Lecture Notes in Networks and Systems 1824

Lei Zhang
Bo Huang
Jin Xu
Jin Liu *Editors*

Proceedings of the 13th International Conference on Information Systems and Computing Technology

Volume 2

Springer

Lecture Notes in Networks and Systems 1824

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Lei Zhang · Bo Huang · Jin Xu · Jin Liu
Editors

Proceedings of the 13th International Conference on Information Systems and Computing Technology

Volume 2

Springer

Editors
Lei Zhang
Tongji University
Shanghai, China

Bo Huang
The University of Hong Kong
Hong Kong, China

Jin Xu
Peking University
Beijing, China

Jin Liu
Shanghai Maritime University
Shanghai, China

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-92-2878-2 ISBN 978-981-92-2879-9 (eBook)
https://doi.org/10.1007/978-981-92-2879-9

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

On behalf of the organizing committee, it is our great pleasure to welcome you to the 2025 13th International Conference on Information Systems and Computing Technology (ISCTech). Held from August 15th to 17th, 2025, in the vibrant city of Shanghai, China, ISCTech 2025 brought together leading researchers, academics, and industry practitioners to share their latest advancements and insights in the dynamic fields of information systems and computing technology.

This year's conference proceedings feature a collection of meticulously selected and peer-reviewed research papers. These contributions span a wide range of cutting-edge domains, including Computer Vision and Image Processing, Artificial Intelligence and Machine Learning, Communication and Network, and Data Science and Analytics. The high quality of these papers reflects the rigor and dedication of our global academic community.

We are deeply grateful to our esteemed keynote speakers for enriching the conference with their profound knowledge and expertise. We were honored to have Prof. Li Chai (Zhejiang University, China), who shared his insights on "Momentum-accelerated Average Consensus of Multi-agent Systems"; Prof. Bo Huang (The University of Hong Kong, China), who discussed "Spatial Intelligence for Urban Environmental Sustainability"; and Prof. Rong Gu (Nanjing University, China), who spoke on "Efficient Task Scheduling and Data Transmission Optimization for Serverless Computing." Their presentations provided invaluable perspectives and sparked engaging discussions among participants.

The success of ISCTech 2025 would not have been possible without the immense effort and dedication of many individuals. We extend our sincere gratitude to all members of the organizing committee and Shanghai Maritime University for their tireless work in making this event a reality. Also, we thank the authors and reviewers who have contributed to this event by preparing papers of high quality. Finally, we would like to express our profound appreciation to the publisher and editorial staff of Springer for accepting and supporting the publication of this proceedings volume. Their partnership is crucial in disseminating the research presented at this conference to a global audience.

We sincerely hope that these proceedings will serve as a valuable resource for all participants and readers, providing scientific inspiration and stimulating new ideas for future research.

The Committee of ISCTech 2025

Conference Organization

General Chairs

Prof. Bo Huang	The University of Hong Kong, China
Prof. Lei Zhang	Tongji University, China
Prof. Jin Xu	Peking University, China
Prof. Srikanta Patnaik	Interscience Institute of Management and Technology, India
Prof. Simon Xu	Algoma University, Canada
Prof. Rigui Zhou	Shanghai Maritime University, Canada

Conference Executive Chairman

Prof. Jin Liu	Shanghai Maritime University, China

Tutorial Committee

Prof. Zhigong Wang	Southeast University, China
Prof. Seppo Sirkemaa	University of Turku, Finland
Prof. Carlos Juiz	Universitat de les Illes Balears, Spain
Asst. Prof. Mowafaq Salem Alzboon	Jadara University, Jordan

Technical Chairs

Prof. Wenlong Yi	Jiangxi Agricultural University, China
Assoc. Prof. Hamed Taherdoost	University Canada West, Canada
Assoc. Prof. Dong Zhang	Sun Yat-sen University, China
Assoc. Prof. Xiaoyan Zhang	Northwestern Polytechnical University, China

Publication Chairs

Prof. Wensheng Zhang	Shangdong University, China
Assoc. Prof. Xu Li	Northwestern Polytechnical University, China

Financial Chair

Assoc. Prof. Jia Shi	Xidian University, China
Assoc. Prof. Wei Liang	Northwestern Polytechnical University, China

Publicity Chair

Assoc. Prof. Huanliang Xiong	Jiangxi Agricultural University, China
Asst. Prof. Frederico Branco	Universidadede Trás-os-Montes e Alto Douro, Portugal

Conference Program Chairs

Prof. Jinhui Tang	Nanjing University of Science and Technology, China
Prof. Wenjian Liu	City University of Macau, China
Prof. Tao Han	National University of Defense Technology, China
Prof. Tao Shen	Kunming University of Science and Technology, China
Prof. Yue Cao	Wuhan University, China
Prof. Sheng Wu	Beijing University of Posts and Telecommunications, China
Prof. Wei Zhang	Zhejiang University, China
Prof. Yong Zhang	Beijing University of Posts and Telecommunications, China

Technical Committee Chairs

Prof. Lixin Li	Northwestern Polytechnical University, China
Prof. Weixiao Meng	Harbin Institute of Technology, China
Prof. Xianhua Cai	Southeast University, China
Prof. Qingwang Wang	Kunming University of Science and Technology, China
Prof. Bei Gong	Beijing University of Technology, China
Prof. Shaoyong Guo	Beijing University of Posts and Telecommunications, China
Prof. Dengyong Zhang	Changsha University of Science and Technology, China

Prof. Eryu Zhu	Beijing Jiaotong University, China
Prof. Guan Yuan	China University of Mining and Technology, China
Prof. Lei Zhang	Chongqing University, China

Technical Program Committee

Prof. Célio Gonçalo Marques	Polytechnic University of Tomar, Portugal
Prof. Paulo Rogerio Perfeito Tome	Polytechnic Institute of Viseu, Portugal
Prof. Wei-Chiang Hong	Asia Eastern University of Science and Technology, China
Prof. Gabriel Ruiz Andrade	Autonomous University of Baja California, Maxico
Prof. Viacheslav M. Tyutyunnik	International Nobel Information Centre, Russia
Assoc. Prof. Donghong Xu	China University of Mining and Technology, China
Assoc. Prof. Chien-Ming Chen	Nanjing University of Information Science and Technology, China
Asst. Prof. Yuri Antonacci	University of Palermo, Italy
Asst. Prof. António Jorge Gonçalves De Gouveia	University of Trás-os-Montes and Alto Douro, Portugal
Asst. Prof. Xianyu Zhang	Shanghai Jiao Tong University, China
Dr. Sandeep Jagtap	Cranfield University, UK
Dr. César Arturo Guerra García	Autonomous University of San Luis Potosi, Mexico
Asst. Prof. Ziyi Yang	Beijing Institute of Technology, China
Assoc. Prof. Xiaoming Wang	Nanjing University of Posts and Telecommunications, China
Dr. Muhammad Yaqub	Hunan University, China
Atif Mehmood	Lecturer Department of Computer Science Liverpool Hope University, UK

Contents

An Efficient Pyramid Transformer for Remote Sensing Image Object Detection

Zheng Wu, Qixuan Wang, Haochen Shi, and Ning Li(✉)

Nanjing University of Aeronautics and Astronautics, Nanjing, China
lnee@nuaa.edu.cn

Abstract. Remote sensing imagery exhibits high spatial resolution, heterogeneous backgrounds, spatially dense objects, and substantial scale variations, posing significant challenges for object detection systems. Object detection, as a critical task in machine vision, is also applied to remote sensing images. Ensuring detection performance is challenging due to the aforementioned characteristics of remote sensing images. To address this challenge, a Transformer-based object detection framework is proposed, which aims to mitigate the issues of high computational cost and insufficient local feature extraction encountered by conventional Transformers when processing high-resolution remote sensing images. Specifically, the model is comprised of an encoder and a decoder. An Efficient Pyramid-structured Transformer encoder (EP Transformer) is employed to extract both coarse-grained and fine-grained features. An Efficient Multi-Head Self-Attention mechanism (EMHSA) is introduced within the encoder to enhance the decoder's sensitivity to detailed features while reducing the computational complexity of the attention mechanism. The decoder merges multi-scale features and transforms them into object detection results, thereby accomplishing the object detection task for remote sensing images. Experimental results demonstrate that the proposed EP Transformer achieves a mAP of 77.12% on the DOTA dataset.

Keywords: Transformer · Object Detection · Remote sensing · Pyramid structure

1 Introduction

Advances in China's remote sensing systems have substantially improved spatial resolution and data acquisition capabilities, enabling high-clarity imagery with enhanced spectral-temporal abundance. As a core subtask of computer vision, object detection is increasingly applied in remote sensing domains. However, most existing methods still suffer from insufficient feature extraction capabilities and high computational complexity.

Recent convolutional neural network (CNN)-based object detection algorithms (e.g., Faster R-CNN [1] and the YOLO [2] series) have substantially improved detection performance. Nevertheless, these CNN-based approaches exhibit limited receptive fields due to convolutional kernel constraints and weak long-range modeling capabilities, resulting in global information loss. To address this, researchers have developed detection

L. Zhang et al. (Eds.): ISCTech 2025, LNNS 1824, pp. 1–8, 2027.
https://doi.org/10.1007/978-981-92-2879-9_1

frameworks based on conventional Transformers [3]. Vision Transformer (ViT) [4], for instance, employs multi-head self-attention (MHSA) mechanisms for global modeling, effectively mitigating insufficient global context and achieving competitive performance in object detection tasks. Despite these merits, ViT outputs single-scale feature maps due to inherent Transformer limitations, leading to deficient detail representation. Furthermore, standard MHSA incurs prohibitive computational overhead.

To overcome these limitations of conventional architectures, we introduce a pyramid-structured Transformer encoder. This hierarchical encoder generates multi-scale feature maps enriched with contextual information for object prediction tasks. Additionally, we propose an Efficient Multi-Head Self-Attention (EMHSA) mechanism that incorporates depthwise convolutional operations to inject inductive biases into attention computation. Compared with traditional MHSA, EMHSA enhances the encoder's sensitivity to fine-grained features while reducing computational complexity.

In summary, the primary contributions of our proposed remote sensing image object detection model are as follows:

Introduction of a pyramid-structured Transformer encoder to optimize multi-scale feature extraction, thereby providing robust support for subsequent decoder predictions.

Proposal of an efficient self-attention mechanism (EMHSA) that augments detail-feature extraction capability through integration with depthwise convolution while reducing computational complexity.

2 Method

2.1 Overall Architecture

We introduce EP Transformer, an object detection model for remote sensing image. Figure 1 illustrates the overall architecture of the proposed model, which is designed based on the Transformer framework. Input images are processed through a pyramid-structured Transformer encoder to extract feature maps at four distinct scales. These multi-scale features are merged via bilinear interpolation and subsequently fed into the decoder for subsequent prediction. The attention mechanism employed in the encoder corresponds to the proposed Efficient Multi-Head Self-Attention (EMHSA).

2.2 Pyramid-Structured Transformer Encoder

The encoder employs a hierarchical FPN structure, partitioning input images (size H $\times$ W $\times$ 3) into WH/4 $\times$ 4 patches of 4 $\times$ 4 resolution, yielding multi-scale feature maps for downstream prediction tasks. Compared to the 16 $\times$ 16 patches used in ViT, smaller patches enhance spatial granularity and benefit dense prediction tasks. These patches are then processed by a hierarchical Transformer encoder, which output feature maps at resolutions of1/4, 1/8, 1/16, and 1/32 of the original image scale. Feature dimensionality at decoder stage i is denoted as C_i.

Figure 2 details the structure of the Transformer module within the encoder. The input first passes through a normalization layer, followed by the EMHSA mechanism. The output of EMHSA undergoes a residual connection with the original input. This is

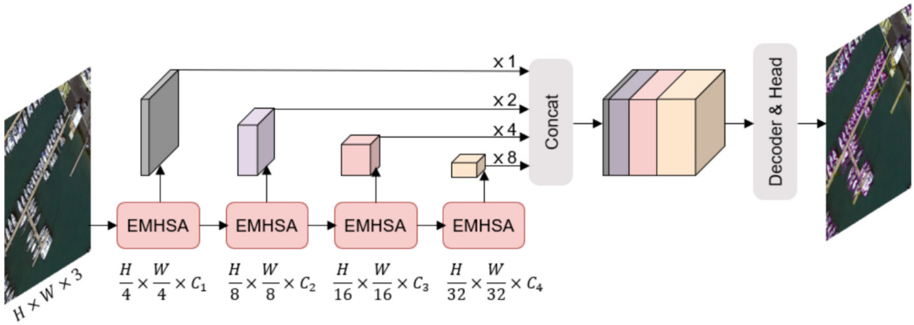

Fig. 1. EP transformer overall network architecture.

then normalized again and processed by a feed-forward network (FFN), with another residual connection applied thereafter. Crucially, Batch Normalization (BatchNorm) replaces Layer Normalization (LayerNorm) to better adapt to 2D image data, as BatchNorm preserves spatial relationships and exhibits superior performance in vision tasks compared to LayerNorm.

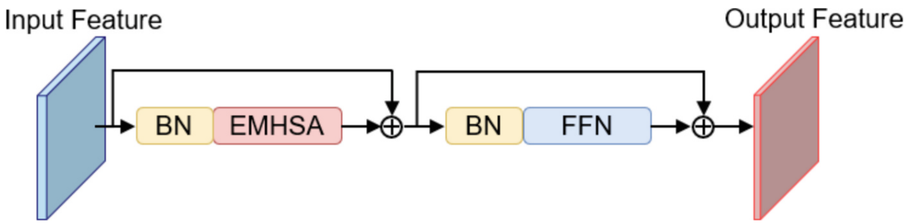

Fig. 2. Transformer module.

The process is formally defined as:

$$X_{att} = X + \text{EMHSA}(\text{BN}(X)) \tag{1}$$

$$X_{out} = X_{att} + \text{FFN}(\text{BN}(X_{att})) \tag{2}$$

where X denotes the original input, X_{att} represents the output of the EMHSA module, X_{out} is the final output of the Transformer block, BN(·) indicates the Batch Normalization operation, and FFN(·) refers to the Feed-Forward Network.

2.3 Efficient Multi-Head Self-Attention

We present the design of the Efficient Multi-Head Self-Attention (EMHSA) mechanism, with its structure depicted in Fig. 3.

EMHSA supplants standard MHSA to alleviate computational bottlenecks in high-resolution processing. While retaining the Query, Key, and Value inputs of standard

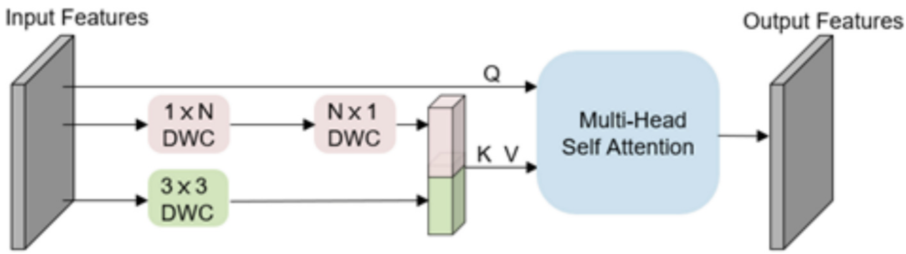

Fig. 3. The structure of the EMHSA.

MHSA, EMHSA introduces two convolutional branches prior to the attention computation. Orthogonal Strip Convolution Branch: The feature maps undergo depthwise convolutions with orthogonal kernels of sizes $1 \times N$ and $N \times 1$ to capture long-range spatial dependencies and enhance object edge continuity. Local Detail Extraction Branch: A 3×3 depthwise convolution extracts local patterns (e.g., textures, corners) while preserving high-frequency information. This process is formalized as:

$$X_1 = \mathrm{DWC}_{N\times 1}(\mathrm{DWC}_{1\times N}(X)) \tag{3}$$

$$X_2 = \mathrm{DWC}_{3\times 3}(X) \tag{4}$$

where X denotes the original input feature map, X_1 represents the output of the first convolutional branch, X_2 corresponds to the output of the second convolutional branch, and $\mathrm{DWC}_{i \times j}$ indicates depthwise convolution with a kernel size of $i \times j$.

The feature maps output by the convolutional operation are flattened and concatenated to form Y. This representation not only encapsulates rich contextual information from the image but also exhibits a shorter sequence length compared to directly flattening the original input X. Consequently, Y can replace X in computing the Key (K) and Value (V) matrices. The computational process is formulated as:

$$(Q, K, V) = \left(XW_q, YW_k, YW_v\right) \tag{5}$$

where W_q, W_k, and W_v denote the weight matrices for computing the Query, Key, and Value projections, respectively. The computational process within the MHSA module for the inputs Q, K, and V is formalized as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where d_k denotes the vector dimension of Q and K. The scaling operation (division by d_k) prevents excessively large magnitudes in the dot product of Q and K, thereby stabilizing the training process.

From the above procedure, it follows that after convolutional processing, the sequence lengths of K and V are reduced by a factor of N compared to their original dimensions. Consequently, the computational complexity of the attention mechanism is reduced by a factor of N^2.

3 Experimental Results and Analysis

3.1 Analysis of Comparative Experimental Results

The DOTA dataset comprises 1,411 training images, 458 validation images, and 937 test images, encompassing 15 categories of remote sensing objects: vehicles, plane, ship, storage tank, baseball diamonds, tennis court, basketball court, ground track field, harbor, etc. To validate the efficacy of EP Transformer, we benchmark it against several current mainstream methods on the DOTA dataset. These methods include Gliding Vertex [6], Oriented R-CNN [7], ReDet [8], RoI Trans [9], Faster R-CNN [1], SCRDet [10]. Tables 1 and 2 presents a quantitative comparison of category-specific precision and mean Average Precision (mAP) between the proposed EP Transformer framework and current mainstream methods on the DOTA benchmark. To ensure experimental fairness, all models were executed under identical hardware configurations and utilized the same training and testing datasets derived from DOTA.

The server configuration used for model training is as follows: Linux operating system, Python3.8 environment, PyTorch1.10.0, and CUDA11.1. We modify the model with a batch size of 2 on RTX3090, using Adam optimizer. The learning rate is set at 1e-3.

Table 1. Comparison of experimental quantitative results (UNIT:%).

Methods	PL	BD	BR	GTF	SV	LV	SH	TC
G Vertex	89.64	**85.00**	52.26	**77.34**	73.01	73.14	86.82	90.74
O Rcnn	89.46	82.12	54.78	70.86	**78.93**	83.00	**88.20**	90.19
ReDet	88.79	82.64	53.97	74.00	78.13	**84.06**	88.04	**90.89**
RoI Trans	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74
F Rcnn	88.44	73.06	44.86	59.09	73.25	71.49	77.11	90.84
SCRDet	**89.98**	80.65	52.09	68.36	68.36	60.32	72.41	90.85
EP Trans	89.84	83.76	**55.66**	74.08	77.64	79.24	87.94	90.77

As demonstrated in Tables 1 and 2, the proposed EP Transformer exhibits significant advantages in the key metric of mean Average Precision (mAP) compared to other object detection methods. Superiority, on Large-Scale Targets with High Aspect Ratios:

EP Transformer outperforms CNN-based models in detecting bridge, ground track field, soccer ball field, and roundabout—targets characterized by large scales and high aspect ratios. This validates Transformer's capability to model long-range spatial dependencies and geometric distortions, overcoming CNN's limited receptive fields.

Advantages on Small-Scale Targets with Indistinct Features: EP Transformer surpasses traditional Transformer-based detectors in identifying small vehicle, swimming pool, and helicopter—targets exhibiting small scales and weak visual features. This highlights the efficacy of the pyramid-structured Transformer encoder and EMHSA in enhancing feature extraction for fine-grained details. The pyramid Transformer encoder

Table 2. Comparison of experimental quantitative results (UNIT:%).

Methods	BC	ST	SBF	RA	HA	SP	HC	mAP
G Vertex	79.02	86.81	59.55	70.91	72.94	**70.86**	57.32	75.0
O Rcnn	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.8
ReDet	87.78	85.75	61.76	60.39	**75.96**	68.07	63.59	76.2
RoI Trans	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.5
F Rcnn	78.94	83.90	48.59	62.95	62.18	64.91	56.18	69.0
SCRDet	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.6
EP Trans	**88.82**	**87.74**	**66.82**	**68.36**	72.53	69.86	**63.73**	**77.1**

captures multi-scale contextual information. EMHSA's depthwise convolutions preserve high-frequency edge features.

Representative detection results on the DOTA dataset are visualized in Fig. 4, demonstrating EP Transformer's robust detection capability across multi-scale and densely clustered targets.

3.2 Analysis of Ablation Experiments

In this study, ablations on DOTA dataset validate EP-Transformer's superiority over component-excised variants. Table 3 presents the encoder depth analysis, while Table 3 details the attention mechanism comparisons.

Table 3. Encoder depth analysis.

Experimental group	Encoder depth	Downsampling rates	mAP(%)
Group A	2	[1/4,1/8]	71.64
Group B	3	[1/4,1/8,1/16]	75.88
Group C	4	[1/4,1/8,1/16,1/32]	**77.12**
Group D	5	[1/4,1/8,1/16,1/32,1/64]	76.25

As demonstrated in Table 3, the model achieves optimal detection performance with 4 encoder layers. The performance degradation observed with 5 layers—reaching only 76.25% mAP compared to 77.12% at 4 layers—is attributed to feature dilution of small targets during the feature merging phase caused by excessive downsampling. Table 4 reveals that the synergistic enhancement from EMHSA's dual-branch convolutional design significantly boosts detection performance. This conclusively validates the critical role of EMHSA in feature extraction.

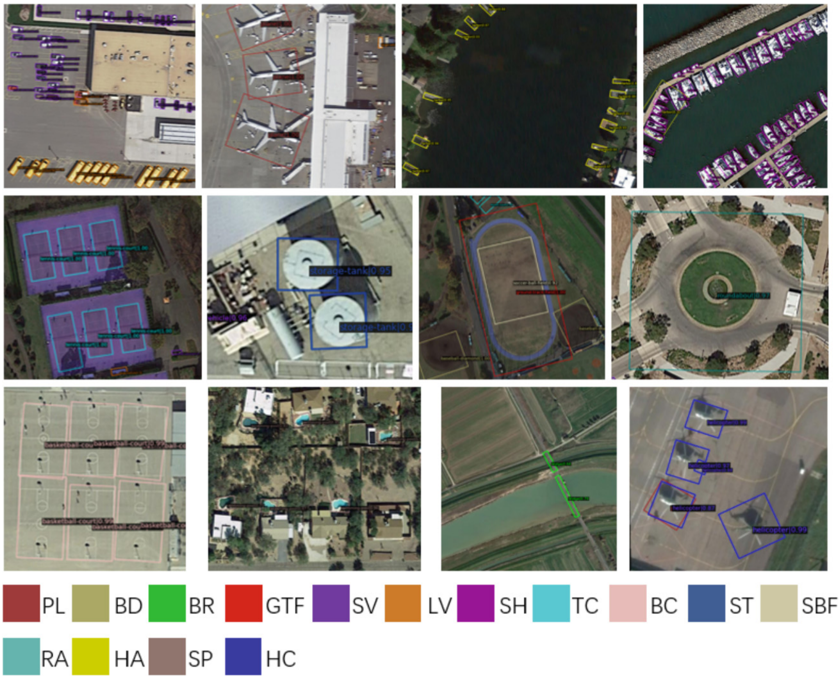

Fig. 4. Detection results of EP transformer on the DOTA dataset.

Table 4. Attention mechanism comparisons.

Experimental group	Attention module	Strip convolution branch	Detail extraction branch	mAP(%)
Group E	MHSA			70.74
Group F	EMHSA	√		73.62
Group G	EMHSA		√	74.92
Group H	EMHSA	√	√	**77.12**

4 Conclusion

This paper proposes EP Transformer, a novel framework for remote sensing image object detection. The architecture comprises a pyramid-structured Transformer encoder and a task-specific decoder. The encoder extracts multi-scale contextual features through hierarchical representations. Crucially, we introduce Enhanced Multi-Head Self-Attention (EMHSA) to replace standard MHSA, which integrates depthwise convolutions with attention mechanisms. This co-design enhances detail sensitivity while reducing computational complexity compared to vanilla Transformers.

Experimental results demonstrate that EP Transformer demonstrated superior detection capabilities on the DOTA benchmark, attaining 77.12% mAP.

Acknowledgement. This study was partially supported by the Z Key Laboratory Development Fund and the ZBA Key Laboratory Open Fund (1022-ZBA24085).

References

1. Ren, S.Q., He, K.M., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
2. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
3. Vaswani, A., Shazeer, N., Parmar, N.: Attention is all you need [A]. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems [C] (2017)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X.H., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
5. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.P.: DOTA: a large-scale dataset for object detection in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)
6. Xu, Y.C., Fu, M.T., Wang, Q.M., Wang, Y.K., Chen, K., Xia, G.S., Bai, X.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
7. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented R-CNN for object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, pp. 3500–3509 (2021). https://doi.org/10.1109/ICCV48922.2021.00350
8. Han, J., Ding, J., Xue, N., Xia, G.S.: ReDet: a rotation-equivariant detector for aerial object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2785–2794 (2021). https://doi.org/10.1109/CVPR46437.2021.00281
9. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.K.: Learning ROI transformer for oriented object detection in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2849–2858 (2019)
10. Yang, X., Yang, J.R., Yan, J.C., Zhang, Y., Zhang, T.F., Guo, Z., Sun, X., Fu, K.: Scrdet: towards more robust detection for small, cluttered and rotated objects. In: International Conference on Computer Vision, pp. 8231–8240 (2019)

Straggler Detection Method Based on Parallelism Communication Model in LLM

Congqi Shen[1(✉)], Shuhan Chen[2], Ningchun Liu[1], Aiyue Wang[1], Jun Zhu[1], and Tao Zou[1]

[1] Zhejiang Lab, Hangzhou, China
shencq@zhejianglab.org
[2] Zhejiang University, Hangzhou, China

Abstract. In this paper, we introduce a novel straggler detection method based on the communication traffic model in the LLM training network, with the aim of identifying the straggler during the large language model (LLM) training. The theoretical communication traffic models for the data parallelism and pipeline parallelism are firstly constructed by analyzing communication characteristics to establish a expectation behavior of normal parallelism. Then, by comparing the captured traffic size with the theoretical model, the abnormal GPU pairs could be determined by using statistic threshold and dynamic variance analysis so as to avoid false alarm due to transient fluctuations. Furthermore, a GPU state matrix is set up to localize stragglers based on anomaly ratios. Simulation results demonstrate that the proposed communication traffic is consistent with the actual communication behavior and the proposed detection method could accurately identify straggler in short time.

Keywords: straggler detection · LLM training · communication traffic model

1 Introduction

In large-scale AI model training, communication synchronization among GPUs is critical for training efficiency. Models with hundreds of billions to trillions of parameters require extensive parallel computation across tens of thousands of GPUs. Therefore, gradient and parameter synchronization among GPUs plays a significant role during the training process.

Inefficient communication synchronization would cause timeout errors in synchronization barriers during collective communication, leading to underutilization of computation resources and significantly prolongs training cycles as well as training failure. For instance, Meta's tests revealed that communication latency can consume up to 57% of training time, equivalent to 35% idle time for GPUs worth millions of dollars [1]. Furthermore, accumulated delays may invalidate gradients or conflict with parameter updates, leading to network failures or even

L. Zhang et al. (Eds.): ISCTech 2025, LNNS 1824, pp. 9–18, 2027.
https://doi.org/10.1007/978-981-92-2879-9_2

training collapse. Meta restarted its OPT-175B training over 40 times in two weeks due to training instability. In clusters exceeding 100,000 GPUs, recovery time from network failures may surpass the model's actual training time [2].

Meanwhile, straggler GPUs caused by hardware variations, network congestion, or software errors frequently occur in practice, especially in large GPU clusters. A joint study by Harvard and Meta further indicated that low-precision optimizations (e.g., Flash Attention) introduce numerical bias, exacerbating output inconsistencies across nodes and causing some nodes to lag by an order of magnitude [2]. For example, Google reported over 20 loss spikes during PaLM model training, partly caused by synchronization delays among nodes5.

However, detecting stragglers faces significant challenges. On the one hand, stragglers caused by hardware faults or transient network congestion may be hard for conventional monitoring. On the other hand, applying a detection method would introduce extra cost and affect the training task.

In order to solve the challenges above, some researchers apply traditional methods in distributed training to achieve straggler mitigation for LLM training, including dynamically adjusting task allocation, optimizing network traffic scheduling, or employing fault-tolerance mechanisms [3]. A hierarchical parameter synchronization algorithm called HiPS and a network traffic scheduling scheme named CEFS for heterogeneous distributed training is proposed in [4]. HiPS optimizes the parameter synchronization process through hierarchical synchronization, reducing bandwidth contention and cumulative communication latency, thereby significantly reducing synchronization time. However, this algorithm does not fully consider the impact of performance differences among nodes on synchronization efficiency. CEFS prioritizes the traffic scheduling of straggler nodes' parameter synchronization, allowing them to trigger forward computation earlier and thus alleviating the system's blockage caused by stragglers. Nevertheless, this scheme may face high scheduling complexity in large-scale distributed systems. A self-adaptive distributed training framework called AntDT is introduced in [5]. AntDT dynamically adjusts task allocation and resource scheduling to improve the computational efficiency of straggler nodes. However, the framework may lack adaptability to network topologies in large-scale distributed systems, and its fault-tolerance capability in multi-node collaborative training needs further validation. A fault-tolerant framework named BAFT for distributed deep neural network training is proposed in [6]. However, in complex network environments with dynamic changes, the framework may incur additional performance overhead due to the complexity of the fault-tolerance mechanism. A universal checkpointing mechanism for efficient and flexible checkpoint management is proposed in [7]. By optimizing the storage and recovery process of checkpoints, this mechanism enhances the system's fault-tolerance capability. However, in large-scale distributed systems, the universal checkpointing mechanism may face issues such as high storage resource consumption and long recovery times.

In fact, LLM training generates many-to-one Incast traffic patterns, which is different from communication pattern in traditional distribution training. When parameter servers simultaneously receive gradients from hundreds of worker nodes, egress ports on leaf switches experience instantaneous micro-bursts that frequently trigger traffic anomalies and training failures. Traditional methods typically complete detection in minute-level timeframes, whereas single iteration timeouts in large model training can cause job failures, necessitating millisecond-level response. During large model training, compute nodes span racks, servers, and accelerators—a topology fundamentally different from traditional data centers. Consequently, conventional methods are [10] ill-suited for LLM training networks. An efficient fault node recovery method in WSNs via Fat-Tree topology reconstruction is proposed in [10]. By dynamically adjusting the network topology, EPRA-FT quickly recovers faulty nodes and enhances the fault tolerance of wireless sensor networks. However, it may face challenges in topology reconstruction efficiency in large-scale networks.

In response, recent research has begun developing specialized straggler detection approaches tailored to LLM training networks. The detection of stragglers in serverless computing by using reinforcement learning is proposed in [8]. Grasshopper dynamically adjusts task allocation based on real-time node performance monitoring using reinforcement learning algorithms, effectively reducing the impact of stragglers on system performance. However, its reliance on complex reinforcement learning models may lead to high computational overhead. However, its reliance on complex reinforcement learning models may lead to high computational overhead. EdgeBERT, an optimization method for on-chip inference in multi-task NLP is proposed in [9]. By optimizing model structure and inference processes, EdgeBERT significantly reduces inference latency. However, further optimization may be needed to improve the model's generality and scalability in complex multi-task scenarios. However, these methods would introduce significant extra cost.

The papers [11,12] explore fault detection for large model training networks on the edge side. However, edge-side fault detection is mostly used for post-failure intervention and lacks sub-second anomaly prediction capability. When congestion is detected, Incast traffic has already caused packet loss, triggering NCCL timeouts and aborting the training. The limitations of these methods stem from the fact that existing approaches do not truly integrate the study of detection methods with the traffic patterns generated during large model training.

In this paper, we investigate the communication patterns in LLM training tasks and propose a real-time straggler detection method for LLM training networks. The main contributions of this paper are summarized as follows.

(1) We construct communication traffic models for pipeline parallelism and data parallelism by analyzing the communication features in LLM.

(2) By capturing traffic size among GPU pairs and utilizing the traffic models we established, we propose a straggler detection method based on both static threshold and spatial variation among GPUs. Simulation results demonstrate that it can detect straggler in LLM training networks at an early stage, thereby avoiding synchronization timeouts and preventing training task failures.

2 Communication Traffic Model

As mentioned above, existing approaches fail to take full advantage of intrinsic communication characteristics during the LLM training procedure. Since these methods only use traditional traffic features for abnormal detection, These methods need to accumulate extensive and high-dimensional historical data of the traffic feature to establish a normal behavior model. As a result, without the communication characteristics, those methods would involve complex computations and result in detection latencies. In fact, the communication pattern during the LLM training procedure is closely related to a specific training task.

According to our theoretical analysis, due to the abundance of repetitive computational operations in LLM training, communication behaviors demonstrate periodic patterns. Besides, with the application of 3D parallelism, *i.e.*, data parallelism (DP), tensor parallelism (TP), and pipeline parallelism (PP), communication traffic across different GPUs shows temporal and spatial correlation characteristics. In other words, whether a training task is normal or not could be identified by analyzing the communication pattern. TP primarily involves intra-node communication, where GPUs within the same server exchange tensor fragments, leveraging high-bandwidth intra-node links with lower latency and less susceptibility to stragglers compared to inter-node communication. In contrast, DP and PP, occurring during inter-node communication, would lead to stragglers due to inter-node latency and network congestion. Thus, we focus on DP and PP in this work, as they are the key bottlenecks for large-scale LLM training efficiency. We first determine the baseline traffic expectations by establishing the communication traffic model of the LLM training procedure. Then, we capture DP and PP traffic features from switches and compare them to the corresponding traffic expectations. Both temporal and spatial comparison would be applied to avoid an ineffective static threshold or a transient burst. Furthermore, a GPU state matrix is set up to identify which GPU is a straggler. In this way, early and precise detection of stragglers could be achieved. In this section, we investigate the temporal communication patterns of DP and PP, and determine the communication traffic model of GPU generated by DP and PP.

2.1 Communication Traffic Model for PP

Pipeline parallelism is a distributed training strategy that partitions deep neural network models into consecutive stages across multiple compute nodes, primarily employed when model size exceeds single-server memory capacity (e.g., GPT-3, LLaMA). PP occurs between GPUs with same number among rails, as blue arrows shown in Fig. 1. Each GPU processes a part of layers of the model, with data transferred across stages. When there are stragglers during PP, some GPUs would experience increased idle time, reduced hardware utilization, and delays in global gradient synchronization, compromising training efficacy. Consequently, a traffic model for PP would enable effective identification of stragglers.

Let P denote the number of model parameters, and each parameter occupies 2 bytes. Let T denote the number of tokens in the dataset, and N denote the

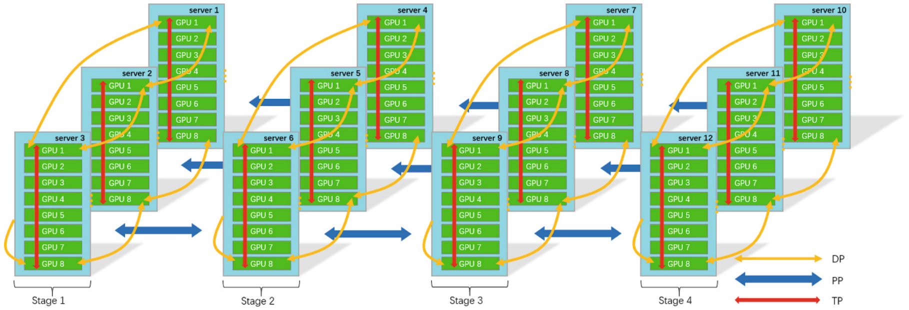

Fig. 1. DP, PP, and TP communication in LLM training

number of batches in each epoch. Assume that the training dataset is uniformly distributed across batches, thus the number of tokens per batch is $\frac{T}{N}$. Let G_{dp}, G_{tp}, and G_{pp} denote the number of GPUs used for DP, TP, and PP, respectively. During the forward propagation, the pipeline is divided into G_{pp} stages and the number of tokens allocated per GPU is $\frac{T}{N \times G_{dp}}$ as we assume that the training dataset is uniformly distributed to G_{dp} GPUs. Meanwhile, the embedding vocabulary would be transmitted among pipeline stages under PP. Let H denote the embedding dimension of the word vector and the average traffic per GPU during a PP communication is calculated as $2 \times \frac{T}{N \times G_{dp}} \times H$. Let t_0 denote the start timestamp of a training epoch, and the moment when a PP occurs could be calculated as $t_m = t_0 + b(\Delta t_{\mathrm{fwd}} + \Delta t_{\mathrm{bwd}}) + \alpha \Delta t_{\mathrm{pp}} + \beta \Delta t_{\mathrm{ppb}}$, where b is the number of completed batches, α and β are the numbers of completed pipeline stages in the forward and backward propagation of the current batch, respectively. The forward propagation timeΔt_{fwd} for each batch is $\Delta t_{\mathrm{fwd}} = \frac{2 \times \frac{T}{N} \times P}{G_{dp} \times G_{tp} \times G_{PP} \times \delta \times \tau}$, where δ is the peak FLOPS of a GPU and τ is the GPU utilization on average. The backward propagation time for each batch Δt_{bwd} is $\Delta t_{\mathrm{bwd}} = 2 \times \Delta t_{\mathrm{fwd}}$. The pipeline parallelism time interval during the forward pass is $\Delta t_{\mathrm{pp}} = \frac{\Delta t_{\mathrm{fwd}}}{G_{PP}}$, and during the backward pass is $\Delta t_{\mathrm{ppb}} = \frac{\Delta t_{\mathrm{bwd}}}{G_{PP}}$. Let $\Pi(m, t)$ represent the communication traffic model for PP described as

$$\Pi(m, t) = \begin{cases} 2 \times \frac{T}{N \times G_{dp}} \times H & \text{if } t = t_{\mathrm{m}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

2.2 Communication Traffic Model for DP

DP is also a distributed training strategy that partitions the training dataset into multiple subsets and distributes them across distinct GPUs. It is primarily applied when processing large-scale datasets within a single server becomes impractical. Each GPU maintains a complete copy of the model parameters, independently computes local gradients, and aggregates these gradients via gradient synchronization (AllReduce) to update the model parameters. DP communication occurs between GPUs with same number within the same rail, as

Assuming that the fixed error is 3km, the fixed error obeys the (0,3) normal distribution and is distributed in the X, Y and Z axes. The random error is 1km, and the fixed error obeys the (0,1) normal distribution, which is also distributed in the X, Y and Z axes. The simulation results are as follows (Figs. 3, 4, 5 and 6).

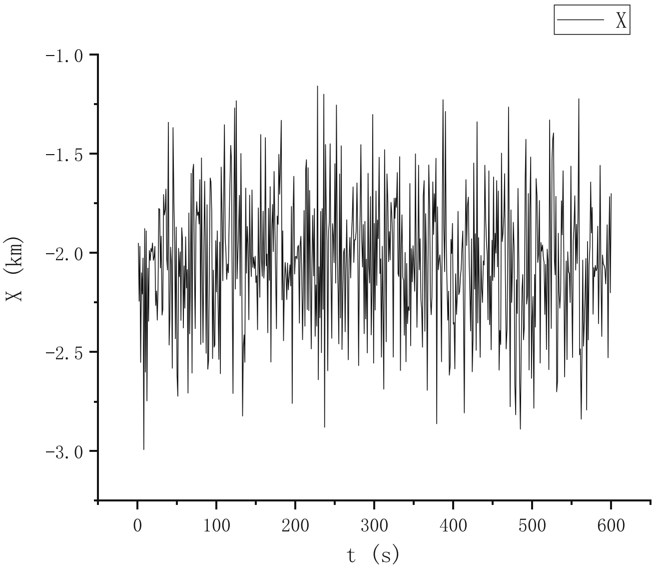

Fig. 3. X direction error

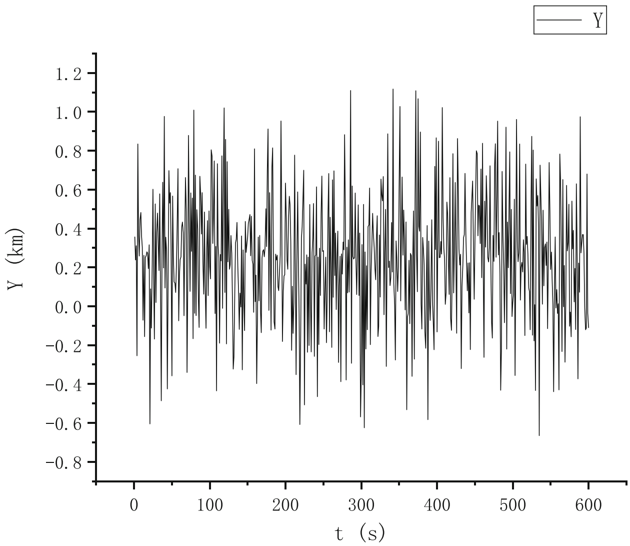

Fig. 4. Error curve of Y direction

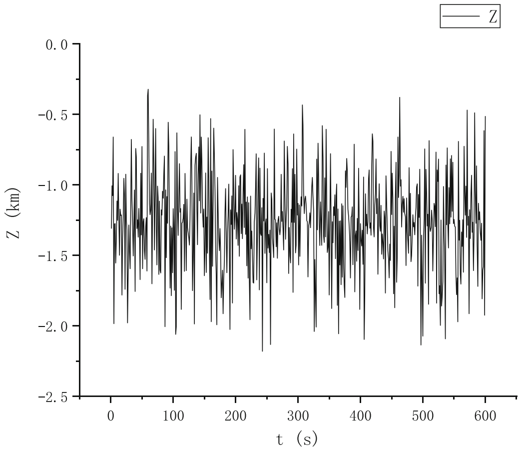

Fig. 5. Error curve of Z direction

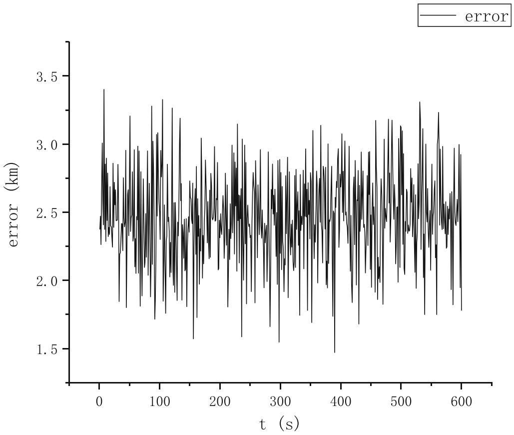

Fig. 6. Positioning error curve

4 Summary

In this paper, based on the space-based remote sensing system, a satellite dynamics model is built to simulate the motion of the satellite. The calculation results are basically consistent with the STK software simulation results. At the same time, this paper designed a set of constellation of space-based remote sensing system, constructed a simple observation error model using the working principle of space-based remote sensor, and carried out simulation calculation on the single moving target motion scene through the proposed multi satellite collaborative target positioning method. The simulation technology research carried out in this paper can evaluate the observation capability of the space-based remote sensing system, and has guiding significance for the setting of space-based remote sensors and load parameters and satellite scheduling.

References

1. Mengyu, Q., Jinlin, T., Yahu, L., Qizhi, X., Shengyang, W.: Infrared dim and small flying target recognition algorithm for space-based surveillance [J]. China Space Sci. Technol. **42**(5), 125–132 (2022)
2. Hamming, R.W., et al.: Space-based infrared system (SBIRS) for missile warning and defense. Lincoln Lab. J. **13**(2), 241–256 (2002)
3. Haiyi, C., Xinwei, Z., Chenguang, Z., Chi, X., Fan, M., Jun, D.: Overall design and technological innovation of XXX satellite [J]. China Space Sci. Technol. **40**(5), 1–9 (2020)
4. Schumacher, P.W., et al.: Overview of missile warning and space surveillance sensors. In: Proceedings of SPIE (Vol. 6239). SPIE (2006)
5. Dongxing, Y., Guangyi, S., Yukun, Z., etc: Research progress of remote sensing image scene classification [J]. Space Return Remote Sens. **45** (4), 124–138 (2024)
6. Roberts, T.J., Wright, S.A.: Next generation overhead persistent infrared (OPIR) architecture and the role of low earth orbit. In: AIAA Scitech 2020 Forum. American Institute of Aeronautics and Astronautics (2020)
7. A review on the developments and space applications of mid-and long-wavelength infrared detection technologies. Frontiers Inf. Technol. Electron. Eng. **25**(8) (2024)
8. Fan, J.-X.: Status quo and trend of infrared system and technologies for America's ballistic missile defense system [J]. Infrared Laser Eng. (5) (2006)
9. Canavan, G.H.: Missile defense for the 21st Century [M]. The Heritage Foundation, 2003. From Earth observation satellite to earth observation brain. Geomatics Inf. Sci. Wuhan Univ. **42**(2) (2017)
10. Hongbo, Z.: Theory and Methods of Spacecraft Orbital Mechanics [M]. Beijing: National Defense Industry Press (2015)
11. Theories and Methods of Spacecraft Orbital Mechanics [M]. Beijing: National Defence Industry Press (2015)
12. Dezert ballistic track initialization from a biosatellite surveillance imaging system [C]. 5th International Conference on Information Fusion. 1015~1022. IEEE (2002)
13. Development status and literature analysis of China's earth observation remote sensing satellites. J. Remote Sens. **24**(5) (2020)
14. Yudong, X., Jigang, H., Boyang, L.: Research on strategy based remote sensing satellite management and control method [J]. China Space Sci. Technol. **40** (4), 97–106 (2020)
15. Psiaki, M.L.: Detection and tracking of hypersonic vehicles on reentry with space-based sensors. J. Guid. Control. Dyn. **43**(2), 217–228 (2020)

Author Index

L. Zhang et al. (Eds.): ISCTech 2025, LNNS 1824, pp. 353–354, 2027.
https://doi.org/10.1007/978-981-92-2879-9