Современные инновации, системы и технологии //
Modern Innovations, Systems and Technologies

2023; 3(2)
eISSN: 2782-2818
https://www.oajmist.com

UDC 004.932
DOI: https://doi.org/10.47813/2782-2818-2023-3-2-0213-0224

EDN: PXVGVR

DETR-crowd is all you need

Weijia Liu[1], Zheng Zishen[2], Fan Ke[3], He Kun[4], Huang Taiqiu[5], Liu Weijia[1], Ke Xianlun[6], Xu Yuming[5]

[1]*Trine University, Phoenix, USA*
[2]*Taiyuan University of Technology, Taiyuan, China*
[3]*Arizona State University, Phoenix, USA*
[4]*Illinois Institute of Technology, Chicago, USA*
[5]*Shenzhen University, Shenzhen, China*
[6]*Yunnan University, Kunming, China*

Abstract. "Crowded pedestrian detection" is a hot topic in the field of pedestrian detection. To address the issue of missed targets and small pedestrians in crowded scenes, an improved DETR object detection algorithm called DETR-crowd is proposed. The attention model DETR is used as the baseline model to complete object detection in the absence of partial features in crowded pedestrian scenes. The deformable attention encoder is introduced to effectively utilize multi-scale feature maps containing a large amount of small target information to improve the detection accuracy of small pedestrians. To enhance the efficiency of important feature extraction and refinement, the improved EfficientNet backbone network fused with a channel spatial attention module is used for feature extraction. To address the issue of low training efficiency of models that use attention detection modules, Smooth-L1 and GIOU are combined as the loss function during training, allowing the model to converge to higher precision. Experimental results on the Wider-Person crowded pedestrian detection dataset show that the proposed algorithm leads YOLO-X by 0.039 in AP50 accuracy and YOLO-V5 by 0.015 in AP50 accuracy. The proposed algorithm can be effectively applied to crowded pedestrian detection tasks.

Keywords: Computer vision; crowded pedestrian detection; DETR.

For citation: Weijia , L., Zheng , Z., Fan , K., He , K., Huang , T., Liu , W., Ke , X., & Xu , Y. (2023). DETR-crowd is all you need. Modern Innovations, Systems and Technologies, 3(2), 0213–0224. https://doi.org/10.47813/2782-2818-2023-3-2-0213-0224

INTRODUCTION

Pedestrian detection is a branch of object detection with important applications in security monitoring, intelligent driving, and traffic monitoring. In crowded pedestrian detection scenarios, the small size of pedestrian targets and the significant occlusion between them pose a challenge for detection. Specifically, crowded scenes result in missing partial feature

© Weijia L., Zheng Z., Fan K., He K., Huang T., Liu W., Ke X., Xu Y., 2023

information due to overlapping targets, and occlusion introduces noise interference and causes difficulty in effectively refining important features. Moreover, high pedestrian density can lead to low resolution and insufficient feature information, which causes missed detections.

Various methods have been proposed to address these challenges. Some methods focus on detecting pedestrian-specific body parts and features, while others use attention modules and residual networks to improve feature extraction and refinement. Additionally, some methods use multi-camera viewpoints to reduce the effects of occlusion. However, these methods still struggle with missed detections, particularly for small targets in dense crowds.

To address these issues, this study proposes three improvements to the DETR attention model. Firstly, an improved EfficientNet backbone network is used for feature extraction, which incorporates a channel spatial attention module to efficiently extract and refine important channel and spatial information in feature maps. Secondly, a deformable attention encoder is used to naturally aggregate multi-scale features and improve the detection of small targets. Thirdly, the Smooth-L1 combined with GIOU is used as the loss function during training to improve efficiency and convergence to higher accuracy.

Experimental results on the Wider Person crowded pedestrian detection dataset show that the proposed improved DETR algorithm has strong detection capabilities for crowded pedestrian scenarios with small targets and occlusions.

DETR-CROWD: IMPROVED DETR DETECTION ALGORITHM

The proposed improved DETR algorithm consists of three parts: an improved attention detection module for pedestrian detection, a neck network that preprocesses and outputs multi-scale feature maps from the backbone network's output features, and an improved EfficientNet backbone network for feature extraction. During pedestrian detection, the backbone network extracts the features from the input RGB image and passes the features from the 6th, 7th, and 8th layers of the network to the neck network. The neck network converts the obtained feature maps into multi-scale feature maps with a channel size of 256 and a fixed size. The attention detection module adds learnable positional encodings to the multi-scale feature maps before sending them to the Deformable Transformer Encoder for attention encoding. The decoding is performed by the Transformer Decoder, which outputs detection anchor boxes. The network architecture of the improved DETR is shown in Figure 1.

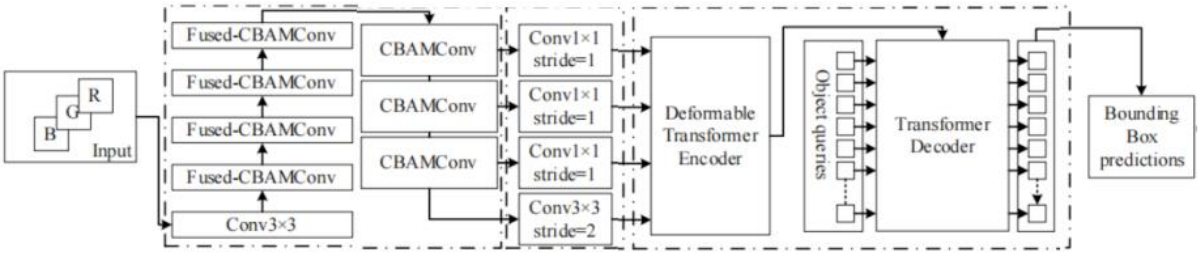


Figure 1. DETR-crowd network structure diagram.

The improved attention detection module

DETR utilizes an attention detection module for detection result output, but the module struggles to effectively utilize multi-scale feature maps that contain a significant amount of information on small targets, leading to low detection efficiency for small pedestrian targets. To address this issue, this study proposes an improvement to the DETR attention detection module by incorporating a Deformable Transformer Encoder. The structure of the DETR attention detection module is shown in Figure 2.

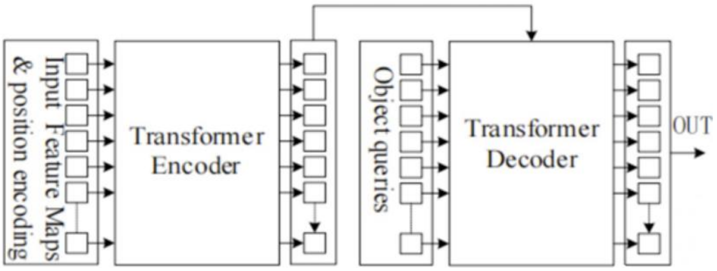


Figure 2. DETR transformer detector.

Transformer Encoder

The attention encoder is a coding structure of the attention detection module that calculates the attention weight between each pixel point and other pixel points in the feature map, creating a global feature map. It is combined with the attention decoder to give the model global modeling capabilities, which can convert the object detection problem into a set prediction problem without outputting redundant prediction boxes. By using the attention detection module, the model focuses on specific important features during the output of detection results, alleviating the interference of noise on object detection and allowing the model to complete the detection of occluded targets even in the absence of some target features. Therefore, the model with the attention detection module is suitable for occluded target detection tasks.

The deformable attention encoder has feature encoding capability and can retain global modeling ability when combined with the decoder. Unlike the attention encoder, it only calculates the attention weight between the sampling point and its nearby pixels (learned by the model), reducing the computational complexity while maintaining performance. The deformable encoder can naturally aggregate different scale feature maps, allowing the model to effectively use the multi-scale feature maps extracted by the backbone network without using the feature pyramid network structure, reducing the loss of semantic information for small targets in the downsampling process, and improving the detection performance of small objects [1-4]. In this paper, the deformable attention encoder is used to improve the attention detection module, allowing the model to effectively use multi-scale feature maps. The attention detection module with the deformable attention encoder is shown in Figure 3.

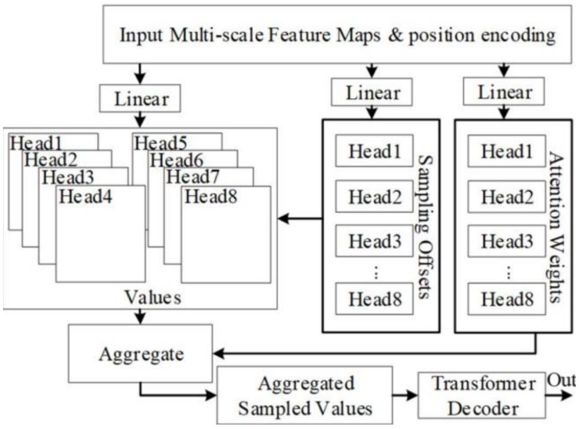


Figure 3. Attention detector with deformable attention encoder.

In Figure 3, the multi-scale feature maps output by the Neck layer are added with learnable positional encoding and then passed to the deformable attention encoder for attention weight updating. The multi-scale feature maps are transformed by three fully connected layers into pixel attention weights $W'mx$, attention offsets $\Delta Pmlqk$, and attention weight coefficients $Amlqk$, respectively. The attention offsets represent the positional displacement between the current reference point and its neighboring pixels, while the attention weight coefficients indicate the attention weight values associated with these pixels. When the deformable attention encoder focuses on a reference point, it calculates the offset of all the pixels related to the current reference point through the attention offsets, and then updates the attention weight value of the current reference point by combining the attention weight and attention weight coefficients of these pixels [5-7]. This completes the global attention weight encoding for all pixels. The feature map with completed global attention weight encoding is directly outputted to the decoder to obtain the detection results. The multi-head deformable attention is defined in equation (1).

$$MSDeformAtten(z_{q,}\hat{p}_q, x) = \sum_{m=1}^{8} W_m \left[\sum_{l=1}^{L} \sum_{k=1}^{K} A_{mqlk} \cdot W'_m x(\varphi_1(\hat{p}_q) + p_{mqlk})\right] \quad (1)$$

In the equation above, z_q represents the original input feature, $p\hat{}_q$ represents the normalized coordinates of the current reference point, *x* represents the feature map index, W_m represents the multi-head attention, where *m* is the attention index (in this paper, *m* has a maximum value of 8), *l* represents the dimension index of the feature pyramid, and *k* represents the current sampling point index.

Improved backbone network

In order to provide the efficient multi-scale feature maps for the DETR algorithm using the deformable attention encoder, this paper replaces the original ResNet-50 backbone network with an improved EfficientNet backbone network for feature extraction.

Basic network ResNet-50 backbone network

DETR, a Transformer-based end-to-end object detection method, uses the classic backbone network ResNet-50. While ResNet-50 introduced residual structures to alleviate the problem of gradient vanishing during training of deep neural networks, it increases the network's performance by adding more layers, resulting in a large number of parameters and a deep stack of layers. This makes it difficult to effectively extract features from partially occluded objects with limited information and provide efficient multi-scale feature maps for subsequent encoding networks.

Basic network EfficientNet backbone network

EfficientNetB0-B7 backbone networks were obtained through Neural Architecture Search (NAS), resulting in a series of efficient networks with significantly fewer parameters compared to ResNet-50. Even the smallest EfficientNet-B0 backbone network has less than a quarter of the parameters of ResNet-50 and achieves a TOP1 accuracy in ImageNet classification task that is about 1% higher than ResNet-50. The core module of the EfficientNet backbone network is the MBConv module, which combines depthwise separable convolution and the squeeze-and-excitation (SE) module and is activated by the Swish function. The MBConv module efficiently purifies important channel features during the feature extraction stage, enabling the EfficientNet backbone network to extract features more effectively [7-10].

Channel Spatial Attention Module

In practical detection tasks, spatial information between objects is crucial for detecting occluded objects. However, during the convolution process, the feature map size gradually decreases, resulting in a loss of spatial information between objects. To improve the detection performance of occluded objects by purifying spatial information, this study replaces the compression and excitation modules in the MBConv module with a Channel Spatial Attention (CSA) module [10-14].

The CSA module is a lightweight convolution attention module that sequentially calculates attention maps for both channel and spatial dimensions. The attention maps are then multiplied with the feature maps for adaptive feature optimization. In a simulation experiment conducted by Zhang Chenjia et al., using five identical radio source signals as the dataset, seven different attention modules (ECA-Net, SE-Net, SK-Net, ResNeSt, CBAM, DANet, and PAFNet) were compared for classification recognition under the same network conditions, and the CSA module was found to perform the best. The structure of the CSA module is shown in Figure 4.

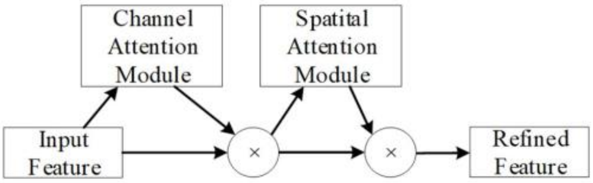


Figure 4. Convolutional block attention module.

EXPERIMENTAL RESULTS AND ANALYSIS

Experimental Dataset

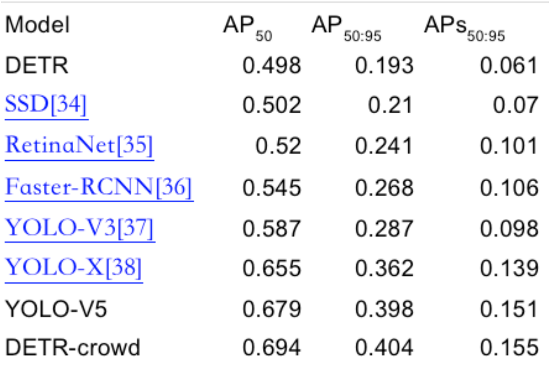
This study validated the proposed model on the publicly available Wider-Person [31] dataset for detecting pedestrians in crowded outdoor scenes and the USC pedestrian detection dataset, which mostly consists of surveillance videos and contains a small number of occluded targets.

The Wider-Person dataset contains 13382 images with a total of 400,000 different levels of occlusion on human targets. This study randomly selected 9000 images with provided labels and divided them into an 8:2 ratio for training and validation experiments. The USC pedestrian detection dataset contains 358 images with a total of 816 pedestrian targets, with a small number of occluded targets. This study re-annotated all 358 images and divided them into an 8:2 ratio for training and validation experiments [15-18].

As the model with the transformer encoder requires a large amount of training resources and takes longer to converge, the ablation experiments were performed on the smaller USC pedestrian detection dataset to validate the effectiveness of each proposed improvement module. The cross-validation experiments were conducted on the larger Wider-Person dataset to validate the superior performance of the proposed algorithm in detecting small and occluded targets.

Horizontal comparison experiment

This study conducted comparative experiments between the proposed algorithm and some commonly used pedestrian detection models. Each model was trained for 200 epochs on the Wider-Person crowded pedestrian dataset [19-23]. Since the Wider-Person dataset contains a large number of occluded and small pedestrian targets, the AP50, AP50:95, and APs50:95 were used as performance metrics. All experimental groups used the Adam optimizer with a learning rate of 0.0001. The experimental results are shown in table on Figure 5.



Model	AP_{50}	$AP_{50:95}$	$APs_{50:95}$
DETR	0.498	0.193	0.061
SSD[34]	0.502	0.21	0.07
RetinaNet[35]	0.52	0.241	0.101
Faster-RCNN[36]	0.545	0.268	0.106
YOLO-V3[37]	0.587	0.287	0.098
YOLO-X[38]	0.655	0.362	0.139
YOLO-V5	0.679	0.398	0.151
DETR-crowd	0.694	0.404	0.155

Figure 5. Performance comparison test results.

According to this table, in dense pedestrian detection scenarios, the proposed algorithm's regular detection accuracy and small target detection accuracy are higher than that of commonly used pedestrian detection algorithms such as YOLO-X and YOLO-V5.

To better compare the detection performance of the proposed algorithm and the original DETR model, this study visualized the detection results of the DETR model and the proposed algorithm in crowded pedestrian scenes in Figure 6. The left side shows the detection results of the DETR model, where there are many missed detections. The right side shows the detection results of the proposed algorithm, with fewer missed detections [24-28]. The proposed algorithm can be well applied to crowded pedestrian detection tasks.

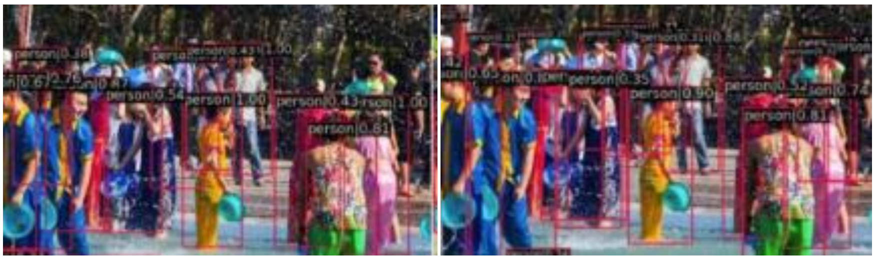


Figure 6. Comparison of testing results.

DISCUSSION

The proposed algorithm achieved higher detection accuracy than commonly used pedestrian detection algorithms such as YOLO-X and YOLO-V5 on the Wider-Person crowded pedestrian detection dataset. Although algorithms using attention detection structures have advantages in detecting occluded and small-scale targets, they require more computational resources. On the single GPU experimental platform and dataset used in this study, training the improved DETR algorithm for 200 epochs required approximately 145.6 hours, which is 5.4 times longer than the time required for training the YOLO-V5 algorithm. Future research should focus on reducing the training resources required for attention detection algorithms and improving their training efficiency.

CONCLUSION

This study proposed an improved object detection algorithm based on DETR. The use of an improved attention detection module enabled the model to aggregate multi-scale feature maps and effectively improve the detection ability of small and occluded targets [29-31]. An improved EfficientNet backbone network was used as the feature extraction network, enhancing the model's ability to extract and refine important features, thereby increasing the detection accuracy. During training, a loss function composed of Smooth-L1 and GIOU was used to further converge the model to higher accuracy [32].

REFERENCES

[1] Hu W., Liu X., Xie Z. Ore image segmentation application based on deep learning and game theory. World science: problems and innovations. 2022: 71-76.

[2] Zhouyi X., Weijun H., Yanrong H. Intelligent acquisition method of herbaceous flowers image based on theme crawler, deep learning and game theory. Kronos. 2022; 7(4(66)): 44-52.

[3] Xie Z., Hu W., Fan Y., Wang Y. Research on multi-target recognition of flowers in landscape garden based on GHOSTNET and game theory. Development of science, technology, education in the 21st century: topical issues, achievements and innovations. 2022: 46-56.

[4] Song Y., Chen B., Liu X., Weijun H., Xiangyu X., Yuqi Y. Audio and video editing system design based on OpenCV. Informatics. Economics. Management. 2022; 1(2): 0101-0120.

[5] Xiaomin L., Yuehang S., Borun C., Xiaobin L., Weijun H. A novel deep learning based multi-feature fusion method for drowsy driving detection. Industry and agriculture. 2022: 34-49.

[6] Hu W., Zheng T., Chen B., Jin J., Song Y. Research on product recommendation system based on deep learning. Basic and applied scientific research: current issues, achievements and innovations. 2022: 116-124.

[7] He W., Hu W., Yang Y., Shen H., Wu Y., Song Y., Liu X. Improved left- and right-hand tracker using computer vision. Student scientific research. 2022: 21-29.

[8] Xie Z., Hu W., Zhu J., Li B., Wu Y., He W., Liu X. Left and right-hand tracker based on convolutional neural network. Topical issues of modern science of education. 2022: 61-67.

[9] He W., Hu W., Wu Y., Sun L., Liu X., Chen B. Development history and research status of convolutional neural networks. Student Scientific Forum. 2022: 28-36.

[10] Yuan C., Liu X., Zhang Z. The Current Status and progress of Adversarial Examples Attacks. Proceedings of the 2021 International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE; 2021:707-711.

[11] Liu X., Liu W., Yi S., Li J. Research on Software Development Automation Based on Microservice Architecture. Proceedings of the 2020 International Conference on Aviation Safety and Information Technology. 2020: 670-677.

[12] Liu X., Xie X., Hu W., Zhou H. The application and influencing factors of computer vision: focus on human face recognition in medical field. Science, education, innovations: topical issues and modern aspects. 2022: 32-37.

[13] Shen G., He K., Jin J., Chen B., Hu W., Liu X. Capturing and analyzing financial public opinion using NLP and deep forest. Scientific research of students and pupils. 2022: 66-71.

[14] Chen B., Song Y., Cheng L., He W., Hu W., Liu X., Chen J. A review of research on machine learning in stock price forecasting. Science and modern education: topical issues, achievements and innovations. 2022: 56-62.

[15] He K., Song Y., Shen G., He W., Liu W. Based on deep reinforcement learning and combined with trends stock price prediction model. Topical issues of modern scientific research. 2022: 156-166.

[16] Ou S., Gao Y., Zhang Z., Shi C. Polyp-YOLOv5-Tiny: A Lightweight Model for Real-Time Polyp Detection. Proceedings of 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA). IEEE; 2021; 2:1106-1111.

[17] Jiajun J., Wanting Y. The use of computer vision technology in intelligent agricultural machinery. Science and education: preserving the past, creating the future. 2022: 9.

[18] Xu J., Chen J., Li B., Li X. Analysis of the bargaining game and buyer's benefit model. Proceedings of the Conference on Modern Scientific Research: Current Issues, Achievements and Innovations. 2022: 91-96.

[19] Wu J., Lee P. P., Li Q., Pan L., Zhang J. CellPAD: Detecting performance anomalies in cellular networks via regression analysis. Proceedings of 2018 IFIP Networking Conference (IFIP Networking) and Workshops. IEEE; 2018: 1-9.

[20] Sun Q., Zhao C., Petrosian O., Li Y. Power allocation in wireless cellular networks: stochastic algorithm-based approach. Management processes and sustainability. 2022; 9(1): 357-362.

[21] Li Q., Bai M., Hu W., He J., He K., Meng L., Xu S. Examination of the optimal solution problem for Go based on Monte Carlo algorithm. Proceedings of XXII International Scientific and Practical Conference.2023, March 26; Melbourne; 2023: 34-41.

[22] Qingyuan L., Wenke D., Weijun H., Kun H., Weijia L., Yanyou W., Penghui L, Alina R. OCTAVE programming for numerical analysis of free vibration of multi-degree-of-freedom structures. Industry and agriculture. 2023: 45-53.

[23] Weijun H., Weilong H., Jipan H., Kun H., Jialun P. Game theoretic method and optimization of electric power companies. Industry and agriculture. 2023: 21-62.

[24] He W., Liu W., He K., Wu Ya. Aircraft target detection vision system based on OpenCV: Proceedings of the II International Scientific and Practical Conference on Fundamental and applied science: topical issues of theory and practice. Penza; 2023: 62-69. EDN OAVTIS.

[25] Jiajun J., Yuehang S., Geya S., Borun C., Kun H., Weijia L., Weijun H. The use of a discrete differential algorithm for deep learning has been the focus of research into the technologies around visual target tracking. Industry and Agriculture. 2022: 6678.

[26] Yu X., Bo L., Xin C. Low light combining multiscale deep learning networks and image enhancement algorithm. Modern Innovations, Systems and Technologies. 2022; 2(4): 0215-0232. https://doi.org/10.47813/2782-2818-2022-2-4-0215-0232

[27] Fan K., Liu W., He K., Wang Z., Ou S., Wu Y. Review: the application of artificial intelligence in distribution network engineering field. Informatics. Economics. Management. 2023; 2(1): 0210-0218. https://doi.org/10.47813/2782-5280-2023-2-1-0210-0218

[28] Ke F., Chen-Yu H., Weijia L., Kun H., Bin S., Yanyou W. Research on computer vision application in industry field: focus on distribution network engineering. Modern Innovations, Systems and Technologies. 2023; 3(1): 0401-0409. https://doi.org/10.47813/2782-2818-2023-3-1-0401-0410

[29] Yuqi Y., Wanting Y., Xin L., Jie X., Lihua L. Studying electronic blood pressure monitor digital recognition algorithm based on computer vision and design. Modern Innovations, Systems and Technologies. 2022; 2(4): 0264–0277. https://doi.org/10.47813/2782-2818-2022-2-4-0264-0277

[30] He K., Zhang L., Liu W., Fan K., Song P., Wu Y. The Review of Application of Deep Detection Network in Distribution Network Engineering field: Proceedings of the Conference on Modern Strategies and Digital Transformations of Sustainable Development of Society, Education and Science; 2023: 139-148.

[31] Fan K., Liu W., He K., Wang Z., Ou S., Wu Y. Review: the application of artificial intelligence in distribution network engineering field. Informatics. Economics. Management. 2023; 2(1): 0210–0218. https://doi.org/10.47813/2782- 5280-2023-2-1-0210-0218

[32] Zheng Z., Bai M., Hu W. Design and implementation of online bookstore: Proceedings of the XXVIII International Part-time Scientific and Practical Conference. Moscow, Empirya; 2023: 62-83.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Weijia Liu, Trine University, Phoenix, United States

Fan Ke, Arizona State University, Phoenix, United States

Liu Weijia, Trine University, Phoenix, United States

Zheng Zishen, Taiyuan University of Technology, Taiyuan, China

He Kun, Illinois State University, Chicago, United States

Ke Xianlun, Yunnan University, Kunming, China

Huang Taiqiu, Shenzhen University, Shenzhen, China

Xu Yuming, Shenzhen University, Shenzhen, China

Статья поступила в редакцию 23.05.2023; одобрена после рецензирования 29.05.2023; принята к публикации 30.05.2023.

The article was submitted 23.05.2023; approved after reviewing 29.05.2023; accepted for publication 30.05.2023.