Cover Article

Parallel Images: Knowledge and Data-Driven Multi-Modal Information Intelligent Analysis

Hui Zhang, Xiaofeng Jia, Guiyang Luo, Yonglin Tian, Shutong Liang, and Dongyang Hong

***Abstract*—Visual perception computing is a crucial area in artificial intelligence, aiming to simulate human vision for the intelligent analysis of complex visual data. However, current methods face several challenges, such as missing data, weak generalization across different scenarios, and difficulties in learning complex patterns, particularly in rare or long-tail situations. The framework of parallel images is reviewed in this paper, which provides new ways to advance visual perception by closely connecting real imaging systems with artificial ones. First, artificial image systems can be built to reflect real environments, enabling both real and artificial images to work together. These artificial systems produce multi-modal data, helping to solve the problem of incomplete data. Second, virtual-to-real model transfer approaches based on multi-view feature fusion are discussed, which support adaptive model improvement and better generalization to new scenarios. Finally, parallel visual models are introduced that combine data from different sources and integrate various types of knowledge, greatly improving performance on diverse visual recognition tasks.**

***Index Terms*—Parallel images, visual perception, multimodal data**

I. Introduction

In recent years, continuous breakthroughs in artificial intelligence (AI), particularly in generative AI and large-scale foundation models, have accelerated the deployment of computer vision technologies as a core pillar of perceptual intelligence across diverse complex scenarios. From autonomous driving and urban security to smart healthcare and industrial quality inspection, the capabilities of visual systems in recognition, understanding, and reasoning are advancing toward higher precision and intelligence. Despite significant progress in algorithmic structures, model capacity, and computational resources, current visual perception systems still face core challenges, such as strong data dependency, limited generalization ability, and high training costs. These bottlenecks severely restrict the scalability and performance of such systems in real-world environments.

The efficacy of vision models hinges on the diversity and annotation quality of training data. However, real-world visual data inherently exhibit a long-tailed distribution: most data correspond to common scenarios (e.g., clear weather, daylight, and routine traffic), while images capturing rare but critical events (e.g., storms, nighttime accidents, and emergencies) remain scarce [1–3]. This data imbalance leads to degraded model performance in high-risk, low-frequency situations, posing substantial safety hazards and misjudgment risks. Moreover, conventional data annotation pipelines heavily rely on manual labor, requiring frame-level, bounding-box, or even pixel-wise labels for tasks, such as image classification, object detection, and semantic segmentation [4–6]. Such processes create bottlenecks in efficiency and introduce quality control challenges.

Although generative AI techniques such as generative adversarial networks (GANs) [7] and diffusion models [8] offer new pathways for data augmentation, most current generative models still focus on general aspects like image quality and style control. They often struggle to generate task-aware, domain-specific visual data with rigorous physical consistency, semantic coherence, and controllability, particularly in complex scenarios. Additionally, while foundation models such as CLIP [9], SAM [10], and GPT-4V [11] demonstrate strong potential in multi-modal alignment and cross-task generalization, their lack of interpretability, uncontrolled generation behavior, and opaque training data sources remain pressing concerns in high-precision and safety-critical visual applications. These issues call for integration with more structured and engineering-oriented technical frameworks for effective calibration and enhancement.

To address these challenges, parallel intelligence [12] introduces a novel paradigm centered on virtual-real interaction, data-driven learning, and computational experimentation. The ACP theory (artificial systems, computational experiments and parallel execution), pioneered in Refs. [13–15], serves as the methodological cornerstone of parallel intelligence. Building upon this framework, the concept of parallel images has emerged, offering a solution that enables the generation of large-scale, diverse, and finely annotated synthetic visual data through the construction of highly realistic and controllable artificial scenes, as shown in Fig. 1. This provides a reproducible, controllable, and low-

Manuscript received: 15 January 2025; revised: 9 March 2025; accepted: 10 May 2025. *(Corresponding author: Guiyang Luo.)*

Citation: H. Zhang, X. Jia, G. Luo, Y. Tian, S. Liang, and D. Hong, Parallel image: Knowledge and data-driven multi-modal information intelligent analysis, *Int. J. Intell. Control Syst.*, 2025, 30(2), 93–107.

Hui Zhang, Shutong Liang, and Dongyang Hong are with School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: huizhang1@bjtu.edu.cn; 24140062@bjtu.edu.cn; 24125326@bjtu.edu.cn).

Xiaofeng Jia is with Department of Data Management, Beijing Big Data Centre, Beijing 100101, China (e-mail: jiaxf@jxj.beijing.gov.cn).

Guiyang Luo is with State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: luoguiyang@bupt.edu.cn).

Yonglin Tian is with State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yonglin.tian@ia.ac.cn).

Digital Object Identifier 10.62678/IJICS202506.10175

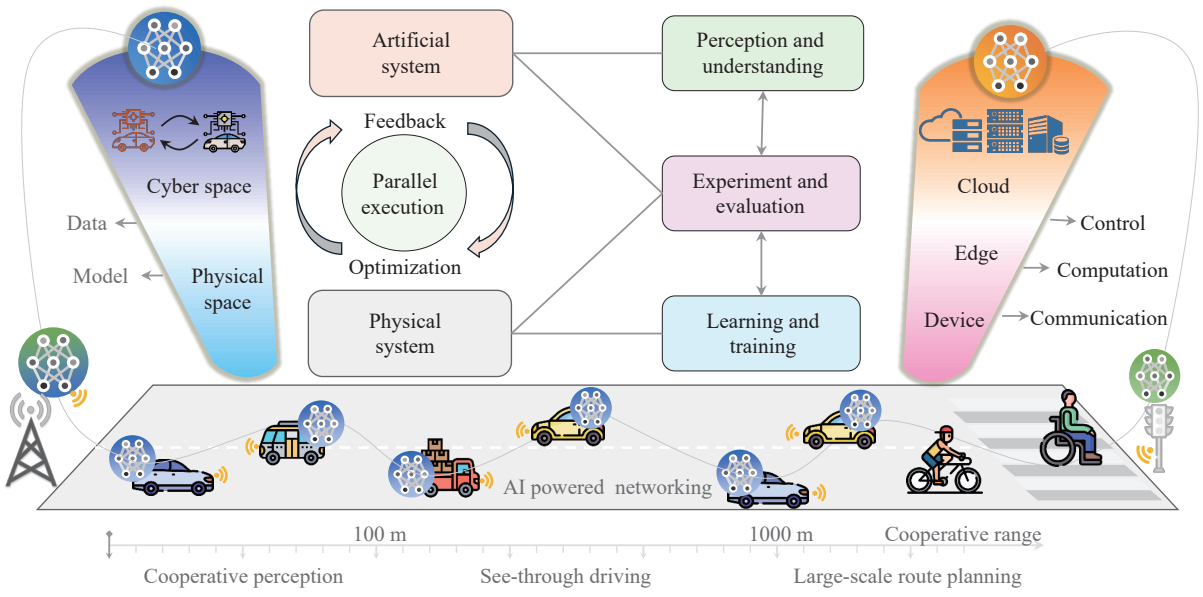


Figure 1 Parallel images methodology.

cost approach to model training and evaluation under data-scarce conditions.

Specifically, the parallel images system is built upon three core components: artificial scene generation, computational experimentation, and parallel execution. First, it utilizes virtual simulation platforms and 3D modeling technologies to construct artificial scenes with high dimensionality and heterogeneity. These scenes are parameterized and controllable, allowing for diverse combinations of variables such as weather conditions, lighting environments, camera viewpoints, and object behaviors. Second, computational experiments are performed using the generated data. These experiments are designed to evaluate and optimize the robustness and generalization capabilities of visual models under complex and variable conditions. Third, feedback signals and performance metrics from the real world are incorporated to iteratively adjust scene parameters. This process enables a closed-loop optimization between the virtual and physical domains and supports the continuous evolution of models. This approach departs from the conventional paradigm of static training followed by one-time deployment. Instead, it promotes a dynamic and adaptive learning framework characterized by virtual training, real-world adaptation, and continuous improvement.

With the rise of multi-modal foundation models, parallel images technology is entering a new phase of deep integration with large-scale models. On the one hand, generative foundation models can act as core engines for virtual scene construction and image synthesis. They are capable of generating high-quality and semantically coherent images based on multi-modal inputs such as text descriptions, sketches, and semantic maps. On the other hand, the parallel images system provides structured, richly annotated, and semantically explicit datasets. These datasets support the fine-tuning and task-specific adaptation of foundation models across various domains. This synergy facilitates the development of next-generation visual perception systems. Such systems are task-driven, guided by semantic constraints, and enhanced through human-machine collaboration. Furthermore, this integration supports a paradigm shift. It enables the transition from purely data-driven intelligence to a more robust framework that incorporates data-driven and knowledge-guided intelligence.

As a systematic framework that integrates AI-based modeling, virtual simulation, and real-world feedback, parallel images technology effectively addresses challenges such as data scarcity, high annotation costs, and limited generalization. Its integration with generative AI and foundation models not only enhances the expressiveness and adaptability of perception systems but also provides a solid foundation for improving the trustworthiness, controllability, and systematicity of AI systems. This paper provides a systematic review of the core concepts, key technologies, and representative applications of parallel images, and explores their research prospects and application potential.

II. Evolution of Parallel Images Technology

With the deep integration and innovative application of artificial intelligence technologies and parallel theory in computer vision, parallel images have evolved into a comprehensive theoretical and technical framework. As illustrated in Fig. 2, this paper systematically categorizes the development of parallel images technologies into four key stages from the perspective of technological evolution: the exploratory experimentation stage, the theoretical formulation stage, the expansion and deepening stage, and the frontier breakthrough stage.

A. Exploratory Experimentation Stage

Parallel images, as an extension of parallel systems in the image domain, emerge in tandem with the development of parallel system theory. To address the challenges of prediction inaccuracies, modeling difficulties, and the limitations of traditional control methods in complex systems research (e.g., socioeconomics and traffic management), parallel systems [13] are introduced as an innovative solution. The approach

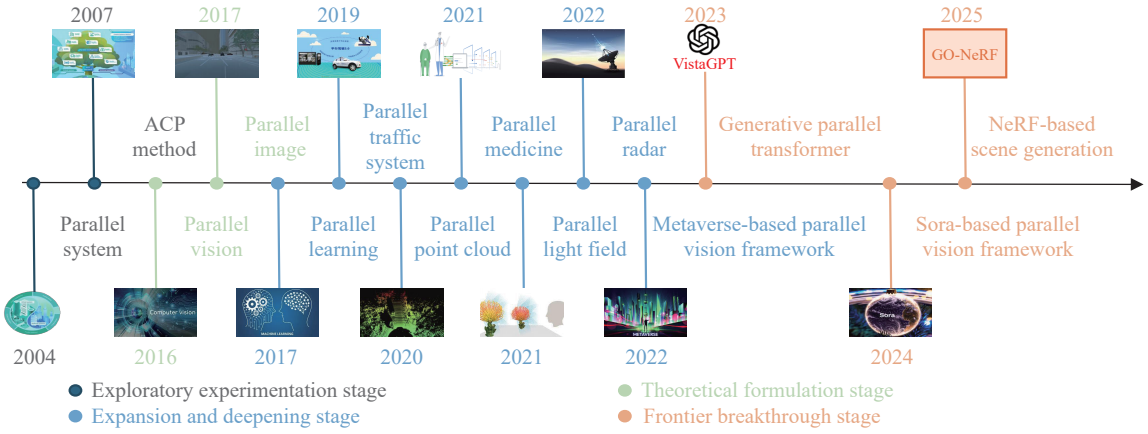


Figure 2 Evolution of parallel images.

constructs artificial systems to facilitate computational experiments and virtual-real interaction, employing a "multiple worlds" perspective to bridge the modeling gap and enhance the proactivity of artificial systems for dynamic control. Subsequently, Wang [16] proposed the ACP method, establishing a comprehensive computational theory and methodological framework for parallel systems. The core concept comprises three key aspects. First, artificial societies are constructed as computational laboratories, where bottom-up modeling is employed to simulate the emergent behaviors of complex systems, thus surpassing the limitations of a conventional "single-world" paradigm and positioning artificial systems as parallel surrogates of reality. Second, the adoption of a "multiple-worlds" perspective enables artificial and real systems to co-evolve in parallel, facilitating controllable, reproducible, and systematic analysis of complex phenomena. Finally, a virtual-real interaction mechanism is established, incorporating dynamic comparison and closed-loop feedback to support experimental validation, decision optimization, and intelligent system control.

Researchers have actively explored the innovative applications of parallel systems and the ACP method across various domains. In the field of traffic management, Wang [17] proposed integrating real-world traffic systems, artificial systems, and training and evaluation modules, achieving an organic combination of computational experiments and virtual-real interaction. When applied in cities like Jinan and Taicang, this system significantly improves emergency response capabilities and strategy optimization in traffic management. Zhu et al. [18] constructed an artificial traffic system using agent-based modeling to simulate the impacts of population structure, activity plans, travel behavior, and adverse weather conditions on traffic decision-making. This approach effectively addresses the limitations of traditional traffic simulations in modeling microscopic behavior and accounting for indirect factors. The ACP method also demonstrates strong adaptability in the realm of public safety. Duan et al. [19] developed an artificial society that integrated multiple models to simulate epidemic transmission processes and assess the effectiveness of various intervention strategies, providing scientific support for pandemic control. At the level of urban governance, Wang et al. [20] applied the ACP method to an intelligent parking system. By simulating and optimizing different management strategies, they achieve dynamic allocation and efficient scheduling of urban parking resources. These research efforts collectively demonstrate that the ACP method, through its closed-loop mechanism of "modeling-experiment-feedback-optimization", offers a highly generalizable, efficient, and sustainable technical pathway for addressing intelligent control problems in complex systems.

B. Theoretical Formulation Stage

Wang et al. [21] systematically proposed parallel vision based on ACP theory for the first time. This method innovatively integrates three core components of parallel systems with computer vision techniques: constructing artificial scenes to address data scarcity, leveraging computational experiments to optimize model performance, and employing parallel execution to facilitate continuous improvement. Then, the theory of parallel images [22] was introduced to tackle the issue of insufficient model generalization caused by difficulties in real-world image data collection and high labeling costs in computer vision. The key breakthroughs include: (1) extending traditional image generation from single data collection to a virtual-real collaborative parallel system model, (2) establishing a dynamic evolutionary learning mechanism that enables models to update themselves in open environments continuously. These innovations provide novel solutions for visual tasks in dynamic scenarios such as autonomous driving and intelligent surveillance, propelling the paradigm shift in computer vision from static analysis to dynamic evolution.

C. Expansion and Deepening Stage

As parallel system theory and the ACP method mature and evolve, this innovative methodology extends to multiple key domains, forming domain-specific technical frameworks. In machine learning, Li et al. [23] introduced the concept of parallel learning, which leveraged the parallel evolution of software-defined artificial data systems and real-world data to address the challenges of model training under small-sample conditions. In the domain of intelligent transportation, parallel transportation emerges through the work of Lv et al. [24]. By constructing artificial transportation systems that interact with

real-world road networks, this framework facilitates the collaborative optimization of congestion prediction and traffic signal control. In the healthcare sector, Wang [25] developed parallel medicine, employing the ACP method to construct virtual medical systems that interacted with real clinical data. This approach transforms clinical "small data" into synthetic "big data" and further distills it into "deep intelligence" for precise diagnosis and treatment, effectively overcoming model training challenges under small-sample conditions, such as in rare disease identification and personalized medicine. These domain-specific advancements validate the general applicability of the ACP method and establish a new paradigm of "virtual-real interaction and parallel evolution" for addressing complex system problems in traffic management, disease diagnosis, and machine learning.

In parallel vision, the technical potential of ACP methods becomes increasingly evident. Zhang et al. [26] focused on the technical implementation of parallel vision, proposing a framework that leveraged synthetic data to train vision models, enhancing model performance through global/local feature alignment and virtual-real interaction. The effectiveness of this approach is validated in object detection and instance segmentation tasks. Given the powerful capabilities of the metaverse in virtual simulation and data generation, Zhang et al. [27] further introduced a metaverse-based parallel traffic vision framework. By constructing a virtual traffic space, implementing computational experiment-driven model learning, and optimizing feedback through virtual-real parallelism, they address challenges in generalization, data scarcity, and complex scene adaptation of environmental perception in intelligent transportation systems, thereby extending the application boundaries of parallel vision.

Meanwhile, parallel images technology continues to achieve in-depth development across multiple domains. In the field of intelligent transportation testing, Li et al. [28] developed a parallel testing system based on virtual-reality interaction, transforming real-world scenario data into virtual extreme scenarios (e.g., adverse weather and emergency events). This system generates diverse test tasks to efficiently and safely evaluate and enhance the intelligence of autonomous vehicles. In pedestrian detection, Zhang et al. [29] employed parallel vision methods to construct virtual scenes for generating synthetic data. By integrating online learning mechanisms, they achieve efficient training and dynamic optimization of pedestrian detection models in fixed-camera scenarios. In medical image analysis, Shen et al. [30] constructed artificial models of pathological tissues and interacted them with real clinical data, enabling high-precision diagnosis under small-sample conditions. To address the limitations of traditional 2D images in expressive dimensions and realism, parallel images technology further integrates 3D reconstruction, light field imaging, and radar simulation, leading to new directions such as parallel point clouds [31], parallel light fields [32], and parallel radar [33]. By combining the light field imaging's ability to capture complete optical information with the spatial modeling advantages of 3D reconstruction, these methods achieve multi-dimensional and high-fidelity digital reconstruction of complex scenes.

D. Frontier Breakthrough Stage

The rise of multi-modal large models injects new momentum into the development of parallel images technology, significantly reducing the cost of constructing highly diverse artificial scenes. By receiving multi-modal inputs, generative large models produce logically coherent, stylistically consistent, and structurally reasonable 2D images and 3D scenes. Tian et al. [34] proposed a generative parallel transformer framework that constructed a modular in-vehicle Transformer federation and an autonomous driving system composition platform based on large language models. The framework enables heterogeneous resource integration and virtual-real interactive optimization across multiple driving scenarios by integrating scenario engineering and federated intelligence technologies. Yu et al. [35] proposed a parallel vision framework based on the video generation model Sora. By integrating Sora with the ACP parallel system theory, they constructed a closed-loop intelligent perception system with virtual-real interaction. This system automatically generates high-fidelity driving scene data from textual descriptions and continuously enhances the perception generalization capability of autonomous vehicles in unknown scenarios through a feedback optimization mechanism. Zhang et al. [36] presented a text-driven 3D scene generation method based on neural radiance field (NeRF). By combining a pre-trained text-to-image diffusion model to generate initial content and geometric priors, and employing a progressive refinement strategy, they achieve multi-view-consistent high-fidelity scene synthesis. Dai et al. [37] proposed a virtual reality content generation method based on NeRF. By generating high-quality 3D objects from text prompts and user-specified regions, and employing compositional rendering and optimization strategies, they achieved seamless integration of objects and scenes.

Leveraging the powerful generalization and knowledge representation capabilities of large models, parallel images can achieve cross-modal and highly consistent synthetic data generation across multiple task domains. The resulting image data not only exhibit high controllability in semantic expression and structural organization but also facilitate fine-grained annotation and reusable datasets. This provides stable, abundant, and high-quality data support for training, validation, and deployment stages of visual models. By deeply integrating with large models, parallel images effectively address the bottlenecks of data acquisition and annotation while establishing a solid foundation for building a new generation of visual perception systems that are explainable, controllable, and evolvable.

III. Overall Framework of Parallel Images

Within the overall framework of parallel images, as illustrated in Fig. 3, artificial scene generation, computational experiment simulation, and virtual-real parallel execution form a cohesive pipeline that connects data construction, model optimization, and closed-loop feedback. Figure 4 illustrates the relevant key technologies. These three components complement each other to establish a continuously self-optimizing intelligent visual computing

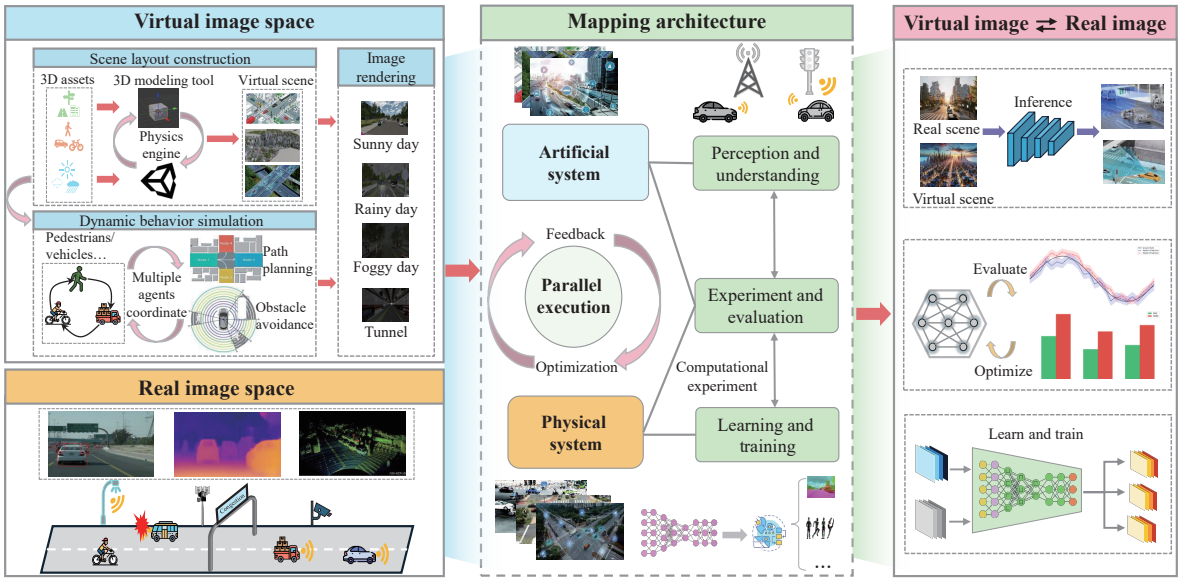


Figure 3 Framework of parallel images.

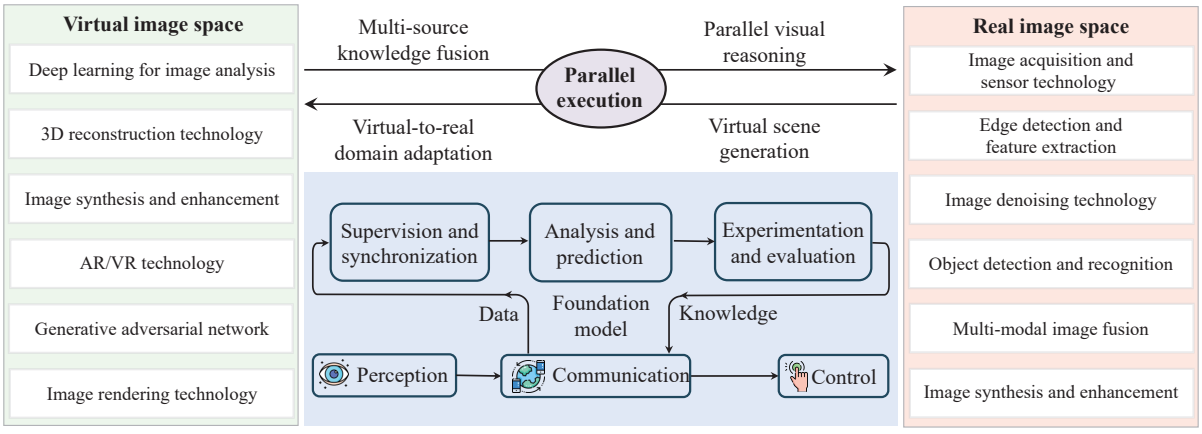


Figure 4 Key technology in parallel images.

system, providing robust support for intelligent perception and understanding in complex environments.

A. Artificial Scene Generation

As a foundational component of the parallel images framework, artificial scene generation focuses on constructing highly realistic and diverse virtual environments using advanced technologies such as computer graphics, virtual reality, and microscopic simulation, aiming to accurately replicate complex and dynamic real-world scenarios.

The process of artificial scene generation within the parallel images framework encompasses three core stages: scene framework construction, dynamic behavior simulation, and image rendering. Scene framework construction establishes a solid physical and geometric foundation for dynamic behavior simulation, while dynamic behavior simulation introduces realistic and complex interactions within the environment. Finally, a programmable rendering system converts the virtual environment into high-fidelity and annotatable visual data. First, scene framework construction relies on 3D modeling tools and physics engines to create scene topology and configure object properties. In practice, game engines, simulation tools, and 3D modeling software are employed to construct scenes, sourcing or creating large-scale 3D models that include static objects (e.g., buildings, roads and vegetation), dynamic objects (e.g., pedestrians and vehicles), and natural environmental elements (e.g., rain, snow and fog). Each scene element is assigned precise physical properties and geometric characteristics. Next, based on this static foundation, dynamic behavior simulation is applied to pedestrians, vehicles, and other dynamic objects. This stage leverages intelligent algorithms for path planning, obstacle avoidance, and other complex behaviors, utilizing real-time communication mechanisms to achieve multi-agent interactive coordination. Finally, a programmable virtual camera system simulates real-world shooting processes to render and generate image data. This system supports multi-angle and multi-parameter controllable image capture and incorporates advanced rendering techniques such as ray tracing to produce high-fidelity image datasets with comprehensive annotation information.

Artificial scenes within the parallel images framework offer

significant advantages in terms of data diversity, annotation accuracy, and scene coverage. They effectively address key challenges in real-world scenarios, such as high data acquisition costs, low annotation efficiency, and insufficient coverage of long-tail scenarios, providing a reliable data foundation for training and validating computer vision models.

B. Computational Experiment Simulation

Computational experimental simulation is the core component of the parallel images framework. It primarily focuses on conducting large-scale computational experiments using the generated synthetic scene data to test, optimize, and validate the performance of visual algorithms.

Artificially generated virtual data can effectively simulate various extreme and long-tail scenarios that are challenging to collect on a large scale in real-world environments, providing diverse testing conditions for robust visual model validation. By integrating virtual data with real-world data, the generalization performance of visual models can be significantly enhanced, leveraging the broad coverage of virtual data and the distributional authenticity of real-world data to improve model adaptability across diverse scenarios. Given the distributional differences between artificial scene data and real-world data, domain adaptation techniques are employed. Models first learn general feature representations from extensive virtual data and then undergo targeted fine-tuning using real-world data, thereby substantially enhancing their adaptability to practical application scenarios.

Selecting appropriate visual models for training is crucial to meet specific application requirements and research objectives. In virtual traffic scenarios, training autonomous driving perception models can effectively address varying traffic densities and pedestrian behaviors by leveraging their capabilities in feature extraction and object localization. For image classification tasks, architectures such as vision transformer [38] and EfficientNet [39] demonstrate strong transfer learning performance in synthetic data training because of their self-attention mechanisms and superior parameter efficiency. By utilizing artificial scene data, these models can more accurately identify image categories. In semantic and instance segmentation tasks, training segmentation models using synthetic building scene data can significantly improve recognition accuracy for elements such as roads, buildings, and vegetation. For pose estimation and behavior analysis, training action recognition models in virtual human motion scenarios can enhance predictive accuracy for complex postures. Throughout the training process, model parameters are continuously adjusted to enable the models to learn key features and patterns from image data, thereby optimizing their performance for specific tasks.

Evaluating model performance across diverse datasets is essential for verifying a model's generalization capability and adaptability to various real-world scenarios. In addition to generalization assessment, computational experiments serve as a robust tool for model comparison and selection. By training and evaluating multiple visual models on a standardized dataset, one can systematically compare performance metrics and identify the most suitable model for practical deployment. Furthermore, computational experiments enable an in-depth analysis of how variations in model architectures and training parameters affect performance. For instance, by adjusting factors such as the number of neural network layers or the learning rate, researchers can observe corresponding performance changes, thereby gaining valuable insights for model optimization and identifying the optimal configuration for specific tasks [6, 40].

C. Virtual-Real Parallel Execution

Virtual-real parallel execution, a pivotal mechanism within the parallel images framework, is essential for realizing its closed-loop optimization structure. This mechanism systematically refines virtual scene models based on explicit feedback from real-world deployments, concurrently leveraging these optimized virtual models to enhance real-world visual models. Such dual-path iterative optimization significantly enhances the effectiveness of intelligent perception and interpretation of complex environments.

The iterative optimization process unfolds through a clearly defined feedback loop. Visual models are initially deployed in real-world scenarios for practical validation, collecting comprehensive performance data that capture discrepancies between predictions and actual environmental observations, alongside variations in model performance across diverse operational contexts. This feedback is systematically processed to inform and optimize virtual scene models. For example, in autonomous driving scenes, in which visual models exhibit elevated error rates under adverse conditions such as foggy weather, real-world observations initiate the creation of an equivalent foggy environment within the virtual space. Critical parameters such as atmospheric scattering coefficients, illumination intensities, and visibility distances, are meticulously adjusted to ensure high fidelity between virtual conditions and real-world environmental characteristics. Dynamic objects and their behaviors within these virtual environments are also calibrated, including adjustments to vehicle speed distributions under compromised visibility and refinement of fog light activation rules to synchronize virtual target behaviors with the observed real-world data.

The refined virtual model, developed through precise real-world feedback, subsequently informs real-world visual systems. Artificial scenes, benefiting from inherent flexibility and replicability, enable the seamless transformation of optimized parameters and refined detection methodologies from virtual environments back to real-world implementations. This iterative and bi-directional feedback mechanism fosters continuous reciprocal enhancement between virtual and real-world models. By repeatedly deploying real-world visual models, collecting precise performance feedback, accordingly refining virtual scenes, and reapplying these refinements to real-world contexts, the visual perception system progressively optimizes. This mechanism ensures highly adaptive, accurate, and reliable visual perception capabilities, directly benefiting complex and practical applications.

IV. KEY TECHNOLOGY OF PARALLEL IMAGES

A. Virtual Scene Generation Driven by Multi-Modal Data

(1) Editable content generation in virtual scenes. Editable content generation in virtual scenes leverages synthetic generation algorithms based on deep learning and virtual reality to construct virtual environments that enable users to perform content editing operations using input data such as text, images, and audio.

In parallel images, software-defined artificial image systems generate extensive synthetic image data based on real-world "small data", thereby constructing a comprehensive big data repository for parallel images. Zhang et al. [29] proposed a parallel vision approach for scene-specific pedestrian detection, addressing the challenges of data scarcity and environmental variability by first pre-training the model on augmented-reality data and then incrementally optimizing it with newly synthesized data as the scene evolves. Similarly, Li and Wang [41] applied the ACP methodology to the field of visual perception for intelligent driving. Their work facilitates the construction of complex artificial driving scenes and the generation of large-scale annotated datasets under challenging imaging conditions, improving both the training and evaluation of vision models in dynamic and adverse scenarios. Tian et al. [31] further strengthened this capability through virtual-real interaction mechanisms for 3D point cloud generation, enhancing the fidelity and editability of object-level modeling within these environments. Sorscher et al. [42] addressed the limitations of traditional data scaling laws by introducing the data pruning techniques to optimize training datasets. Their method reduces redundancy while enhancing the relevance of training data, leading to improved neural scaling and reduced computational costs in content-editable environments.

Traditional 3D reconstruction and rendering methods primarily rely on geometric modeling and texture mapping, generating realistic 3D scenes by constructing detailed polygonal meshes and applying real-world textures. However, these methods face challenges such as data acquisition difficulties, high modeling complexity, and substantial rendering costs, especially when dealing with dynamic scenes or complex materials. To address these challenges, Mildenhall et al. [43] introduced neural radiance field, a method for synthesizing high-fidelity 3D reconstructions and view-consistent renderings from 2D images by representing scenes as continuous volumetric functions. NeRF effectively preserves geometric and visual consistency under transformations, making it essential for generating editable content with realistic appearance and structure. Based on this concept, Tian et al. [44] explored how learning visual features from models rather than data can impact the quality of generated scenes. Their insights provide valuable guidance for optimizing rendering system training strategies, striking a balance between realism and flexibility in scene manipulation.

The NeRF model represents 3D scenes as radiance fields, enabling continuous view rendering and detailed reconstruction. However, its generation process relies on dense viewpoint data and suffers from slow rendering speed, heading to real-time applications challenging. Additionally, NeRF's implicit representation modeling lacks direct support for semantic control, limiting scene editing and content manipulation capabilities. In contrast, diffusion models utilize a stepwise perturbation and restoration process to effectively capture data distributions, demonstrating strong generative capabilities while seamlessly integrating multi-modal information (e.g., text and images), thus enabling text-guided scene editing and generation. Rombach et al. [45] demonstrated the capabilities of latent diffusion models (LDMs) in generating high-resolution images that served as detailed textures and backgrounds in virtual environments, achieving state-of-the-art performance while significantly reducing computational costs compared with pixel-based diffusion models. Lee et al. [46] extended this concept to 3D by introducing a triplane-based diffusion model for synthesizing real-world outdoor scenes. Their approach encodes 3D scenes as compact triplane feature representations and leverages sinusoidal embeddings to predict semantic classes, enabling the generation of semantically rich and structurally coherent 3D scenes with high-quality rendering and detailed annotations. To further enhance scene compositional flexibility and support user-guided 3D generation, Zhou et al. [47] introduced generative 3D Gaussians with layout-guided control (GALA3D), a framework that leveraged large language models (LLMs) to generate initial layouts and employed layout-guided 3D Gaussian representations with adaptive geometric constraints. The framework integrates conditioned diffusion for instance-scene optimization, enabling realistic scene generation with consistent geometry, texture, and object interactions while maintaining high-fidelity object-level details.

To address the challenge of generating realistic and diverse driving scenes with consistent geometry and semantics, Li et al. [48] introduced a unified framework named UniScene. This method leverages an occupancy-centric representation to bridge 3D geometry and semantics. It jointly learns multi-object layout, scene-level occupancy, and fine-grained geometry generation using a transformer-based architecture. By unifying scene understanding and generation, UniScene enables controllable, structured, and coherent driving scene synthesis from a single latent code. To enable realistic and controllable street view synthesis, Yan et al. [49] proposed StreetCrafter, a video diffusion framework tailored for urban scenes. It incorporates a multi-scale layout-to-video generation pipeline, which is guided by semantic and instance layouts. The model introduces a controllable generation module that allows editing and manipulation of scene elements across time. By leveraging spatial-temporal consistency and layout conditioning, StreetCrafter achieves high-quality and editable street view video synthesis.

(2) Style attribute transformation of virtual scenes. The style attribute transformation of virtual scenes refers to the application of various techniques to alter the visual or artistic style of a scene, transitioning it from one stylistic state to another.

Early methods based on convolutional neural networks (CNNs) utilize multi-layer convolutional operations to

effectively capture texture, color, and other fine-grained features in images. Gatys et al. [50] proposed a CNN-based image style transfer method that utilized a pre-trained VGG-19 network to extract feature representations from both content and style images. The content features are represented by the activations in higher layers, capturing the overall structural information of the image, while the style features are represented by the Gram matrices of feature maps across multiple layers, capturing the statistical properties of texture and color distributions. During the optimization process, a randomly initialized noise image is iteratively updated by minimizing the weighted sum of content loss and style loss, resulting in a synthesized image that preserves the structure of the content image while adopting the artistic style of the style image. This method effectively separates and recombines content and style, producing artistic-style images with high perceptual quality.

Generative adversarial networks (GANs) are also widely applied in image style transfer tasks. A GAN consists of a generator and a discriminator, both of which are trained adversarially to improve performance. In style transfer applications, the generator is responsible for producing images with a specific style, while the discriminator assesses whether the generated images align with the target style and appear realistic. This adversarial mechanism enables GANs to generate high-quality stylized images. Zhu et al. [51] proposed a method for unpaired image-to-image translation that learnt a mapping between a source domain X and a target domain Y without requiring paired training data. By incorporating adversarial loss to align distributions and cycle consistency loss to preserve content structure, the method enables tasks such as style transfer, object transfiguration, and photo enhancement in the absence of paired examples. Recently, Wang et al. [52] proposed a single sample-based traffic generative adversarial network (SST-GAN), a method for generating realistic images of scarce driving scenes based on a single sample, utilizing style transfer through transition retraining and content generation guided by a structural similarity index loss. The method effectively expands rare scene datasets for deep learning-based vision algorithms, enhancing their adaptability to extreme weather and traffic conditions.

Controllable and semantically guided style transformation methods achieve visual consistency and flexible control over details such as lighting, material, and texture by explicitly separating geometric structure from style attributes. Compared with the traditional methods, they effectively reduce data requirements while enhancing the editability and scalability of the generated content. Recent advancements have increasingly focused on controllable and semantically guided style transformations in virtual environments. Zhang et al. [53] introduced reference-based non-photorealistic radiance field (Ref-NPR), a controllable framework that applied a single stylized reference view to guide the stylization of 3D scenes. By decoupling scene geometry from style attributes such as color and brushstroke patterns through radiance field modeling, the framework enables targeted and high-fidelity artistic rendering of virtual scenes. Zhang et al. [54] proposed a novel approach for synthesizing realistic 3D natural scenes from a single semantic mask by leveraging generative models and view-dependent rendering. Their framework utilizes a semantic field as an intermediate representation, enabling precise control over style attributes such as lighting, material, and texture, thereby enhancing the realism and consistency of generated scenes. Additionally, in the process of virtual scene construction and style transformation, 3D reconstruction techniques extract geometric structure information from multi-view images or depth data to reconstruct the three-dimensional shape and texture mapping of a scene, laying the foundation for subsequent stylization operations. Höllein et al. [55] proposed a style transfer method specifically designed for indoor 3D reconstructed scenes. The method achieves view-independent and style-consistent 3D stylization by optimizing the explicit textures of the reconstructed meshes. Specifically, depth-aware and angle-aware optimization strategies are employed, incorporating surface normals and depth information to maintain texture consistency across the entire scene. Furthermore, multi-view image optimization is applied to prevent stretching and distortion issues commonly seen in traditional 2D style transfer when adapted to 3D scenes.

To address the limitations of low-fidelity and open-loop simulation in autonomous driving, Yan et al. [56] proposed DrivingSphere, a 4D simulation framework with realistic sensor modeling and dynamic interactions. It reconstructs a high-fidelity 4D environment by integrating multi-agent trajectories, dynamic objects, and sensor data over time. The framework enables closed-loop simulation with perception-action feedback, supporting both visual realism and behavioral accuracy for training and evaluating autonomous driving systems. To overcome the inefficiency of standard diffusion models in real-time driving tasks, Liao et al. [57] proposed DiffusionDrive, a novel end-to-end driving framework. It introduces a truncated denoising diffusion process to generate future driving actions efficiently. By learning from expert demonstrations, the model predicts trajectories in a coarse-to-fine manner, significantly reducing the inference steps. DiffusionDrive balances generation quality and efficiency, enabling accurate and responsive autonomous driving in complex urban environments.

B. Multi-View Feature Fusion and Virtual-Real Domain Transfer

(1) Multi-task feature fusion and semantic information interaction. The aim of multi-task feature fusion and semantic information interaction is to efficiently integrate features from parallel images captured from different viewpoints, conditions, or time points, while facilitating semantic information exchange among multiple related tasks.

Recently, research in multi-task learning has increasingly emphasized cross-task and cross-modal feature fusion, focusing on leveraging semantic interactions to boost overall performance and robustness. Liu et al. [58] proposed SegMiF, a multi-interactive feature learning architecture designed for image fusion and segmentation. The method leverages dual-task correlation to enhance the performance of both tasks. SegMiF introduces a hierarchical interactive attention module and a dynamic weighting factor, effectively balancing feature

interactions between the fusion and segmentation networks to generate visually appealing fused images. Similarly, Wang et al. [59] proposed a multiple enhancement network (MENet) for salient object detection, which was designed to simulate human perception by iteratively aggregating multi-scale boundary and region features through a dual-branch decoder. The framework is guided by a multi-level hybrid loss function, effectively enhancing segmentation accuracy in complex scenes.

References [60, 61] focus on enhancing specific components within the multi-task visual processing pipeline, particularly in the domains of semantic encoding and decoding. Huang et al. [60] proposed a reinforcement learning-based semantic bit allocation framework, where semantic concepts, defined by class, spatial, and visual attributes, were encoded via a convolutional semantic encoder. The bit allocation strategy is dynamically optimized through reinforcement signals linked to semantic task performance. For decoding, a GAN-based semantic reconstruction module integrates both local and global contextual features via attention mechanisms, achieving high-fidelity semantic restoration. Wan et al. [61] addressed human-object interaction detection by introducing a multi-level semantic recognition model. Their approach employs attention mechanisms to dynamically emphasize human pose-related regions, thereby enhancing the model's capacity for fine-grained interaction understanding. Transformer-based architectures leverage efficient semantic feature selection and modeling. Chen et al. [62] proposed a local attention transformer that performed adaptive down-sampling by learning to retain task-relevant pixels. Their model introduces a point-based attention block supported by balanced clustering and learnable neighborhood merging, yielding efficient representations for pixel-level segmentation tasks.

In the context of large multi-modal models, unified semantic interaction across tasks and domains has become a critical research direction. Zhang et al. [63] proposed OMG-LLaVA, a unified framework that integrated pixel-level visual understanding with large language models. By aligning visual inputs, perception priors, and visual prompts with text instructions, OMG-LLaVA enables LLMs to generate both language and segmentation outputs. The framework employs a unified token generation mechanism, incorporating text, pixel, and object tokens to flexibly handle diverse vision-language tasks. Xin et al. [64] introduced multi-modal alignment prompt (MmAP), a method designed for cross-domain multi-task learning by aligning text and visual modalities during the fine-tuning process. The framework leverages task grouping and task-specific prompts to facilitate efficient semantic transfer across tasks while maintaining modality-specific feature alignment and achieving significant performance gains with minimal parameter tuning.

To bridge the gap between vision-language understanding and autonomous driving, Wang et al. [65] proposed OmniDrive, a comprehensive dataset and framework. It incorporates multi-modal data, images, driving states, and natural language descriptions, augmented with counterfactual reasoning scenarios. The method enables models to learn causal relationships by comparing real cases with counterfactual cases. OmniDrive supports tasks such as action prediction, reasoning, and instruction following, fostering deeper semantic understanding for decision-making in driving environments. To model complex spatial-temporal dynamics in autonomous driving, Zhao et al. [66] proposed DriveDreamer4D, a world model-based framework for 4D scene understanding. It leverages latent dynamics modeling to learn from sequential multi-view and multi-agent data. The method encodes past observations into a compact latent space and predicts future scenes via a learned dynamics prior. DriveDreamer4D enables efficient, scalable, and predictive representation of driving environments for planning and simulation.

(2) Multi-modal collaborative optimization for virtual-real domain adaptation. Multi-modal cooperative optimization in virtual-real domain adaptation focuses on addressing how to effectively process and optimize parallel data from real-world scenes and data generated from virtual scene construction. The objective is to enable models to effectively adapt to the discrepancies between virtual and real domains, encompassing differences in image features, scene semantics, and other aspects.

To address the challenges of domain discrepancies between virtual and real environments, recent studies have explored various strategies for multi-modal collaborative optimization in domain adaptation. Zhang et al. [67] proposed a coarse-to-fine domain adaptation framework designed for traffic object detection. The method initially applies a coarse alignment module to mitigate domain discrepancies at a global feature level, followed by a fine-grained adaptation module that refines region-level features through adaptive feature calibration, enhancing detection accuracy under domain shifts. Similarly, Zhou et al. [68] proposed a multi-granularity alignment domain adaptation framework for object detection, which aligned features at three levels: image, instance, and pixel. By integrating global feature alignment, instance-level feature refinement, and pixel-wise adaptation, the method effectively reduces domain discrepancies and improves detection robustness in cross-domain scenarios. Lu et al. [69] proposed MLNet, a mutual learning network with neighborhood invariance for universal domain adaptation. The framework employs two collaborative networks to learn complementary domain-invariant features while enforcing neighborhood consistency across source and target samples, thereby enhancing feature alignment and improving adaptation performance under diverse domain shifts.

Beyond domain adaptation in object detection, several works have addressed related issues in multi-modal and cross-domain perception tasks, such as aligning heterogeneous data representations and ensuring robust performance across varying environmental conditions and sensor configurations. Wang et al. [70] proposed a parallel vision framework aimed at addressing long-tail regularization in autonomous driving scenarios by leveraging the IVFC testing environment. The method integrates virtual and real-world visual data to mitigate data imbalance, employing a two-stage training process that includes feature re-weighting for tail classes and

domain adaptation to align synthetic and real data distributions, effectively enhancing detection accuracy for rare objects and challenging scenarios. Gao et al. [71] proposed a comprehensive benchmark and a novel model for light field saliency detection, addressing the challenges of extracting depth and focusing cues effectively. The method leverages a dual-branch architecture, consisting of a spatial branch for foreground-background segmentation and a depth branch for salient region refinement.

From the perspective of remote sensing, Huang [72] proposed a framework for efficient remote sensing that combined harmonized transfer learning and modality alignment to address cross-domain discrepancies in multi-modal remote sensing data. The method employs a dual-stream network to separately process optical and SAR data, followed by a modality alignment module that bridges the feature gaps through adversarial learning. Additionally, a transfer learning strategy is utilized to adapt pre-trained models to the target domain, enhancing classification accuracy while minimizing computational overhead. In the domain of image fusion, Bai et al. [73] proposed a learning-based image fusion framework that leveraged reconstruction with a learnable loss function via meta-learning. The method employs a meta-learning strategy to optimize the fusion network by dynamically adjusting the loss function, thereby enhancing the fusion quality and preserving essential information from source images across diverse scenarios.

C. Parallel Execution with Heterogeneous Data and Knowledge Fusion

(1) Structured information extraction and representation from heterogeneous data. In the parallel images domain, the extraction and representation of structured information from heterogeneous data serve as crucial components for achieving virtual-real interaction and scene modeling. Different types of data, such as point clouds, RGB images, depth images, and semantic labels, exhibit significant variations in data structure, resolution, and noise characteristics. Effectively integrating these diverse data modalities and constructing a unified representation framework are essential for comprehensive scene perception and accurate modeling of real-world environments.

To address the challenges of extracting structured information from heterogeneous data sources, recent studies have proposed a range of methods that emphasize multi-modal fusion, cross-domain alignment, and efficient feature representation. Fan et al. [74] proposed a spatial contextual feature learning framework that designed for large-scale point cloud segmentation. It leverages a multi-scale contextual aggregation module to capture spatial dependencies across varying scales, while a context-aware attention mechanism refines feature representations by emphasizing critical spatial regions, thereby enhancing segmentation accuracy in the complex scenes. Lu et al. [69] proposed a mutual learning network designed to address universal domain adaptation by incorporating neighborhood invariance. The method employs a dual-branch architecture, in which one branch focuses on domain-invariant feature extraction while the other emphasizes neighborhood consistency. The network jointly learns mutual representations, enabling robust adaptation across diverse domain shifts. Hemker et al. [75] proposed HEALNet, a hybrid early-fusion attention learning network designed for multi-modal biomedical data integration. HEALNet combines modality-specific and shared parameter spaces within an iterative attention-based architecture, enabling the model to preserve structural information unique to each modality while capturing cross-modal interactions in a shared latent space. This design allows effective handling of missing modalities during both training and inference and facilitates model interpretability by operating directly on raw data inputs.

High-resolution remote sensing images often suffer from spectral distortions during the fusion process because of discrepancies between spectral and spatial resolutions. To address this issue, Li et al. [76] proposed an image fusion method based on image segmentation to reduce spectral distortions in the high-resolution remotely-sensed imagery. By segmenting the panchromatic (PAN) image, the method identifies mixed pixels (MPs) near object boundaries. These MPs are then fused using the spectral information of pure pixels within the same segment, enhancing spectral fidelity and spatial detail. This approach effectively sharpens object boundaries and improves the overall quality of the fused image. Image manipulation localization is crucial for detecting forgeries and preserving image integrity. However, existing methods often rely on pre-trained networks, which may introduce biases and limit generalization to unseen manipulation types. Zhou et al. [77] proposed a pre-training-free framework for image manipulation localization using non-mutually exclusive contrastive learning. This method leverages contrastive loss to learn discriminative features between manipulated and authentic regions without relying on pre-trained networks, enabling robust detection of localized image alterations across diverse manipulation types. In the domain of unsupervised visual understanding, Li et al. [78] proposed text-aided clustering (TAC), an externally guided image clustering method that integrated textual semantics from WordNet. TAC constructs a text space by selecting discriminative nouns that distinguish image semantic centers. Each image is then paired with a retrieved noun, forming a text counterpart. To enhance clustering performance, TAC employs cross-modal mutual distillation, aligning neighborhood structures between image and text modalities. This approach effectively leverages external knowledge to improve clustering accuracy across various benchmarks.

Unregistered infrared-visible image fusion remains a challenging task because of the inherent modality discrepancies and lack of accurate alignment between the infrared images and visible images. Traditional methods often treat registration and fusion as separate stages, leading to cumulative errors and suboptimal fusion results. Li et al. [79] proposed a novel framework for unregistered infrared-visible image fusion that integrated implicit registration and fusion into a single stage. It employs a shared shallow feature encoder and a learnable modality dictionary to align cross-modal features, enhancing consistency and reducing modality

discrepancies. Additionally, a correlation matrix captures pixel relationships between modalities, facilitating effective feature alignment and improving fusion quality. To address the challenge of effective multi-agent sensor fusion under noisy radar signals, Huang et al. [80] proposed V2X-R, which was a cooperative perception framework. It introduces a denoising diffusion model to fuse LiDAR and 4D radar data across vehicles. The model learns to denoise and align radar features with LiDAR representations, enabling robust and complementary multi-modal feature fusion for accurate 3D object detection in V2X scenarios. To tackle the challenge of reconstructing dynamic driving scenes from incomplete or noisy observations, Ni et al. [81] proposed ReconDreamer, which was a world model framework with online restoration capability. It integrates a latent dynamics model with a restoration module that progressively refines incomplete inputs during rollout. By jointly modeling scene dynamics and performing online correction, ReconDreamer enables accurate 4D reconstruction of driving environments, enhancing perception and planning in real-world autonomous driving scenarios.

(2) Vision perception driven by multi-source data fusion.

In the domain of vision perception, the fusion of multi-source data has emerged as a critical technique to address the limitations of single-modality inputs and enhance the model robustness, especially in complex real-world scenarios characterized by long-tail data distributions. Researchers have proposed various strategies to exploit the complementary strengths of heterogeneous sensors and data modalities, thereby improving perceptual accuracy and generalization [70, 82]. Wang et al. [70] proposed a framework named long-tail regularization (LoTR) to address the challenges posed by rare scenarios in autonomous driving perception systems. They introduce the parallel vision actualization system (PVAS), which employs closed-loop optimization and virtual-real interaction to generate and test large-scale long-tail driving scenarios. Implemented within the intelligent vehicle future challenge (IVFC), PVAS effectively mitigates long-tail effects, enhancing the robustness of autonomous vision systems in complex environments. Zhao et al. [82] proposed fusion via vision-language model (FILM), a novel image fusion framework that integrated textual semantics into the fusion process. FILM generates semantic prompts from input images using techniques like image captioning and dense captioning, which are then processed by ChatGPT to produce detailed textual descriptions. These descriptions are encoded using a frozen BLIP2 model and fused to guide the extraction and fusion of visual features through cross-attention mechanisms. This approach enhances contextual understanding and achieves superior results across various fusion tasks, including infrared-visible, medical, multi-exposure, and multi-focus image fusion.

Reliable eye-state recognition, precise eye-center localization, and accurate gaze estimation are fundamental for applications such as driver-monitoring systems, AR/VR interaction, and human-computer interfaces. Most prior work tackles these three subtasks in isolation, causing error accumulation and limiting robustness when facing exhibit occlusions, large head poses, or diverse eye appearances. Motivated by the need for a unified and resilient solution, Zhu et al. [83] proposed a joint cascaded regression framework for simultaneous eye state, eye center, and gaze estimation. This method iteratively refines eye landmark positions and openness probabilities using shape and appearance features. By modeling the eye state as a continuous openness probability, the framework enhances robustness to occlusions and varying eye appearances. Additionally, it employs a learning-by-synthesis strategy and combines real and synthetic data to improve training efficiency and accuracy. Song et al. [84] proposed a lightweight high-definition mapping method based on multi-source data fusion perception for autonomous driving. This approach constructs local semantic maps (LSMs) by integrating data from multiple onboard sensors. These LSMs are then uploaded to a cloud server, where multiple maps of the same road section and collected through crowdsourcing are fused to generate high-definition maps. The method employs an improved two-stage semantic alignment algorithm for multi-trajectory pose optimization, enhancing mapping efficiency and accuracy while reducing costs and update latency.

The integration of diverse image modalities has also become a focal point of research. Liu et al. [58] proposed a multi-interactive feature learning framework for image fusion and segmentation. This cascaded architecture integrates a fusion sub-network and a segmentation sub-network, facilitating mutual enhancement between tasks. A hierarchical interactive attention block ensures fine-grained feature mapping, while a dynamic weighting factor automatically balances task contributions. Additionally, they introduced the FMB benchmark, a full-time multi-modality dataset with 15 annotated categories, to support comprehensive evaluation. Building on this foundation, Luo et al. [85] proposed hierarchical attention and parallel filter fusion network for multi-source data classification, particularly focusing on hyperspectral image (HSI) and synthetic aperture radar (SAR) data. The hierarchical attention module integrates global, spectral, and local features to provide comprehensive feature representations. Additionally, the parallel filter fusion module enhances cross-modal feature interactions by operating in the frequency domain, facilitating the effective fusion of multi-source data. This approach addresses the challenges of exploiting abundant features simultaneously, leading to improved classification performance on multi-source remote sensing datasets.

Another development focuses on the challenge of image misalignment. Li H et al. [79] propose a unified framework for unregistered infrared-visible image fusion that integrates cross-modality alignment and fusion in a single stage. It employs a shared shallow feature encoder and a learnable modality dictionary to align cross-modal features, enhancing consistency and reducing modality discrepancies. Additionally, a correlation matrix captures pixel relationships between modalities, facilitating effective feature alignment and improving fusion quality.

V. Integration Trend between Parallel Images and Generative Foundation Model

A. Background and Motivation

Recent advancements in generative artificial intelligence and large-scale multi-modal foundation models have significantly reshaped the landscape of visual computing. Techniques such as diffusion models [86], GANs [87], and neural radiance fields [88] have achieved notable progress in image quality, semantic consistency, and multi-modal intelligent perception capabilities. Concurrently, models such as CLIP [9], DALL·E [89], SAM [10], and GPT-4V [11] have demonstrated strong abilities in cross-modal alignment, task transfer, and general semantic modeling, propelling visual systems toward general intelligent perception.

Despite these advancements, foundational models face several limitations in specific task scenarios: (1) lack of physical controllability and realism in generated scenes; (2) opacity in training data sources, leading to generalization and robustness issues in high-security-sensitive domains; and (3) high degrees of model opacity, lacking rapid adaptation mechanisms based on real-world feedback.

B. Fusion Pathway

To address these challenges, integrating parallel images technology with generative AI presents a promising approach. This integration primarily manifests in the following aspects.

(1) Generative models for virtual scene construction. Utilizing generative models to drive virtual scene content generation enhances image semantics and style control. By incorporating text-driven diffusion models and pre-trained language-vision models (e.g., CLIP-guided NeRF [90]), it becomes feasible to generate high-fidelity images, videos, and even 3D scenes from natural language, high-level semantic maps, and behavioral descriptions. This capability significantly enriches the diversity and customizability of artificial scenes in parallel images.

(2) Structured synthetic data for model fine-tuning and evaluation. Parallel images technology can produce structured, multi-modal, and well-annotated synthetic datasets to support fine-tuning and evaluation of large models. Given that current large models often exhibit instability in specific tasks (e.g., small object detection and extreme weather perception). Parallel images can synthesize high-quality training data covering rare scenarios and multi-modal sensor views. Combined with virtual-real interaction mechanisms, this data can be used for targeted reinforcement training and pre-deployment simulation assessments of large models.

(3) Real-world feedback for adaptive learning. Establishing a feedback learning mechanism based on large models enables cross-scene adaptation and evolutionary optimization. Through parallel execution mechanisms, performance data collected from large models deployed in real environments can serve as feedback signals to the virtual world. This process facilitates scene reconstruction, parameter fine-tuning, and knowledge inversion based on generative models, forming a closed loop of "generation-simulation-feedback-optimization". Consequently, visual systems can possess real-time evolution and autonomous adaptation capabilities.

C. Future Outlook

The integration of parallel images technology with generative AI and large-scale multi-modal models not only enhances task specificity and controllability in data generation, but also addresses semantic blind spots in model training. Moreover, it establishes a closed-loop mechanism supporting continuous learning and feedback optimization. As multi-modal models with reasoning capabilities continue to mature, parallel images are poised to transition from "perception assistance" to "knowledge-driven" cognitive intelligence. It becomes a critical supporting technology for constructing general-purpose visual systems.

VI. Conclusion

As a novel image generation and modeling framework developed from the theory of parallel systems, parallel images constructs artificial scene environments, conducts computational experiments, and enable parallel execution between virtual and real spaces. Through this process, it forms a closed-loop mechanism of "modeling-training-feedback-optimization", providing high-quality, diverse, and structured synthetic data to support visual perception systems. This paper systematically reviews the theoretical foundations and developmental path of parallel images, with a particular focus on recent research progress and practical applications in key areas such as virtual scene generation, virtual-real domain adaptation, and heterogeneous knowledge-driven parallel execution.

Acknowledgment

This work was supported by the Natural Science Foundation of China (Nos. 62203040 and 62472048) and the Beijing Natural Science Foundation (No. L242081).

References

[1] G. Luo, C. Shao, N. Cheng, H. Zhou, H. Zhang, Q. Yuan, and J. Li, EdgeCooper: Network-aware cooperative LiDAR perception for enhanced vehicular awareness, *IEEE J. Select. Areas Commun.*, 2024, 42(1), 207–222.

[2] G. Luo, H. Zhang, Q. Yuan, and J. Li, Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception, in *Proc. 30th ACM International Conference on Multimedia*, Lisboa, Portugal, 2022, 3578–3586.

[3] G. Luo, H. Zhou, N. Cheng, Q. Yuan, J. Li, F. Yang, and X. Shen, Software-defined cooperative data sharing in edge computing assisted 5G-VANET, *IEEE Trans. Mob. Comput.*, 2021, 20(3), 1212–1229.

[4] G. Luo, Q. Yuan, J. Li, S. Wang, and F. Yang, Artificial intelligence powered mobile networks: From cognition to decision, *IEEE Network*, 2022, 36(3), 136–144.

[5] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F. -Y. Wang, Mask SSD: An effective single-stage approach to object instance segmentation, *IEEE Trans. Image Process.*, 2020, 29, 2078–2093.

[6] X. Xue, X. Yu, D. Zhou, X. Wang, C. Bi, S. Wang, and F.-Y. Wang, Computational experiments for complex social systems: Integrated design of experiment system, *IEEE/CAA J. Autom. Sin.*, 2024, 11(5), 1175–1189.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, *Commun. ACM*, 2020, 63(11), 139–144.

[8] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, Diffusion models in vision: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023, 45(9), 10850–10869.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., Learning transferable visual models from natural language supervision, in *Proc. 38th International Conference on Machine Learning*, 2021, 8748–8763.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., Segment anything, in *2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023, 3992–4003.

[11] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. Ma, Y. Li, L. Xu, D. Shang et al., On the road with GPT-4V (Ision): Early explorations of visual-language model on autonomous driving, arXiv preprint arXiv: 2311.05332, 2023.

[12] J. Yang, X. Wang, Y.-T. Wang, Z.-M. Liu, X.-S. Li, and F.-Y. Wang, Parallel intelligence and CPSS in 30 years: An ACP approach, *Acta Autom. Sin.*, 2023, 49(3), 614–634, (in Chinese).

[13] F.-Y. Wang, Parallel system methods for management and control of complex systems, *Control Decis.*, 2004, 19(5), 485–489, (in Chinese).

[14] F.-Y. Wang, X. Wang, L. Li, and L. Li, Steps toward parallel intelligence, *IEEE/CAA J. Autom. Sin.*, 2016, 3(4), 345–348.

[15] F.-Y. Wang, J. J. Zhang, and X. Wang, Parallel intelligence: Toward lifelong and eternal developmental AI and learning in cyber-physical-social spaces, *Front. Comput. Sci.*, 2018, 12(3), 401–405.

[16] F.-Y. Wang, Toward a paradigm shift in social computing: The ACP approach, *IEEE Intell. Syst.*, 2007, 22(5), 65–67.

[17] F.-Y. Wang, Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications, *IEEE Trans. Intell. Transport. Syst.*, 2010, 11(3), 630–638.

[18] F. Zhu, F.-Y. Wang, R. Li, Y. Lv, and S. Chen, Modeling and analyzing transportation systems based on ACP approach, in *Proc. 14th International IEEE Conference on Intelligent Transportation Systems*, Washington, DC, USA, 2011, 2136–2141.

[19] W. Duan, Z. Cao, Y. Wang, B. Zhu, D. Zeng, F.-Y. Wang, X. Qiu, H. Song, and Y. Wang, An ACP approach to public health emergency management: Using a campus outbreak of H1N1 influenza as a case study, *IEEE Trans. Syst. Man Cybern. Syst.*, 2013, 43(5), 1028–1041.

[20] F.-Y. Wang, L.-Q. Yang, J. Yang, Y. Zhang, S. Han, and K. Zhao, Urban intelligent parking system based on the parallel theory, in *Proc. International Conference on Computing, Networking and Communications*, Kauai, HI, USA, 2016, 1–5.

[21] K.-F. Wang, C. Gou, and F.-Y. Wang, Parallel vision: An ACP-based approach to intelligent vision computing, *Acta Autom. Sin.*, 2016, 42(10), 1490–1500, (in Chinese).

[22] K. Wang, Y. Lu, Y. Wang, Z. Xiong, and F.-Y. Wang, Parallel imaging: A new theoretical framework for image generation, *Pattern Recognit. Artif. Intell.*, 2017, 30(7), 577–587, (in Chinese).

[23] L. Li, Y.-L. Lin, D.-P. Cao, N.-N. Zheng, and F.-Y. Wang, Parallel learning—A new framework for machine learning, *Acta Autom. Sin.*, 2017, 43(1), 1–8, (in Chinese).

[24] Y. Lv, Y. Chen, J. Jin, Z. Li, P. Ye, and F. Zhu, Parallel transportation: Virtual-real interaction for intelligent traffic management and control, *Chin. J. Intell. Sci. Technol.*, 2019, 1(1), 21–33, (in Chinese).

[25] F.-Y. Wang, Parallel medicine: From warmness of medicare to medicine of smartness, *Chin. J. Intell. Sci. Technol.*, 2021, 3(1), 1–9, (in Chinese).

[26] H. Zhang, X. Li, and F. Wang, The basic framework and key algorithms of parallel vision, *J. Image Graphics*, 2021, 26(1), 82–92, (in Chinese).

[27] H. Zhang, G. Luo, Y. Li, and F.-Y. Wang, Parallel vision for intelligent transportation systems in metaverse: Challenges, solutions, and potential applications, *IEEE Trans. Syst. Man Cybern. Syst.*, 2022, 53(6), 3400–3413.

[28] L. Li, X. Wang, K. Wang, Y. Lin, J. Xin, L. Chen, L. Xu, B. Tian, Y. Ai, J. Wang et al., Parallel testing of vehicle intelligence via virtual-real interaction, *Sci. Robot.*, 2019, 4(28), eaaw4106.

[29] W. Zhang, K. Wang, Y. Liu, Y. Lu, and F.-Y. Wang, A parallel vision approach to scene-specific pedestrian detection, *Neurocomputing*, 2020, 394, 114–126.

[30] T. Shen, C. Gou, J. Wang, J. Huang, Y. He, H. Xue, Z. Jin, and F.-Y. Wang, Parallel medical imaging: An ACP-based approach for intelligent medical image recognition with small samples, in *Proc. IEEE 1st International Conference on Digital Twins and Parallel Intelligence*, Beijing, China, 2021, 226–229.

[31] Y.-L. Tian, Y. Shen, Q. Li, and F.-Y. Wang, Parallel point clouds: Point clouds generation and 3D model evolution via virtual-real interaction, *Acta Autom. Sin.*, 2020, 46(12), 2572–2582, (in Chinese).

[32] F.-Y. Wang, X. Meng, S. Du, and Z. Geng, Parallel light field: The framework and processes, *Chin. J. Intell. Sci. Technol.*, 2021, 3(1), 110–122, (in Chinese).

[33] Y. Liu, Y. Shen, L. Fan, Y. Tian, Y. Ai, B. Tian, Z. Liu, and F.-Y. Wang, Parallel radars: From digital twins to digital intelligence for smart radar systems, *Sensors*, 2022, 22(24), 9930.

[34] Y. Tian, X. Li, H. Zhang, C. Zhao, B. Li, X. Wang, and F.-Y. Wang, VistaGPT: Generative parallel transformers for vehicles with intelligent systems for transport automation, *IEEE Trans. Intell. Veh.*, 2023, 8(9), 4198–4207.

[35] H. Yu, X. Liu, Y. Tian, Y. Wang, C. Gou, and F.-Y. Wang, Sora-based parallel vision for smart sensing of intelligent vehicles: From foundation models to foundation intelligence, *IEEE Trans. Intell. Veh.*, 2024, 9(2), 3123–3126.

[36] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao, Text2NeRF: Text-driven 3D scene generation with neural radiance fields, *IEEE Trans. Vis. Comput. Graph.*, 2024, 30(12), 7749–7762.

[37] P. Dai, F. Tan, X. Yu, Y. Peng, Y. Zhang, and X. Qi, GO-NeRF: Generating objects in neural radiance fields for virtual reality content creation, *IEEE Trans. Vis. Comput. Graph.*, 2025, 31(5), 3087–3097.

[38] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu et al., A survey on vision transformer, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, 45(1), 87–110.

[39] M. Tan and Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in *Proc. 36th International Conference on Machine Learning*, Long Beach, CA, USA, 2019, 6105–6114.

[40] X. Xue, X.-N. Yu, D.-Y. Zhou, C. Peng, X. Wang, Z.-B. Zhou, and F.-Y. Wang, Computational experiments: Past, present and perspective, *Acta Autom. Sin.*, 2023, 49(2), 246–271, (in Chinese).

[41] X. Li and F. Wang, Parallel visual perception for intelligent driving: Basic concept, framework and application, *J. Image Graphics*, 2021, 26(1), 67–81, (in Chinese).

[42] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos, Beyond neural scaling laws: Beating power law scaling via data pruning, in *Proc. 36th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2022, 1419.

[43] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, *Commun. ACM*, 2022, 65(1), 99–106.

[44] Y. Tian, L. Fan, K. Chen, D. Katabi, D. Krishnan, and P. Isola, Learning vision from models rivals learning vision from data, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2024, 15887–15898*.

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, 10674–10685.

[46] J. Lee, S. Lee, C. Jo, W. Im, J. Seon, and S.-E. Yoon, SemCity: Semantic scene generation with triplane diffusion, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, 28337–28347.

[47] X. Zhou, X. Ran, Y. Xiong, J. He, Z. Lin, Y. Wang, D. Sun, and M.-H. Yang, GALA3D: Towards text-to-3D complex scene generation via layout-guided generative Gaussian splatting, in *Proc. 41st International Conference on Machine Learning*, Vienna, Austria, 2024, 2570.

[48] B. Li, J. Guo, H. Liu, Y. Zou, Y. Ding, X. Chen, H. Zhu, F. Tan, C. Zhang, T. Wang et al., UniScene: Unified occupancy-centric driving scene generation, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 11971–11981.

[49] Y. Yan, Z. Xu, H. Lin, H. Jin, H. Guo, Y. Wang, K. Zhan, X. Lang, H. Bao, X. Zhou et al., StreetCrafter: Street view synthesis with controllable video diffusion models, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 822–832.

[50] L. A. Gatys, A. S. Ecker, and M. Bethge, Image style transfer using convolutional neural networks, in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, 2414–2423.

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proc. 2017 IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 2242–2251.

[52] J. Wang, Y. Wang, Y. Tian, X. Wang, and F.-Y. Wang, SST-GAN: Single sample-based realistic traffic image generation for parallel vision, in *Proc. 2022 IEEE 25th International Conference on Intelligent Transportation Systems*, Macau, China, 2022, 1485–1490.

[53] Y. Zhang, Z. He, J. Xing, X. Yao, and J. Jia, Ref-NPR: Reference-based non-photorealistic radiance fields for controllable scene stylization, in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 4242–4251.

[54] S. Zhang, S. Peng, T. Chen, L. Mou, H. Lin, K. Yu, Y. Liao, and X. Zhou, Painting 3D nature in 2D: View synthesis of natural scenes from a single semantic mask, in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 8518–8528.

[55] L. Höllein, J. Johnson, and M. Nießner, StyleMesh: Style transfer for indoor 3D scene reconstructions, in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, 6198–6208.

[56] T. Yan, D. Wu, W. Han, J. Jiang, X. Zhou, K. Zhan, C.-Z. Xu, and J. Shen, DrivingSphere: Building a high-fidelity 4D world for closed-loop simulation, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 27531–27541.

[57] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang et al., DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 12037–12047.

[58] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation, in *Proc. 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023, 8081–8090.

[59] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, Pixels, regions, and objects: Multiple enhancement for salient object detection, in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 10031–10040.

[60] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, Toward semantic communications: Deep learning-based image semantic coding, *IEEE J. Sel. Areas Commun.*, 2023, 41(1), 55–71.

[61] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, Pose-aware multi-level feature network for human object interaction detection, in *Proc. 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), 2019, 9468–9477.

[62] Z. Chen, K. Patnaik, S. Zhai, A. Wan, Z. Ren, A. Schwing, A. Colburn, and F. Li, AutoFocusFormer: Image segmentation off the grid, in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 18227–18236.

[63] T. Zhang, X. Li, H. Fei, H. Yuan, S. Wu, S. Ji, C. C. Loy, and S. Yan, OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding, in *Proc. 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024, 2291.

[64] Y. Xin, J. Du, Q. Wang, K. Yan, and S. Ding, MmAP: Multi-modal alignment prompt for cross-domain multi-task learning, in *Proc. 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2024, 16076–16084.

[65] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 22442–22452.

[66] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang et al., DriveDreamer4D: World models are effective data machines for 4D driving scene representation, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 12015–12026.

[67] H. Zhang, G. Luo, J. Li, and F.-Y. Wang, C2FDA: Coarse-to-fine domain adaptation for traffic object detection, *IEEE Trans. Intell. Transp. Syst.*, 2022, 23(8), 12633–12647.

[68] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. Wu, Multi-granularity alignment domain adaptation for object detection, in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, 9571–9580.

[69] Y. Lu, M. Shen, A. J. Ma, X. Xie, and J.-H. Lai, MLNet: Mutual learning network with neighborhood invariance for universal domain adaptation, in *Proc. 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2024, 3900–3908.

[70] J. Wang, X. Wang, T. Shen, Y. Wang, L. Li, Y. Tian, H. Yu, L. Chen, J. Xin, X. Wu et al., Parallel vision for long-tail regularization: Initial results from IVFC autonomous driving testing, *IEEE Trans. Intell. Veh.*, 2022, 7(2), 286–299.

[71] W. Gao, S. Fan, G. Li, and W. Lin, A thorough benchmark and a new model for light field saliency detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023, 45(7), 8003–8019.

[72] T. Huang, Efficient remote sensing with harmonized transfer learning and modality alignment, arXiv preprint arXiv: 2404.18253, 2024.

[73] H. Bai, Z. Zhao, J. Zhang, Y. Wu, L. Deng, Y. Cui, B. Jiang, and S. Xu, Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning, *Int. J. Comput. Vis.*, 2025, 133(5), 2547–2567.

[74] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, and F.-Y. Wang, SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation, in *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, 14499–14508.

[75] K. Hemker, N. Simidjievski, and M. Jamnik, HEALNet: Multimodal fusion for heterogeneous biomedical data, in *Proc. 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024, 2057.

[76] H. Li, L. Jing, Y. Tang, and L. Wang, An image fusion method based on image segmentation for high-resolution remotely-sensed imagery, *Remote Sens.*, 2018, 10(5), 790.

[77] J. Zhou, X. Ma, X. Du, A. Y. Alhammadi, and W. Feng, Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning, in *Proc. 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023, 22289–2229.

[78] Y. Li, P. Hu, D. Peng, J. Lv, J. Fan, and X. Peng, Image clustering with external guidance, in *Proc. 41st International Conference on Machine Learning*, Vienna, Austria, 2024, 1117.

[79] H. Li, Z. Yang, Y. Zhang, W. Jia, Z. Yu, and Y. Liu, MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025, 47(5), 3673–3690.

[80] X. Huang, J. Wang, Q. Xia, S. Chen, B. Yang, X. Li, C. Wang, and C. Wen, V2X-R: Cooperative LiDAR-4D radar fusion with denoising diffusion for 3D object detection, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 27390–27400.

[81] C. Ni, G. Zhao, X. Wang, Z. Zhu, W. Qin, G. Huang, C. Liu, Y. Chen, Y. Wang, X. Zhang et al., ReconDreamer: Crafting world models for driving scene reconstruction via online restoration, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 1560–1569.

[82] Z. Zhao, L. Deng, H. Bai, Y. Cui, Z. Zhang, Y. Zhang, H. Qin, D. Chen, J. Zhang, P. Wang et al., Image fusion via vision-language model, in *Proc. 41st International Conference on Machine Learning*, Vienna, Austria, 2024, 60749–60765.

[83] J. Zhu, M. Li, Y. Yu, and C. Gou, A joint cascaded framework for simultaneous eye state, eye center, and gaze estimation, in *Proc. 2022 26th International Conference on Pattern Recognition*, Montreal, Canada, 2022, 770–776.

[84] H. Song, B. Hu, Q. Huang, Y. Zhang, and J. Song, A lightweight high definition mapping method based on multi-source data fusion perception, *Appl. Sci.*, 2023, 13(5), 3264.

[85] H. Luo, F. Gao, J. Dong, and L. Qi, Hierarchical attention and parallel

filter fusion network for multisource data classification, *IEEE Geosci. Remote Sens. Lett.*, 2024, 21, 5508905.

[86] L. Zhang, A. Rao, and M. Agrawala, Adding conditional control to text-to-image diffusion models, in *Proc. 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023, 3813–3824.

[87] R. Mehmood, R. Bashir, and K. J. Giri, Text conditioned generative adversarial networks generating images and videos: A critical review, *SN Comput. Sci.*, 2024, 5(7), 935.

[88] O. Gordon, O. Avrahami, and D. Lischinski, Blended-NeRF: Zero-shot object generation and blending in existing neural radiance fields, in *Proc. 2023 IEEE/CVF International Conference on Computer Vision Workshops*, Paris, France, 2023, 2933–2943.

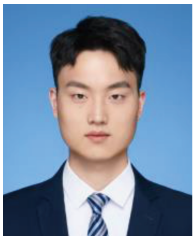
[89] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, Hierarchical text-conditional image generation with CLIP latents, arXiv preprint arXiv: 2204.06125, 1(2): 3, 2022.

[90] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, CLIP-NeRF: Text-and-image driven manipulation of neural radiance fields, in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, 3825–3834.

Hui Zhang received the PhD degree in control theory and control engineering from University of Chinese Academy of Sciences, China, in 2020. From August 2018 to October 2019, she was supported by University of Chinese Academy of Sciences as a joint-supervision PhD student at University of Rhode Island, USA. She is currently an associate professor at School of Computer Science and Technology, Beijing Jiaotong University, China. Her research interests include multi-agent collaboration, multimodal perception, and embodied intelligence.

Xiaofeng Jia is the director at Department of Data Management of Beijing Big Data Center, China, a professor-level senior engineer. His research interests include complex systems, federated intelligence, and deep relational learning.

Guiyang Luo is an associate professor at Computer Science Department, Beijing University of Posts and Telecommunications, China. He worked as a postdoctoral fellow at Computer Science Department, Beijing University of Posts and Telecommunications, China, from 2020 to 2022. His research interests include multi-agent systems and machine-type communications.



Yonglin Tian received the PhD degree in control science and engineering from University of Science and Technology of China, China, in 2022. He is currently an assistant researcher at State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China. His research interests include parallel intelligence, autonomous driving, and intelligent transportation systems.

Shutong Liang received the BS degree from Shanxi University, China, in 2023. She is currently pursuing the MS degree at School of Computer Science and Technology, Beijing Jiaotong University, China. Her research interests include multi-agent collaboration and embodied intelligence.

Dongyang Hong received the BS degree from Southwestern University of Finance and Economics in 2024. She is currently pursuing the MS degree at School of Computer Science and Technology, Beijing Jiaotong University, China. Her research interests include multi-agent collaboration and computer vision.