the explosive development of autonomous driving.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62101471) and partially supported by the Shenzhen Research Institute of City University of Hong Kong, the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CityU 21201420), Shenzhen Science and Technology Funding Fundamental Research Program (No. 2021Szvup126), National Natural Science Foundation of Shandong Province (No. ZR2021LZH010), Changsha International and Regional Science and Technology Cooperation Program (No. kh2201023), Chow Sang Sang Group Research Fund sponsored by Chow Sang Sang Holdings International Limited (No. 9229062), CityU MFPRC (No. 9680333), CityU SIRG (No. 7020057), CityU APRC (No. 9610485), CityU ARG (No. 9667225), and CityU SRG-Fd (No. 7005666).

References

[1] M. Carlier, Electric vehicles worldwide—Statistics and facts, https://www.statista.com/topics/1010/electric-mobility/\#dossierKeyfigures, 2023.

※ Moonsoon Power Monitor, https://www.msoon.com/online-store, accessed on Nov 11th, 2022.

[2] N. Marinello, M. Proesmans, and L. V. Gool, TripletTrack: 3D object tracking using triplet embeddings and LSTM, in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 4499–4509.

[3] C. Luo, X. Yang, and A. Yuille, Self-supervised pillar motion learning for autonomous driving, in *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 3182–3191.

[4] H. Qiu, F. Ahmad, F. Bai, M. Gruteser, and R. Govindan, AVR: Augmented vehicular reality, in *Proc. 16th Annual International Conference on Mobile Systems, Applications, and Services*, Munich, Germany, 2018, pp. 81–95.

[5] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, LaserNet: An efficient probabilistic 3D object detector for autonomous driving, in *Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 12669–12678.

[6] C. Finn, P. Abbeel, and S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *Proc. 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 1126–1135.

[7] W. Huang, W. Li, L. Tang, X. Zhu, and B. Zou, A deep learning framework for accurate vehicle yaw angle estimation from a monocular camera based on part arrangement, *Sensors*, vol. 22, no. 20, p. 8027, 2022.

[8] Q. Khan, P. Wenzel, and D. Cremers, Self-supervised steering angle prediction for vehicle control using visual odometry, in *Proc. 24th International Conference on Artificial Intelligence and Statistics*, Virtual Event, 2021, pp. 3781–3789.

[9] D. Roy, Y. Li, T. Jian, P. Tian, K. R. Chowdhury, and S. Ioannidis, Multi-modality sensing and data fusion for multi-vehicle detection, *IEEE Transactions on Multimedia*, doi: 10.1109/TMM.2022.3145663.

[10] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, Multi-view 3D object detection network for autonomous driving, in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 6526–6534.

[11] Q. Pan, J. Wei, H. Cao, N. Li, and H. Liu, Improved DS acoustic–seismic modality fusion for ground-moving target classification in wireless sensor networks, *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2419–2426, 2007.

[12] T. Gong, Y. Kim, J. Shin, and S. -J. Lee, MetaSense: Few-shot adaptation to untrained conditions in deep mobile sensing, in *Proc. 17th Conference on Embedded Networked Sensor Systems* (*SenSys*), New York, NY, USA, 2019, pp. 110–123.

[13] J. Zhang, Z. Chen, C. Luo, B. Wei, S. S. Kanhere, and J. Li, MetaGanFi: Cross-domain unseen individual identification using WiFi signals, *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–21, 2022.

[14] G. Yin, J. Zhang, G. Shen, and Y. Chen, FewSense, towards a scalable and cross-domain Wi-Fi sensing system using few-shot learning, *IEEE Transactions on Mobile Computing*, doi: 10.1109/TMC.2022.3221902.

[15] R. Xiao, J. Liu, J. Han, and K. Ren, OneFi: One-shot recognition for unseen gesture via COTS WiFi, in *Proc. 19th ACM Conference on Embedded Networked Sensor Systems* (*SenSys*), Coimbra, Portugal, 2021, pp. 206–219.

[16] G. Lan, B. Heit, T. Scargill, and M. Gorlatova, GazeGraph: Graph-based few-shot cognitive context sensing from human visual behavior, in *Proc. 18th Conference on Embedded Networked Sensor Systems* (*SenSys*), Virtual Event, 2020, pp. 422–435.

[17] P. Sirinam, N. Mathews, M. S. Rahman, and M. Wright, Triplet fingerprinting: More practical and portable website fingerprinting with n-shot learning, in *Proc. 2019 ACM SIGSAC Conference on Computer and Communications Security* (*CCS*), London, UK, 2019, pp. 1131–1148.

[18] M. Chen, Y. Wang, H. Xu, and X. Zhu, Few-shot website fingerprinting attack, *Computer Networks*, vol. 198, p. 108298, 2021.

[19] C. Wang, J. Dani, X. Li, X. Jia, and B. Wang, Adaptive fingerprinting: Website fingerprinting over few encrypted traffic, in *Proc. 11th ACM Conference on Data and Application Security and Privacy*, Virtual Event, 2021, pp. 149–160.

[20] X. Chu, L. Chen, and W. Yu, NAFSSR: Stereo image super-resolution using NAFNet, in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 1238–1247.

[21] Y. Cao, A. Dhekne, and M. Ammar, ITrackU: Tracking a pen-like instrument via UWB-IMU fusion, in *Proc. 19th Annual International Conference on Mobile Systems, Applications, and Services*, Virtual Event, WI, USA, 2021, pp. 453–466.

[22] J. Chen, P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh, A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter, *Remote sensing of Environment*, vol. 91, nos. 3&4, pp. 332–344, 2004.

[23] P. Cronin, X. Gao, C. Yang, and H. Wang, Charger-Surfing: Exploiting a power line side-channel for smartphone information leakage, in *Proc. 30th USENIX Security Symposium*, Boston, MA, USA, 2021, pp.

681–698.

[24] R. Ning, C. Wang, C. Xin, J. Li, and H. Wu, DeepMag: Sniffing mobile app. in magnetic field through deep convolutional neural networks, in *Proc. 2018 IEEE International Conference on Pervasive Computing and Communications* (*PerCom*), Athens, Greece, 2018, pp. 1–10.

[25] A. S. L. Cour, K. K. Afridi, and G. E. Suh, Wireless charging power side-channel attacks, in *Proc. 2021 ACM SIGSAC Conference on Computer and Communications Security* (*CCS*), Virtual Event, 2021, pp. 651–665.

[26] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, Deep learning for time series classification: A review, *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[27] F. -J. Wu and G. Solmaz, CrowdEstimator: Approximating crowd sizes with multi-modal data for Internet-of-Things services, in *Proc. 16th Annual International Conference on Mobile Systems, Applications, and Services*, Munich, Germany, 2018, pp. 337–349.

[28] Y. Soneda, Y. Matsuda, Y. Arakawa, and K. Yasumoto, M3B corpus: Multi-modal meeting behavior corpus for group meeting assessment, in *Adjunct Proc. 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proc. 2019 ACM International Symposium on Wearable Computers*, London, UK, 2019, pp. 825–834.

[29] G. Zhao, G. Ben-Yosef, J. Qiu, Y. Zhao, P. Janakaraj, S. Boppana, and A. R. Schnore, Person re-ID testbed with multi-modal sensors, in *Proc. 19th ACM Conference on Embedded Networked Sensor Systems*, Coimbra, Portugal, 2021, pp. 526–531.

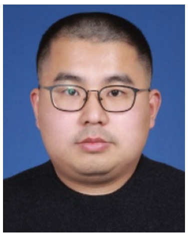
[30] T. Li, J. Huang, E. Risinger, and D. Ganesan, Low-latency speculative inference on distributed multi-modal data streams, in *Proc. 19th Annual International Conference on Mobile Systems, Applications, and Services*, Virtual Event, 2021, pp. 67–80.

[31] C. Zhu, K. Li, Q. Lv, L. Shang, and R. P. Dick, IScope: Personalized multi-modality image search for mobile devices, in *Proc. 7th International Conference on Mobile Systems, Applications, and Services*, Krakow, Poland, 2009, pp. 277–290.

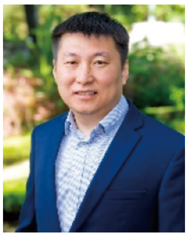
[32] D. Li, J. Xu, Z. Yang, Q. Zhang, Q. Ma, L. Zhang, and P. Chen, Motion inspires notion: Self-supervised visual-LiDAR fusion for environment depth estimation, in *Proc. 20th International Conference on Mobile Systems, Applications, and Services*, Portland, OR, USA, 2022, pp. 114–127.

[33] A. Pandey and D. Wang, A new framework for CNN-based speech enhancement in the time domain, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[34] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren, Wavoice: A noise-resistant multi-modal speech recognition system fusing mmWave and audio signals, in *Proc. 19th ACM Conference on Embedded Networked Sensor Systems*, Coimbra, Portugal, 2021, pp. 97–110.



Tao Ni received the master degree from Australian National University in 2021 and the bachelor degree from Shanghai Jiao Tong University in 2018. He is currently pursuing the PhD degree at the Department of Computer Science, City University of Hong Kong. His research interests include cyber-physical system security, contactless side channels, and low-power wireless sensing.



Linqi Song received the PhD degree in electrical engineering from University of California, Los Angeles (UCLA), USA in 2015, and the BS and MS degrees from Tsinghua University, China in 2006 and 2009, respectively. He is currently an assistant professor at the Department of Computer Science, City University of Hong Kong. He was a postdoctoral scholar at the Department of Electrical and Computer Engineering, UCLA. His research interests include information theory, machine learning, and big data.

Keqi Song received the bachelor degree from Southwest Jiaotong University in 2020. He is currently pursuing the MPhil degree at the Department of Computer Science, City University of Hong Kong. His research interests include cyber-physical systems and wireless sensing.

Weitao Xu received the PhD degree from University of Queensland in 2017 (advised by Prof. Neil Bergmann and Dr. Wen Hu). He is an assistant professor at the Department of Computer Science, City University of Hong Kong. Before that, he was a postdoctoral research associate at the School of Computer Science and Engineering (CSE), University of New South Wales from June 2017 to August 2019. His research areas include mobile computing, sensor network, and IoT.