Deep Image Keypoint Detection Using Cascaded Depth Separable Convolution Modules

Rui Deng
Network Information Center, Jilin Vocational College of Industrial and Technology, Jilin, 132013, China

***Abstract*—Depth images have become an important data source for human bone keypoint detection due to their three-dimensional information. To optimize the efficiency of keypoint detection in depth images, a depth image keypoint detection model that combines cascaded depth separable convolution modules is constructed. The model first performs data cleaning and preprocessing on the image, replacing traditional convolutional layers with depthwise separable convolutional modules. The Faster OpenPose network is introduced to replace the traditional convolutional network structure with the lighter MobileNetV1 for detecting keypoints in the image. When the dataset size was 4000, the Faster OpenPose model had an accuracy of 0.97 and a mean square error of 0.03. The recognition rates for four different images were 0.91, 0.87, 0.89, and 0.93, respectively. The processing times were 0.32, 0.31, 0.28, and 0.27, respectively. The method of depth image keypoint detection combined with cascaded depth separable convolution modules has good practicality and excellent detection performance for various images, providing new ideas for future keypoint detection technology research.**

Keywords*—Depth image; DWCA; key point detection; OpenPose; cascade depth

I. Introduction

Deep Image (DI) keypoint detection is important in 3D object reconstruction, facial recognition, gesture recognition, and robot navigation. These tasks typically require accurate extraction of key points of objects or scenes from DI for further analysis or processing. However, traditional methods rely heavily on manually designed feature extractors, which often exhibit significant limitations in complex scenes. Compared with traditional two-dimensional images, DI provide rich 3D spatial information, making the understanding and analysis of the scene more accurate. In human pose estimation, DI can more accurately capture the skeletal structure of the human body, reducing the impact caused by lighting changes, occlusion, and other issues [1]. However, how to effectively extract key points of the human body from DI remains a challenge. The current keypoint detection technology, such as the OpenPose based network architecture, can achieve relatively accurate pose estimation on two-dimensional images. However, there are still certain limitations in DI processing, such as high computational complexity of the model, sensitivity to background information interference, and the need to improve keypoint localization accuracy [2]. The limitation of existing models is that they cannot effectively handle noise and background interference in depth images, resulting in a decrease in the accuracy of keypoint detection. Many traditional models rely on high-quality input data, and in practical applications, depth images often have varying degrees of noise, which can affect model performance. In addition, existing models often lack sensitivity to changes in human posture and complex background responses, limiting their application in multi human environments. The computational complexity of the model is also a problem, and classical convolutional neural networks may face performance bottlenecks when processing real-time data. Therefore, this study proposes a DI keypoint detection method that combines cascaded depth separable convolution modules. This method replaces the traditional VGG-19 network structure with a more lightweight MobileNetV1 by introducing the Faster OpenPose network, and designs a Depthwise Separable Convolutional Module (DWCA) based on this to reduce computational complexity while maintaining the model's prediction capability. The main reason for choosing a model based on bilateral filtering and Faster OpenPose network is its excellent noise processing ability and efficient 3D information preservation. Bilateral filtering can effectively remove noise in depth images while preserving edge features, ensuring the accuracy of keypoint localization. The Faster OpenPose network significantly improves computational efficiency and meets real-time detection requirements by replacing VGG-19 with MobileNetV1. The introduction of depthwise separable convolution modules and feature fusion mechanisms enhances the ability to extract key features and improves the performance of multi person pose estimation. The model can effectively extract regions of interest related to the human body, filter out irrelevant backgrounds, and further improve detection accuracy and efficiency. The contribution of the research lies in the proposed bilateral filtering based deep image processing model, which effectively removes noise from deep images while preserving key edge features, significantly improving the accuracy of keypoint localization. Secondly, by introducing the Faster OpenPose network and replacing the traditional VGG-19 with MobileNetV1, the computational efficiency of the model has been improved, making real-time keypoint detection possible and adapting to the needs of various application scenarios. The deep separable convolution module based on feature fusion introduced in the study enhances the ability to extract important features from depth images and improves the performance of keypoint detection. In addition, the model effectively extracts regions of interest related to the human body, filters out irrelevant background information, and further improves the accuracy and efficiency of keypoint localization. The innovation lies in optimizing the performance of DI keypoint detection by improving the OpenPose network structure and introducing more efficient convolution modules. The research aims to provide new technological paths for the field of Computer Vision (CV) and offer better solutions for

keypoint detection in practical applications.

The research content is as follows: An examination of the research topics of other scholars in the field is given in Section II. An overview of the principal methodologies employed in Section III. The results of the model experiment is presented in Section IV. Discussion is given in Section V and finally, the paper is concluded in Section VI.

II. Related Works

Under the development of computer technology, CV is becoming increasingly important. Zhou K et al. built an improved codebook pattern model to improve the processing efficiency of rapid action vide. The combination of this method with the CV approach had the potential to enhance the accuracy of feature recognition in fast-action sequences, facilitate the effective processing of fast-motion videos, and improve the feature recognition effect [3]. Zhou H et al. discovered that despite the current advancements in video surveillance, autonomous vehicles, and other related fields, there was still a significant opportunity for further development in predicting the future trajectory of pedestrians. A spatiotemporal graph neural network based on attention interaction perception had been proposed, which demonstrated effective capability [4]. Lee J et al. put forth a new YOLO model to handle the real-time object detection problem of YOLO. This architecture maintained YOLO's high accuracy and ease of use. The proposed method model could effectively solve problems related to real-time processing [5]. Jiang X et al. proposed a blockchain based model sharing method to address the issues of autonomous driving object detection. This method combined mobile edge computing technology and YOLOv2 model to reduce regional differences, and its effectiveness and reliability were superior to the reference model [6].

Yang Y et al. constructed a traffic recognition method with deep convolutional neural networks, which was able to detect and classify the input images, thereby obtaining more clear traffic information. The algorithm demonstrated superior accuracy in traffic image classification, offering a more optimal solution for smart traffic monitor [7]. Chen D et al. proposed an underwater ship detection model with optimized YOLOv3, which enhanced feature extraction capabilities in various environments by introducing an attention module. This model had good recognition and detection capabilities, proving its superiority in ship detection in water transportation [8]. Tumrani et al. proposed a decentralized and multi-attribute learning network, which adopted a vehicle keypoint detection model based on local attention for regions with more DIcriminative information. This method could improve the ability and robustness of vehicle recognition [9].

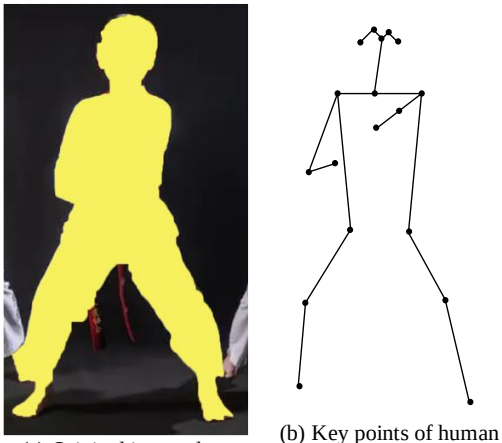
In conclusion, a substantial body of research has been conducted in the field of computer imaging, yielding notable findings, but have not conducted more in-depth studies on DI. Moreover, various studies have focused on specific domain problems and rely on specific datasets, resulting in insufficient generalization ability of models in other datasets or practical applications. This study proposes a DI keypoint detection method that combines DWCA. This method replaces the traditional VGG-19 network structure with a more lightweight MobileNetV1 by introducing the Faster OpenPose network, and uses a DWCA on this basis to reduce computational complexity while maintaining the model's prediction capability.

III. Methods

In the first section, a DI processing model based on Bilateral Filtering (BF) is proposed to address the issue of noise in DI. In the second section, OpenPose network is adopted to detect keypoints in images and improve the model for defects.

A. DI Processing Model Based on BF

A DI is an image used to represent the distance from each pixel in a scene to the camera. Unlike traditional two-dimensional images, it contains three-dimensional information in the scene and can be used to more accurately understand and analyze the spatial structure of the scene [10-11]. An image is typically a two-dimensional matrix consisting of three color channels, with the value of each channel representing the color intensity of that pixel. Each pixel in a color image provides color information that can describe the appearance of objects in the scene. DI is a two-dimensional matrix, but each pixel value represents the distance from that pixel to the camera or sensor. DI reflects the 3D structural information of the scene, not the color, thus protecting privacy. The study extracts key points of human bones, as shown in Fig. 1.



(a) Original image data (b) Key points of human skeleton

Fig. 1. Extraction of key points in the human body.

In Fig. 1, firstly, the resolution of the image is converted, and secondly, public tools are used to process the data. The joints of the human body on each screen are estimated, and the coordinates of each joint point are recorded. The DI body surface keypoint localization algorithm mainly consists of six steps, and its process is shown in Fig. 2.

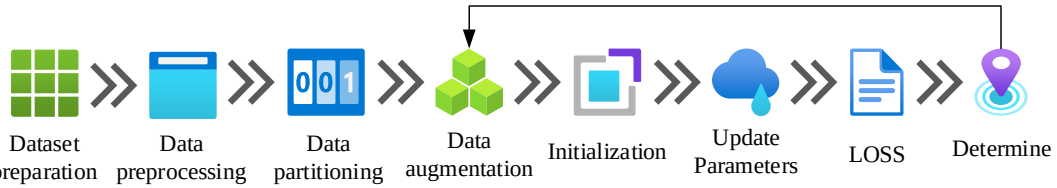


Fig. 2. Body surface key point localization process.

In Fig. 2, during the model training process, the dataset is first prepared, including collecting, organizing, and organizing data to provide a foundation for model training [12-13]. In the data preprocessing stage, the data is cleaned, normalized, and processed to ensure data quality and consistency. Subsequently, the dataset is partitioned into three sets, thereby ensuring that the model exhibits the requisite capacity for generalization. Then, the dataset is augmented to enhance the discrepancy of the data and reinforce the model's robustness. After the model parameters are initialized, the model begins to be trained; The error predicted by the loss function model is evaluated and updated based on the parameters of the error model. Then the loss function is recalculated and checked for convergence. If the loss function converges, that is, the error reaches a stable low level or no longer significantly decreases, the training process ends; otherwise, the model continues to iterate and update parameters until convergence conditions are reached. During the model testing process, the trained model parameters are first loaded, and then the test dataset is preprocessed to ensure consistency with the training data [14]. The preprocessed test data is input to generate a heatmap of key points. Next, the specific coordinates of the key points are extracted as the final output of the model. The performance of this process model on new data can be measured by the accuracy of key points that need to be ensured, and ultimately the entire testing process ends.

In the collection of DI, a lot of background information will also be collected. These background information include objects, walls, floors, etc. that are unrelated to the human body, and these additional pieces of information can interfere with keypoint localization tasks. It is necessary to use appropriate methods to filter out unnecessary background information and focus on processing Region of Interest (ROI) related to the human body [15-16]. The method's importance is to use breadth first search to identify foreground objects, and filter out background parts based on depth and height information to obtain background removed DI. After extracting the ROI, it still cannot meet the requirements, so median filtering is used to process the image, and its expression is illustrated in Eq. (1).

$$depth_{median}(x,y) = median_{(s,t)\in N(x,y)}[depth(s,t)] \quad (1)$$

In Eq. (1), $depth_{median}(x,y)$ represents the depth value after median filtering operation; $median$ represents the function for calculating the median; N is the neighborhood of the pixel; $depth(s,t)$ is the depth value of the pixel at position (s,t) [17-18]. Although median filtering can remove noise, the details of the image are also severely lost. Therefore, the BF method is used to smooth the image, and its expression is shown in Eq. (2).

$$depth_{double}(x,y) = \sum_{(s,t)\in N(x,y)} W(s,t) \times depth(s,t) \quad (2)$$

In Eq. (2), $depth_{double}(x,y)$ represents the depth value; $W(s,t)$ represents the normalized weight, and the aforementioned expression is demonstrated in Eq. (3).

$$W(s,t) = exp\left(-\frac{(s-x)^2+(t-y)^2}{2\sigma_d^2} - \frac{(depth(s,t)-depth(x,y))^2}{2\sigma_r^2}\right) \quad (3)$$

In Eq. (3), σ_d and σ_r respectively represent the parameters of spatial distance and pixel difference size. Before preprocessing, normalization is performed, and the weights of the neighborhood are first calculated, as shown in equation (4).

$$sum = \sum_{(s,t)\in N(x,y)} W(s,t) \quad (4)$$

In Eq. (4), sum is the sum of weights in the neighborhood. Then, the weights of each pixel are equal to the index value divided by the sum of weights, as expressed in Eq. (5).

$$W(s,t) = W(s,t) / sum \quad (5)$$

In Eq. (5), $W(s,t)$ represents the neighborhood weight, and the image after BF is shown in Fig. 3.

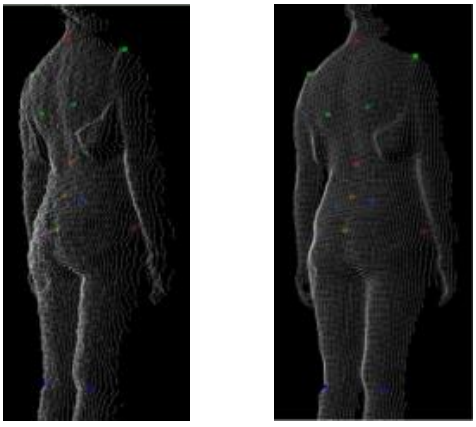


Fig. 3. Comparison of key points before and after BF.

In Fig. 3, the BF operation can effectively remove image noise while preserving the edge features of the image. BF can reduce noise interference in the image, making the ROI clearer and improving the accuracy of the Faster OpenPose network in keypoint detection. Through BF, the background information in the image is smoothed, while the key points and edge features related to the human body are preserved, promoting to focus on identifying the key points of the human body.

B. DI Keypoint Detection Model Combined with Cascaded DWCA

After completing the data cleaning, OpenPose network is adopted to detect key points in the image. OpenPose is a powerful open-source library specifically designed for real-time detection and estimation of keypoints on multiple people's bodies, faces, hands, and feet. The core process includes extracting features from input images or video frames, generating part affinity fields and keypoint heatmaps, and connecting these keypoints through post-processing to form a complete human pose skeleton [19-20]. The classic OpenPose has obvious flaws, so research has improved OpenPose and proposed a Faster OpenPose. VGG-19 in OpenPose has been replaced with MobileNetV1, improving the efficiency of the model. The DWCA based on feature fusion mechanism is introduced, and its structure is shown in Fig. 4.

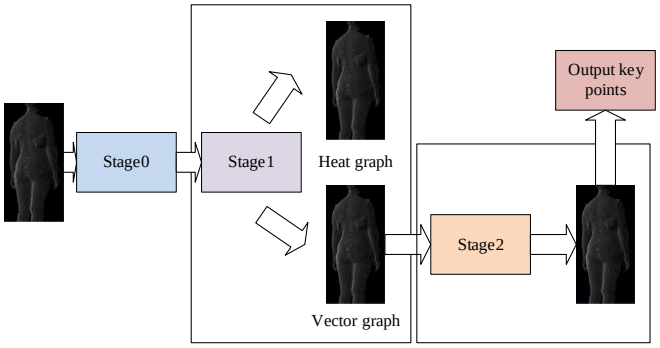


Fig. 4. Faster OpenPose structure.

In Fig. 4, the first DI is received as input, which captures the distance information between the object and the camera, providing a three-dimensional representation of the scene. Next, Stage 0 is responsible for feature extraction, processing the input DI through convolutional neural networks to extract relevant features that can be used for subsequent pose estimation. In Stage 1, heatmaps are generated, which represent the probability distribution of key point positions in the image. Partial correlation vector maps are generated, encoding the directions and associations between different body parts, helping to determine the connection relationships between detected joint points and forming a complete skeleton [21-22]. Then it enters Stage 2, where the heatmap interpolation is amplified to match the original resolution of the input DI before the final output. This step ensures that the detected keypoints are aligned correctly with the original image size. Finally, the keypoints are output, which represent the coordinates of the body joints in the input image and can be used for further tasks. In the Stage0 structure, an alternating structure of DWCA and regular modules is adopted, as shown in Fig. 5.

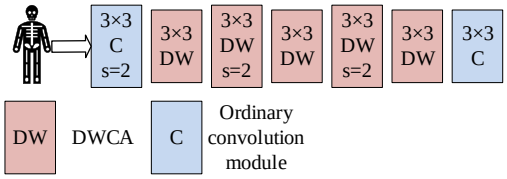


Fig. 5. Schematic diagram of stage0 structure.

In Fig. 5, it contains a regular convolution module and DWCA. In the initial stage of the process, the spatial dimension of the input image is reduced by a factor of two. Next, multiple DWCAs are adopted. These modules divide convolution operations into two steps: In deep convolution, a distinct convolution kernel is applied to each input channel. In contrast, point-by-point convolution employs a 1×1 convolution to integrate the channels of the deep convolution output. This structure maintains the accuracy of the model while reducing computational complexity. The entire MobileNetV1 network gradually increases the number of channels, ultimately generating a smaller high-dimensional Feature Map (FM) for subsequent classification or other tasks. In the process of deep convolution, for an input FM, the computational complexity of traditional convolution is shown in Eq. (6).

$$J = h' \times w' \times D_{out} \times (k \times k \times D_{in}) \tag{6}$$

In Eq. (6), $h' \times w' \times D_{out}$ represents the size of the output FM; $k \times k$ represents the size of the filter. In deep convolution, the convolution kernel only acts on each input channel, and its computational complexity is shown in Eq. (7).

$$J_m = h' \times w' \times D_{in} \times (k \times k) \tag{7}$$

Then point by point convolution is used to mix all input channels and generate an output FM, which is expressed as Eq. (8).

$$J_n = h' \times w' \times D_{in} \times D_{out} \tag{8}$$

Stage1 is the initialization stage, Stage2 is the refinement stage, and their structures are shown in Fig. 6.

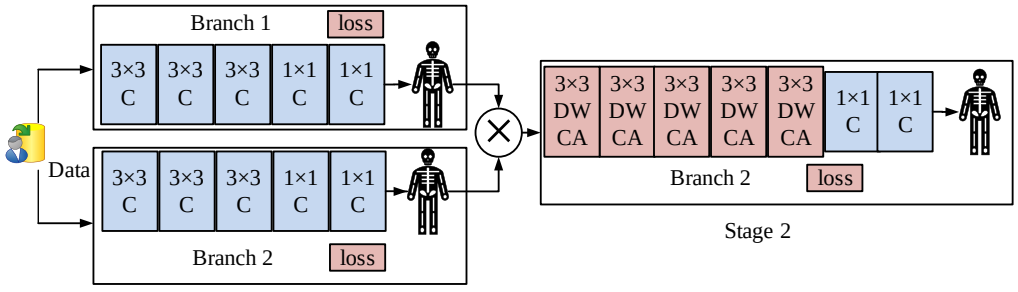


Fig. 6. Structure of stage1 and stage2.

In Fig. 6, Stage 1 includes Branch 1 and Branch 2. Branch 1 adopts multi-layer convolution operations, including 3×3 standard convolution and 1×1 convolution. After these convolution operations, the FM's size is output, and then the loss function is used to calculate the loss. Branch 2 also uses a series of convolution operations, but the output FM's size is h×w, and the loss is calculated through another loss function. The output S1 of Stage 1 is a combination of the outputs of two branches, which will be passed on to the next Stage 2. In Stage 2, the complexity of the network further increases, and Branch 2 performs different operations in Stage 2. DWCA is introduced in branch 2, and the final FM size is output. The output of this branch is also calculated through a loss function. The final output of Stage 2 is represented by S2. The loss function in Stage 1 is shown in Eq. (9).

$$\begin{cases} L_1 = f_1(S_1^1) \\ L_2 = f_2(S_2^2) \end{cases} \tag{9}$$

In Eq. (9), L_1 and L_2 represent the loss functions of the output FM of branch 1 and branch 2, respectively; S_1^1 and S_2^2 represent the FM of branch 1 and branch 2, respectively. The final loss function of the model is shown in Eq. (10).

$$L_t = L_1 + L_2 + L_3 \tag{10}$$

In Eq. (10), L_t represents the total loss function; L_1, L_2, and L_3 respectively represent the loss functions of the three stages.

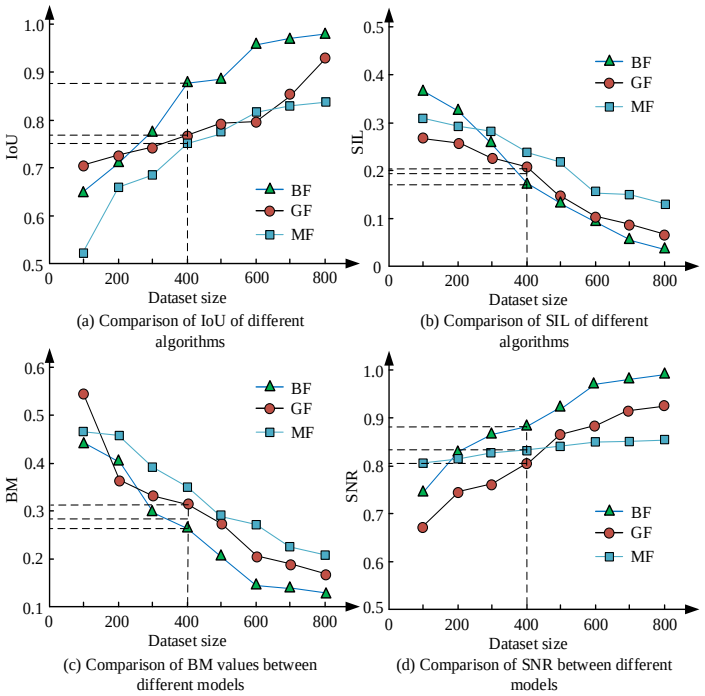
IV. Results

In the first section, image processing models based on Gaussian Filtering (GF) and Mean Filtering (MF) were introduced as comparison models for comparison. In the second section, the performance of the DI detection model combined with cascaded DWCA was analyzed.

A. Localization Performance based on DI Surface Keypoints

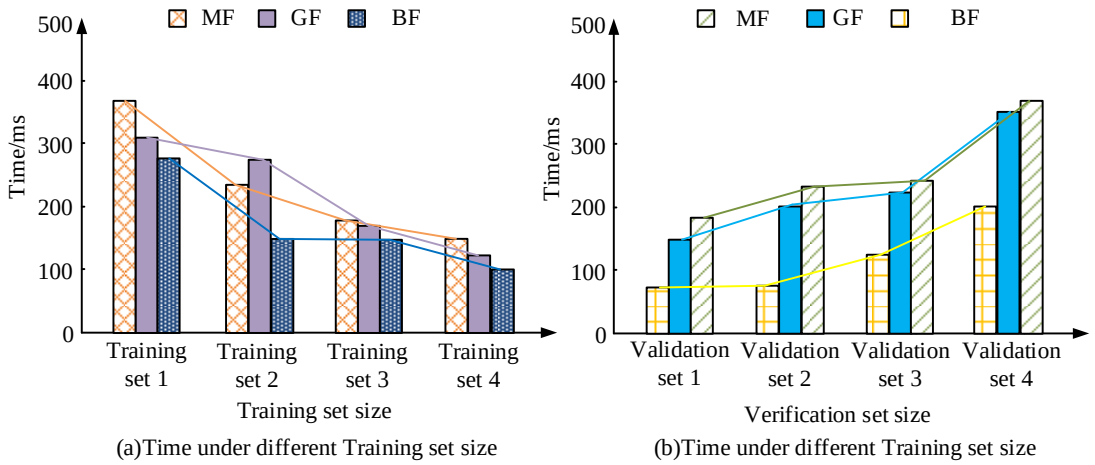
The COCO public dataset was utilized, which comprises over 100,0000 images, encompassing 80 distinct categories of objects, including humans, animals, vehicles, and furniture, as well as various scenes and environments. Each image has detailed annotation information, including the category of the object, the position and size of the bounding box, and the keypoint information of the object, providing standard evaluation metrics. The CPU model used was Intel (R) Core (TM) i7-9700F, with a frequency of 3.00GHz. The graphics processor model was NVIDIA GeForce GTX 1660 Ti, with 8GB of video memory. The operating system was Windows 10. This study introduced GF based image processing models and MF based image processing models as comparative models for comparison. The results are shown in Fig. 7.

Fig. 7, 7(b), 7(c), and 7(d) illustrate the comparison results of Intersection over Union (IoU), Structural Information Loss (SIL), Bonferroni Mean (BM), and Signal to Noise Ratio (SNR) of images processed by various algorithms. As illustrated in Fig. 7, an increase in the training set size was accompanied by a corresponding rise in the IoU of each model after processing images. As the training set size was up to 800, the IoU of BF, GF, and MF were 0.83, 0.91, and 0.96. The SIL was 0.10, 0.06, and 0.03. The BM was 0.24, 0.15, and 0.08. The SNR was 0.81, 0.91, and 0.99. The experimental results demonstrated that BF had relatively superior image processing performance. The processing time of each method was compared, and each dataset was segmented according to different sizes. The sizes of Dataset 1 to Dataset 4 were 100, 200, 400, and 800. Fig. 8 presents the results.



(a) Comparison of IoU of different algorithms
(b) Comparison of SIL of different algorithms
(c) Comparison of BM values between different models
(d) Comparison of SNR between different models

Fig. 7. Performance comparison of various image processing algorithms.



(a)Time under different Training set size (b)Time under different Training set size

Fig. 8. Comparison of model recognition time.

In Fig. 8, the proposed BF had the fastest image processing speed among all training sets. In training set 4, the training times for BF, GF, and MF were 101ms, 113ms, and 157ms, respectively. In Fig. 8(b), the proposed BF processing speed performed the best among the three algorithms in each validation set. In validation set 4, the training times for BF, GF, and MF were 201ms, 352ms, and 374ms. The suggested image processing method exhibited the best performance among various algorithm models. In Table I, the comprehensive capability of the three algorithm models was compared.

TABLE I. COMPARISON OF IMAGE PROCESSING PERFORMANCE OF VARIOUS ALGORITHMS

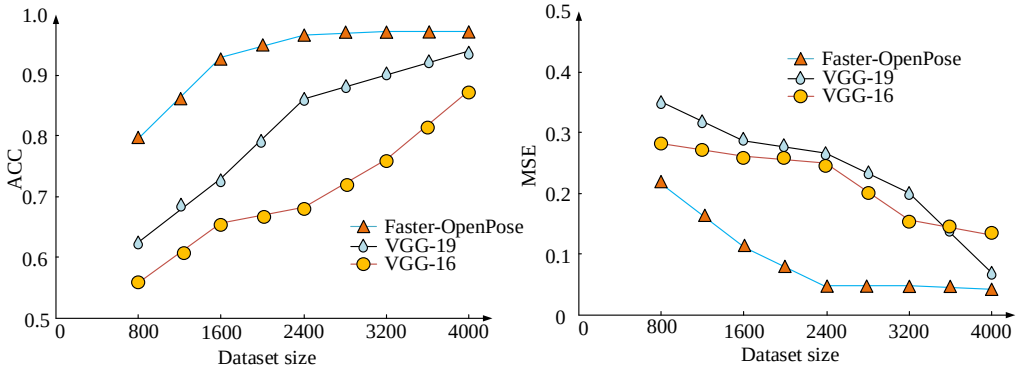
Image type	MF		GF		BF	
	IoU	*SIL*	*IoU*	*SIL*	*IoU*	*SIL*
Image type 1	0.82	0.35	0.87	0.19	0.95	0.13
Image type 2	0.75	0.33	0.84	0.17	0.91	0.11
Image type 3	0.89	0.31	0.94	0.15	0.99	0.09
Image type 4	0.81	0.33	0.85	0.17	0.96	0.11
Image type 5	0.90	0.37	0.96	0.21	0.98	0.15

In Table I, the recognition IoUs of MF for various types of images were 0.82, 0.75, 0.89, 0.81, and 0.90, respectively. The IoUs of GF model for various types of images were 0.87, 0.84, 0.94, 0.85, and 0.96, respectively. The IoU values of BF for various types of images were 0.95, 0.91, 0.99, 0.96, and 0.98, respectively. Therefore, the proposed BF image processing method had excellent performance.

B. *Performance Analysis of DI Detection Model Combined with Cascaded DWCA*

Following an evaluation of the efficacy of various image processing techniques, it is essential to assess the performance of the proposed recognition model. Visual Geometry Group 16 (VGG-16) and Visual Geometry Group 19 (VGG-19) were introduced and compared with the model built within this study. Fig. 9 clearly illustrates the results.

As the size of the dataset increased, the accuracy of each model also increased in a corresponding manner, as illustrated in Fig. 9(a). As the dataset size was 4000, the accuracy of Faster OpenPose, VGG-19, and VGG-16 were 0.97, 0.91, and 0.84, respectively. In Fig. 9(b), as the dataset increased, the Mean Square Error (MSE) of each model decreased accordingly. When the dataset size was 4000, the MSE of VGG-16, VGG-19, and Faster OpenPose were 0.11, 0.09, and 0.03, respectively. The analysis of different types of DI is shown in Fig. 10.



(a) Comparison of accuracy of different algorithm models (b) Comparison of MSE of different algorithm models

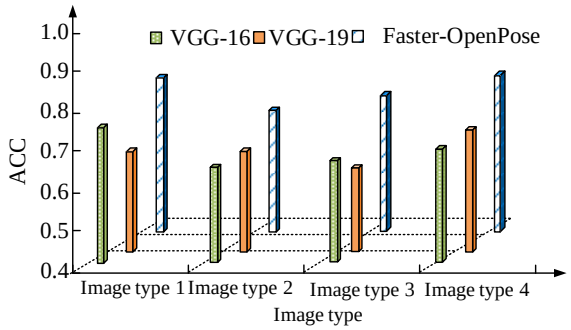
Fig. 9. Performance comparison results of various models.

In Fig. 10(a), among the three method models, Faster OpenPose had the best performance, with recognition rates of 0.91, 0.87, 0.89, and 0.93 for four different images, respectively. In Fig. 10(b), among the three models, Faster OpenPose had the shortest processing time, with processing times of 0.32, 0.31, 0.28, and 0.27 for the four types of images, respectively. Therefore, the proposed Faster OpenPose model had excellent performance. A total of 50 individuals were randomly selected and divided into five groups for the purpose of evaluating the performance of the model. Table II clearly presents the results.

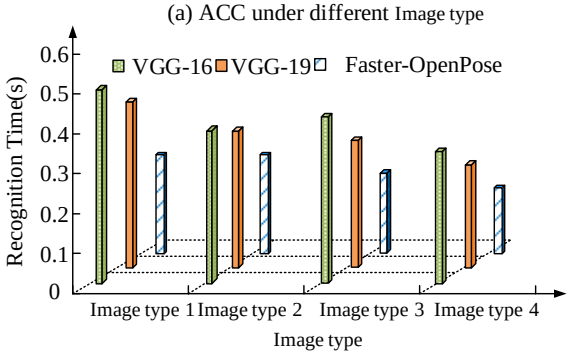
TABLE II. USER EVALUATION FORM

Type	Group 1	Group 2	Group 3	Group 4	Group 5	AVG
Faster-OpenPose	93.5	93.4	88.6	81.6	87.2	88.5
VGG-19	77.9	85.2	75.4	74.9	83.3	79.2
VGG-16	74.6	81.9	70.5	68.5	81.4	75.2

In Table II, the five evaluation groups rated Faster OpenPose at 93.5, 93.4, 88.6, 81.6, 87.2, and 88.5, respectively, and rated VGG-16 at 74.6, 81.9, 70.5, 68.5, 81.4, and 75.2, respectively. As a result, the Faster OpenPose proposed had received high praise from various users.



(a) ACC under different Image type



(b) Model recognition time under Image type

Fig. 10. Analysis of recognition performance of different images.

V. DISCUSSION

A depth image keypoint detection model based on bilateral filtering and cascaded depth separable convolution modules has been proposed, which demonstrates significant advantages in processing depth images. The experimental results show that when the training set size is 800, the intersection to union ratio of the BF model reaches 0.96, which shows better image processing performance compared to the 0.91 and 0.83 of GF and MF. In addition, the signal-to-noise ratio of the bilateral filtering model is 0.99, while GF and MF are 0.91 and 0.81, respectively. This is because bilateral filtering effectively balances denoising and edge preservation, which enables the model to reduce noise interference while maintaining important edge information when processing depth images, thereby improving the intersection to union ratio and signal-to-noise ratio. This is similar to the research results of Smith J et al. [23]. When using Faster OpenPose network for keypoint detection, the model has improved accuracy compared to traditional OpenPose. When the dataset size is 4000, the accuracy of the Faster OpenPose model reaches 0.97, while VGG-19 and VGG-16 are 0.91 and 0.84, respectively, and the MSE is significantly reduced. Faster OpenPose is 0.03, VGG-19 and VGG-16 are 0.09 and 0.11, which is similar to the research results of Jones M et al. [24]. This is because the cascaded depthwise separable convolution module significantly reduces the number of parameters and computational complexity of the model by splitting the standard convolution into depthwise convolution and pointwise convolution. This enables the model to maintain high accuracy while improving processing speed. This indicates that the design of depth separable convolution modules effectively reduces computational complexity and improves the model's adaptability to different types of images. However, there are also some shortcomings in the research. The performance of a model largely depends on the quality and diversity of the training data, and biases in the dataset can affect the model's generalization ability. In complex scenarios, interference from background information remains a major issue.

VI. CONCLUSION

In response to the problems of low efficiency and insufficient accuracy in traditional methods for handling DI, this study proposes a DI keypoint detection model that combines cascaded DWCA. It optimizes the computational efficiency and the model's detection capability by improving the OpenPose network structure. When the training set size was 800, the IoU of BF, GF, and MF were 0.83, 0.91, and 0.96; The SIL were 0.10, 0.06, and 0.03; The BM were 0.24, 0.15, and 0.08; The SNRs were 0.81, 0.91, and 0.99. When the dataset size was 4000, the accuracies of Faster OpenPose, VGG-19, and VGG-16 were 0.97, 0.91, and 0.84, respectively, with MSE of 0.11, 0.09, and 0.03, respectively. Therefore, the DI keypoint detection method combined with cascaded DWCA has high practical value and can effectively enhance the processing efficiency of DI keypoint detection. However, the research also has certain limitations. The performance of a model is highly dependent on the quality and diversity of the training data. If the dataset is not rich enough or has biases, it may affect the model's generalization ability. Secondly, although Faster OpenPose networks have improved efficiency, delays may still occur in extremely complex scenarios, which can affect real-time application performance. Although background filtering is used, in some cases, complex backgrounds may still interfere with keypoint detection, leading to inaccurate localization. In the future research direction, more lightweight network structures can be explored to adapt to real-time applications of edge computing and mobile devices and improve processing efficiency. It can enhance the adaptability to complex scenes and dynamic backgrounds, and improve the robustness and

accuracy of the model by integrating more types of data. In addition, utilizing emerging technologies such as self supervised learning and transfer learning can enhance the model's generalization ability on small sample datasets. Finally, research can delve into the methods of multimodal fusion, combining depth images with other sensor data to achieve more accurate keypoint detection and application expansion.

REFERENCES

[1] Li K, Liu Z, Zhou J, Dai Y, Liu Q, An J, Liu Y. Detection Algorithm of the Seabed Man-made Elongated Target Based on Synthetic Aperture Sonar Image. Basic & clinical pharmacology & toxicology, 2020, 127(1):117-118.

[2] Xiong Y, Yang L. Asian international students' help-seeking intentions and behavior in American Postsecondary Institutions. International Journal of Intercultural Relations, 2020, 80(2021):170-185.

[3] Zhou K, Zhang Z, Yuan R, Chen E. A deep learning algorithm for fast motion video sequences based on improved codebook model. Neural Computing and Applications, 2023, 35(6): 4353-4368.

[4] Zhou H, Ren D, Xia H, Fan M, Yang X, Huang H. AST-GNN: An Attention-based Spatio-temporal Graph Neural Network for Interaction-aware Pedestrian Trajectory Prediction. Neurocomputing, 2021,445(20):298-308.

[5] Lee J, Hwang K. YOLO with adaptive frame control for real-time object detection applications. Multimedia Tools and Applications, 2022, 81(25): 36375-36396.

[6] Jiang X, Yu F R, Song T, Ma Z, Zhu D. Blockchain-Enabled Cross-Domain Object Detection for Autonomous Driving: A Model Sharing Approach. IEEE Internet of Things Journal, 2020, 7(5):3681-3692.

[7] Yang Y. A Vehicle Recognition Algorithm Based on Deep Convolution Neural Network. Traitement du Signal, 2020, 37(4):647-653.

[8] Chen D, Sun S, Lei Z, Shao H, Wang Y.Ship Target Detection Algorithm Based on Improved YOLOv3 for Maritime Image. Journal of Advanced Transportation, 2021, 21(10):212-223.

[9] Tumrani S, Deng Z, Lin H, Shao J.Partial attention and multi-attribute learning for vehicle re-identification. Pattern Recognition Letters, 2020, 138(10):290-297.

[10] Kikuchi T, Fukuda T, Yabuki N. Diminished reality using semantic segmentation and generative adversarial network for landscape assessment: evaluation of image inpainting according to colour vision. Journal of Computational Design and Engineering, 2022, 9(5): 1633-1649.

[11] Li G, Ji Z, Qu X, Zhou R, Cao D. Cross-domain object detection for autonomous driving: A stepwise domain adaptative YOLO approach. IEEE Transactions on Intelligent Vehicles, 2022, 7(3): 603-615.

[12] Lee J, Hwang K. YOLO with adaptive frame control for real-time object detection applications. Multimedia Tools and Applications, 2022, 81(25): 36375-36396.

[13] Liang S, Wu H, Zhen L, Hua Q, Garg S, Kaddoum G. Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 25345-25360.

[14] Huang L, Ye L, Li R, Zhang S, Qu C, Li S. Dynamic human retinal pigment epithelium (RPE) and choroid architecture based on single-cell transcriptomic landscape analysis. Genes & Diseases, 2023, 10(6): 2540-2556.

[15] Wang X, Sun X, Wang Z. Construction of visual evaluation system for building block night scene lighting based on multi-target recognition and data processing. IET Circuits, Devices & Systems, 2023, 17(3): 149-159.

[16] Hui, Peng, Yifan, Zhang, Sen, Yang, Bin, Song. Battlefield Image Situational Awareness Application Based on Deep Learning. IEEE Intelligent Systems, 2019, 35(1):36-43.

[17] Laroca R, Zanlorensi L A, Gonalves G R, Todt E, Menotti D. An efficient and layout¨Cndependent automatic license plate recognition system based on the YOLO detector. IET Intelligent Transport Systems, 2021, 15(4):483-503.

[18] Feng H, Jie S, Hang M, Wang R, Fang F, Zhang G. A novel framework on intelligent detection for module defects of PV plant combining the visible and infrared images. Solar Energy, 2022, 236(4):406-416.

[19] Hu G X, Hu B L, Yang Z, Huang L, Li P. Pavement Crack Detection Method Based on Deep Learning Models. Wireless Communications and Mobile Computing, 2021, 32(1):1-13.

[20] Fitzpatrick B R, Berends M, Ferrare J J, Waddington RJ. Virtual Illusion: Comparing Student Achievement and Teacher and Classroom Characteristics in Online and Brick-and-Mortar Charter Schools. Educational Researcher, 2020, 49(3):161-175.

[21] Soffer T, Cohen A. Students' engagement characteristics predict success and completion of online courses. Journal of Computer Assisted Learning, 2019, 35(3):378-389.

[22] Pal S, Roy A, Shivakumara P, Pal U. Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images. Artificial Intelligence and Applications, 2023, 1(3), 145-154.

[23] Smith J, Brown A, Johnson L. Robust Feature Extraction for Keypoint Detection in Complex Environments. Journal of Computer Vision, 2022, 45(3): 123-135

[24] Jones M, Li Y. Multimodal Fusion Techniques for Enhanced Detection Performance. International Conference on Image Processing, 2023, 12(2): 45-60.