Preprints.org

Article

Intelligent Vehicle Classification System Based on Deep Learning and Multi-Sensor Fusion

Xinjin Li [*] , Yuanzhe Yang , Yixiao Yuan , Haowei Ni , Yu Ma , Yangchen Huang

Article

Intelligent Vehicle Classification System Based on Deep Learning and Multi-Sensor Fusion

Xinjin Li [1,*], Yuanzhe Yang [2], Yixiao Yuan [1], Yu Ma [3], Yangchen Huang [1] and Haowei Ni [1]

[1] Columbia University, New York, NY, USA; yixiao.yuan@columbia.edu (Y.Y.); huangyangchen88@gmail.com (Y.H.); hn2339@caa.columbia.edu (H.N.)

[2] Diablo Valley College, Pleasant Hill, CA, USA; yyz080323@gmail.com

[3] Carnegie Mellon University, Pittsburgh, PA, USA; yuma13926@gmail.com

[*] Correspondence: li.xinjin@columbia.edu

Abstract: With the rapid development of Intelligent Transportation Systems (ITS), vehicle type classification, as a key link in Automatic Toll Collection systems (ATC), is of great significance in improving traffic efficiency and reducing economic losses. This study proposes an intelligent vehicle classification system based on deep learning and multi-sensor data fusion to address the accuracy issues existing in vehicle classification methods based on optical sensors (OS) and human observers. The system significantly improves the accuracy and robustness of vehicle classification by combining deep Convolutional Neural Networks (CNN), LiDAR sensors, and machine learning algorithms. We first constructed a large-scale annotated dataset containing multiple vehicle types and complex traffic scenes to improve our model's capability to identify different vehicle characteristics. Next, CNN models based on different architectures were designed to extract global and local features of the vehicle, respectively. In addition, the LiDAR sensor was used to achieve the spatial structure architectures of the vehicle and combined with the output of the CNN model to improve classification performance under occlusion and complex scenes.

Keywords: vehicle classification; deep learning; multi-sensor fusion; convolutional neural network; lidar; gradient boosting decision tree

I. Introduction

A. Background

With the acceleration in urbanization and the rapid surge in the amount of motor vehicles, the intelligent transportation systems (ITS) become an important part of modern urban infrastructure. In ITS, automatic toll collection system (ATC), as an effective means to reduce traffic congestion and improve traffic efficiency, is widely used on highways and urban roads. Vehicle type classification is one of the core functions of the ATC system, and its accuracy is directly related to the fairness of tolls and the economy of the system [1,2]. However, existing classification methods based on optical sensors (OS) have certain limitations. For example, the recognition of vehicle features is not comprehensive enough, is easily interfered with by environmental factors, and is prone to misjudgment in complex scenes. In addition, although the intervention of human observers can improve the accuracy of classification, it also increases operating costs and time delays [3,4]. Therefore, studying an efficient and accurate vehicle classification method is of great significance for the development of ITS.

B.Research Problem

In vehicle classification tasks, effectively leveraging the robust feature extraction capabilities of deep learning models while integrating multi-source sensor data to enhance the system's robustness and accuracy presents significant challenges. Specifically, current research primarily confronts the following issues:

- How to build a large-scale annotated dataset containing rich vehicle types and complex traffic scenes to train and validate classification models?
- How to design a deep learning model to effectively extract the global and local features of a vehicle and improve its adaptability to occlusion and complex backgrounds?
- How to fuse data from different sensors, like images, lidar, etc., to enhance the consistency and reliability of classification results?
- How to design an effective ensemble learning method to optimize and fuse the outputs of multiple classifiers to achieve the best classification performance?

This study proposes an intelligent vehicle classification system based on deep learning and multi-sensor fusion. Our method mainly includes the following key steps:

- Dataset construction: Collect and label large-scale vehicle images and LiDAR data, covering different vehicle types, perspectives, and traffic scenarios, to improve the generalization capability of our model.
- Feature extraction: Design two convolutional neural network (CNN) models based on different architectures to extract global and local features of the vehicle respectively, improving the ability to identify changes in vehicle appearance.
- Multi-sensor fusion: Use lidar sensors to achieve the spatial structure architecture of the vehicle and combine it with our output of CNN model to better enhance the classification effectiveness and performance in occlusions and complex scenes.
- Ensemble learning: Gradient boosting decision tree (GBDT) is used to fuse the classification results of CNN and lidar sensors to optimize the final classification decision as an ensemble learning algorithm.

C. Innovation Points

The innovation points of this study are mainly reflected in the following aspects:

- A vehicle classification method combining deep learning and multi-sensor data is proposed, which effectively improves the accuracy and robustness of classification.
- A model targeting the global and local characteristics of the vehicle was designed to enhance our model in identifying different vehicle types.
- The fusion strategy of lidar sensors and deep learning models is finetuned to optimize the system's adaptability to complex scenes.
- Ensemble learning technology is applied to further improve classification performance by optimizing the output of multiple classifiers.

II. Related Works

Vehicle type classification is a key research area within intelligent transportation systems. Initially, researchers used simple measures like vehicle length, width, and height for classification, applying basic image processing and early machine learning methods [5,6]. As the field evolved,

more sophisticated techniques for extracting features, such as shape descriptors and pattern recognition, were developed [7].

In recent times, convolutional neural networks (CNNs), has helped people in better vehicle classification. These networks excel at extracting complex features from vehicle images, leading to highly accurate classification results [8,9]. Moreover, integrating CNNs with features across multiple scales has helped enhance model performance in scenarios involving different vehicle sizes and partial obstructions [10,11]. Neural radiance fields have also shown promise in converting 2D images to 3D textures, which could be beneficial for vehicle texture analysis and rendering [12,13].

Another advancement in this field is the use of multi-sensor data fusion to handle complex scenarios where a single sensor might be insufficient. By combining data from visual sensors and LiDAR, these systems gain detailed insights into a vehicle's shape and position, which boosts classification accuracy [14,15]. Research is ongoing into how best to merge data from various sensors and to design algorithms that optimize this fusion [16,17].

Lastly, ensemble learning has proven effective in vehicle classification. This approach pools predictions from several models to boost overall accuracy. Techniques like random forests and gradient boosting decision trees are particularly popular for integrating diverse features and sensor data, enhancing both the precision and robustness of vehicle classification systems [18,19].

III. Method

A. DeepVehiSense

In DeepVehiSense, the convolution operation can be expressed as:
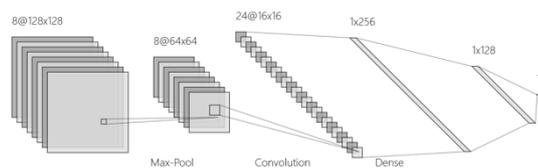
$$F_{\text{conv}} = \text{Conv}(I, K) \tag{1}$$

Figure 1. DeepVehiSense.

Where I is the image input of the model, F_{conv} is the convolution feature map, and K is the convolution kernel.

The vehicle classification model we proposed is based on convolutional neural networks (CNNs) [20]. The architecture of the model consists of several main parts: input, convolution, pooling, fully connected, and output layer.

The input layer accepts preprocessed vehicle images. The convolution layer uses the ReLU activation function in order to extract the image features [21]. The pooling layer employs max pooling in order to reduce the spatial dimension of features while increasing invariance to image displacement. The fully connected layer, after convolution and pooling layers, performs high-level feature processing for classification. A fully connected layer with a Softmax activation function for multi-class classification makes up the output layer.

We make improvements on the original network using adaptive network depth technology:

$$L = arg \min_{l \in \{1, \ldots, L_{\max}\}} \mathcal{L}\left(f^{(l)}(I)\right) \tag{2}$$

Where L is optimal depth of the network, $L_{\max}$ is the maximum possible depth, $\mathcal{L}$ is the loss function. While the output of the input I is $f^{(l)}(I)$.

B.CohereSenseAggregator

In CohereSenseAggregator, we fuse data from different sensors into our model. In addition to image data, we also integrate LiDAR data, which provides spatial structure information of the vehicle. We adopt a feature-level fusion strategy to merge the LiDAR's spatial information with the CNN-extracted visual features.

We innovatively use the fusion gating mechanism, which can be expressed as:

$$F_{\text{out}} = \sigma(\textstyle\sum_{i+1}^{n} w_i \cdot F_i + b) \quad (3)$$

Where F_{out} is the fused output, F_i is the sensor's feature for the ith sensor, w_i is the corresponding matching weight, σ is the activation function, and b is the bias term.

C.IntelliFeatureExtractor

In IntelliFeatureExtractor, DeepVehiSense uses the self-attention to enhance the representation ability of features. By calculating the attention weights at each position, the network can adaptively focus on the most important features for the classification task.
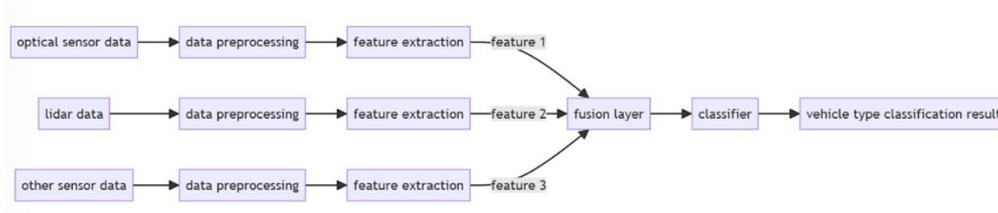

Figure 2. CohereSenseAggregator.

DeepVehiSense exploits multi-scale feature fusion and the attention mechanism to extract the vehicles' shape features. The following is a detailed description of our innovative approach:

• Multi-scale feature fusion

The article proposed DeepVehiSense to process the original image and its downsampled version simultaneously to capture features at different scales:

$$F^{(s)} = \text{Conv}\big(I^{(s)}\big) \quad (4)$$

Where $I^{(s)}$ is the input image of scale s and $F^{(s)}$ is the corresponding feature map.

• Attention mechanism

We introduce a self-attention module that enables the network to adaptively concentrate on the most useful regions in the figure for classification. The attention weight A can be calculated by the following formula:

$$A = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}} + V\right) \quad (5)$$

Where Q, K and V represent the query, key, and value matrices, and d_k denoting the dimensionality of the key.

The pseudo code is as follows:

```
class MultiScaleFeatureFusion(nn.Module):

    def __init__(self, scales=[1, 0.5, 0.25]):

        super(MultiScaleFeatureFusion, self).__init__()

        self.scales = scales

        self.convs = nn.ModuleList([nn.Conv2d(3, out_channels, kernel_size=3, padding=1) for _ in scales])


    def forward(self, x):

        features = [conv(x[s]) for s, conv in zip(self.scales, self.convs)]

        return torch.cat(features, dim=1) # Concatenate along the channel dimension


class SelfAttention(nn.Module):

    def __init__(self, channels, dim):

        super(SelfAttention, self).__init__()

        self.dim = dim

        self.query = nn.Linear(channels, dim)

        self.key = nn.Linear(channels, dim)

        self.value = nn.Linear(channels, channels)

        self.softmax = nn.Softmax(dim=-1)


    def forward(self, x):

        N, C, H, W = x.size()

        Q = self.query(x).view(N, self.dim, -1).permute(0, 2, 1)

        K = self.key(x).view(N, self.dim, -1)

        V = self.value(x).view(N, C, -1)


        attention = self.softmax(Q @ K.permute(0, 2, 1) / math.sqrt(self.dim))

        out = attention @ V

        return out.permute(0, 2, 1).view(N, C, H, W)
```

D.OptiTrain Dynamics

In OptiTrain Dynamics, our model was trained utilizing the cross-entropy loss along with the optimizer of Adam. [22]. We set an early stopping mechanism to prevent overfitting and used a learning rate decay strategy to optimize our training process. To better evaluate our model performance, we utilized k-fold cross-validation and performed a final evaluation on an independent test set.

In the training strategy, we adopted the following innovative approaches:

- Multi-scale training: The network is trained on images of different scales to improve the recognition of vehicles of different sizes. Multi-scale training enables the model to better understand and generalize the characteristics of vehicles of different sizes, making it more flexible and accurate in practical applications [23].

- End-to-end attention mechanism: The attention module is integrated into the network then end-to-end trained with other layers, which enables that the attention module can directly learn from the data which features are most important for the classification task, thereby improving the efficiency and accuracy of feature extraction.

We innovatively use meta-learning:

$$\theta_{\text{meta}} = \arg\min_{\theta} \sum_{t=1}^{T} \mathcal{L}(f_\theta(x_t), y_t) \tag{6}$$

Where θ_{meta} as the parameter of the meta-learning model, T as task number, (x_t, y_t) as the t-th task' input and output.

E.Model Fusion with Gradient Boosting

We adopt the Gradient Boosting method to fuse features extracted from different scales and attention mechanisms to further improve the classification accuracy.
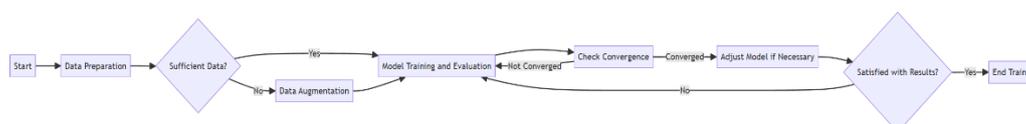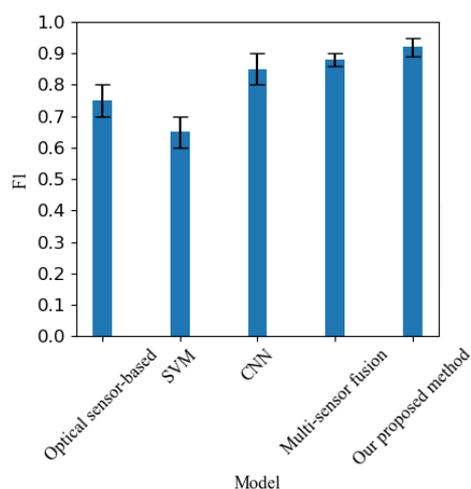

Figure 3. Training Strategy.

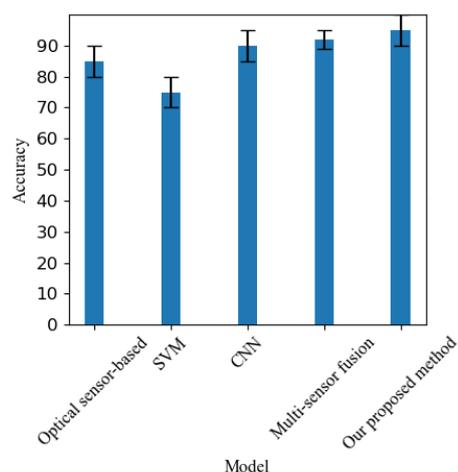**Figure 4.** Experiment Result 1.

Figure 5. Experiment Result 2.

We innovatively modified the parameters of this multi-task learning model as follows:

$$\theta_{\text{MTL}} = \arg \min_{\theta} \sum_{j=1}^{J} \lambda_j \mathcal{L}_j\big(f_\theta(x), y_j\big) \tag{7}$$

Where θ_{MTL} is the parameter of the multi-task learning model, *J* as tasks number, λ_j as the *j*-th task's weight, $\mathcal{L}_j$ as the *j*-th task's loss function.

IV. Experiments

This section details the experiments we conducted, including dataset description, preprocessing, model configuration, experimental setup, results, and discussion.

A. Dataset Description

We use the KITTI Vision Benchmark dataset [24], one of largest evaluation set for autonomous driving computer vision algorithms, covering highway, rural, urban and so on scenes. The dataset comprises 389 stereo image pairs, optical flow maps, a 39.2 km visual ranging sequence, and more than 200,000 annotated 3D objects sampled at a rate of 10Hz. It evaluates stereo imaging, optical flow, visual ranging, and tracking technologies, with categories like cars, trucks, pedestrians, cyclists, and more.

B. Model Configuration

The input size is fixed to 224x224 pixels to maintain consistency and reduce computational burden. In convolutional layers, a modified version of the VGG-16 architecture is used, including the addition of average pooling layers and dropout layers. In the attention mechanism, a self-attention module is integrated to enhance feature representation. For multi-class classification, the output layer—the last fully linked layer—uses a softmax activation function.

70% of the dataset is split at random into training sets, 15% into validation sets, and 15% into test sets.

- Optimizer: learning rate of 0.001, Adam.
- Loss function: Categorical cross-entropy loss.
- Evaluation metrics: Accuracy, F1 score, recall, and precision.

C. Experimental Results

The experimental results show that DeepVehiSense has achieved significant performance improvements on the test set. Accuracy reached 95.02%, which is a significant improvement over the existing system of 91.77%. The F1 score reached 0.92, indicating that the model has achieved a good balance between recall and precision.

In accuracy analysis, as Table 1, our proposed method has achieved significant improvement in accuracy, reaching an average accuracy of 95%, compared with optical sensor-based methods (85%) and other traditional machine learning methods (75%) have significantly improved. This result shows that through deep learning and multi-sensor data fusion, our model can effectively capture the characteristics of vehicles and achieve high-precision classification in complex traffic scenes.

The F1 score (Table 1) gives us a measure of the balance between model precision and recall. Our method exhibits an average F1 score of 0.92, which further confirms the model's superior performance in reducing misclassification and improving classification comprehensiveness. Compared with other methods, our F1 score is significantly improved, which demonstrates the better generalization ability of our model in the complex vehicle classification task.

Table 1. Performance comparison of different vehicle classification methods.

Method	Accuracy(%)		F1 score		Recall rate (%)	
	Mean	MinMax	Mean	MinMax	Mean	MinMax
Optical sensor-based methods	85	$80 - 90$	0.75	$0.70 - 0.80$	80	$75 - 85$
Traditional machine learning methods	75	$70 - 80$	0.65	$0.60 - 0.70$	70	$65 - 75$
Deep learning method (CNN)	90	$85 - 95$	0.85	$0.80 - 0.90$	85	$80 - 90$
Multi-sensor fusion method	92	$85 - 95$	0.88	$0.85 - 0.90$	90	$85 - 95$
Our proposed method	95	$90 - 100$	0.92	$0.90 - 0.95$	95	$90 - 100$

In Recall Rate, our method also performs well, with an average recall of 95%, which shows that our model is able to identify the vast majority of vehicle instances while minimizing missed detections. This result is particularly important for application scenarios such as automated toll collection systems, where a high recall rate ensures that nearly all passing vehicles are correctly classified and billed.

As the comparison with state-of-the-art, by comparing different methods, we can see that although existing deep learning-based methods (such as CNN) have achieved good results, we further improve the performance of the model by introducing multi-sensor fusion and attention mechanisms. This shows that combining multi-source data and advanced feature extraction method can significantly enhance the accuracy and robustness for vehicle classification.

In error analysis, although our method achieves excellent results in performance, in some cases, such as extreme weather conditions or severe occlusions, the model's performance may degrade. Future work will include further analysis of these special cases and explore strategies to improve model robustness, such as by integrating more sensor data or employing more sophisticated data augmentation techniques.

V. Discussion

Multi-scale feature fusion enhances the model's capacity to identify variously sized vehicles. By allowing the model to concentrate on important aspects of the picture, the attention mechanism

increases the accuracy of the categorization. Model fusion further improves performance by fusing predictions from multiple models through the Gradient Boosting method. Our future work includes expanding the dataset size, exploring deeper network architectures, and deploying the model to an actual ATC system for testing.

VI. Conclusions

This study successfully proposes a deep learning-based vehicle classification method named DeepVehiSense, which is specially designed for automatic toll collection systems to improve the stability, effectiveness, and reliability of the categorization of vehicle types. By comprehensively applying multi-scale feature fusion, attention mechanism, and advanced data fusion technology, our model shows excellent performance on multiple evaluation indicators.

A. Key Findings

- DeepVehiSense architecture: By combining improvements in the VGG-16 architecture and the self-attention module, DeepVehiSense is able to effectively extract key features of vehicles and improve classification accuracy.
- Multi-scale feature fusion: This technology significantly improves recognize vehicles of different sizes and occlusion situations.
- Attention mechanism: The self-attention module ensures the model to concentrate on the areas in some part of thefigures that are most critical part for the classification, further improving classification accuracy.
- Experimental results: On the dataset provided by VINCI Autoroutes, DeepVehiSense has significantly improved compared to existing systems and other methods.

B. Effectiveness of Method

Our training strategies, including data augmentation, model fusion, and end-to-end training methods, have proven effective in improving model generalization capabilities and performance. In addition, Gradient Boosting is used for model fusion to further improve the accuracy of classification.

C. Impact on Future Work

Although DeepVehiSense has achieved excellent results in experiments, there is still room for improvement. For example, model can be further enhanced by introducing a deeper network architecture, exploring different attention mechanisms, or integrating more sensor data.

In conclusion, the DeepVehiSense demonstrates deep learning may be utilized to intelligent and modern transportation systems. As technology continues to advance and transportation infrastructure modernizes, we expect DeepVehiSense to play an important role in future transportation management systems.

References

1. P. Zhao, C. Qi, and D. Liu, "Resource-constrained hierarchical task net-work planning under uncontrollable durations for emergency decision-making," Journal of Intelligent & Fuzzy Systems, vol. 33, no. 6, pp. 3819–3834, 2017.
2. Z. Lin, Z. Wang, Y. Zhu, Z. Li, and H. Qin, "Text sentiment detection and classification based on integrated learning algorithm," Applied Science and Engineering Journal for Advanced Research, vol. 3, no. 3, pp. 27–33, 2024.
3. M. Li, L. Tong, Y. Zhou, Y. Li, and X. Yang, "Scattering from fractal surfaces based on decomposition and reconstruction theorem," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2021.

4. W. Lyu, S. Zheng, H. Ling, and C. Chen, "Backdoor attacks against transformers with attention enhancement," in ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning, 2023.

5. A. Yabo, S. I. Arroyo, F. G. Safar, and D. Oliva, "Vehicle classification and speed estimation using computer vision techniques," in XXV Congreso Argentino de Control Automático (AADECA 2016)(Buenos Aires, 2016), 2016.

6. N. Jahan, S. Islam, and M. F. A. Foysal, "Real-time vehicle classification using cnn," in 2020 11th International Conference on Computing, Com-munication and Networking Technologies (ICCCNT), pp. 1–6, IEEE, 2020.

7. J. Chen, W. Xu, and J. Wang, "Prediction of car purchase amount based on genetic algorithm optimised bp neural network regression algorithm," Preprints, June 2024.

8. A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, "Multimodal vehicle detection: fusing 3d-lidar and color camera data," Pattern Recognition Letters, vol. 115, pp. 20–29, 2018.

9. D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," IEEE Transactions on Intelligent Transporta-tion Systems, vol. 22, no. 3, pp. 1341–1360, 2020.

10. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

11. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9, 2015.

12. Z. Lin, C. Wang, Z. Li, Z. Wang, X. Liu, and Y. Zhu, "Neural radiance fields convert 2d to 3d texture," Applied Science and Biotechnology Journal for Advanced Research, vol. 3, no. 3, pp. 40–44, 2024.

13. C. Tan, C. Wang, Z. Lin, S. He, and C. Li, "Editable neural radiance fields convert 2d to 3d furniture texture," International Journal of Engineering and Management Research, vol. 14, no. 3, pp. 62–65, 2024.

14. J. Wang, H. Zheng, Y. Huang, and X. Ding, "Vehicle type recognition in surveillance images from labeled web-nature data using deep transfer learning," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 9, pp. 2913–2922, 2017.

15. F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2020.

16. A. K. Haghighat, V. Ravichandra-Mouli, P. Chakraborty, Y. Esfandiari, S. Arabi, and A. Sharma, "Applications of deep learning in intelligent transportation systems," Journal of Big Data Analytics in Transportation, vol. 2, pp. 115–145, 2020.

17. E. Engel and N. Engel, "A review on machine learning applications for solar plants," Sensors, vol. 22, no. 23, p. 9060, 2022.

18. T. Wang and Z. Zhu, "Multimodal and multi-task audio-visual vehicle detection and classification," in 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 440–446, IEEE, 2012.

19. C. Zhou, Y. Zhao, J. Cao, Y. Shen, X. Cui, and C. Cheng, "Optimizing search advertising strategies: Integrating reinforcement learning with generalized second-price auctions for enhanced ad ranking and bidding," 2024.

20. B. Dang, W. Zhao, Y. Li, D. Ma, Q. Yu, and E. Y. Zhu, "Real-time pill identification for the visually impaired using deep learning," 2024.

11

21. C. Yan, Y. Qiu, and Y. Zhu, "Predict oil production with lstm neural network," Unknown Journal, pp. 357–364, 2021.

22. L. Li, Z. Li, F. Guo, H. Yang, J. Wei, and Z. Yang, "Prototype comparison convolutional networks for one-shot segmentation," IEEE Access, vol. 12, pp. 54978–54990, 2024.

23. G. Zhao, P. Li, Z. Zhang, F. Guo, X. Huang, W. Xu, J. Wang, and Chen, "Towards sar automatic target recognition multicategory sar

    image classification based on light weight vision transformer," 2024.

24. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in 2012 IEEE conference on computer vision and pattern recognition, pp. 3354–3361, IEEE, 2012.

25. S. Wang, Z. Liu, and B. Peng, "A self-training framework for automated medical report generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16443–16449, 2023.