

Metaverse Meets Intelligent Transportation System: An Efficient and Instructional Visual Perception Framework

Junfan Wang, *Student Member, IEEE*, Yi Chen, Xiaoyue Ji, *Member, IEEE*, Zhekang Dong, *Senior Member, IEEE*, Mingyu Gao, and Chun Sing Lai, *Senior Member, IEEE*

Abstract—The combination of the Metaverse and intelligent transportation systems (ITS) holds significant developmental promise, especially for visual perception tasks. However, the acquisition of high-quality scene data poses a challenging and expensive endeavor. Meanwhile, the visual disparity between the Metaverse and the physical world poses an impact on the practical applicability of the visual perception tasks. In this paper, a Metaverse Intelligent Traffic Visual Framework, MITVF, is developed to guide the implementation of visual perception tasks in the physical world. Firstly, a two-stage metadata optimization strategy is proposed that can efficiently provide diverse and high-quality scene data for traffic perception models. Specifically, an element reconfigurability strategy is proposed to flexibly combine dynamic and static traffic elements to enrich the data with a low cost. A diffusion model-based metadata optimization acceleration strategy is proposed to achieve efficient improvement of image resolution. Secondly, a Meta-Physical adaptive learning method is proposed, and further applied to visual perception tasks to compensate for the visual disparity between the Metaverse and the physical world. Experimental results show that MITVF achieves a 10× acceleration in optimization speed, ensuring the image quality and reconstructing diverse. Further, MITVF is applied to the traffic object detection task to verify the effectiveness and validity. The performance of the model trained with 5k real data exceeded that of the model trained with 200k real data, with AP₅₀ reaching 67.7%.

Index Terms—adaptive learning method, metadata optimization strategy, MITVF, Metaverse.

I. INTRODUCTION

THE ITS is a typical complex system that integrates various visual perception tasks, including object detection [1, 2], object tracking [3, 4], and object segmentation [5]. Deep learning-based visual perception models are extensively applied in ITS, playing a pivotal role in ensuring its safe and stable development.

The efficient and secure execution of perception tasks within extant ITS confronts significant challenges. Firstly, deep learning-based perception models necessitate extensive data for

This work was supported in part by the Key R&D Project of Hangzhou under grants No. 2022AIZD0009, 2022AIZD0022, the Key Research and Development Program of Zhejiang Province grant No. 2022C01062 (*Corresponding author: Zhekang Dong*).

J. Wang and M. Gao are with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou, China, 310018, and also with the Zhejiang Provincial Key Lab of Equipment Electronics, Hangzhou, China, 310018, (e-mail: wangjunfan@hdu.edu.cn, mackgao@hdu.edu.cn).

Y. Chen with the Ocean College, Zhejiang University, Hangzhou, China, 310027, (e-mail: morningone@126.com).

LIST OF ABBREVIATIONS

Abbreviation	Definition
MITVF	Metaverse Intelligent Traffic Vision Framework
ITS	Intelligent Transportation System
SAR	Society of Automotive Engineers
MAdiff	Diffusion Model-based Metadata Optimization Acceleration Strategy
DAdet	Domain Adaptive-based Traffic Object Detector
RPN	Region Proposal Network
RoI	Region of Interest
EMA	Exponential Moving Average
GRL	Gradient Reversal Layer
GAN	Generate Adversarial Network
mAP	Mean Average Precision
FID	Fréchet Inception Distance
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity
IS	Inception Score

training and testing, incurring substantial costs associated with the creation of high-quality datasets. Secondly, rich and diverse data are needed to cope with complex traffic scenarios in practical applications. Furthermore, the availability of high-fidelity simulation scenarios is constrained, impeding targeted testing for specific issues or conditions and diminishing overall testing efficiency.

The emergence of the Metaverse [6, 7] offers a promising solution to the visual perception of ITS. The six levels of autonomous driving, as defined by the SAR, range from Level 0 (no automation) to Level 5 (full automation) [8]. Metaverse can exert certain advantages at every level, such as providing rich data and a safe testing environment. By constructing the Metaverse visual perception framework, it can provide massive and diverse traffic scene data, improve the generalization of the model, and reduce the cost of model research and development. However, directly combining the Metaverse and ITS will bring certain challenges: a) Efficiently optimizing data for the

X. Ji with the Department of Precision Instrument, Tsinghua University, Beijing, China, (email: jixiaoyue@mail.tsinghua.edu.cn).

Z. Dong is with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou, China, 310018, and also with the College of Electrical Engineering, Zhejiang University, Hangzhou, China, 310027, (e-mail: englishp@hdu.edu.cn).

C. S. Lai is with the Department of Electronic and Computer Engineering, Brunel University London, London, UB8 3PH, UK and also with the School of Electronic and Electrical Engineering, Guangdong University of Technology, Guangzhou, China 510006 (email: chunsing.lai@brunel.ac.uk).

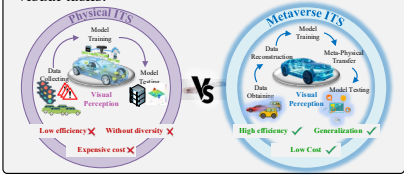
Reference	Research Gaps	Motivations	Contributions
[10-15]	<ol style="list-style-type: none"> 1. Perception models based on deep learning require large amounts of data for training and testing. Creating a complete high-quality dataset is time- and labor-intensive, and collection involves certain risks. 2. The visual perception model in the intelligent transportation system (ITS) requires targeted training for certain weather, emergencies, special scenes, etc. during the development and testing process. It is difficult for existing datasets to meet this training task flexibly and efficiently; 3. The safety of ITS requires high-fidelity testing of visual perception models. The actual test site is limited and it is impossible to simulate some harsh special scenarios, and the test process is dangerous. 	<p>The Metaverse can create an artificial dimensional space parallel to the real world, customize and construct highly realistic simulation scenes, and provide rich data for the physical world. automotive companies have already harnessed the Metaverse for the design and verification processes of complete vehicles. By integrating visual perception models developed within the Metaverse with ITS, we can effectively address existing challenges related to data availability and quality.</p>	<p>A Metaverse Intelligent Traffic Visual Framework (i.e., MITVF) is proposed, enabling efficient implementation of visual perception tasks in ITS. As our best knowledge, it is the first framework capable of combining the Metaverse and traffic visual tasks.</p> 
<p style="color: red; font-weight: bold;">↓ Challenges faced by combining Metaverse and visual perception in ITS ↓</p>			
[16-19]	<p>The image quality of the virtual traffic scenes generated by the Metaverse cannot be compared with the quality captured by real cameras, which will affect the training performance of the model. And simply applying Metaverse virtual data cannot meet the targeted training needs of the model.</p>	<p>The image optimization method optimizes the low-resolution images generated by the Metaverse and weighs the relationship between optimization speed and image quality.</p>	<p>A two-stage metadata optimization strategy is proposed, comprising an element reconfigurability strategy and a diffusion model-based metadata optimization acceleration strategy, with the aim of expeditiously generating comprehensive and high-quality metadata.</p>
[9, 20-22]	<p>There are certain visual differences between custom-built data in the Metaverse and real data, which will cause the performance of models trained with Metaverse data to degrade in actual scenarios.</p>	<p>Domain adaptation methods can reduce the feature gap between metaverse data and real data and improve model performance.</p>	<p>A domain adaptive learning method is proposed to reduce the visual disparity between the Metaverse and the physical world. It improves the generalization ability of the perception model and can be reliably applied to the physical world</p>

Fig. 1. Based on the researches [9-22], the research gaps in environmental perception of ITS is summarized and corresponding contributions is proposed.

generation a comprehensive and high-quality metadata (virtual data from Metaverse) is a critical challenge [19, 23]. Low-resolution virtual images struggle to fulfill the long-distance sensing requirements in practical applications, and the optimization of extensive metadata incurs significant costs. Moreover, the metadata optimizing needs to be flexible to cover complex and diverse traffic scenarios. b) Exploring the nonlinear mapping relationship between the virtual and real-world domains is essential. Visual disparity [24, 25] between the Metaverse and the physical reality inevitably introduce feature deviations during the training process, thereby impacting the applicability in the physical world.

In this paper, the MITVF is proposed to guide the efficient performance of visual perception tasks in the physical world.

A two-stage metadata optimization strategy is proposed for the challenges of low metadata quality and limited scene diversity. At the first stage, through the traffic element reconfigurability strategy, the physical world is divided into dynamic and static elements to efficiently customize and reconfigure a variety of complex traffic scenarios and improve the generalization of the training model. At the second step, an optimization acceleration strategy is proposed to simplify the backward reasoning in the diffusion model [18] by designing a non-Markov process, improve the processing speed and ensure the quality of metadata.

To deal with the inherent visual disparity, a domain adaptive learning method is proposed. The adversarial training is applied between the Metaverse and physical domains to improve model performance. The image-level self-attention feature alignment module and the instance-level feature aggregation mechanism are designed to reduce the feature space distance between the Metaverse domain and the physical domain, enabling the perception model in Metaverse to be applied to the real world

without distinction.

On the whole, the main contributions of our work and the research gaps as shown in Fig. 1. The remainder of this paper is structured as follows: Section II summarizes the existing Metaverse applications and related technologies. Section III provides an overview of the MITVF framework. In Section IV and Section V, the two-stage metadata optimization strategy and domain adaptive learning method are introduced in detail. The experimental results and conclusion are respectively shown in Section VI and Section VII.

II. RELATIVE WORK

In this section, we provide an overview of the current applications and pertinent technologies within the Metaverse, and analyze the shortcomings of existing Metaverse applications in various fields

In terms of combining the Metaverse with ITS, [26-29] combined the Metaverse with vehicles to optimize existing systems such as smart cockpits, traffic flow management analysis and road maintenance system. [30] provides an evaluation solution for metaverse and autonomous driving algorithm testing. Wang et al. [21] proposed a video analysis system in the metaverse environment, combining virtual reality with artificial intelligence to build fully intelligent video analysis to improve system detection performance. Gilles et al. [15] used virtual data provided by the Metaverse to improve existing training data.

Lee et al. [31] used an exploratory approach to analyze current qualitative data characterizing the state of the business of meta-boundary services for healthcare and to learn from the current business trends in meta-boundary services to derive applicable strategies. In other fields, Contreras et al. [32]

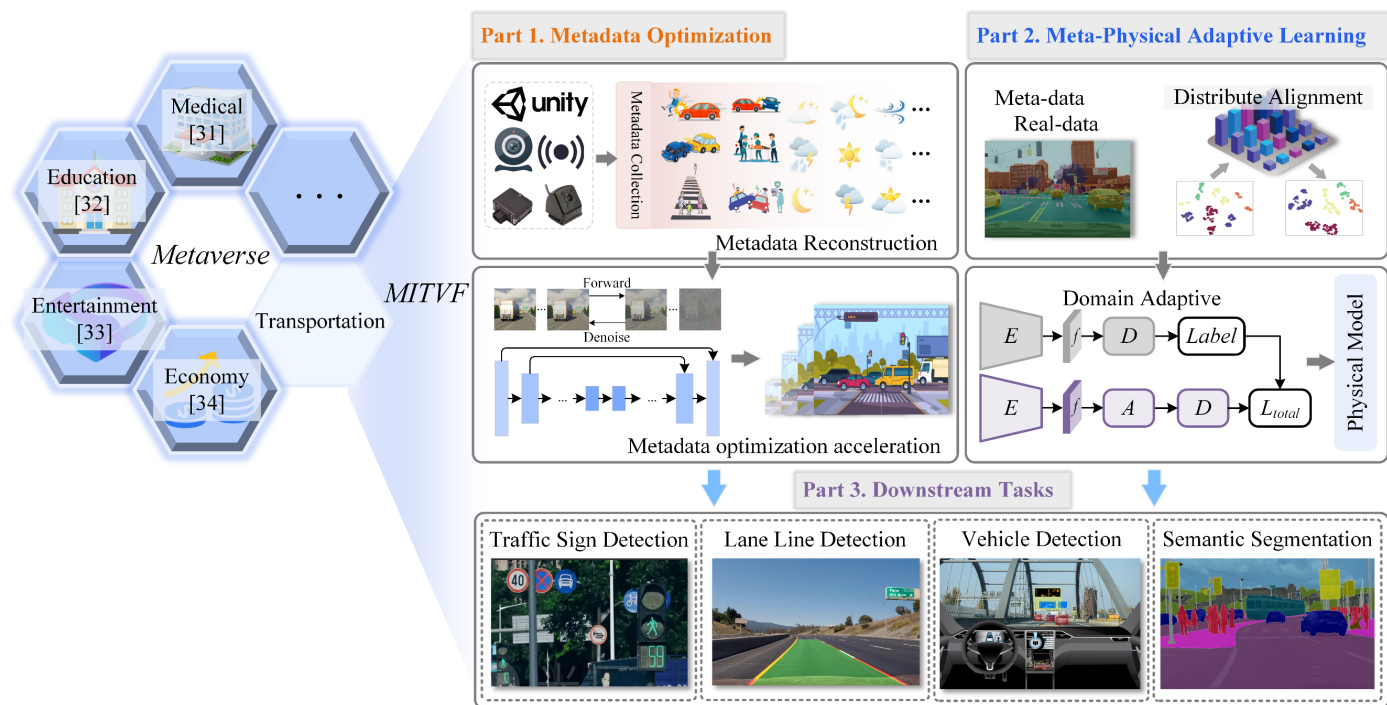


Fig. 2. Theoretical framework of the MITVF: Metadata Optimization, Meta-Physical Adaptive Learning, and Downstream tasks.

suggested that by utilizing the metaverse, educational institutions can provide students and staff with a 360° experience that offers greater flexibility and adaptability to unforeseen events. Li et al. [33] explore the changes and development of smart home entertainment scene experience design in the metaverse perspective. Jeong et al. [34] proposed a new e-commerce platform innovative business model, which utilizes the metaverse technology to combine live commerce with meta-virtual, and overcomes the limitations of the existing online shopping.

The above researches provide participants with an immersive experience by constructing corresponding virtual worlds. However, they ignore the importance of data optimization during the construction process. Immersive experience has high requirements for the resolution and realism of scene data. Secondly, there is a certain degree of visual disparity between the virtual world and the physical world, and the knowledge learned in the Metaverse is difficult to applied to the physical world. For example, in the medical field, metaverse technology is used for medical research and simulation. However, due to the difference in images between the virtual and real worlds, it is difficult to fully apply the research results to the real world. And, using the Metaverse to provide students with virtual laboratories also have certain safety risks due to the visual disparity.

Data quality and visual disparity issues also exist in the integration of ITS with the Metaverse. The proposed MITVF provides solutions for the efficient performance of visual perception tasks in the Metaverse through data optimization strategies and domain adaptive learning.

III. METAVERSE INTELLIGENT TRAFFIC VISION FRAMEWORK

The visual perception model based on MITVF not only relies on

the training and testing with high-definition massive metadata but also requires the ability to apply the learned knowledge to real-world scenarios. Fig. 2 shows the applications of the Metaverse in multiple fields. At present, its technical application in transportation is insufficient. MITVF combines the Metaverse and visual perception technology to promote the safe and efficient development of ITS. The MITVF contains three parts: Metadata Optimization, Meta-Physical Adaptive Learning and Downstream Tasks.

Part 1. Metadata Optimization in Fig. 2 includes two stages: metadata reconstruction and metadata optimization acceleration strategies. In the metadata reconstruction, virtual sensors simulate the physical parameters to generate different virtual data sequences. The entire phase automatically ensures the automatic generation of virtual data that is not only accurately annotated but also versatile enough to be applied to a wide array of visual perception tasks. Furthermore, the metadata reconstruction process is designed to enhance the diversity of metadata, allowing for customized scene construction through the free combination of elements. This flexibility facilitates the creation of tailored virtual environments for specific research or application needs. However, the low resolution of the virtual sensor acquisition frame has an impact on the detection and training of visual perception models, and the simulator requires extremely high cost to render high-quality images. To address these challenges, the metadata optimization acceleration strategy emerges as a crucial solution. This strategy is centered around the enhancement of image quality with minimal computational and time costs. By employing a non-Markov process, the strategy significantly boosts the image optimization efficiency of the diffusion model. This innovative approach ensures the provision of high-quality data for visual tasks, thereby

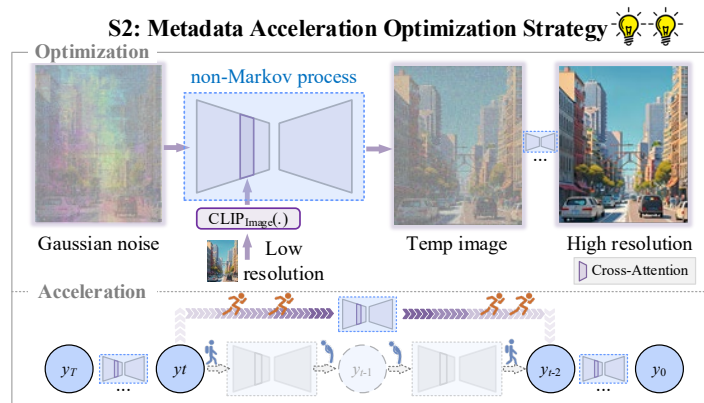
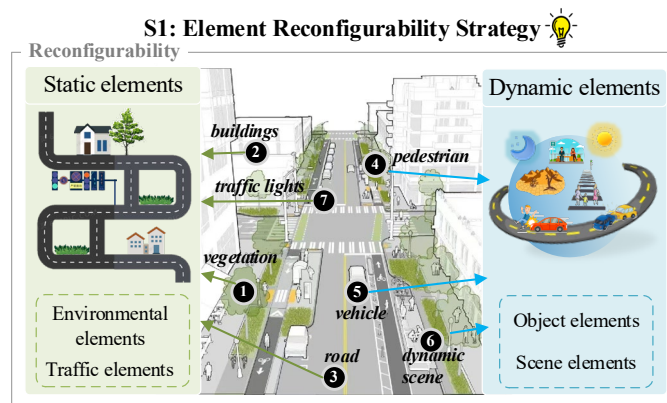


Fig. 3. Two-stage Metadata optimization strategy.

overcoming the constraints imposed by low-resolution virtual sensor data and high rendering costs.

Part 2. Meta-Physical Adaptive Learning in Fig. 2 provides a powerful method for adapting Metaverse visual perception to the real world. There are certain distribution differences between virtual and physical data, which may be caused by various factors, such as variations in lighting conditions, object appearance, camera view, etc. These differences may lead to degraded performance of the completed models trained under the metaverse when tested in the physical world. In our approach, varying proportions of metadata from the Metaverse and real-world data are concurrently fed into a domain adaptation model, which primarily comprises an encoder (E), a decoder (D), and a self-attention feature alignment module (A). This domain adaptation model is founded on a knowledge distillation architecture, wherein both the metaverse model and the physical model adhere to an identical network structure. Through the process of feature alignment, the physical model is able to utilize the features learned by the metaverse model. This enables the virtual-to-real transfer of model performance, ensuring that the knowledge acquired in the virtual domain of the metaverse is effectively applied to enhance the accuracy and robustness of the model in real-world scenarios.

Part 3 of Fig. 2 illustrates the diverse range of downstream tasks supported by our proposed MITVF. As a comprehensive framework that merges the Metaverse with visual perception tasks in ITS, MITVF offers a versatile platform for various applications, including traffic object detection, lane line detection, and semantic segmentation. The integration of ITS with the Metaverse not only enriches the available data for model training but also provides a safe and controlled environment for testing and refining visual perception algorithms, making MITVF a powerful tool in advancing ITS capabilities and promoting safer and more efficient transportation systems.

IV. TWO-STAGE METAVERSE OPTIMIZATION STRATEGY

This section proposes a two-stage Metadata optimization strategy (as shown in Fig. 3), including an element reconfigurability strategy and a diffusion model-based metadata optimization acceleration strategy. The specific optimization strategy description is provided below.

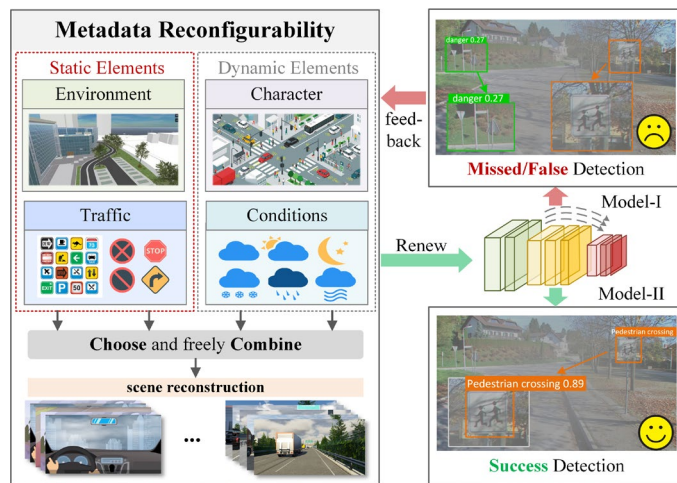


Fig. 4. Element reconfigurability strategy for metadata.

A. Element reconfigurability strategy

Metaverse can be used to generate rich data to simulate different driving scenarios, which is faster and more flexible than data collection in real-world. According to the composition of the physical world, all the components can be roughly divided into two parts: the static and dynamic elements. Static elements include environmental elements (buildings, vegetation, roads, etc.) and traffic elements (traffic signs, traffic lights, etc.); dynamic elements include object elements (pedestrians, vehicles, etc.) and scene elements (light, weather, accidents, etc.). According to these elements, on the basis of digital restoration of real traffic scenes, the construction of customized traffic scenes can be realized, such as bad weather (rain, snow, haze, etc.), unexpected accidents (collision, congestion, fire, etc.), and extreme scenes (wilderness, mountainous areas, etc.).

Customized reconstruction of metadata can be achieved through the above element division. Real scenes are often the result of the joint action of multiple factors, and it is difficult to train a visual perception model individually for a certain scene or problem. Through customized reconstruction of metadata, the efficiency of visual perception model development and testing can be improved, facilitating efficient perception under diverse conditions.

To improve training and testing efficiency, metadata can be optimized based on test results in the physical world. The

metadata-trained model is tested on physical driving scenarios, and the metadata is updated by re-simulating underperforming scenarios. Through this feedback loop mechanism, metadata is in a continuous process of updating and optimization to improve the robustness and generalization of visual model learning. The above process is shown in Fig. 4.

B. Diffusion model-based metadata optimization acceleration strategy

In this paper, the MAdiff is proposed, which combines the conditional probability diffusion model to achieves fast optimization of quality images, named.

In the training phase, the model input includes low-resolution image x , noisy image y_t , and high-resolution image y_0 . y_t is generated by y_0 through the diffusion and noise process. In the inference stage, the input is the low-resolution image x , and the output is a high-resolution generation image.

Taking x as the conditional input of the generative model, a high-resolution image is obtained by performing reverse denoising on y_t and repeating iteratively T times. The objective function of this process is as follows:

$$E_{(x,y)} E_{z \sim \mathcal{N}(0,I)} \left\| f_\theta \left(x, \sqrt{\gamma_t} y_0 + \sqrt{1-\gamma_t} z \right) - z \right\|_2^2 \quad (1)$$

where (x,y) is sampled from the training dataset, z denotes Gaussian distribution sampling noise, I represents the identity matrix, $\mathcal{N}(0, I)$ represents the standard normal distribution, and f_θ denotes the noise prediction model, here is U-Net. $\gamma_t = \prod_{i=1}^t \alpha_i$ and $\alpha_i / \alpha_{1:T}$ are both hyper-parameters, subject to $0 < \alpha_i < 1$, determining the variance of the noise added at each iteration. The objective function minimizes the loss of the constrained model by computing the square of the 2-norm between f_θ and z . And E represents the Expect.

In general, the inference process is trained by finding an inverse Markov transformation that maximizes the likelihood of the training data.

$$p_\theta(y_{t-1} | y_t, x) = \mathcal{N}(y_{t-1} | \mu_\theta(x, y_t, \gamma_t), \sigma_t^2 I) \quad (2)$$

where p_θ represents the probability distribution predicted by the diffusion model, μ_θ represents the mean value of the prediction noise, and σ_t^2 represents the variance value of the prediction noise. The mean value of p_θ can be obtained by parametric solution.

$$\mu_\theta(x, y_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left(y_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(x, y_t, \gamma_t) \right) \quad (3)$$

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(y_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(x, y_t, \gamma_t) \right) + \sqrt{1-\alpha_t} z_t \quad (4)$$

A non-Markovian inference process is used to increase the inference speed of the diffusion model, enabling rapid optimization of metadata. Considering the non-Markovian inference cannot directly calculates $p_\theta(y_{t-1}|y_t, y_0)$, we assume it conforms to a special distribution, which is no longer restricted by the Markov chain but it needs to ensure $y_t = \sqrt{\gamma_t} y_0 + \sqrt{1-\gamma_t} z$. Only in this way can the optimization goal of the diffusion model remain unchanged during the forward propagation process. Therefore, we can solve it through the undetermined coefficient method:

$$q_\sigma(y_{t-1} | y_t, y_0) = \mathcal{N}(\mu_t, \sigma_t^2 I) \quad (5)$$

$$\mu_t = \sqrt{\alpha_{t-1}} y_0 + \sqrt{\frac{\sigma_{t-1}^2}{\sigma_t^2} - 1} \cdot (y_t - \sqrt{\alpha_t} y_0)$$

where the μ_t represents the expectation of the normal distribution. The size of σ controls the randomness of the forward process. When σ tends to 0, the sampling process will no longer be random.

We make a preliminary prediction of x_0 through a straightforward denoising process.

$$g_\theta^t(y_t) := (y_t - \sqrt{1-\alpha_t} \cdot f_\theta^t(x, y_t)) / \sqrt{\alpha_t} \quad (6)$$

where f_θ^t represents the noise prediction result. Taking $p(x_T) = \mathcal{N}(0, I)$ as a priori condition, the process of predicting y_{t-1} through y_t is as follows:

$$p_\theta^{(t)}(y_{t-1} | y_t) = \begin{cases} \mathcal{N}(g_\theta^1(y_t, x), \sigma_t^2 I) & \text{if } t = 1, \\ q_\sigma(y_{t-1} | y_t, g_\theta^t(y_t, x)) & \text{otherwise,} \end{cases} \quad (7)$$

According to (7), the sample y_{t-1} is generated by y_t and x :

$$y_{t-1} = \psi_t \cdot f_\theta^t(x, y_t) + \sqrt{\frac{\alpha_{t-1}}{\alpha_t} \sigma_t^2} \cdot z_t \quad (8)$$

$$\psi_t = \sqrt{\sigma_{t-1}^2 - \sigma_t^2} - \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot \sigma_t^2$$

where $z_t \sim \mathcal{N}(0, I)$ is the standard Gaussian distribution noise. Different σ will lead to different generation processes. When $\sigma = \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)} \sqrt{1-\alpha_t/\alpha_{t-1}}$, this forward process will become a primitive Markov process.

This paper obtains latent variables $\{x_{\tau_1}, \dots, x_{\tau_S}\}$ by setting a subsequence τ , which is an increasing subsequence of length S from $\{1, \dots, T\}$. Executing (8) on these latent variables in the reverse inference stage, when the length of the subsequence is much smaller than T , we can achieve a significant improvement in computational efficiency due to the iterative nature of the sampling process.

V. DOMAIN ADAPTIVE LEARNING-VIRTUAL TO REALITY

To intuitively reflect the effectiveness of the domain adaptive learning method in solving the visual disparity between Metaverse and physical world, the domain adaptive learning method is applied to the traffic object detection task. In this section, the DAdet is designed, which consists of three main components: physical-metaverse feedback optimization strategy, image-level self-attention feature alignment, and instance-level feature aggregation mechanism. The source domain (labeled metadata N_S) and the target domain (unlabeled real images N_t) are defined correspondingly. Within the labeled metadata, each image is represented as $D_S = \{X_S, B_S, C_S\}$, whereas the unlabeled real images are denoted as $D_T = \{X_t\}$.

According to the knowledge distillation, the training process of DAdet comprises the metaverse model and the physical model, as shown Fig. 5. Both models share identical structures. The feature encoder and detector are initialized through the source domain data and used as the initial models of the metaverse and the physical world. To generate accurate pseudo-labels for real traffic images, a strong-weak enhancement mechanism [35] is introduced. The unlabeled images with weak

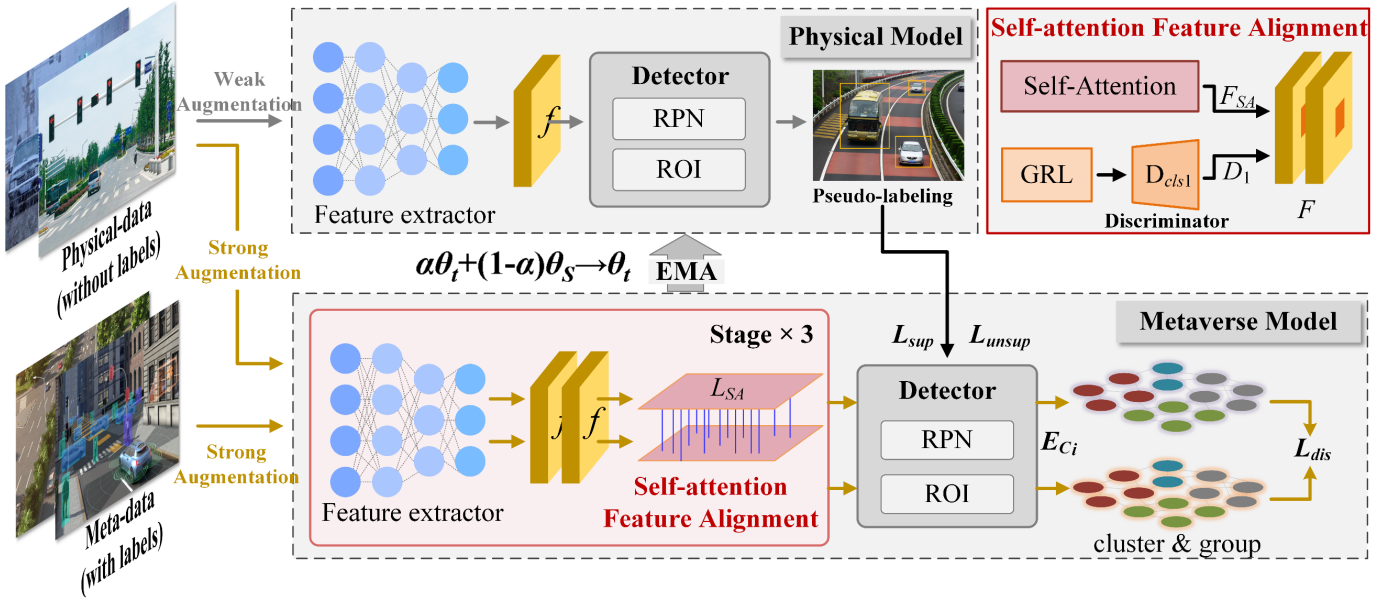


Fig. 5. Domain adaptive-based traffic object detector.

augmentation are the input to the metaverse model, and labeled and unlabeled images with strong augmentation are jointly used as the input to the physical model. The weak augmentation only processes the image through Horizontal Flip operation [36]. For strong augmentation, operations such as GaussianBlur and Cutout are used to increase the diversity of images to better improve the generalization ability of the training model [1]. The physical model minimizes the difference between its pseudo-label outputs and outputs of the metaverse model through a loss function, aiming to improve its performance in the real-world. The total loss of the DAdet is as follows:

$$L = L_{sup} + \lambda_{unsup} \cdot L_{unsup} + \lambda_{dis} \cdot L_{dis} + \lambda_{SA} \cdot L_{SA} \quad (9)$$

where λ is used to control the corresponding loss weighting.

A. Physical-metaverse feedback optimization

The initialized model is trained supervised by the labeled data with the following loss function expressions:

$$\mathcal{L}_{sup}(X_S, B_S, C_S) = \mathcal{L}_{cls}^{rpn} + \mathcal{L}_{cls}^{roi} + \mathcal{L}_{reg}^{rpn} + \mathcal{L}_{reg}^{roi} \quad (9)$$

This loss function includes classification loss and regression loss in both RPN and RoI. The cross-entropy loss is used for classification and the L1 loss is used for regression.

The physical world model is trained based on the real data with pseudo-labels provided by the metaverse model after initialization operation, with the following losses:

$$\mathcal{L}_{unsup}(X_S, \hat{C}_S) = \mathcal{L}_{cls}^{rpn}(X_S, \hat{C}_S) + \mathcal{L}_{cls}^{roi}(X_S, \hat{C}_S) \quad (10)$$

where \hat{C}_S denotes the pseudo-label of the real image generated under the metaverse. Since the metaverse model only generates the confidence of the object class, instead of the location of the bounding box, the regression loss calculation is skipped here.

EMA [37] is used to implement the feedback optimization of the physical-metaverse model. The physical model feeds the parameters back to the metaverse model by adversarial training of source domain and target domain, resulting in better detection performance of the metaverse model for unlabeled real images. Here the feedback optimization of the weights can

be defined as:

$$\varepsilon \cdot W_T + (1 - \varepsilon) \cdot W_S \rightarrow W_T \quad (11)$$

where $W_{\{T,S\}}$ denotes parameters in the teacher (metaverse) /student (physical world) model, the parameter ε represents the EMA degree.

B. Self-attention feature alignment

The self-attention feature alignment module implements feature alignment of focal regions by means of adversarial learning and self-attention mechanism. The feature map f_1 is combined with a domain classifier and a self-attention module to generate a new domain-invariant feature map. The domain classifier is trained in an adversarial learning manner using L_1 loss, with the expression as in (13).

$$L_1 = \sum_{i=1}^{N_S} \log(D_1(F_i^S))^2 + \sum_{i=1}^{N_T} \log(1 - D_1(F_i^T))^2 \quad (12)$$

where F_i^S and F_i^T denote the features extracted from the metaverse image and the real image, respectively. D_1 serves as the pixel-wise probability for the domain classifier to generate the source and target domains, N_S and N_T represent the metaverse and physics features number of a batch respectively. The GRL allows the gradient of the domain classification loss to be automatically inverted during backpropagation, which in turn enables an adversarial loss similar to that of GAN.

The f processed by the domain classifier D_{cls1} and the self-attention mechanism are merged and calculated as follows:

$$F = [\text{softmax}(QK^T) \cdot V] \cdot D_{cls1}(f_1) \quad (13)$$

where the generated feature F is again merged with f_1 to obtain $f_2 = F_2(F \cdot f_1)$. As the feature extractors are stacked, the features are mined from shallow to deep. The shallow features focus on the color and texture of the object, and the deeper features have richer semantic information. The self-attention feature alignment module is applied in two stages to achieve multi-level feature alignment. The loss of this process is represented by the L_{SA} .

C. Feature aggregation mechanism

The imaginary and real-world images in Backbone are aligned at the attention level and the final output feature map f_3 is fed to the RPN used in the Faster-RCNN. Several proposals and their fixed feature vectors $f_r \in \mathbb{R}^{N \times m}$ are generated in the RPN and then hierarchical aggregation clustering is used to cluster the main features in f_r to obtain main feature group. Each proposal is treated as a cluster, and the two closest clusters are merged together using cosine distance as the merging metric. When the intra-cluster dissimilarity exceeds the cluster radius parameter, the merging stops. The above process is specified as follows:

$$\text{dist}(a, b) = 1 - \frac{a \cdot b}{\|a\| \|b\|} \quad (14)$$

$$\text{MaxLink}(A, B) = \max \{ \text{dist}(a, b) : a \in A, b \in B \} \quad (15)$$

where A and B denote the features of two groups in two clusters, a and b denote the feature embedding of a single proposal. $\text{dist}(\cdot)$ denotes the cosine distance, and $\text{MaxLink}(\cdot)$ denotes the merging process.

When the clustering is completed, the instances assigned to each cluster are pooled to construct a representative embedding:

$$E_{c_i} = \frac{\sum_{i=0}^{N_{c_i}} e_i}{N_{c_i}} \quad (16)$$

where c_i denotes the i -th cluster and N_{c_i} denotes the number of assigned instances. Finally, E_{c_i} is fed to the discriminator to execute the instance-level alignment:

$$L_{dis} = -d \log(D(E_{c_i})) - (1-d) \log(1 - D(E_{c_i})) \quad (17)$$

where $d = \{0, 1\}$ denotes the metaverse image and the real image respectively.

VI. EXPERIMENT AND ANALYSIS

Vehicles and traffic signs are important detection objects for visual perception tasks in ITS. Vehicle is characterized by large size, fast movement and large number, which requires fast detection of all targets in the screen; traffic signs are small in size and their variety is rich with certain similarities. Therefore, this section takes the detection of vehicles and traffic signs as an example to verify the effectiveness of MITVF for vision tasks.

A. Datasets

There have been researches related to exploring virtual datasets as metadata and reviewed publicly available driving datasets and virtual test environments. Based on the above, several typical traffic datasets are selected for experimental verification, as follows:

CURE-TSD [14]: The video sequences in the CURE-TSD are divided into two categories: real data and metadata synthesized by the simulator. The real sequences are processed using 12 different types of effects and 5 different challenge levels. The virtual sequences are processed using 11 different types of effects and 5 different challenge levels. With the virtual/real data and the synthesized harsh/extreme scenarios, CURE-TSD can be used to study the robustness of traffic sign recognition algorithms in challenging environments.

51world Synthetic Dataset [38]: The dataset camera and LIDAR related data generated by the autonomous driving simulation test platform 51Sim-One. In this paper, images from the dataset and annotation information are processed as one of the source domains for the experiment, containing a total of 8888 images.

BDD100K [39]: The dataset consists of 100k videos captured in the US, covering different weather conditions (sunny, cloudy, rainy, etc.) and different times of the day (day, night) with diverse traffic scenarios. Some of the multi-scene all-weather data from BDD100K is taken as one of the target domains for the experiment, containing a total of 1000 images. The same traffic element categories in 51world synthetic dataset and BDD100K were selected, with a total of 10 categories.

KITTI [40] and **VKITTI** [41]: VKITTI contains 50 high-resolution monocular videos (21,260 frames). This dataset is a virtual video dataset synthesized from simulated images, with videos generated from 5 different virtual worlds in an urban environment under different imaging and weather conditions. The virtual world is created using the Unity game engine and a novel real-to-virtual cloning method. This virtual scene has a corresponding physical world scene in the KITTI dataset.

B. Experiments Details

Metrics. Nine metrics are selected in this paper to evaluate the performance of the model in terms of detection accuracy. The details are as follows:

$$\text{mAP} = \frac{\sum_{i=1}^C \text{AP}_i}{C} \quad (18)$$

where C is the total number of categories detected and AP is the detection accuracy of individual categories.

FID: A measure used to calculate the distance between the real image and the feature vector of the generated image. The smaller the FID value, the higher the similarity.

PSNR: The mean square error between the original image and the processed image. The larger the value, the better the image quality.

SSIM: Measure the structural similarity of two images, the larger the similarity, the higher the degree of similarity.

IS: Measure the quality and variety of images

AP₅₀: AP at IoU=0.5

AP_S: AP for small objects: $\text{area} < 32^2$

AP_M: AP for medium objects: $32^2 < \text{area} < 96^2$

AP_L: AP for large objects: $\text{area} > 96^2$

Implementation details. The training and test environment is as follows: Linux 4.15.0-142-generic Ubuntu 18.04, with Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GH, 8×32GB DDR4 and 8×TITAN Xp, 12GB video memory, the batch size is set to 32.

C. Two-Stage Metadata Optimization Strategy

In this section, we verify the two-stage metadata optimization strategy. Firstly, five common weather elements are selected to verify the effectiveness of the element reconfigurability strategy. Fig. 6 illustrates the capability of our framework to reconstruct customized traffic scenes with varying weather conditions. In this demonstration, we have selected five weather elements, namely Snowy Landscape, Clouds, Fog, Snowflakes, and Rain, along with four scene elements to represent different

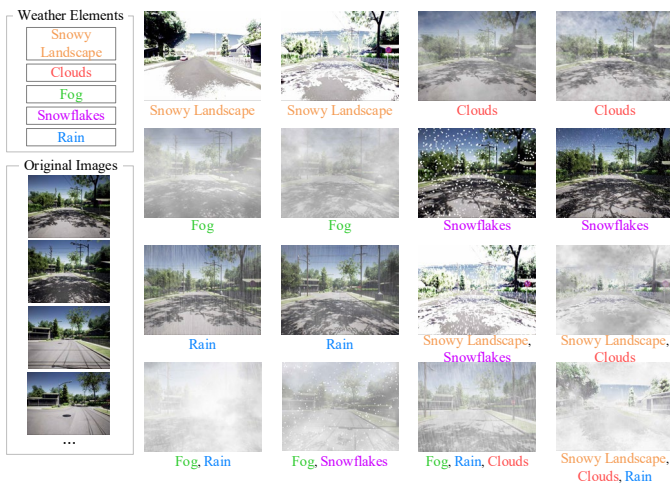


Fig. 6. Samples of customized traffic scene reconstruction

TABLE I
VALIDATION OF ELEMENT RECONFIGURABLE STRATEGIES
BASED ON CURE-TSD DATASET

CURE-TSD (Ours)			
Meta data (%)		Reco. Ratio δ	mAP
Clean data	Reco. data*		
80	0	-	0.515
60	20	0.25	0.520
40	40	0.50	0.534
20	60	0.75	0.551
12	68	0.85	0.568
8	72	0.90	0.553
4	76	0.95	0.547
0	80	1.00	0.491

* Reco. data means reconstruction data.

TABLE II
THE ACCURACY OF THE MODEL UNDER DIFFERENT SR IMAGE
RECONSTRUCTION ALGORITHMS

Method	Dataset	mAP
Bicubic	Data-B	32.8%
SRGAN	Data-S	45.5% (+12.7%)
ESRGAN	Data-E	50.7% (+17.9%)
MAdiff	Data-M	51.2% (+18.4%)

TABLE III
RESULTS OF IMAGE QUALITY UNDER DIFFERENT ϕ

ϕ	L_r	FID↓	IS↑	PSNR↑	SSIM↑	Time	Δt
0.01	10	24.48	115.6	10.27	0.28	0.367s	36.5s
0.02	20	10.48	164.4	16.86	0.66	0.734s	36.1s
0.05	50	7.39	173.9	22.10	0.65	1.847s	35.0s
0.10	100	6.62	184.2	23.32	0.72	3.653s	33.2s
0.50	500	6.48	185.1	26.12	0.74	18.33s	18.5s
1.00	1000	6.35	186.5	27.36	0.78	36.82s	0.0s

aspects of the traffic environment. By strategically combining these elements, we can generate a diverse array of traffic scenes under different weather conditions. For instance, the right part

of Fig. 6 showcases examples of road scenes in rainy, snowy, and foggy weather, respectively. The flexibility of our framework allows for the superimposition of multiple elements, thereby enriching the complexity and realism of the traffic scenes. This ability to customize and reconstruct traffic scenes is crucial for the development and testing of vision-based detection models. By providing a controlled environment with various weather conditions, our framework enables the evaluation of model robustness and performance in scenarios that closely mimic real-world conditions. This, in turn, facilitates the development of more resilient and accurate detection systems for traffic scene analysis. Customized traffic scenarios build rich traffic scenario data and provide conditions for targeted training and testing. Table I shows the impact of the reconfigurability strategy on detection accuracy. The training data consists of 20% real data and 80% metadata. The metadata in the training data is divided into reconstructed data (Reco. data) and unreconstructed data (Clean data), and the ratio of the two (Reco. Ratio δ) is adjusted to verify the detection accuracy of the detection model in complex scenarios. As the proportion of reconstructed data increases, the ability to cope with complex environments (generalization) continues to improve. When $\delta=0.85$, the optimal mAP=56.8% under complex working conditions.

Secondly, for the diffusion model-based optimization acceleration strategy, the optimize the performance is verified at first. Fig. 7 shows the optimized results of the MAdiff, SRGAN [42] and SRGAN [43] for low-resolution images. In this experiment, the Bicubic interpolation method is employed to generate low-resolution versions of the original images, which serves as a baseline for our comparisons. The low-resolution images processed by Bicubic interpolation exhibit a significant loss of detail, making it challenging to distinguish specific features. Both SRGAN and ESRGAN are capable of enhancing the resolution of the images, resulting in improved clarity and detail. However, our proposed MAdiff strategy demonstrates a superior ability to restore image quality. The comparative analysis of clarity changes in the figure indicates that MAdiff yields images with higher clarity than both SRGAN and ESRGAN.

51world Synthetic Dataset is used as training data, which uses Bicubic [44] to generate low-resolution images (480×270), and the original images are used as high-resolution images (1920×1080). From the figure, our method has a better ability to optimize the backlit vehicles and road traffic information in the distance. MAdiff generates images with different levels of detail during the diffusion process, which helps to remove noise and preserve fine details and textures in high-resolution images. And GAN-based method has a limited receptive field and cannot capture the long-range dependencies in the image, which may cause the generated image to appear blurred or artifacts.

The image quality will affect the subsequent detection results. Table II shows the detection accuracy after different optimization processes. Bicubic is used to generate the low-resolution dataset Data-B as a baseline. SRGAN, ESRGAN, and MAdiff process the Data-B to generate datasets Data-M, Data-E, and Data-M of the same number and size, respectively. It can be seen that MAdiff has the highest detection accuracy, which is 18.4% higher than that without optimization.



Fig. 7. Super-resolution reconstruction results of different methods on the 51world dataset, where Low-quality represents the result of 4 times bicubic interpolation down-sampling.

Further, the optimization speed is verified. Table III shows the relationship between image quality and optimization speed. $\phi = \tau / T$ is used to control the sampling interval length of our method, thereby controlling the optimization speed. L_τ represents the length of the sampling iteration interval. FID, PSNR, SSIM and IS are used to measure image quality. Time represents the optimization time. Δt represents the time reduced after adopting the acceleration strategy. The inference speed of the diffusion model has a linear relationship with the number of denoising iterations. When the number of denoising iterations is less, the inference speed of the model is faster; and vice versa. When ϕ decreases, L_τ decreases accordingly, and the speed of MAdiff decreases linearly. $\phi=1$ means that no acceleration strategy is adopted, and sampling is not performed at this time, and the sampling speed is 36.82 seconds per image.

Since different optimization speeds have different image qualities, and the image quality will affect the detection performance, we need to strike a balance between optimization speed, image quality and detection results. Generally, optimization speed is inversely proportional to image quality and detection accuracy, and image quality is directly proportional to detection accuracy. When the image is optimized to a certain extent, the detection accuracy has little effect on the quality improvement. To better illustrate this relationship, Fig. 8 plots the image quality (as evaluated by PSNR) and optimization time on the horizontal and vertical coordinates, respectively. The size of the circle represents the mAP. When Time=3.682s, mAP=46.11%, and the balance between the two reaches the best, at this time $\phi=0.15$.

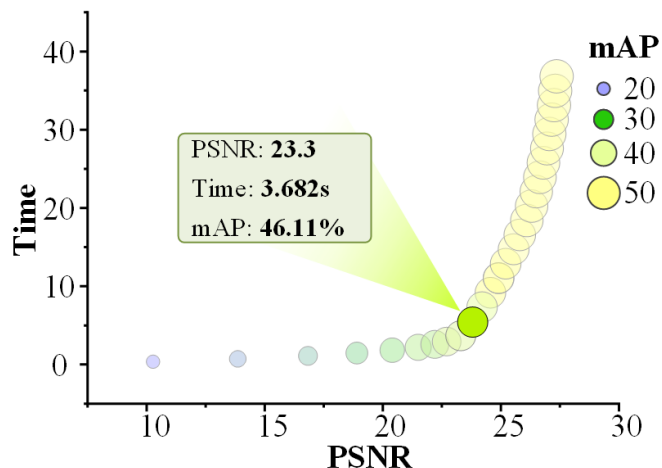


Fig. 8. The relationship between image quality (PSNR), optimization time and detection accuracy.

D. MITVF for Detection

The experiments were conducted on two pairs of datasets: 1) 51world as metadata and the BDD100K as real data; 2) the synthetic and real data in CURE-TSD. The effectiveness of the domain adaptive learning method in the object detection task in this section is verified. It can overcome the visual disparity between the Metaverse and the physical world, and facilitate the application of visual perception tasks to the physical world without distinction.

The comparison methods as follows: Object detection methods: DETR [45], M2Det [46], YOLOv5 [1], YOLOX [47],

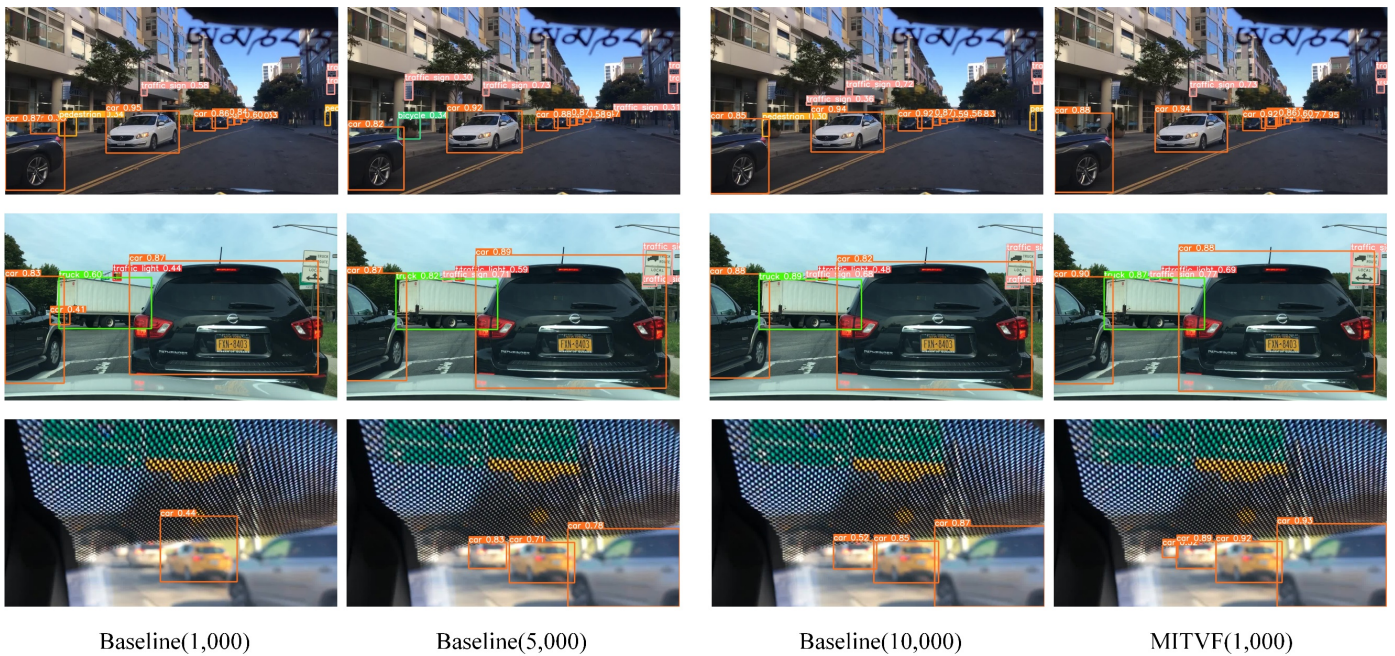


Fig. 9. Vehicle detection sample with or without MITVF on the BDD100K dataset: Baseline is a general object detection method that does not use MITVF, here is Faster-RCNN, which uses different amounts of real data for training, and the number of real datasets in parentheses in the figure; MITVF is trained with a large amount of metadata, and uses 1,000 real images in the adaptive learning stage to complete the performance migration between the virtual and real worlds.

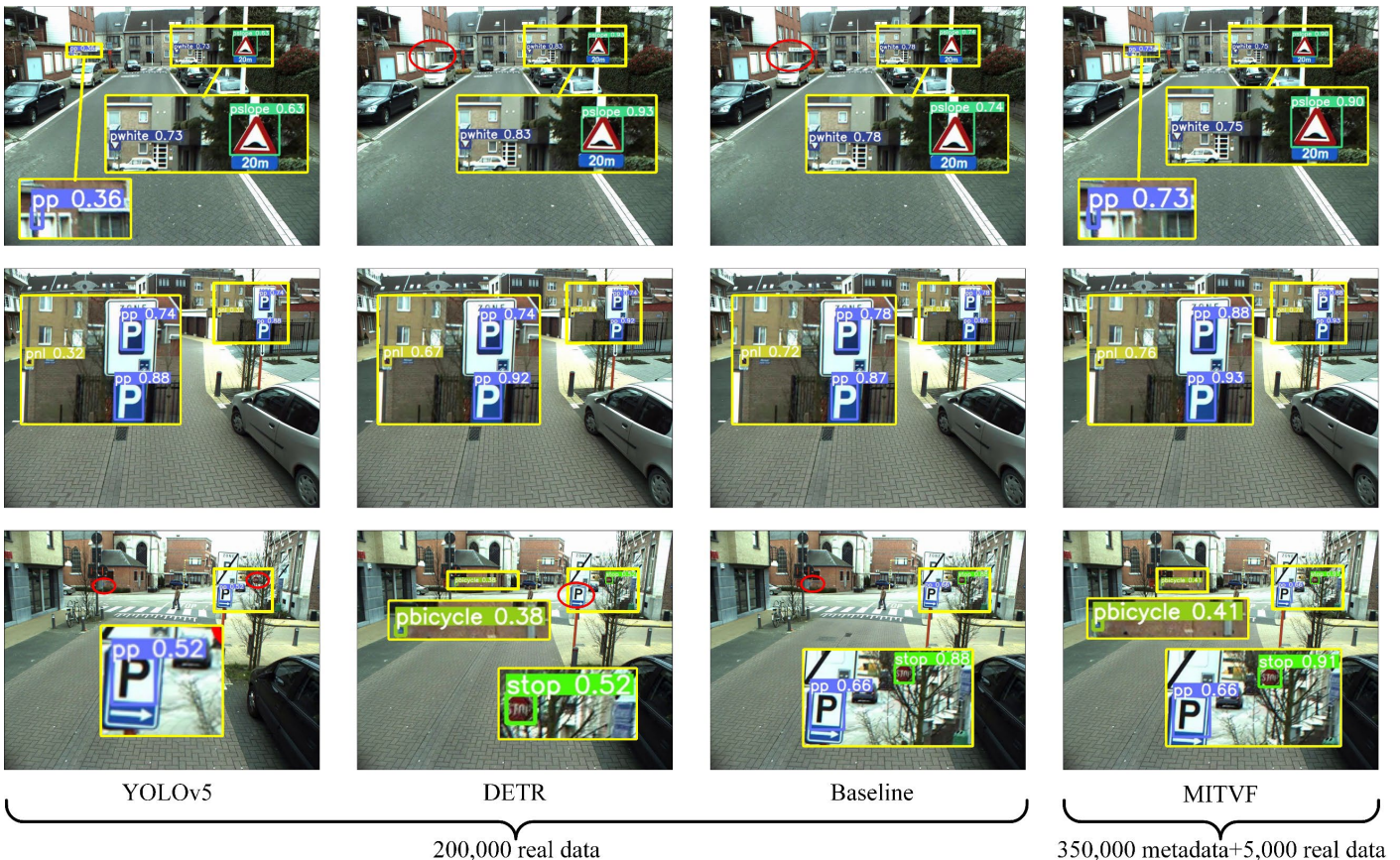


Fig. 10. The detection results of MITVF and other competitors on the traffic sign task on the CURE-TSD dataset.

TABLE IV
RESULTS (AP₅₀) OF 51WORLD SYNTHETIC DATASET→BDD100K.

51 world Synthetic Dataset → BDD100k							
Method	Number of physical world images provided for training						
	100	500	1k	2k	5k	10k	20k
DETR	0.078	0.187	0.214	0.287	0.365	0.445	0.534
M2det	0.097	0.148	0.198	0.304	0.358	0.425	0.458
EfficientDet	0.141	0.204	0.245	0.341	0.394	0.491	0.525
YOLOX	0.112	0.225	0.268	0.357	0.401	0.487	0.497
YOLOv7	0.104	0.169	0.204	0.334	0.378	0.454	0.484
YOLOv5	0.139	0.178	0.220	0.273	0.333	0.416	0.468
VATSD	0.160	0.240	0.279	0.341	0.403	0.541	0.580
CenterNet	0.158	0.234	0.251	0.324	0.386	0.532	0.564
Cascade R-CNN	0.155	0.237	0.268	0.338	0.394	0.543	0.572
BANet	0.112	0.189	0.241	0.305	0.376	0.492	0.533
Fast-Det	0.102	0.164	0.218	0.292	0.352	0.487	0.550
CSIM	0.137	0.179	0.239	0.309	0.371	0.432	0.527
MTSDet	0.161	0.207	0.261	0.313	0.382	0.501	0.536
Baseline	0.157	0.232	0.262	0.340	0.388	0.534	0.552
DA-Detect	0.263	0.387	0.470	0.511	0.512	/	/
FUDA	0.229	0.359	0.452	0.498	0.513	0.515	/
MITVF	0.481	0.531	0.536	0.582	/	/	/

EfficientDet [48], YOLOv7 [49], VATSD [50], Faster R-CNN [51], CenterNet [52], Cascade R-CNN [53], BANet [54], Fast-Det [55], CSIM [56], MTSDet [57]. Domain adaptive learning method: DA-Detect [22] and FUDA [58]. The selected comparison method has a high number of citations and is the latest popular method in the relevant field. Faster-RCNN is used as the baseline, and achieves performance-indistinguishable detection in the physical world by feature alignment between the two domains. The above methods have relatively reliable detection performance, high citations, and have been widely used in traffic object detection tasks.

Fig. 9 illustrates the experimental results of our proposed MITVF framework on the vehicle detection task. Baseline represents the common detection model without domain adaptive learning. The numbers in parentheses indicate the training data utilized for each model, here is the real scene data. MITVF employs domain adaptive learning, leveraging both metadata and a subset of 1,000 real-world images (denoted as '1,000 real data') for training. The comparative results in Fig. 9 clearly demonstrate that MITVF significantly outperforms the Baseline model in terms of detection accuracy. This enhancement is particularly noteworthy given that the training dataset for MITVF is relatively small, comprising only 1,000 real-world images. The superior performance of MITVF can be attributed to its effective utilization of domain adaptive learning, which enables the model to generalize better to real-world scenarios. By integrating metadata and leveraging a limited amount of real scene data, MITVF achieves a more robust and discriminative detection capability.

Fig. 10 presents the comparative results of our MITVF framework on the traffic sign detection task. Notably, MITVF achieves comparable performance to other state-of-the-art methods, such as YOLOv5 and DETR, with only 5,000 real-world images (denoted as '5,000 real data') for training, whereas the other methods require a substantially larger dataset of 200,000 real-world images (denoted as '200,000 real data').

This demonstrates the efficiency and effectiveness of MITVF in leveraging a smaller dataset to achieve competitive results. Traffic signs, compared to vehicles, are smaller and more homogeneous in appearance, which poses additional challenges for detection algorithms. Despite this, other methods like YOLOv5 and DETR, even when trained on the extensive 200,000 real dataset, still exhibit certain limitations, including missed detections and false positives, as indicated by the red circles in Fig. 10. In contrast, MITVF incorporates domain adaptive learning, demonstrates superior detection performance on traffic signs. By leveraging the features learned in the Metaverse and applying them to the physical world, MITVF shows a remarkable reduction in both missed detections and false positives. This enhanced performance underscores the potential of domain adaptive learning in bridging the gap between virtual and real-world data. Furthermore, the ability of MITVF to achieve competitive performance with significantly less real-world data highlights the potential benefits of using virtual data from the Metaverse for training. This approach can substantially reduce the costs associated with data collection and annotation in the real world, while also increasing the diversity and variability of the training dataset. Overall, the results presented in Fig. 10 emphasize the advantages of our MITVF framework in terms of efficiency, effectiveness, and cost reduction for traffic sign detection tasks.

Further, a quantitative analysis is conducted of the proposed methods to quantify the accuracy of each method on the detection task. Table IV and Table V show the performance migration changes of the model trained through metadata on the BDD100K and the CURE-TSD. Based on the domain adaptive learning method, the detection model can achieve better detection performance with a small amount of real data. Table VI shows the accuracy evaluation on the CURE-TSD dataset. The comparative methods are trained on 200,000 real data, and MITVF is trained on 5,000 real data and sufficient metadata. It can be seen that MITVF has better performance in all metrics

TABLE V
RESULTS (AP_{50}) OF CURE-TSD-VIRTUAL \rightarrow CURE-TSD-REAL.

Method	CURE-TSD-Virtual \rightarrow CURE-TSD-Real						
	Number of physical world images provided for training						
	1k	2k	5k	10k	20k	50k	200k
DETR	0.221	0.297	0.387	0.487	0.524	0.535	0.621
M2det	0.199	0.301	0.364	0.429	0.471	0.514	0.543
EfficientDet	0.247	0.354	0.412	0.487	0.498	0.524	0.564
YOLOX	0.271	0.364	0.421	0.517	0.524	0.574	0.591
YOLOv7	0.207	0.344	0.387	0.480	0.497	0.541	0.583
YOLOv5	0.218	0.297	0.314	0.425	0.469	0.524	0.554
VATSD	0.267	0.387	0.471	0.524	0.564	0.635	0.669
CenterNet	0.214	0.309	0.378	0.478	0.522	0.568	0.632
Cascade R-CNN	0.269	0.362	0.442	0.538	0.559	0.634	0.658
BANet	0.201	0.329	0.398	0.499	0.531	0.628	0.632
Fast-Det	0.211	0.327	0.392	0.502	0.538	0.601	0.626
CSIM	0.247	0.358	0.426	0.525	0.569	0.622	0.630
MTSDet	0.228	0.336	0.399	0.489	0.513	0.593	0.603
Baseline	0.272	0.367	0.431	0.541	0.561	0.621	0.651
DA-Detect	0.603	0.631	0.645	0.647	0.648	/	/
FUDA	0.626	0.642	0.651	0.655	0.654	/	/
MITVF	0.667	0.671	0.677	/	/	/	/

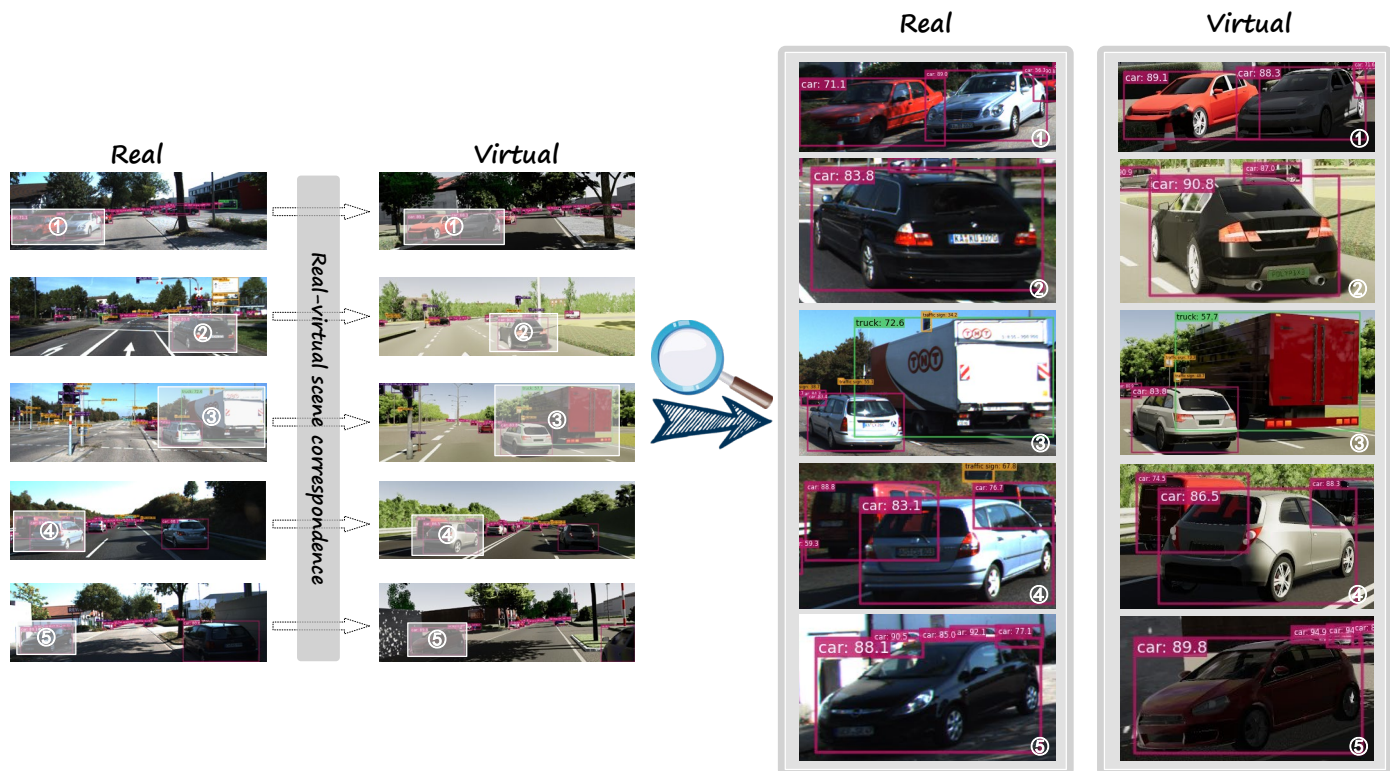


Fig. 11. Detection results on generated virtual scenes and real scenes. The selected datasets are KITTI and VKITTI. “Real” represents the results on KITTI dataset. “Virtual” represents the results on VKITTI dataset.

and outperforms other methods in AP_L , AP_{50} , Recall and other metrics. The above results prove that MITVF has a good prospect and potential for the development of ITS.

E. Virtual and Real Scene Matching

The models trained through the BDD100K and 51world were tested on the KITTI and VKITTI datasets, which proves the matching relationship between the generated metadata and the real scene. VKITTI is the virtual data in KITTI corresponding

to the real scene. The detection results are shown in Fig. 11. The vehicle detection results $AP@50$ were 64.27% (KITTI) and 66.22% (VKITTI) respectively. It can be seen that the accuracy difference between MITVF on the real data and the virtual data is very small, and both can achieve better detection performance. Moreover, this reflects the authenticity of the generated data, and shows that MITVF has good detection performance in both metadata and real scenes.

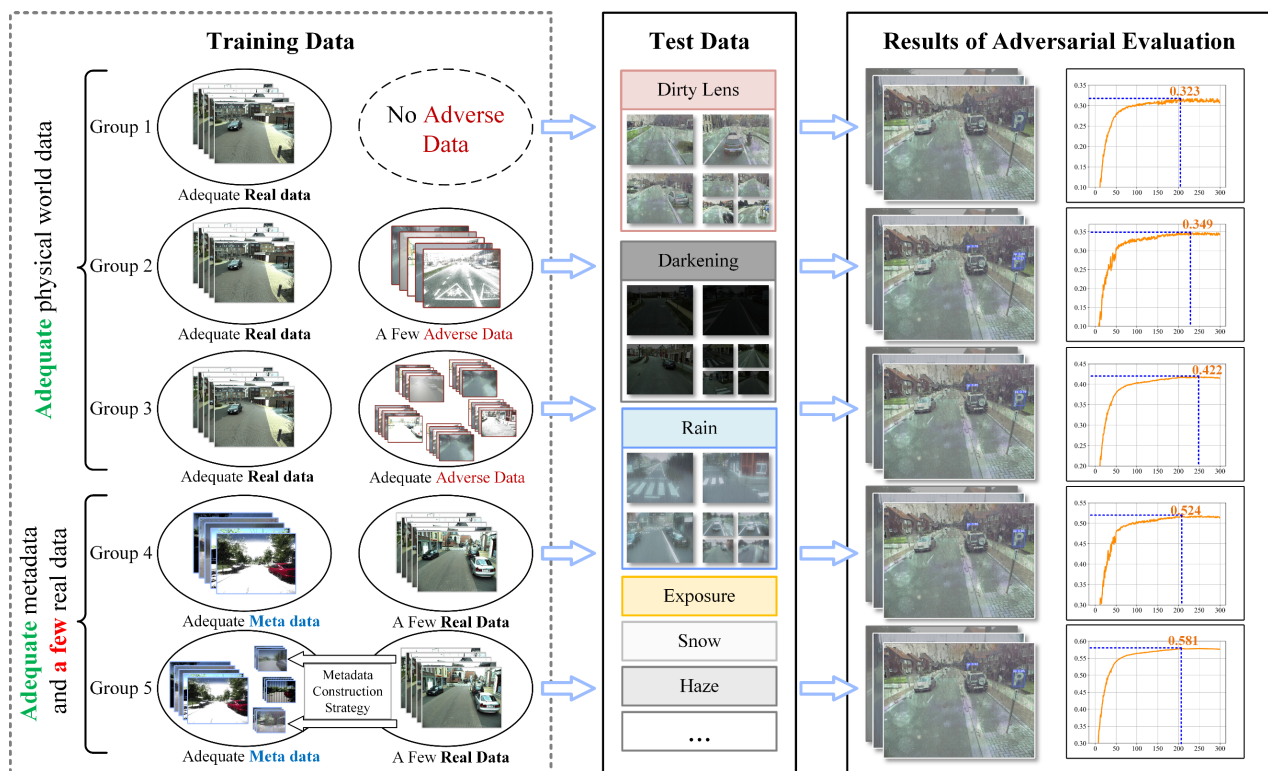


Fig. 12. Visualization process of MITVF robustness experiments for challenging traffic scenarios

TABLE VI
OBJECT DETECTION ACCURACY COMPARISON ON THE CURE-TSD DATASET

Method	mAP	AP ₅₀	AP _S	AP _M	AP _L
DETR	0.563	0.621	0.387	0.524	0.708
M2det	0.501	0.543	0.364	0.471	0.659
Efficientdet	0.547	0.564	0.412	0.498	0.692
YOLOX	0.631	0.591	0.421	0.521	0.712
YOLOv7	0.577	0.583	0.387	0.497	0.681
YOLOv5	0.498	0.554	0.314	0.469	0.661
VATSD	0.662	0.672	0.462	0.509	0.730
CenterNet	0.643	0.651	0.440	0.486	0.693
Cascade R-CNN	0.598	0.660	0.456	0.526	0.738
Baseline	0.594	0.651	0.431	0.501	0.708
MITVF	0.687	0.677	0.471	0.566	0.739

TABLE VII
COMPARES THE COMPUTATIONAL COMPLEXITY AND RUNTIME OF THE ALGORITHM.

Model	Model size	Params.	FLOPs	FPS _{b1}
DETR	159M	41M	225G	20
EfficientDet	15.15M	3.752M	55.0G	26
YOLOX	69.0M	9.010M	26.8G	59
YOLOv7	74.8M	37.34M	120G	87
YOLOv5	14.6M	7.193M	16.7G	125
VATSD	15.1M	7.86M	17.2G	100
CenterNet	730M	32.164M	44.496G	45
Cascade R-CNN	338M	69.395M	85.258G	12
Baseline	159M	41.753M	57.62G	23
MITVF	165.2M	42.263M	59.754G	22

*Params. represents the number of parameters of the model, and FPS_{b1} represents the inference speed of the model when batch-size=1. (On a TITAN Xp graphics card).

TABLE VIII
ROBUSTNESS ANALYSIS OF NOISE EFFECT

Model	Noise	mAP	AP ₅₀	AP _S	AP _M	AP _L
MITVF	Gaussian noise	0.575	0.580	0.351	0.522	0.652
	Pepper noise	0.571	0.576	0.334	0.526	0.661
	Speckle noise	0.556	0.568	0.349	0.509	0.649
Baseline	Gaussian noise	0.501	0.522	0.319	0.487	0.562
	Pepper noise	0.492	0.518	0.302	0.428	0.545
	Speckle noise	0.479	0.503	0.327	0.430	0.535

F. Computational Complexity Analysis

Computational complexity is an important metric in model evaluation. In order to illustrate the computational complexity and running time of the proposed method, we evaluate MITVF from five perspectives: Model size, parameter amount (Params.), FLOPs, inference speed (FPS_{b1}) and O(h). Table VII shows the comparison results between MITVF and other methods. From the table, MITVF has a certain gap compared with existing lightweight networks in terms of Model size and Params. Most of these are caused by the Baseline used, which is to ensure the detection accuracy of various types of traffic objects. Despite this, MITVF still ensures the real-time requirements needed for traffic object detection, and FPS_{b1} reaches 22.

Meanwhile, the overall computational complexity of the model is measured by analysing the $O(h)$ of each module of MITVF. Its main components are as follows: The feature extraction network, the region proposal network, classification and regression layers and the self-attention module. The overall computational complexity of MITVF can be roughly expressed as $O\left(HW + \frac{HW}{s^2} + ND + n^2d\right)$. The H and W represent the

height and width of the input image, respectively. s is the downsampling rate of the feature extraction network. N represents the number of candidate regions, and the feature dimension of each region is D . n represents the sequence length, and d represents the vector dimension.

G. Robustness Analysis

In this section, Gaussian noise, Pepper noise and Speckle noise are used to simulate the device noise interference during image acquisition. The robustness of MITVF was validated and analyzed on the CURE-TSD dataset, as shown in Table VIII. Among them, Physical-Model represents the general detection model under the non-metaverse, the training data is 200,000 real images, and MITVF is trained through 350,000 metadata and 5,000 real images. It can be seen that Metaverse improves the anti-interference ability and robustness of the model by providing a variety of rich training data, and has a good performance in various indicators.

In addition, to further illustrate that MITVF can adapt to challenging traffic scenarios with low cost and high efficiency, we conduct experiments with challenging scenarios and data volume as variables. Fig. 12 illustrates the results of five control experiments conducted to evaluate the performance of our MITVF framework on the traffic sign detection task using the CURE-TSD dataset. The experiments are designed as follows:

Group 1-3: These groups are trained exclusively on varying proportions of real datasets from CURE-TSD, with an increasing proportion of challenging scenes in the training data for each successive group.

Group 4: This group is trained on virtual data from CURE-TSD that encompasses rich and challenging environments, supplemented by a small subset of real data to facilitate cross-domain detection.

Group 5: Similar to Group 4, this group is trained on virtual data with challenging environments and a small subset of real data. However, Group 5 employs a feedback loop mechanism, where the test results are used to optimize the training dataset.

All groups are evaluated on real data from challenging scenarios within CURE-TSD. The right side of Fig. 12 presents the accuracy change curves during the training process for each group.

The results indicate that MITVF significantly enhances visual perception in challenging scenarios through metadata training. Specifically, Group 5 incorporates the feedback loop mechanism, achieves a detection accuracy of 58.1%. This is notably higher than Group 3, which is trained with a sufficient amount of real data containing challenging scenes, yet only reaches an accuracy of 42.2%.

Moreover, the comparison highlights the advantages of using virtual data for training. Acquiring diverse and challenging real-world data can be expensive and risky. In contrast, creating

virtual data in the Metaverse is more convenient and cost-effective, while still providing the necessary diversity and complexity for training robust detection models. This underscores the potential of virtual data and metadata training in improving the performance of detection models in real-world scenarios

VII. CONCLUSION

In view of the challenges of insufficient data quantity, poor scene diversity, and low testing efficiency in existing ITS visual perception tasks, the MITVF is proposed to provide a promising solution to the above challenges. Firstly, a two-stage metadata optimization strategy is proposed to efficiently construct diverse and high-quality metadata. Implement custom construction of metadata through reconfigurable elements to increase data diversity. And the diffusion model-based metadata acceleration optimization strategy expeditiously improves the resolution of low-quality images and provides high-fidelity scenes for visual perception tasks. Secondly, a domain adaptive learning method is proposed to overcome the problem of visual disparity between the Metaverse and the physical world, allowing visual perception tasks under the Metaverse to be efficiently performed in the physical world. In the future, our objective is to augment the capacity of Metaverse for simulating environmental conditions, enhancing its realism in mirroring real-world scenarios. Furthermore, we aim to refine the proposed domain adaptation algorithm to bolster the robustness and generalization of our model by extracting features that remain invariant across different domains

REFERENCES

- [1] J. Wang, Y. Chen, Z. Dong, and M. Gao, "Improved YOLOv5 Network for Real-time Multi-scale Traffic Sign Detection," *Neural Computing and Applications*, vol. 35, pp. 7853–7865, 2022, doi: 10.1007/s00521-022-08077-5.
- [2] R. Bi, J. Xiong, Y. Tian, Q. Li, and X. Liu, "Edge-Cooperative Privacy-Preserving Object Detection Over Random Point Cloud Shares for Connected Autonomous Vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24979–24990, 2022, doi: 10.1109/TITS.2022.3213548.
- [3] B. Yan *et al.*, "Towards Grand Unification of Object Tracking," in *Computer Vision – ECCV 2022*, Tel Aviv, Israel, 2022, pp. 733–751.
- [4] Z. Cao, J. Li, D. Zhang, M. Zhou, and A. Abusorrah, "A Multi-Object Tracking Algorithm With Center-Based Feature Extraction and Occlusion Handling," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4464–4473, 2023, doi: 10.1109/TITS.2022.3229978.
- [5] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, Louisiana, 2022, pp. 4248–4257.
- [6] A. M. Al-Ghaili *et al.*, "A Review of Metaverse's Definitions, Architecture, Applications, Challenges, Issues, Solutions, and Future Trends," *Ieee Access*, vol. 10, pp. 125835–125866, 2022, doi: 10.1109/ACCESS.2022.3225638.
- [7] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, and F. Y. Wang, "Parallel Transportation Systems: Toward IoT-Enabled Smart Urban Traffic Control and Management," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4063–4071, 2020, doi: 10.1109/TITS.2019.2934991.
- [8] Y. Wiseman, "Autonomous Vehicles," in *Research Anthology on Cross-Disciplinary Designs and Applications of Automation*, vol. 2, I. R. M. Association, Ed., 2022, pp. 878–889.
- [9] D. Kumar and N. Muhammad, "Object Detection in Adverse Weather for Autonomous Driving through Data Merging and YOLOv8," *Sensors-Basel*, vol. 23, no. 20, doi: 10.3390/s23208471
- [10] H. Zhang, G. Luo, Y. Li, and F. Y. Wang, "Parallel Vision for Intelligent Transportation Systems in Metaverse: Challenges, Solutions, and

- Potential Applications," *Trans. Syst., Man, Cybern. A, Syst., Humans*, pp. 1-14, 2022, doi: 10.1109/TSMC.2022.3228314.
- [11] K. Wang, C. Gou, N. Zheng, J. M. Rehg, and F.-Y. Wang, "Parallel vision for perception and understanding of complex scenes: methods, framework, and perspectives," *Artificial Intelligence Review*, vol. 48, no. 3, pp. 299-329, 2017/10/01, 2017, doi: 10.1007/s10462-017-9569-z.
- [12] P. Zhou *et al.*, "Vetaverse: A survey on the intersection of Metaverse, vehicles, and transportation systems," *arXiv preprint arXiv:2210.15109*, 2022.
- [13] J. N. Njoku, C. I. Nwakanma, G. C. Amaizu, and D.-S. Kim, "Prospects and challenges of Metaverse application in data-driven intelligent transportation systems," *IET Intelligent Transport Systems*, vol. 17, no. 1, pp. 1-21, 2023, doi: 10.1049/itr2.12252.
- [14] D. Temel, M. H. Chen, and G. AlRegib, "Traffic Sign Detection Under Challenging Conditions: A Deeper Look into Performance Variations and Spectral Characteristics," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3663-3673, 2020, doi: 10.1109/TITS.2019.2931429.
- [15] M. Gilles *et al.*, "MetaGraspNetV2: All-in-One Dataset Enabling Fast and Reliable Robotic Bin Picking via Object Relationship Reasoning and Dexterous Grasping," *IEEE Transactions on Automation Science and Engineering*, pp. 1-19, 2023, doi: 10.1109/TASE.2023.3328964.
- [16] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, and D.-S. Kim, "Artificial intelligence for the metaverse: A survey," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105581, 2023/01/01, 2023, doi: 10.1016/j.engappai.2022.105581.
- [17] M. Xu *et al.*, "A Full Dive Into Realizing the Edge-Enabled Metaverse: Visions, Enabling Technologies, and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 656-700, 2023, doi: 10.1109/COMST.2022.3221119.
- [18] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850-10869, 2023, doi: 10.1109/TPAMI.2023.3261988.
- [19] A. Cortés, C. Rodríguez, G. Vélez, J. Barandiarán, and M. Nieto, "Analysis of Classifier Training on Synthetic Data for Cross-Domain Datasets," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 190-199, 2022, doi: 10.1109/TITS.2020.3009186.
- [20] W. Li, X. Liu, and Y. Yuan, "SIGMA++: Improved Semantic-Complete Graph Matching for Domain Adaptive Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9022-9040, 2023, doi: 10.1109/TPAMI.2023.3235367.
- [21] D. Wang and T. Zhang, "Establishment and Optimization of Video Analysis System in Metaverse Environment," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 14, no. 10, 2023, doi: 10.14569/IJACSA.2023.0141006.
- [22] J. Li, R. Xu, J. Ma, Q. Zou, J. Ma, and H. Yu, "Domain Adaptive Object Detection for Autonomous Driving under Foggy Weather," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 612-622.
- [23] J. Y. Kim and J. M. Oh, "Opportunities and Challenges of Metaverse for Automotive and Mobility Industries," in *Int. Conf. ICT Convergence*, Jeju Island, Korea, Republic of, 2022, pp. 113-117.
- [24] M. Ragab *et al.*, "Adversarial Multiple-Target Domain Adaptation for Fault Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-11, 2021, doi: 10.1109/TIM.2020.3009341.
- [25] J. Song, Y. Chen, J. Ye, and M. Song, "Spot-Adaptive Knowledge Distillation," *IEEE Transactions on Image Processing*, vol. 31, pp. 3359-3370, 2022, doi: 10.1109/TIP.2022.3170728.
- [26] L. U. Khan, A. Elhagry, M. Guizani, and A. E. Saddik, "Edge Intelligence Empowered Vehicular Metaverse: Key Design Aspects and Future Directions," *IEEE Internet of Things Magazine*, vol. 7, no. 1, pp. 120-126, 2024, doi: 10.1109/IOTM.001.2300078.
- [27] W. Li *et al.*, "Intelligent Cockpit for Intelligent Vehicle in Metaverse: A Case Study of Empathetic Auditory Regulation of Human Emotion," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2173-2187, 2023, doi: 10.1109/TSMC.2022.3229021.
- [28] L. Fan, D. Cao, C. Zeng, B. Li, Y. Li, and F. Y. Wang, "Cognitive-Based Crack Detection for Road Maintenance: An Integrated System in Cyber-Physical-Social Systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 6, pp. 3485-3500, 2023, doi: 10.1109/TSMC.2022.3227209.
- [29] A. Danylec, K. Shahabaddar, H. Dia, and A. Kulkarni, "Cognitive Implementation of Metaverse Embedded Learning and Training Framework for Drivers in Rolling Stock," *Machines*, vol. 10, no. 10, doi: 10.3390/machines10100926
- [30] D. Pamucar, M. Deveci, I. Gokasar, M. Tavana, and M. Köppen, "A Metaverse Assessment Model for Sustainable Transportation Using Ordinal Priority Approach and Aczel-Alsina Norms," *Technological Forecasting and Social Change*, vol. 182, p. 121778, 2022/09/01, 2022, doi: 10.1016/j.techfore.2022.121778.
- [31] C. W. Lee, "Application of Metaverse Service to Healthcare Industry: A Strategic Perspective," *International Journal of Environmental Research and Public Health*, vol. 19, no. 20, doi: 10.3390/ijerph192013038
- [32] G. S. Contreras, A. H. González, M. I. S. Fernández, C. B. Martínez, J. Cepa, and Z. Escobar, "The importance of the application of the metaverse in education," *Modern Applied Science*, vol. 16, no. 3, pp. 1-34, 2022, doi: 10.5539/mas.v16n3p34.
- [33] Y. Li and X. Song, "Toward a Metaverse Era: A Study on the Design of Smart Home Entertainment Scene Experience for Empty-Nest Youth," in *Proceedings of the Tenth International Symposium of Chinese CHI*, Guangzhou, China, 2024, pp. 62-71.
- [34] H. Jeong, Y. Yi, and D. Kim, "An innovative e-commerce platform incorporating metaverse to live commerce," *International Journal of Innovative Computing, Information and Control*, vol. 18, no. 1, pp. 221-229, 2022, doi: 10.24507/ijic.18.01.221.
- [35] J. Yu *et al.*, "MTTrans: Cross-domain Object Detection with Mean Teacher Transformer," in *Computer Vision—ECCV 2022*, Tel Aviv, Israel, 2022, pp. 629-645.
- [36] K. Zhang, Z. Cao, and J. Wu, "Circular Shift: An Effective Data Augmentation Method For Convolutional Neural Network On Image Classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1676-1680.
- [37] Y. J. Li *et al.*, "Cross-Domain Adaptive Teacher for Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, New Orleans, LA, USA, 2022, pp. 7571-7580.
- [38] 51Sim-One. (2022). *51WORLD Synthetic Dataset* [Online]. Available: <https://github.com/51WORLD/SyntheticDataset>.
- [39] F. Yu *et al.*, "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, Seattle, WA, USA, 2020, pp. 2633-2642.
- [40] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "VirtualWorlds as Proxy for Multi-object Tracking Analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4340-4349.
- [41] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231-1237, 2013
- [42] X. Wang *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, Munich, Germany, 2018.
- [43] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 105-114.
- [44] Z. Huang and L. Cao, "Bicubic interpolation and extrapolation iteration method for high resolution digital holographic reconstruction," *Optics and Lasers in Engineering*, vol. 130, p. 106090, 2020, doi: 10.1016/j.optlaseng.2020.106090.
- [45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, Glasgow, UK, 2020, pp. 213-229.
- [46] T. S. Q. Zhang, Y. Wang, Z. Tang, Y. Chen, L. Cai, H. Ling, "M2Det A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network," in *AAAI Conf. Artif. Intell.*, Hawaii, USA, vol. 33, 2019, pp. 9259-9266.
- [47] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," 2021, *arXiv:2107.08430*.
- [48] R. P. M. Tan, Q.V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, Seattle, WA, USA, 2020, pp. 10781-10790.
- [49] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
- [50] J. Wang, Y. Chen, X. Ji, Z. Dong, M. Gao, and C. S. Lai, "Vehicle-Mounted Adaptive Traffic Sign Detector for Small-Sized Signs in Multiple Working Conditions," *IEEE Trans. Intell. Transp. Syst.*, early access, 2023, doi: 10.1109/TITS.2023.3309644.
- [51] K. H. Shaoqing Ren, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE*

>

- > *Trans. Pattern Anal. Machine Intell.*, vol. 39, pp. 1137-1149, 2017, doi: 10.1109/TPAMI.2016.2577031.
- [52] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic Two-stage Detection," 2021, *arXiv:2103.07461*.
- [53] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154-6162.
- [54] S.-y. Wang, Z. Qu, C.-j. Li, and L.-y. Gao, "BANet: Small and multi-object detection with a bidirectional attention network for traffic scenes," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105504, 2023/01/01/, 2023, doi: 10.1016/j.engappai.2022.105504.
- [55] Y. Chen, Y. Shi, C. Xie, C. Lin, Q. Hu, and Z. Chen, "Fast object detector with center localization confidence based on FCOS for environment perception in urban traffic scene," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, p. 09544070231153199, 2023, doi: 10.1177/09544070231153199.
- [56] Y. F. Lu, J. W. Gao, Q. Yu, Y. Li, Y. S. Lv, and H. Qiao, "A Cross-Scale and Illumination Invariance-Based Model for Robust Object Detection in Traffic Surveillance Scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 6989-6999, 2023, doi: 10.1109/TITS.2023.3264573.
- [57] H. Wei, Q. Zhang, Y. Qian, Z. Xu, and J. Han, "MTSDet: multi-scale traffic sign detection with attention and path aggregation," *Applied Intelligence*, vol. 53, no. 1, pp. 238-250, 2023/01/01, 2023, doi: 10.1007/s10489-022-03459-7.
- [58] Y. Zhu, R. Xu, C. Tao, H. An, Z. Sun, and K. Lu, "An Object Detection Method Based on Feature Uncertainty Domain Adaptation for Autonomous Driving," *Applied Sciences*, vol. 13, no. 11, doi: 10.3390/app13116448