Dean&Francis
ISSN 2959-6157

Methods and Analysis of Accelerating Edge Computing with Photonic Neural Networks in the Context of Autonomous Driving

Yaojun Chen[1, *]

[1]Stony Brook Institute, Anhui University, Hefei, 230031, China
*Corresponding author: R12314086@stu.ahu.edu.cn

Abstract:

This paper reviews the latest research progress and key technologies in accelerating edge computing with optical neural networks, set against the backdrop of autonomous driving. It first outlines the context and research significance of autonomous driving, highlighting the technical limitations faced by traditional electronic computing. Subsequently, it provides a detailed analysis of the latest research achievements in optical neural networks and accelerated computing both domestically and internationally. This includes innovative applications of optical flow estimation methods (FocusFlow, CSFlow), introductions to methods for edge computing and autonomous driving collaboration (OVEAP, DVEAP), and innovative approaches to underlying computations (optical convolution, Monet channels). Through comparative analysis, it reveals current re-search hotspots and trends, analyzes existing short-comings and potential solutions, and explores possible future development directions. This paper aims to systematically explore how optical computing technology can overcome the limitations of traditional electronic computing to build a new generation of high-performance, low-power intelligent computing systems.

Keywords: Edge computing; optical neural networks; autonomous driving; optical flow estimation methods; underlying computations.

1. Introduction

Autonomous vehicles are intelligent cars that achieve driverless operation through computer systems. Their development spans decades and is currently at a critical stage for the implementation of Level 4/Level 5 applications [1]. Their implementation faces multiple technical challenges, including achieving <10ms ultra-low latency dependent on 5G V2X communication, iterative shadow mode operation within data closed-loop systems, and strict power consumption constraints for onboard computing hardware, which

ISSN 2959-6157

must be limited to under 50W [2].

The current limitations stem from inherent bottlenecks in traditional electronic architectures: the "memory wall" under the von Neumann architecture (where data movement consumes over 60% of pow-er); the "computing power wall"; the "power wall" (existing platforms achieve 5-10 TOPS/W energy efficiency; achieving L5-level 1000+ TOPS computing power requires over 500W power consumption, far exceeding automotive-grade thermal management capabilities). Furthermore, traditional GPUs in-efficiently process heterogeneous data streams like lidar [3], and suffer from massive computational redundancy by processing all sensor data rather than just effective information (which may constitute less than 10% of the data) [4].

Optical neural networks (ONNs) offer a break-through solution to the aforementioned challenges through their physical-layer parallelism, light-speed transmission, and ultra-low power consumption [5]. By employing wavelength division multiplexing (WDM) and Mach-Zehnder interferometer (MZI) arrays, ONNs fundamentally circumvent the "memory wall," reducing matrix operation latency to the nanosecond range while cutting power consumption to one-five-hundredth that of traditional GPUs [6]. Research demonstrates that ONNs can perform high-speed preprocessing and fusion of heterogeneous data such as lidar point clouds and millimeter-wave radar echoes, compressing latency from milliseconds to microseconds while reducing energy consumption by two orders of magnitude [7]. ONNs inherently align with neural network inference, as their high robustness to computational noise enables low-precision (e.g., 8-bit) analog op-tical computation [8].

However, its in-vehicle deployment still faces three core challenges: First, environmental sensitivity, particularly temperature drift, can degrade performance, necessitating the development of online calibration algorithms [9]. Second, the complexity of optoelectronic co-design, including the implementation of reconfigurable photoactivation functions and system integration. Third, system-level validation and commercialization bottle-necks, such as the lack of real-world scenario testing and scalable manufacturing processes [10].

This paper aims to systematically explore how optical computing technology can overcome the limitations of traditional electronic computing to build a new generation of high-performance, low-power intelligent computing systems. Key areas of focus include: 1) leveraging the compute-in-memory characteristics of optical computing to break through the "memory wall" and "power wall"; 2) utilizing the parallel nature of light to re-duce multi-sensor fusion latency to <1ms; 3) achieving nanosecond-level dynamic reconfiguration of computing architectures through programmable photonic devices; 4) Reducing multi-modal fusion error rates through wavelength encoding. Through in-depth research on ONNs architecture, energy efficiency, robustness, and integration challenges, this work provides a theoretical foundation and technical roadmap for constructing next-generation autonomous driving edge computing platforms.

2. Comparison and Analysis of Representative Technologies

2.1 optical flow estimation methods

2.1.1 FocusFlow Framework

Optical flow estimation is a classic problem in computer vision, aiming to estimate the motion relationships between consecutive video frames. It plays a crucial role in environmental perception for autonomous driving. FocusFlow is an innovative optical flow estimation framework whose core idea is to enhance the accuracy of local motion estimation through explicit modeling of keypoints, thereby improving visual processing performance in applications such as autonomous driving.

Figure 1 illustrates the core improvements of the FocusFlow framework. Figures 1(a) and 1(b) visualize the distribution of point features in the embedding space via PCA dimensionality reduction, where gray points and orange points represent random points and keypoints, respectively. In the traditional RAFT model, the Euclidean distance Lc be-tween the centroids of these two points feature classes is 4.58, indicating the model struggles to effectively encode keypoints into a unified feature space. FocusRAFT, integrated with the FocusFlow frame-work, reduces Lc to 2.58, significantly improving keypoint feature representation. Figures 1(c) and (d) compare real-world performance in autonomous driving scenarios from the KITTI dataset, demonstrating FocusRAFT's notable enhancements in keypoint matching and optical flow estimation. This validates the framework's ability to enhance local accuracy while maintaining global understanding through its conditional control mechanism.

Specifically, the FocusFlow framework comprises the following components:

(a) Traditional optical flow estimation methods treat each image point as an independent sample to estimate flow between image pairs. FocusFlow proposes a point-based, modeling approach that treats each point as an independent sample, and it learns its prior distribution for keypoints. By learning prior information about keypoints, FocusFlow can accurately captures local flow variations and

assigns greater weight to keypoints during optimization. In this way, it significantly enhances the flow estimation precision.

(b) FocusFlow introduces an innovation in loss function design by incorporating the Conditional Point Control Loss (CPCL). In contrast to conventional photometric loss functions that apply uniform weighting across all spatial points, CPCL employs a non-uniform weighting strategy. The system dynamically modulates the weight assigned to each point according to its Euclidean distance from identified keypoints. This distance-based weighting mechanism allows the model to prioritize key-point-oriented optical flow estimation, thereby substantially enhancing the precision of keypoint estimation processes.

Additionally, CPCL is integrated with conventional photometric loss terms to construct a hybrid loss function architecture. This combinatorial framework enables the model to maintain an optimal balance between keypoint-specific estimation accuracy and comprehensive frame-wide estimation performance throughout the optimization process.

(c) The Conditionally Controlled Encoder (CCE) framework comprises two fundamental components: a Frame Feature Encoder (FFE) and a Conditional Feature Encoder (CFE). The CFE module operates by receiving keypoint mask information as input signals. These mask signals serve to provide directional guidance to the FFE during the feature extraction process, specifically enabling the prioritization of keypoint-associated regions within the spatial domain.

This conditional control mechanism creates a flexible attention system, which allows the network architecture to adjust its focus dynamically. The network focuses on keypoint locations during computation. This attention strategy improves keypoint optical flow estimation accuracy. The proposed architecture works well with different optical flow network frameworks and this compatibility feature enables performance improvements. The architecture also maintains the original network design principles. The system preserves existing architectural characteristics while enhancing performance.

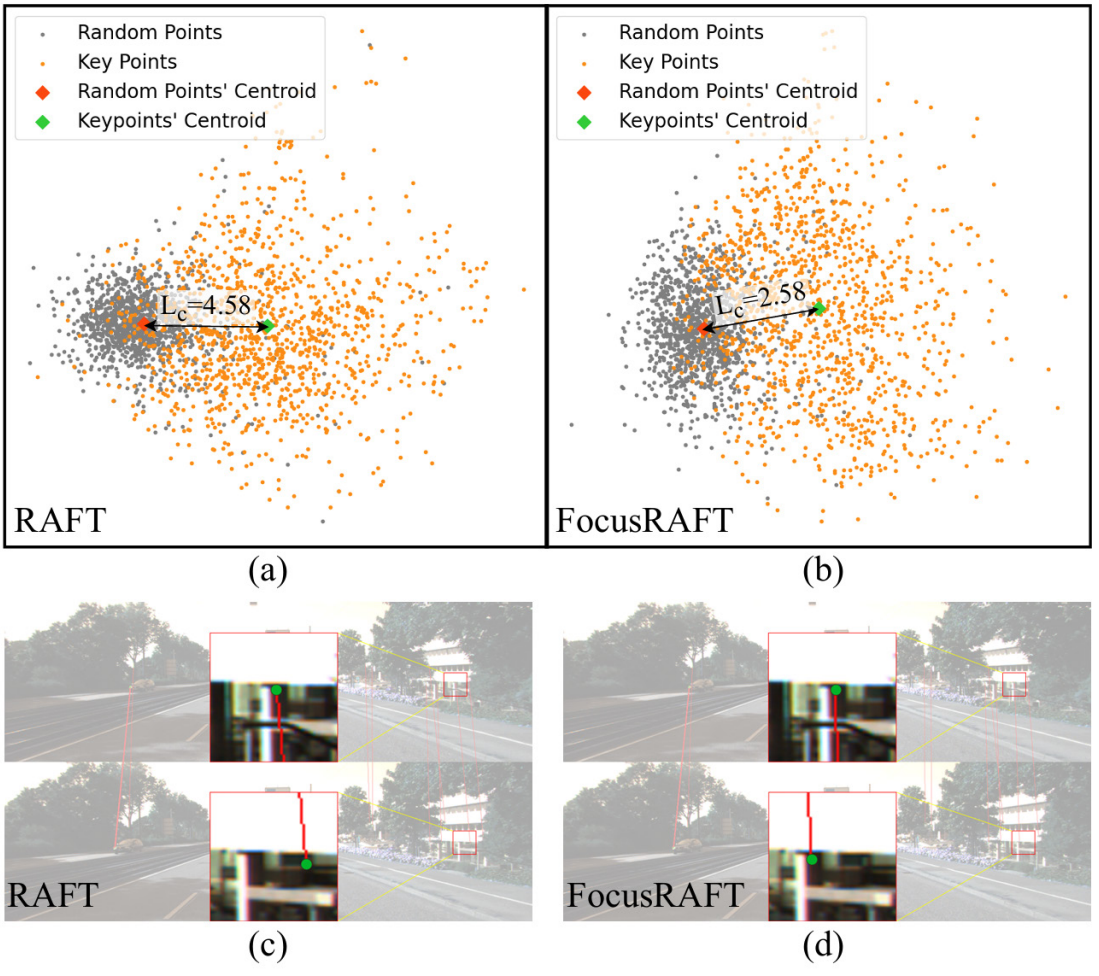

Fig. 1 Core improvements of the Focusflow framework [11]

ISSN 2959-6157

2.1.2 CSFlow Framework

CSFlow provides a solution to address computational challenges in optical flow estimation. These challenges occur in autonomous driving scenarios. The framework aims to improve estimation precision and accuracy. The system works in complex urban traffic environments. The system uses orthogonal decomposition techniques as its main computational approach. This decomposition method enables the capture of global contextual information. The method processes information across the entire visual field. The architecture also maintains high computational efficiency during processing. This design ensures real-time operational capabilities.

Figure 2 illustrates the CSFlow architecture. Its core innovation lies in introducing the Cross-Stripe Correlation (CSC) module and Correlation Regression Initialization (CRI) module. The CSC module encodes global contextual information into the correlation volume through horizontal and vertical stripe operations, and simultaneously reduces computational complexity from O(H×W×H×W) to O(H×W×(H+W)). The CRI module utilizes orthogonal correlation information to generate an initial optical flow field without parameters, providing a superior starting point for subsequent iterative optimization. Through three stages—feature extraction, cost volume computation, and iterative refinement—the entire system effectively addresses the challenges of optical flow estimation in autonomous driving scenarios involving texture-deficient and similar-texture regions. It significantly improves estimation accuracy while maintaining computational efficiency [12].

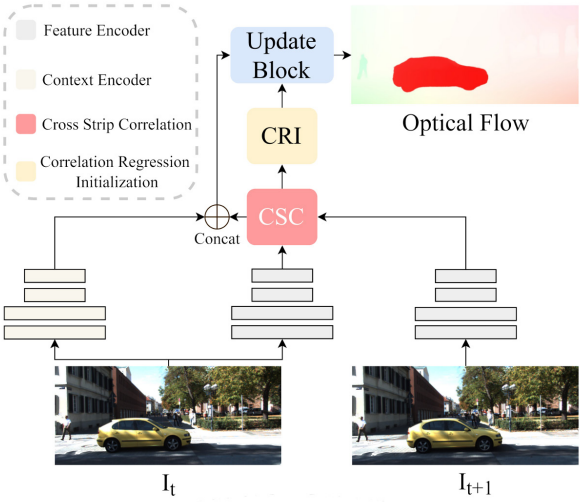

Fig. 2 CSFlow framework [12]

Figure 3 illustrates the architecture of the CSC module. This module generates query matrices (Qv, Qh) and key matrices (Kv, Kh) from input features F1 and F2 via 1×1 convolutions. The key matrices are aggregated through vertical and horizontal strip operations into orthogonal global key matrices Kv and Kh. The query matrices are transposed and dot-multiplied with their corresponding key matrices to generate vertical correlation volumes Cv and horizontal correlation volumes Ch. These are ultimately concatenated with the full-pair correlation volume C to form the aggregated volume Ĉ. This design reduces computational complexity from O(H×W×H×W) to O(H×W×(H+W)), encoding global contextual information while maintaining efficiency.

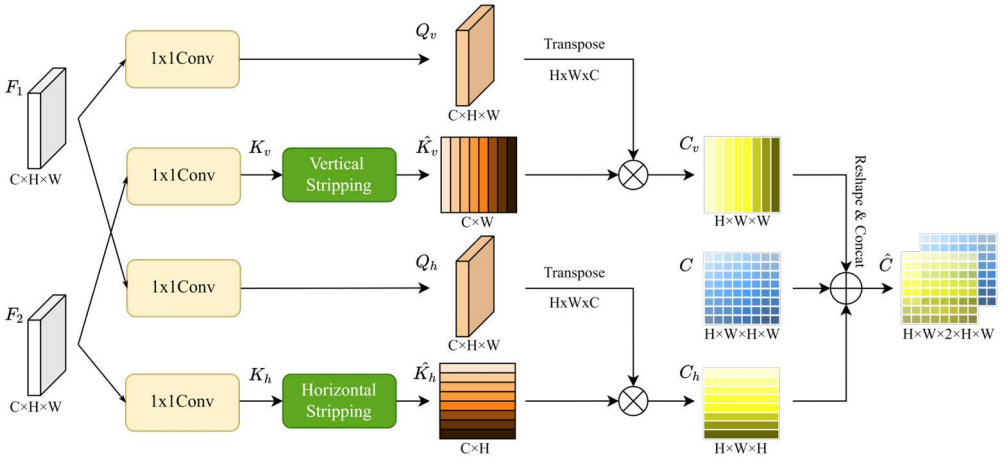

Fig. 3 Architecture of the CSC Module [12]

Figure 4 illustrates the parameter-free initialization process of the CRI module. The vertical and horizontal correlation volumes Cv and Ch are each passed through a softmax activation function before being element-wise multiplied with the original volume. This generates the orthogonal optical flow components v0 and u0, which are concatenated to form the initial optical flow field V0. This design leverages global similarity information to provide

robust initial values without requiring additional parameters. Consequently, subsequent iterative optimization can focus on resolving ambiguous regions, effectively enhancing the accuracy of optical flow estimation.

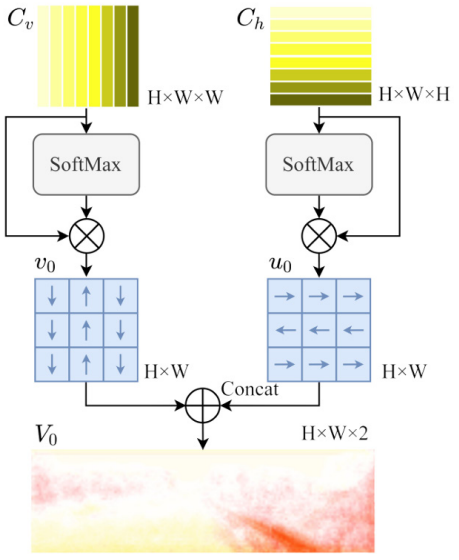

Fig. 4 Parameter-Free initialization process for CRI module [12]

2.2 Edge Computing and Autonomous Driving Synergy

Collaborative resource allocation strategies for edge computing and autonomous driving primarily include OVEAP (Integer Linear Programming, ILP) and DVEAP (Deep Reinforcement Learning, DRL). OVEAP primarily addresses optimal association matching between vehicles and edge servers in static environments, allocating optimal computational resources for autonomous driving tasks such as perception, path planning, and control. For dynamic environments, the resource allocation problem is modeled as a Markov Decision Process (MDP). Deep Q-Network (DQN) is employed to dynamically adjust offloading strategies in real-time, adapting to vehicle movement, network fluctuations, and server load variations.

OVEAP models resource allocation problems as an integer linear programming (ILP) problem, with the primary variables and constraints as follows:

(1) Decision variables: x_{ij}¡Ê0,1 ,

x_{ij}=1 , indicates that VNF i assigned to edge servers j ,

x_{ij}=0 , indicates that VNF i not assigned to edge servers j .

(2) $x^{mec}_{ea,vnf}$¡Ê0,1 :Indicates whether to offload Virtual Network Functions (VNF) from the Autopilot Chain to edge servers.

(3) y^{mec}¡Ê0,1 :Indicates edge server mec is whether activated for use.

(4) The objective function can be expressed as [13]

$$\max\left(\sum_{ea}\sum_{mec} x^{mec}_{ea,L_{ea}} 0 \sum_{mec} y^{mec}\right) \quad (1)$$

To address OVEAP's scalability issues, DVEAP em-ploys a Deep Q-Network (DQN) to transform the resource allocation problem into a Markov Decision Process (MDP):

(1) Current resource utilization status of edge servers (remaining CPU/GPU/storage ca-pacity).

(2) Resource requirements for VNFs awaiting offloading (e.g., perception modules de-mand high GPU resources, control modules require low latency).

3. Challenges in Underlying Computing Hardware

3.1 Implementation of Optical Convolution

In traditional electronic computing architectures, convolution operations form the core processing of deep neural networks (particularly convolutional neural networks, CNNs), yet their energy consumption and processing speed are increasingly encountering bottlenecks. To overcome the physical limitations of electronic devices, optical processing units (OPUs) have been proposed to perform convolution operations. Leveraging the parallelism, low latency, and high bandwidth advantages of optical signals, OPU significantly enhances computational efficiency in cloud computing scenarios.

The discrete convolution operation between the convolution kernel and the input data is defined as [13]:

$$y(q)=\sum_{p} x(p)?\omega(p0q) \quad (2)$$

In optical systems, this operation is achieved through the following steps: First, the volume hologram "ω" is decomposed into positive and negative parts, which are separately loaded onto two FSR wavelength segments of the optical frequency comb. The input data "x(p)x(p)" is modulated onto the corresponding wavelength optical signals through MZM array modulation. Each MZM corresponds to one data point, and the optical intensity after modulation is "x(p)·Ip(λ)x(p)·Ip(λ)". The modulated optical signals are then input into the AWGR. According to the wavelength routing rules, signals with wavelengths " λp, qλp, q", q will reach output port q. At output port q, the optical intensity is the sum of contributions from all input ports p [13]:

ISSN 2959-6157

$$Iq=¡Æp¡Ænx(p)?\omega n(p0q) \quad (3)$$

$$Iq=p¡Æn¡Æx(p)?\omega n(p0q) \quad (4)$$

where n represents the FSR period.

As shown in Figure 5 during the weight and data loading phase, the system employs wavelength division multiplexing (WDM) to modulate computational data onto optical carriers, enabling multi-dimensional parallel data input. This design fully leverages the inherent advantages of optical systems. Compared to the serial data readout mode of traditional electronic computing, optical systems can simultaneously process different weights and input data across multiple wavelength channels, funda-mentally enhancing data processing efficiency. Serving as the computational core, the AWGR path leverages the wavelength routing properties of the arrayed waveguide grating to perform optical convolution operations. Through the periodic transmission characteristics of the AWGR, the system completes the entire convolution kernel computation within a single clock cycle. This parallel processing capability significantly surpasses traditional optical computing methods based on delay lines. More importantly, the AWGR supports remote weight loading, enabling physical separation between computational nodes and data storage nodes. This lays the technical foundation for true cloud-based optical computing. During the parallel computation phase, multiple OPU units collaborate to process complex AI tasks. The system decomposes large-scale convolutional neural networks into sub-tasks, with each OPU independently handling specific computations, achieving an efficient distributed computing architecture.

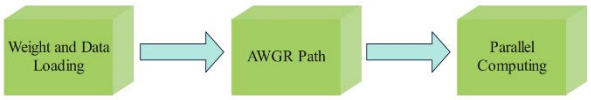

Fig. 5 Optical Convolution Flowchart

3.2 Multi-channel Optical Neural Network Architecture of "Monet"

Current optical computing architectures are constrained by limitations in optical interconnects and interaction operators, often confined to single-channel structures with processing capabilities restricted to simple tasks such as handwritten digit classification and saliency detection 14]. The computational limitations and scalability constraints of single-channel optical neural networks (ONNs) severely hinder the optical implementation of advanced machine vision tasks. To address this, researchers propose Monet—a multi-channel optical neural network architecture based on a projection-interference prediction framework. Monet sup-ports general-purpose multi-input, multi-channel optical computation, with intra-channel and inter-channel connections realized through optical interference and diffraction mechanisms.

Figure 6 illustrates the schematic diagram of the Monet channel. Within the Monet architecture, input data and model weights are encoded across multiple modes within the optical spectrum, with signal modulation achieved using optical modulators such as Mach-Zehnder modulators (MZM). Each modulation channel corresponds to a specific frequency within the optical frequency comb, enabling each data unit to be assigned to an independent optical frequency band. The modulated optical signal and weight information undergo parallel processing through the combination of an arrayed waveguide grating (AWGR) and micro-ring resonators.

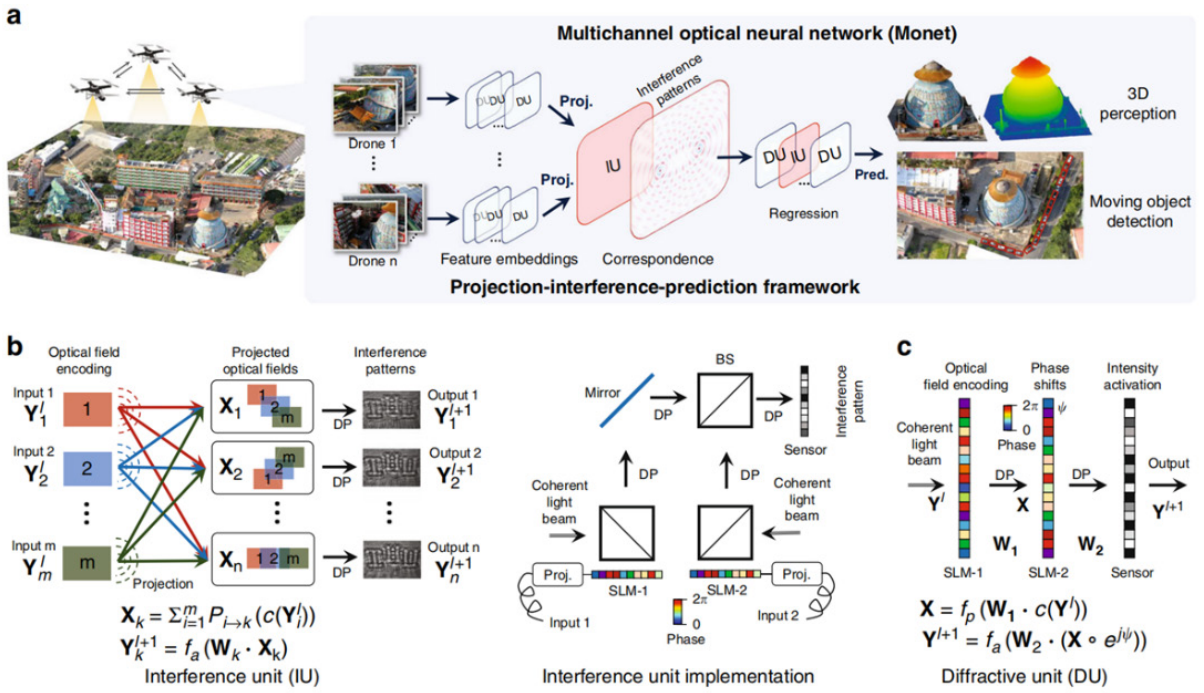

Fig. 6 Schematic Diagram of the Monet Channel [14]

Specifically, Monet channels implement the Arrayed Waveguide Grating Router (AWGR) architecture for wavelength-division routing of optical signals. The system exploits wavelength selectivity of optical pathways to perform convolution computations. Input signals and convolutional kernel weights undergo mapping to distinct wavelengths via optical waveguide arrays. Following modulation and weighting procedures, signals are systematically routed to respective processing units within the AWGR infrastructure.

Micro-ring resonators provide filtering and separation functionality for optical signals to complete convolution operations. Throughout this computational process, both mathematical operations and signal transmission occur exclusively within the optical domain. This approach effectively circum-vents bandwidth limitations and power consumption issues inherent in conventional electronic computing methodologies.

Monet Channel achieves highly efficient optical convolutional computing through multimodal parallel processing of optical spectra, integrating advanced arrayed waveguide grating and micro-ring resonator technologies. This significantly enhances computational speed and energy efficiency in generative AI tasks. Its core innovation lies in integrating optical communications with deep learning computation, pioneering a new paradigm of low-power, high-parallel optical convolutional computing. This approach holds significant promise for future applications in large-scale optical computing platforms and edge computing domains.

4. Comprehensive Analysis, Comparison, and Discussion

Create a table to conduct a horizontal comparison of all technical solutions in this chapter based on the following dimensions:

To systematically examine the application potential and limitations of different technical approaches in autonomous driving edge computing, this section conducts a comparative analysis of the aforementioned representative solutions across five dimensions: core innovations, bottleneck break-throughs, distinctive advantages, inherent short-comings, and development maturity (as shown in Table 1). This comparison aims to clarify the trade-offs between power consumption, efficiency, real-time performance, and feasibility for each method, thereby enhancing understanding of the current coexistence of multiple technical pathways in this field.

Comparisons reveal that various technologies attempt to address common challenges in autonomous driving edge computing—such as storage, power consumption, latency, and computational capacity bottlenecks—from different angles, yet each faces distinct limitations. For instance, while FocusFlow and CSFlow demonstrate high accuracy and efficiency in optical flow estimation tasks, they remain constrained by environmental adaptability and system complexity. OVEAP and edge computing focus

ISSN 2959-6157

on optimizing resource allocation and response latency, yet their scalability and device management issues remain unresolved. Optical Convolution and Monet leverage photonic properties to achieve ultra-high parallelism and energy efficiency, yet remain in early exploratory stages regarding computational accuracy, process implementation, and system integration.

Table 1. Comparative Analysis of Technologies

Technical Solution	Core Innovation Points	Target Bottlenecks	Key Advantages	Inherent Limitations	Technology Maturity
FocusFlow	Photoelectric Heterogeneous, Dynamic Scheduling	Power Wall, Efficiency Wall	High Energy Efficiency, Real-time Performance	High System Complexity	Research Prototype
CSFlow	CSC Global Encoding, CRI Regression Initialization	Efficiency Wall, Computing Power Wall	Low Computational Complexity, High Accuracy	Environmental Adaptability to be Validated	Algorithm Simulation
OVEAP	ILP Precise Modeling	Resource Utilization	Optimal Solution Solving	Poor Scalability (NP-Hard)	Simulation Validation
Edge Computing	Data-near processing	Excessive network latency	Ultra-low latency response	Complex device management	Relatively mature infrastructure
Optical Convolution	Optical components replacing electronic computation	Electronic computation power bottleneck	Ultra-high parallel processing capability	Limited computational precision	Early experimental stage
Monet	Multi-channel optical interconnect	Throughput, parallelism	High parallel bandwidth	Precision limitations, manufacturing difficulty	Laboratory demonstration

Overall, no single technology comprehensively addresses all challenges. Multi-technology fusion and cross-layer optimization may emerge as the dominant future direction. For instance, combining optical computing's high parallelism with edge computing's low-latency architecture, or employing lightweight algorithms to reduce accuracy demands in optical computing, could yield more competitive solutions. This comparison not only summarizes the current technological landscape but also points to complementary and integrated development paths for future research.

5. Shortcomings and Potential Solutions

5.1 Limitations of Optical Flow Estimation Al-gorithms and Their Solutions

Current optical flow estimation algorithms face multiple technical challenges and performance bottlenecks in practical applications. The FocusFlow framework increases system complexity and computational overhead due to its high dependency on keypoint information. Its application limitations primarily target keypoint-dense autonomous driving scenarios, with limited generalization capabilities for other scenarios. Although the CSFlow framework reduces computational complexity, the fixed nature of its cross-strip operations lacks adaptability, making it prone to feature confusion in texture-similar regions. Core challenges shared by these frameworks include limited handling of com-plex scenes, prominent computational inefficiency, strong dependence on training data, and insufficient environmental robustness.

To address these challenges, lightweight net-work architectures (e.g., MobileNet, EfficientNet) can be adopted, combined with knowledge distillation and neural architecture search to optimize computational efficiency. Concurrently, establishing adaptive computation mechanisms—such as early termination, region-adaptive computation, and dynamic network depth adjustment—can enhance system responsiveness and generalization capabilities.

5.2 Shortcomings and Solutions

Current edge computing and autonomous driving collaboration lacks effective fault-tolerance mechanisms for network latency fluctuations, and end-to-end hard real-time assurance capabilities remain immature. To address this, Time-Sensitive Networking (TSN) and 5G Ultra-Reliable Low-Latency Communication (URLLC) technologies can be integrated into the network to achieve microsec-

ond-level latency and ultra-high reliability. Resource allocation and management can be optimized through proximity-based edge node deployment, load prediction algorithms, hybrid critical scheduling, and dynamic task offloading. At the algorithmic level, strategies such as model com-pression, incremental inference, and task partitioning can reduce computational overhead by 50–90%, thereby enhancing system execution efficiency.

5.3 Limitations and Solutions for Optical Computing Technologies

Optical computing remains constrained by physical-layer challenges including insufficient de-vice precision, nonlinear distortion, and channel fluctuations. A hybrid-precision computing architecture can be proposed, integrating an optical acceleration layer (low precision), a hybrid correction layer (medium precision), and an electronic precision layer (high precision). Through intelligent task partitioning and optoelectronic interface optimization, this approach achieves an effective balance between computational accuracy and speed, offering a new development pathway for AI acceleration ap-plications.

6. Application and Future Outlook

6.1 Applications and Prospects of Optical Flow Estimation

The future application prospects for optical flow estimation are exceptionally broad. It will not only continue playing a vital role in core tasks like autonomous driving but also expand into broader intelligent systems. In autonomous driving and intelligent transportation, high-precision optical flow estimation provides foundational support for vehicle environmental perception, further enhancing the reliability of object detection, trajectory prediction, and anomaly behavior recognition, thereby driving the overall optimization of intelligent transportation systems. In robot perception and navigation, optical flow enables stable obstacle avoidance, path planning, and autonomous localization for drones, service robots, and warehouse systems. Particularly in complex dynamic environments, sparse keypoint-based optical flow methods significantly enhance adaptability. Concurrently, within augmented reality and virtual reality applications, optical flow estimation facilitates more natural user interaction and virtual object tracking, thereby elevating immersion and interactive experiences. In medical image analysis, optical flow technology demonstrates immense potential, such as dynamic tracking of heartbeats, blood flow, or tissue motion, providing support for clinical diagnosis and minimally invasive surgery. Optical flow also holds extensive value in video editing and intelligent surveillance, serving both visual enhancement tasks like video frame interpolation and motion deblurring, as well as enabling anomaly detection of abnormal motion patterns in security monitoring. Overall, future optical flow estimation will continually seek a balance between global scene understanding and local fine-grained tracking to accommodate diverse application demands.

6.2 Applications and Outlook for Edge Compu-ting and Autonomous Driving Synergy

From a technological architecture perspective, future edge computing systems will develop toward enhanced intelligence and adaptability. 6G network technologies will mature and bring new capabilities. These technologies include terahertz communication and reconfigurable smart surfaces. They will pro-vide better bandwidth and connectivity for edge computing. This improvement will enable edge nodes to handle complex perception fusion tasks. The systems will achieve true distributed intelligence.

Edge computing architectures will change from static deployments to dynamic self-organizing net-works. Mobile edge computing nodes will deploy flexibly. They will achieve seamless coverage for autonomous vehicles. This dynamic architecture adapts to high-speed vehicle mobility. It also adjusts computational resource distribution based on traffic flow changes. This approach enhances overall system efficiency.

Artificial intelligence technologies will integrate deeply into this field. This integration will become a core driver for advancement. Current re-source optimization methods use deep reinforcement learning. These methods show promising ap-plications. They will evolve toward multi-agent reinforcement learning approaches.

Federated learning technology will be introduced into autonomous driving systems. This technology enables systems to share learning experiences while protecting privacy. It accelerates algorithmic iteration and optimization. Transfer learning will use large-scale pre-trained models. This approach will enhance the cognitive capabilities of edge computing nodes. The nodes will handle more diverse driving scenarios.

6.3 Future Development Directions of Optical Computing

Optical computing is developing toward three main directions. These directions are architectural diversification, integrated modularity, and application versatility.

Computational architectures are becoming more diverse. Systems are shifting from centralized designs to distrib-

ISSN 2959-6157

uted designs. They are also changing from single-channel to multi-channel configurations. Traditional approaches emulate existing electronic architectures. New approaches use native de-signs that leverage optical properties. This diversification enables optical computing to identify optimal technical solutions for different application scenarios.

System integration is continuously improving. Photonic integration technology is becoming more mature. Future optical computing systems will develop toward on-chip integration. They will also adopt modular design approaches. These developments will reduce system complexity significantly. They will also decrease manufacturing costs.

Application domains are expanding deeply. Optical computing is entering new fields and applications. This expansion broadens the practical im-pact of optical computing technologies. Moving be-yond current task-specific validation toward universal computing platforms, optical computing is poised to play a vital role in broader fields such as big data processing, scientific computing, and artificial intelligence.

From a longer-term perspective, advancements in the underlying technologies of optical computing may bring about a fundamental shift in computational paradigms. Amid the rapid advancement of emerging computing technologies like quantum computing and neuromorphic computing, optical computing is poised to play a unique role within this new computational ecosystem, particularly in ap-plications demanding high-speed, low-power, and high-bandwidth processing. Simultaneously, the deep integration of optical computing with other emerging technologies will foster novel application paradigms and commercial opportunities, providing more efficient and sustainable computational sup-port for the intelligent society of the future.

7. Conclusion

This paper systematically reviews the key technological pathways and current development status of optical neural networks accelerating edge computing in the context of autonomous driving. Through in-depth analysis of representative solutions—including the FocusFlow and CSFlow optical flow estimation algorithms, the OVEAP and DVEAP edge-cooperative architectures, optical convolutions, and the Monet multi-channel optical neural network—it reveals the technological progress and innovative approaches in overcoming the traditional electronic computing "memory wall," "computational power wall," and "power consumption wall." Research indicates that optical neural networks, leveraging their physical-layer parallelism, light-speed transmission, and ultra-low power consumption, offer breakthrough solutions for meeting autonomous driving systems' stringent real-time requirements (<50ms latency) and power constraints (automotive-grade <50W). Theoretically, they can reduce matrix operation latency to the nanosecond range and power consumption to 1/500th that of traditional GPUs.

However, current technological development still faces multiple challenges. Optical flow estimation algorithms exhibit significant shortcomings in complex scene adaptability, computational efficiency optimization, and environmental robustness. Edge computing collaborative architectures require refinement in handling network latency fluctuations and ensuring hard real-time guarantees. Optical computing technology faces physical constraints such as device precision, environmental sensitivity, and system integration complexity. These technical bottlenecks indicate that a single technological approach cannot comprehensively address the complex demands of autonomous driving edge computing. Multi-technology convergence and cross-layer collaborative optimization will become the dominant direction for future development.

Looking ahead, optical computing will evolve toward architectural diversification, integrated modularity, and application generalization. With the rapid advancement of emerging technologies like quantum computing and neuromorphic computing, optical computing is poised to play a unique role in new computational ecosystems, particularly in scenarios demanding high-speed, low-power, and high-bandwidth processing. Simultaneously, the deep integration of technologies like 6G networks, Time-Sensitive Networking (TSN), and federated learning will propel edge computing from static deployment toward dynamic, self-organizing net-works, enabling true distributed intelligence. By constructing an integrated opto-electro-computational in-vehicle computing platform and conducting system-level validation in real-world complex scenarios, a critical leap from laboratory prototypes to industrial applications can be achieved.

This research provides a theoretical foundation and technical roadmap for constructing next-generation high-performance, low-power intelligent computing systems for autonomous driving. It holds significant theoretical and practical value for advancing the comprehensive upgrade of intelligent transportation systems. Future research should prioritize interdisciplinary collaboration and industry-academia-research integration. Through continuous technological innovation and system integration, the ultimate goal is to achieve large-scale industrial application of optical computing technology in critical domains like autonomous driving.

References

[1] Maurer M, Gerdes J C, Lenz B, et al. Autonomous driving: technical, legal and social aspects. Springer Nature, 2016.

[2] Ibn-Khedher H, Laroui M, Mabrouk M B, et al. Edge computing assisted autonomous driving using artificial intelligence. In 2021 International Wireless Communications and Mobile Computing (IWCMC), 2021: 254-259.

[3] Han Song, Mao Huizi, Dally W J, et al. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprintarXiv:1510.00149, 2015.

[4] Fu Tingzhao, Zhang Jianfa, Sun Run, et al. Optical neural networks: progress and challenges. Light: Science & Applications, 2024, 13(1): 263.

[5] Shen Yichen, Harris N C, Skirlo S, et al. Deep learning with coherent nanophotonic circuits. Nature Photonics, 2017, 11(7): 441-446.

[6] Tait A N. Quantifying power in silicon photonic neural networks. Physical Review Applied, 2022, 17(5): 054029.

[7] Feldmann J, Youngblood N, Karpov M, et al. Parallel convolutional processing using an integrated photonic tensor core. Nature, 2021, 589(7840): 52-58.

[8] Hamerly R. The future of deep learning is photonic: reducing the energy needs of neural networks might require computing with light. IEEE Spectrum, 2021, 58(7): 30-47.

[9] Shastri B J, Tait A N, de Lima T F, et al. Photonics for artificial intelligence and neuromorphic computing. Nature Photonics, 2021, 15(2): 102-114.

[10] Zhang Weipeng, Huang Chaoran, Peng Hsuan-Tung, et al. Silicon microring synapses enable photonic deep learning beyond 9-bit precision. Optica, 2022, 9(5): 579-584.

[11] Ibn-Khedher H, Laroui M, Moungla H, et al. Next-generation edge computing assisted autonomous driving based artificial intelligence algorithms. IEEE Access, 2022, 10: 53987-54001.

[12] Xing Sizhe, Sun Aolong, Wang Chengxi, et al. Seamless optical cloud computing across edge-metro network for generative AI. Nature Communications, 2025, 16(1): 6097.

[13] Yi Zhonghua, Shi Hao, Yang Kailun, et al. FocusFlow: boosting key-points optical flow estimation for autonomous driving. IEEE Transactions on Intelligent Vehicles, 2023, 9(1): 2794-2807.

[14] Xu Zhihao, Xiaoyun Yuan, Zhou Tiankuang, Fang Lu. A multichannel optical computing architecture for advanced machine vision. Light: Science & Applications, 2022, 11(1): 255.