Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.047057

ARTICLE

DGConv: A Novel Convolutional Neural Network Approach for Weld Seam Depth Image Detection

Pengchao Li[1,2,3,*], Fang Xu[1,2,3,4], Jintao Wang[1,2], Haibing Guo[4], Mingmin Liu[4] and Zhenjun Du[4]

[1]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China

[2]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China

[3]University of Chinese Academy of Sciences, Beijing, 100049, China

[4]Shenyang SIASUN Robot & Automation Co., Ltd., Shenyang, 110168, China

*Corresponding Author: Pengchao Li. Email: lipengchao@sia.cn

Received: 23 October 2023 Accepted: 08 December 2023 Published: 27 February 2024

ABSTRACT

We propose a novel image segmentation algorithm to tackle the challenge of limited recognition and segmentation performance in identifying welding seam images during robotic intelligent operations. Initially, to enhance the capability of deep neural networks in extracting geometric attributes from depth images, we developed a novel deep geometric convolution operator (DGConv). DGConv is utilized to construct a deep local geometric feature extraction module, facilitating a more comprehensive exploration of the intrinsic geometric information within depth images. Secondly, we integrate the newly proposed deep geometric feature module with the Fully Convolutional Network (FCN8) to establish a high-performance deep neural network algorithm tailored for depth image segmentation. Concurrently, we enhance the FCN8 detection head by separating the segmentation and classification processes. This enhancement significantly boosts the network's overall detection capability. Thirdly, for a comprehensive assessment of our proposed algorithm and its applicability in real-world industrial settings, we curated a line-scan image dataset featuring weld seams. This dataset, named the Standardized Linear Depth Profile (SLDP) dataset, was collected from actual industrial sites where autonomous robots are in operation. Ultimately, we conducted experiments utilizing the SLDP dataset, achieving an average accuracy of 92.7%. Our proposed approach exhibited a remarkable performance improvement over the prior method on the identical dataset. Moreover, we have successfully deployed the proposed algorithm in genuine industrial environments, fulfilling the prerequisites of unmanned robot operations.

KEYWORDS

Weld image detection; deep learning; semantic segmentation; depth map geometric feature extraction

1 Introduction

Welding seam polishing is a critical procedure in the manufacturing industry to ensure the durability and aesthetic appeal of the final product. It involves removing the irregularities and rough surfaces on the welding seam to make it smooth and consistent. This process can generate fine dust that poses health hazards to workers. Therefore, the use of robots for welding seam polishing is becoming

increasingly prevalent [1]. Robots can work for extended periods without the need for breaks, ensuring high productivity and reducing the risk of worker injury [2,3].

However, to enable a robot to polish welding seams accurately, it must first perceive the welding seam shape, size, and position. This requires perception devices that can collect data on welding seams and analyze their category and location. Line scan cameras are commonly used in such scenarios due to their ability to capture continuous images of the object. Line scan cameras use the principle of orthogonal projection to generate data and create depth maps that contain 3D geometric information [4].

In contemporary research, segmentation algorithms for RGB and gray images have reached a high level of maturity. However, there exists a noticeable gap in the development of segmentation algorithms tailored for depth images. The primary objective of this study is to address the suboptimal segmentation outcomes encountered by current deep learning algorithms when applied to weld depth data. To overcome this challenge, our research aims to introduce a novel neural network methodology designed to augment the feature extraction capabilities of depth images. The ultimate goal is to significantly enhance the segmentation effectiveness within the weld area [5]. However, the purpose and means of welding seam detection technology are similar, so this paper will introduce the welding seam detection methods in the above scenario. These methods target the characteristics of line structure light. Some work [6,7] processed each frame of line scanning data and uses its gradient, intersection, or intercept features to complete welding seam detection. These methods can perform well in the case where the geometric shape of the welding interface is simple and consistent, which is also the case for most welding seam tracking systems currently used.

Traditional 2D segmentation methods are well-established in the field of computer vision. However, applying these methods to the depth maps generated by line scan cameras poses significant challenges. Depth maps contain little texture information and primarily consist of height information [8]. Existing 2D segmentation methods cannot efficiently extract the geometric information contained in depth maps, leading to poor segmentation accuracy [5].

Currently, the best deep neural networks for image segmentation are specifically designed to process pixels, using Convolutional Neural Network (CNN) to extract features from each pixel and its neighboring pixels [9]. This method effectively extracts features from most colored and textured images. Another method is to use the transform module to extract pixel context features, while attention mechanisms selectively focus on the most relevant and prominent features in the image, improving the ability to extract salient features. These methods improve the performance of the model by extracting local features of the image [10,11]. However, these methods directly ignore the geometric relationships between the elements of the depth map and have structural limitations on the extraction of local geometric features [12].

To address these issues, we collected data from actual industrial robot polishing sites and created the Standardized Linear Depth Profile (SLDP) dataset for weld seam depth maps under orthogonal projection to help promote research in this field. A detailed description of this dataset is provided in Section 4 of this paper. We propose a new feature extraction operation for depth maps called Deep Geometry Convolution (DGConv), which can effectively extract the geometric features contained in the depth map. Instead of directly processing pixel values, DGConv extracts features by analyzing the vector features between neighboring points. Since it focuses on learning the vector relationships between neighboring points, DGConv can be seen as a deep learning implementation of traditional hand-designed features, making it possible to group points in both Euclidean and semantic spaces.

DGConv is easy to implement and can be integrated into existing deep learning models to improve its ability to extract features from depth maps. We designed a depth geometric feature extraction network module based on DGConv to extract the features of the depth map more effectively. The schematic diagram of the processing flow of this module is shown in Fig. 1. At the same time, we integrated it into the FCN8 [13] network and added the detection head module. Experiments are carried out to prove that the proposed network achieves the most advanced performance on the SLDP dataset. In addition, the compatibility experiments with the module and the detection head module show that DGConv can improve the performance of the existing network.

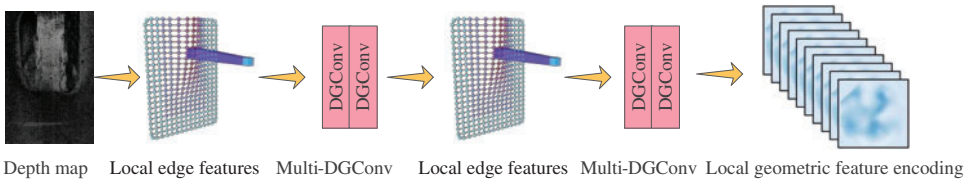


Figure 1: Deep local geometric feature question module. Local edge features convert depth maps into point cloud information and calculate edge features of neighboring and central points. Multi DGConv is a module composed of multi-layer deep geometric convolutions. Local geometric feature encoding is a local geometric feature map extracted from a depth map through two layers of Local edge features and Multi DGConv processing

We summarize the key contributions of our work as follows:

• We propose a method for weld seam feature extraction on depth images. Our method, called DGConv, effectively extracts local geometric features from depth images by leveraging the power of convolutional neural networks.

• Based on the depth geometric convolution, we introduce a novel seam segmentation network, which achieves better performance on industrial-level datasets.

• We constructed and developed a realistic dataset for industrial weld grinding, which serves as an evaluation benchmark for visual detection by robots during the weld grinding process.

• To facilitate the application of artificial intelligence in industrial settings, we release the source code and dataset to the public.

2 Related Work

Welding detection is a common problem in robot operations, and traditional detection of welds usually relies on manual extraction of features based on human experience. With the excellent performance of deep neural networks in segmentation problems, we can apply their methods to the localization of welds in industrial scenes. Therefore, this section will elaborate on welding seam detection and the application of deep learning in segmentation.

2.1 Weld Seam Detection

Weld seam detection is a critical task in welding automation that involves the detection of the weld seam location and orientation in an image [6]. Several approaches have been proposed in the literature, including template matching, Hough transform, and deep learning-based methods. One approach [14] employed the main characteristics of the gray gradient in the weld image to propose an improved Canny edge detection algorithm, enabling edge detection, as well as extraction of seam

and pool characteristic parameters. The paper of [15] presented Otsu's method for automatic image segmentation and then used an improved Hough algorithm for line detection. Finally, based on the characteristics of the detection line in image analysis, weld edge detection was completed by [16]. Median filtering was observed by Banafian et al. [5] to enhance the connectivity of flat weld edge detection, whereas the Laplace operator improved accuracy and compatibility with different types of welds [7]. An efficient algorithm for weld seam detection in butt joint welding was presented in [17], which involved locating a pair of weld seam edges in a local area and iteratively detecting and linking the remnant edge from the endpoints of each edge. The basic idea of the approach is to find a pair of weld seam edges in the local area first. Then, starting from the two endpoints of each edge, search for the remnant edge by iterative edge detection and edge linking. During the welding process, factors such as spatter residues and the complex texture of the workpiece can result in noisier and lower contrast image data. In order to improve the quality of the data obtained, prior knowledge of welding seam features has been utilized for edge detection by [18]. Additionally, edge detection operators [19] were utilized to generate a gray-level feature factor and membership function for welding images. This information is then inputted into a decision logic constructed using fuzzy theory. The decision logic utilizes linear regression to determine the correct coordinates of the welding seam, thereby improving the accuracy of the overall welding image analysis process.

2.2 Semantic Segmentation by Deep Learning

Deep learning has shown good performance in semantic segmentation tasks, so deep learning methods can be used to complete the segmentation of weld depth maps. FCN8 is a segmentation network that uses only convolutional layers to generate pixel-wise predictions. FCN8 uses Vgg16 [20] as the feature coding module. The proposed method addresses the problem of generating dense predictions for semantic segmentation. UNet uses skip connections to propagate the low-level feature maps to the decoder module [21]. The proposed method addresses the problem of capturing both low-level and high-level features for semantic segmentation. UNet includes skip connections to fuse the low-level and high-level features and a refinement module to improve the segmentation accuracy. UNet++ [22] integrates U-Net models from different layers of the U-Net network, which can capture features from different levels and integrate them through feature superposition, improving the detection ability for target objects of different sizes. UNet+++ [23] removes the dense convolutional blocks of UNet++ and proposes a full-size skip connection method, which adds deep feature map content to the low-level model structure, improving its feature extraction ability at the full scale. E-Net includes a bottleneck module to reduce the number of parameters and a residual module to improve the segmentation accuracy [24]. SegNet [25] is a segmentation network that uses a decoder module to recover the spatial information lost during the pooling and downsampling operations. Reference [26] shows a proposed solution for simultaneously addressing the challenges of localization and classification in segmentation tasks. Moreover, it introduces a novel Boundary Refinement block aimed at enhancing the algorithm's ability to accurately identify and detect object boundaries. To extract features at different scales, a pyramid pooling module was utilized by [27]. The proposed method effectively tackles the challenge of capturing multi-scale contextual information for semantic segmentation. Deeplab V3+ [28] is an extension of the Deeplab V3 [29] architecture which uses atrous convolution and spatial pyramid pooling to extract features from images. Deeplab V3+ is designed to solve the problem of semantic segmentation in images with diverse object scales and is particularly effective in segmenting small objects. The proposed method uses a feature pyramid network with atrous convolution to extract multi-scale features, which are then combined with a spatial pyramid pooling module to obtain multi-resolution context information. A Dense Upsampling Convolution

[30] successfully captures and decodes more detailed information concerning the loss of bilinear upsampling. At the same time, the proposed Hybrid Dilated Convolution effectively expands the network's perception field for aggregating global information and reduces the "grid issue" caused by standard dilation convolution. UperNet [31] uses a top-down pathway to incorporate high-level semantic information into low-level feature maps. The proposed method addresses the problem of fusing multi-resolution features effectively and achieving high-resolution segmentation results. It uses a top-down pathway with skip connections to fuse multi-scale features and also includes a refinement module to improve the segmentation accuracy.

3 Approach

We present a new method for extracting local features from point clouds that are inspired by traditional point cloud feature descriptors. Unlike hand-designed features, we do not pre-specify the basic geometric features extracted from point pairs based on explicit constraints. Instead, we use graph neural networks to process point pairs and extract deeper features. In this chapter, we provide evidence of the effectiveness of our approach. In comparison to graph convolutions that deal with spatially unordered data, our method constructs neighboring graphs by referencing CNN and selecting neighboring points using a depth map. Additionally, after each convolutional layer, the graph is updated to ensure that the local features of each layer are propagated to the next layer and computed based on the embedding sequence from the previous layer. As a result, the local features are more stable, consequently ensuring better feature extraction.

In this section, we provide a comprehensive explanation of the proposed deep-geometric convolutional network (DGConv). We elaborate on the principles of the deep geometric convolution module and its ability to enhance feature extraction efficiency and improve weld seam detection. Additionally, we describe the structure of the weld seam detection network that comprises DGConv and its associated modules.

3.1 Deep Geometric Convolutional Network

Weld seam detection requires algorithms to identify the region where the weld seam is located. This task can be classified as semantic segmentation in computer vision, which provides probability scores for the semantic information corresponding to each pixel. As mentioned in related work, two major challenges need to be addressed: classification and localization. However, current segmentation algorithms struggle to effectively extract features from depth data for weld seam detection, resulting in limited predictive performance. Additionally, we have observed that existing networks exhibit coupling between classification and segmentation during feature decoding, which negatively impacts algorithm performance.

Based on the aforementioned challenges combined with the characteristics of weld seam data, we propose a deep geometric convolutional network, as shown in Fig. 2. Our network is divided into three modules: the deep geometric convolutional network module, the feature processing module, and the network detection head module, which consists of the category detection head and mask detection head. The Deep Geometric Feature Extraction Moudle (DGCM) is designed specifically for the characteristics of line array data, allowing for effective extraction of the geometric features of line array data. The feature processing module directly utilizes the backbone module of the FCN8 network. In our experiments, we noticed the exceptional ability of FCN8 in extracting depth map features, so we adopted it as the backbone for feature processing, targeting the features extracted by the deep convolutional module. The network detection head module is integrated with the Fused

Feature Map module of the FCN8 module while removing the original detection head of FCN8. The network detection head module separates category detection and mask detection, with each branch focusing on its respective detection objectives.

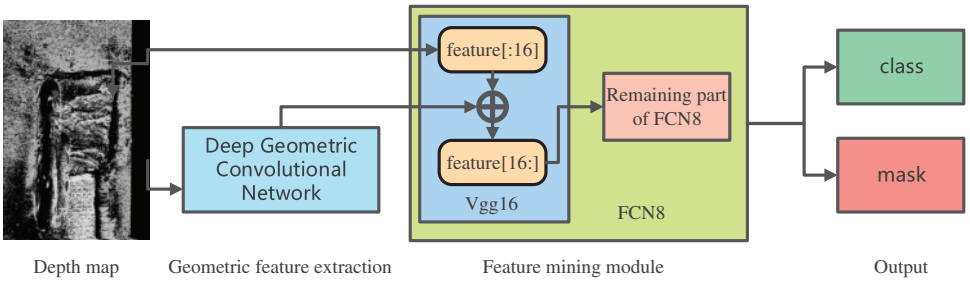


Figure 2: Overall framework of the network. The network is divided into three modules in total. The geometric feature extraction module extracts the geometric features of each point using the geometric features of the depth map, and then deeply mines the geometric features of the points through the FCN8 network. Afterwards, input the information into the class and mask detection heads respectively to obtain the final prediction results. Feature[:16] and feature[16:] respectively represent the first 16 layers of Vgg16 and the networks after 16 layers

The overall workflow of the network is as follows: the camera captures the depth map, and then inputs the data into the local depth geometric convolution module and the first 16 layers of the Vgg16 module in the FCN8 network, and fuses the features obtained by the two modules to obtain the local geometric features. After further processing of the remaining part of FCN8, the characteristic graph is obtained. Finally, the feature map is input into the detection module to determine the type of weld and generate the pixel position mask.

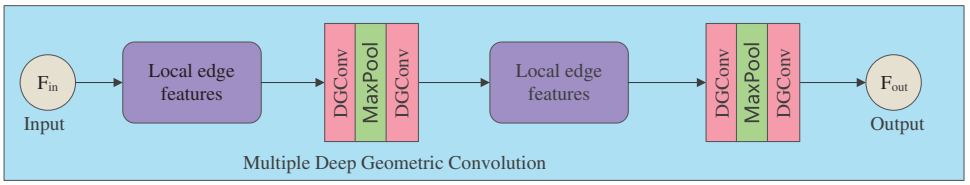
3.2 Deep Geometric Feature Extraction Moudle

The depth pixels contain geometric shape information of the perceived objects, which is very different from the color information contained in RGB. Many studies have been conducted on the rich color texture information in RGB and various segmentation methods have been developed. However, there are few studies on depth maps without color texture. Linear array scanning equipment is widely used in practical industrial applications. Applying the RGB segmentation algorithm directly to these images will produce poor results (as shown in the experiments in the following chapter). Directly applying RGB segmentation algorithms to these images yields poor results (as demonstrated in the experiments in the next chapter).

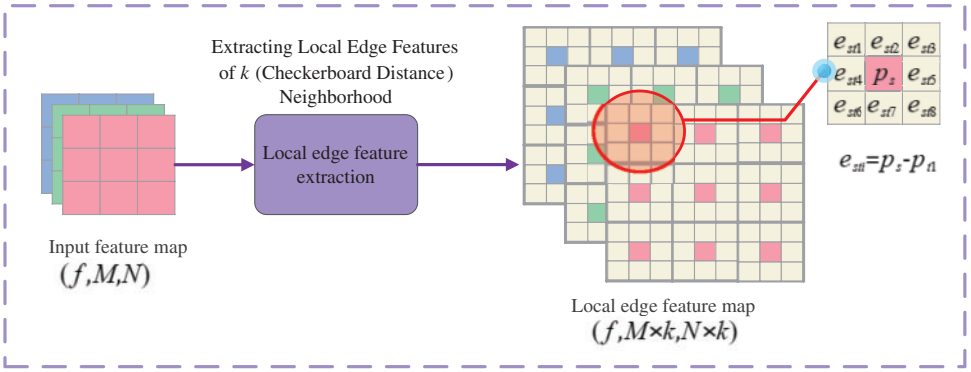
A deep geometric convolution for feature extraction from depth maps is proposed to address this issue. This convolution was inspired by traditional manual point cloud feature descriptors during design, utilizing deep learning features to enhance feature extraction capabilities and improve performance for extracting depth map features that do not contain color information. Before the emergence of deep neural networks, manually designed features were often used to solve visual problems in point clouds. Manually designed geometric feature descriptors are generally based on the neighborhood information of the center point, using the vector relationship between the center point and the neighborhood point coordinates to construct basic geometric features, and then using the statistics of the feature histogram to construct feature descriptors. For example, surface normal vectors and curvature are common local features of each point in a point cloud. A large number of manual features are constructed based on the relative relationships between these features to

construct feature descriptors. Previous studies have utilized the histogram distribution of the normal and curvatures of each point, as well as the spatial relationship between the normal and curvatures of adjacent points, such as angles and distances, to construct high-dimensional features. The high-dimensional hyperspace provided by histograms provides a measurable information space for feature representation. It is invariant to the 6-dimensional pose of the corresponding surface of the point cloud and robust to changes in sampling density and neighborhood noise. For example, classic point cloud feature descriptors include PFH [32] and FPFH [33]. When extracting geometric features, normal and two-point coordinates are the basic geometric features of points, and manually designed features also utilize this feature for feature extraction. However, due to common computational methods, it is not possible to extract high-dimensional nonlinear features. And deep neural networks have proven to be able to extract deeper-level features.

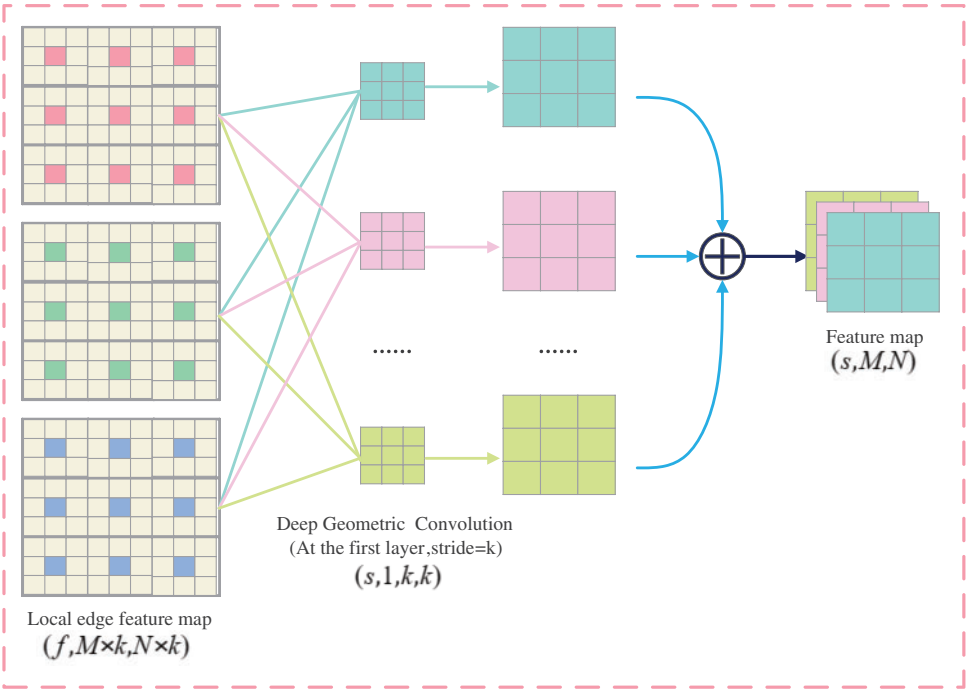
Inspired by traditional manually designed point cloud geometric feature extraction methods, combined with the powerful high-dimensional feature extraction ability of deep learning. We propose local geometric feature convolution for depth maps and design a local geometric feature extraction module based on this convolution. In Fig. 3a, it can be observed that the Local Geometric Feature Extraction Module consists of two sets of Local Edge Feature Extraction Modules and multiple layers of Deep Geometric Convolutions. This module is responsible for extracting geometric features from the depth map. Figs. 3b and 3c respectively depict the schematic processes of the Local Edge Feature Extraction Module and the Deep Geometric Convolution. In Fig. 3b, when extracting local edge features within a k (chessboard distance) neighborhood, assuming the input size of the local edge feature extraction module is, where represents the number of feature channels, and represent the width and height of the features, after processing by the local edge feature module, the size of the local edge feature map becomes. From Fig. 3b, it can be seen that the local edge feature map surrounds each pixel in the original input feature map with a k $\times$ k neighborhood of edge features. In other words, the k-neighbors of each pixel are constructed as vector features, with the center point element coordinates in the original image and the coordinates of neighboring points in the k $\times$ k neighborhood of the original image forming the vector features. Fig. 3c describes the schematic principle of our proposed DGConv for the first layer of processing on the depth map. The size of this first layer of DGConv is, where represents the number of convolution kernels, and k corresponds to the k-neighborhood of the local edge feature extraction module. The convolution kernels have a step size of k when sliding on the feature map. The resulting feature map after this DGConv layer is. In this description, it can be seen that we differ from traditional CNN. In the general design process of convolution (such as convolutional neural networks), to obtain feature maps of the same size, the size of the convolution kernel should be, while DGConv is. This is because the DGConv we designed convolves the relationships between local edge vector features, extracting the relationships between the central point and neighboring points in the form of vectors through convolution. This is influenced by traditional handcrafted features. Handcrafted feature descriptors have limitations when designing features using edge feature vectors, whereas here we utilize the regression capability of deep learning to train more effective geometric features through backpropagation. In CNN convolution operations, each convolution kernel parameter on each channel is different. Only the first layer of the DGConv network, or the DGConv that processes the local edge feature map, must be set with a step size of k. In subsequent layers, the parameters can be set as needed, similar to regular convolutions. Since we are convolving edge vectors, DGConv has only a single channel within the same convolution kernel.



(a) Structure diagram of deep geometric convolutional network



(b) Schematic diagram for extracting local edge features of k (chessboard distance) neighborhood



(c) Schematic diagram of deep geometric feature convolution processing

Figure 3: (a) The structure diagram of the deep local geometric convolution feature extraction module in our proposed algorithm network is described. (b) Detailed description of the local edge feature extraction process and feature map size transformation in the deep local geometric convolution feature extraction module. (c) The process of processing the output feature map using deep local geometric convolution and the size of each encoded information during the process are described in detail

Fig. 4 depicts a schematic diagram of feature extraction using DGConv. For the input feature map, assuming that its k (k = 3) neighborhood obtains neighboring points $P = \{p_{t1}, p_{t2}, \ldots, p_{t8}\}$, and local edge features $E = \{e_{sti} | p_s - p_{ti}\}$ are acquired through local edge feature extraction. To compute the local geometric features of the query point P_S, the specific calculation process of DGConv is as follows:

$$F_s = \sigma\left(p_s \cdot w_{k*k} + \sum_{i=1}^{k*k-1} (p_s - p_{ti}) \cdot w_i\right) \tag{1}$$

where F_s represents the feature encoding extracted by DGConv, W_i represents the weight values of DGConv, P_s represents the coordinates of the central point, P_{ti} represents the coordinates of neighboring points. $e_{sti} = p_s - p_{ti}$ represents the edge features, and σ represents the activation function.

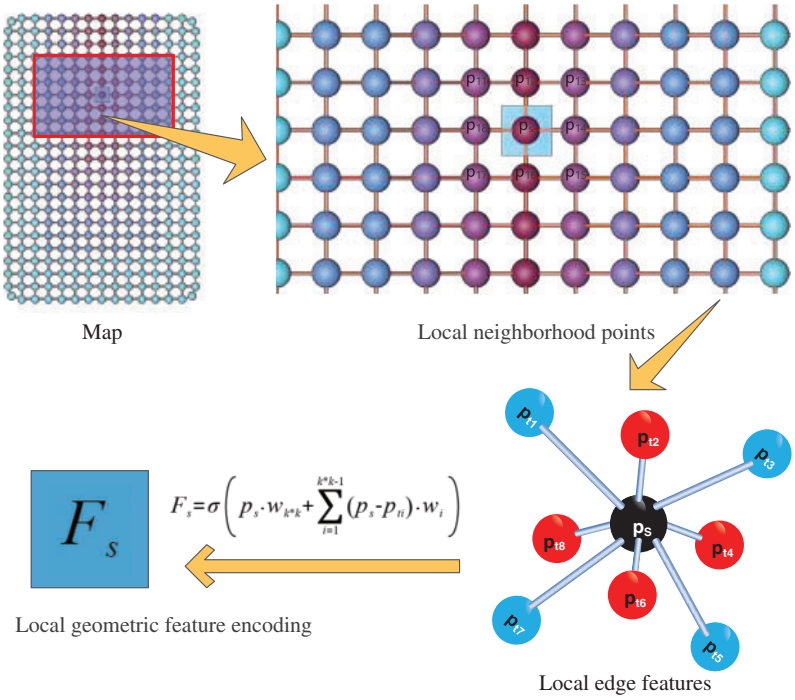


Figure 4: Schematic diagram of deep geometric convolution feature extraction. P_S represents the center point, based on the size of the depth map, select its k (expected distance from the center point is less than or equal to k) neighborhood point, and convert it into spatial three-dimensional coordinates. Among them, F_S is the calculated feature, σ is the activation function, n represents the number of neighboring points in its neighborhood, and w is the weight of deep geometric convolution

According to Eq. (1), it can be quantitatively observed that the fundamental difference between the DGConv convolution operation we propose and the CNN lies in the fact that CNN assigns different weights to each pixel value, while we treat the central point and neighborhood vectors as a whole and assign them weight values. In other words, DGConv convolves the edge vectors formed by the central point and neighborhood points, while CNN performs convolution operations on pixel points to extract features. Therefore, DGConv is more targeted towards geometric features, and the reduction in the number of parameters compared to CNN accelerates its convergence and better targets

the extraction of local geometric features, thus better meeting the real-time requirements of industrial applications.

To broaden the breadth and depth of feature extraction and enhance feature robustness, we use a two-layer local edge feature extraction module and multiple layers of DGConv to construct a local geometric feature extraction module, as shown in the specific structure in Fig. 3a. In theory, this module can regressively fit all cases where traditional handcrafted features are constructed using curvature and normal, and it can also uncover deeper-level features.

3.3 Fully Convolutional Networks

FCN has emerged as a powerful paradigm for image segmentation tasks in computer vision. These networks offer several key advantages, including end-to-end learning, preservation of spatial information, adaptability to variable input sizes, and efficient integration of multi-scale features. Notably, FCN8 maintains pixel-wise predictions, obviating the need for manual feature engineering or intermediate steps, thereby simplifying the segmentation workflow. Furthermore, the incorporation of skip connections facilitates the fusion of both low-level and high-level features, enabling fine-grained detail preservation while leveraging high-level semantic information. FCN8 also supports real-time inference, making it suitable for applications like autonomous vehicles and robotics. Additionally, transfer learning with pre-trained FCN8 models can significantly reduce data requirements for segmentation tasks. The Vgg16 module and the remaining parts of FCN8 in Fig. 5 together form the entire FCN8 module. In sum, FCN8 offers a robust and versatile approach to image segmentation, affording a semantic understanding of scenes, and finding utility across a wide array of academic and practical domains in computer vision research.

3.4 Detection Head Module and Loss Function

The proposed detection head module in this paper consists of two branches: one for category prediction and the other for mask prediction. Fig. 5 depicts the detailed network structure of the detection head module. Since shallow feature maps contain limited semantic information, when solving classification problems, it is common to extract high-dimensional deep features for object classification. However, high-dimensional low-resolution features harm distinguishing between foreground and background boundaries, as they tend to ignore the detection of fine edges. To address this issue, we design the task as two separate branches that do not interfere with each other. The category prediction branch extracts the geometric features of each dimension extracted by the FCN8 network module and uses pooling fusion (pooling shallow feature maps and connecting them with the next feature map along the channel) at each layer to merge the detailed information brought by shallow features into high-dimensional space. Subsequently, a fully connected MLP module is used for classification detection. On the other hand, the mask prediction branch directly performs binary recognition on the geometric feature extracted by the FCN8 network module, distinguishing target and non-target regions based on the overall category of the image.

We utilized the cross-entropy loss function to supervise the classification and mask tasks.

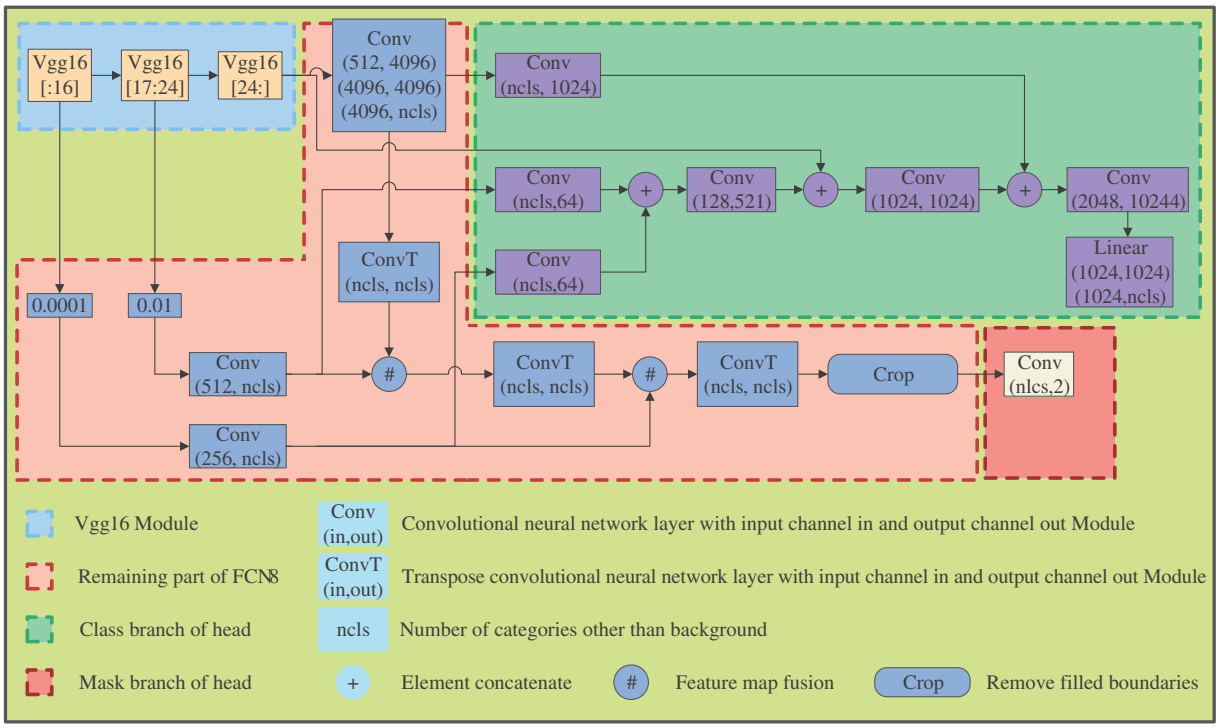


Figure 5: Detailed network structure of DGCM and detection head module. Different colors represent different modules, and modules with the same color in Figs. 2 and 5 correspond to each other

4 Experiments

4.1 Self-Collected Weld Seam Dataset

In this study, we introduce a meticulously curated self-collected dataset tailored for evaluating the proposed approach in the context of robot polishing applications. Data acquisition involved the use of a line structured light camera affixed to the robot's sixth axis, conducting controlled scans across an authentic industrial site under ambient light conditions typical of indoor workshop factories. The specific scene is shown in Fig. 6. The dataset encapsulates the nuances of welding seams in a real robot industry environment, distinguishing between three distinct seam types, as visually illustrated in Fig. 7. Captured in depth maps with an orthogonal projection format, the dataset aligns with the intricacies of practical robot polishing processes. Annotation of the depth maps was carried out using labelme software, incorporating ground truth labels for each pixel to denote welding seam categories. Notably, no preprocessing steps were applied to maintain the authenticity of the data. The dataset comprises unaltered raw data generated directly by the line structured light camera, ensuring fidelity to real-world industrial conditions. Three welding seam types constitute the dataset, each further divided into a training set (633 samples) and a validation set (161 samples), maintaining an 8:2 ratio. Both sets are equipped with ground truth masks indicating spatial positioning and ground truth labels for each specific seam type. Furthermore, the dataset includes essential parameters of the scanning camera, facilitating the conversion of depth images into point clouds. This comprehensive dataset, unprocessed and in its raw form, serves as an ideal resource for validating the proposed segmentation algorithm, providing a robust foundation for real-world applicability and algorithmic efficacy assessment.

Figure 6: Real data collection and algorithm testing application site

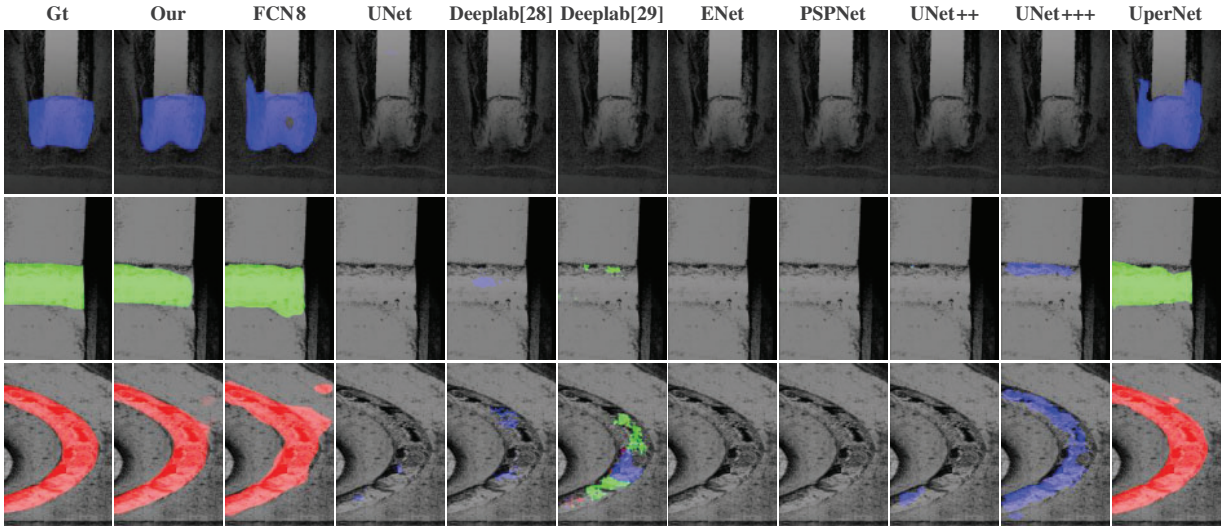


Figure 7: Detection results of the algorithm on the depth map. The blue color represents the identification of the first type of weld seam, the green color represents the identification of the second type of weld seam, and the red color represents the identification of the third type of weld seam

4.2 Implementation Details

The actual photo size collected is 1280×1300. When performing local edge feature extraction, the image will be enlarged K times in both width and height, which significantly consumes GPU memory. Therefore, we sample the image at regular intervals in both width and height and divide it into four images. The hyper-parameter for local edge feature extraction is set to $K = 11$. DGConv is tuned according to the experimental conditions. We use four RTX2080Ti graphics cards for training, with a batch size of 24, SGD as the optimizer, an initial learning rate of 0.001, cosine annealing as the learning rate decay strategy, and momentum set to 0.9.

4.3 Evaluation Metrics

In this task, the proportion of the background is relatively large, to better evaluate the detection capability of this dataset. This article evaluates various algorithms using common metrics in semantic segmentation, including Class Pixel Accuracy (CPA), mean Pixel Accuracy (mPA), Intersection over

Union (IoU), and mean Intersection over Union (mIoU). To establish consistent results and facilitate optimal convergence, 200 epochs of training were conducted for each algorithm, including both the control experiment and the proposed algorithm. Subsequently, the best-performing model was chosen for comparison.

4.4 Experimental Results

Our proposed method is compared with seven algorithms that perform well in semantic segmentation. The results are shown in Table 1, which displays the scores of various evaluation metrics on the SLDP dataset.

Table 1: Precision and Intersection over Union (IoU) Statistics for weld seam detection onSLDP dataset

	CPA				mPA	IoU				mIoU
	0	1	2	3		0	1	2	3	
FCN8	0.969	0.809	0.814	0.923	0.879	0.944	0.659	0.673	0.838	0.779
UNet	0.972	0.067	0.000	0.000	0.260	0.833	0.048	0.000	0.000	0.220
Deeplab [28]	0.983	0.304	0.000	0.000	0.322	0.872	0.212	0.000	0.000	0.271
Deeplab [29]	0.993	0.242	0.445	0.025	0.427	0.892	0.222	0.325	0.025	0.366
ENet	0.979	0.003	0.067	0.000	0.262	0.835	0.003	0.047	0.000	0.221
PSPNet	0.995	0.144	0.068	0.005	0.303	0.873	0.112	0.064	0.005	0.263
UNet++	0.995	0.024	0.000	0.000	0.255	0.847	0.021	0.000	0.000	0.217
UNet+++	0.996	0.018	0.000	0.000	0.253	0.851	0.016	0.000	0.000	0.217
UperNet	0.979	0.787	0.832	0.971	0.893	0.952	0.700	0.733	0.904	0.822
Ours	0.976	**0.865**	**0.903**	**0.963**	**0.927**	**0.959**	**0.749**	**0.820**	**0.896**	**0.856**

Our proposed method achieved the highest scores for mPA and mIoU on the SLDP dataset, while also obtaining the highest precision and Intersection over Union (IoU) scores in the detection of the first and second types of weld seams. Our method also outperformed other segmentation methods in terms of AP and IoU scores for several object classes.

To better evaluate the algorithm's speed, a subset of the original images was sampled at intervals and resized to 640 $\times$ 750 dimensions. The sample set included four times the number of photos compared to the original dataset. We present the computational speed of each algorithm in Table 2, where the performance is measured in frames per second (fps). Although our algorithm is not the fastest solely in terms of speed, it exhibits the fastest speed among the mPA > 0.8 algorithms (including FCN8 and UperNet) when considering performance in both accuracy and speed. A key factor contributing to this is our proposed convolution method, DGConv, which demonstrates considerably lower computational complexity than traditional CNN operating on equivalent neural structures.

We also provide qualitative results of our proposed method on the SLDP datasets. The data statistical results in Table 1 and the schematic detection results in Fig. 7 show that the FCN8 and UperNet algorithms exhibit superior feature detection capabilities for depth maps generated from line-array data, which demonstrates better feature detection capabilities for depth maps obtained using orthogonal projection. UNet, Deeplab [28] and Deeplab [29] can to some extent detect the third type of weld seam. Apart from these results, under the same conditions, the performance of the remaining

algorithms on this dataset is poor. Based on this, our algorithm is an improved version of FCN8, where we have added the DGCMand the Detection head module on the network, achieving the best results on this dataset. To validate the effectiveness of the modules we propose, we have also conducted ablation studies.

Table 2: Computation speed comparisons [fps]

Method	Ours	FCN8	UNet	Deeplab [28]	Deeplab [29]	ENet	PSPNet	UNet++	UNet+++	UperNet
Fps	**15.105**	12.370	18.275	21.018	12.948	10.617	8.146	9.977	13.776	8.814

4.5 Ablation Studies

This paper proposes a DGCM and a detection head module to improve existing algorithm networks for better suitability to the dataset in this study. To evaluate the effectiveness of the proposed modules, we conducted compatibility research experiments. We compared the performance of the complete network, networks with both proposed modules, a network with only the DGCM, a network with only the detection head module, and a network without point sampling and feature refinement modules. The results of Table 3 show that both proposed modules have improved the segmentation accuracy of the network, with the complete network achieving the highest accuracy.

Table 3: Compatibility study on the SLDP dataset. "Checkmark" indicates the corresponding modules added to the original for our proposed modules

Method	DGCM	Detection head module	mPA	mIoU
FCN8			0.879	0.779
	√		0.912	0.839
		√	0.893	0.822
Ours	√	√	0.927	0.856

5 Conclusion

In this article, we propose a novel approach to address the weld seam segmentation and classification problem in the current industrial weld seam depth images. The introduced DGCM is designed to more effectively extract local geometric features from depth images, with the geometric convolution module performing feature extraction better than CNN for depth images. The detection head module enhances the detection capabilities of both classification and segmentation by splitting them into two branches to tackle their respective issues, resolving the correlation between them. Extensive experimental results demonstrate that our method has achieved better performance on the SLDP dataset and exhibits advantages in terms of effectiveness and efficiency. To test the generalization performance of the algorithm, future work will be extended to the study of data sets in other fields. Currently, our dataset obtains depth images through orthogonal projection. The next step in future research is to investigate the algorithmic detection performance of data collected using perspective projection depth cameras.

Acknowledgement: We would like to express our sincere gratitude to Liu Yaoqi, Lu Saichao, and Zhang Jiaxin for their invaluable contributions to the dataset collection and production.

Funding Statement: This work was supported by the National Natural Science Foundation of China (Grant No. U20A20197).

Author Contributions: Study conception and design: Pengchao Li; data curation: Pengchao Li, Fang Xu, Haibing Guo; analysis and interpretation of results: Pengchao Li, Fang Xu, Jintao Wang; draft manuscript preparation: Pengchao Li, Mingmin Liu, Zhenjun Du. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data set used in this study was collected and produced by ourselves on the industrial site. In the future, this data set will be published to promote the application of AI technology in real industrial scenes.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] X. Ke, Y. Yu, K. Li, T. Wang, B. Zhong *et al.,* "Review on robot-assisted polishing: Status and future trends," *Robotics and Computer-Integrated Manufacturing*, vol. 80, pp. 102482, 2023. https://doi.org/10.1016/j.rcim.2022.102482

[2] A. Rout, B. B. V. L. Deepak and B. B. Biswal, "Advances in weld seam tracking techniques for robotic welding: A review," *Robotics and Computer-Integrated Manufacturing*, vol. 56, pp. 12–37, 2019. https://doi.org/10.1016/j.rcim.2018.08.003

[3] J. Liu, T. Jiao, S. Li, Z. Wu and Y. F. Chen, "Automatic seam detection of welding robots using deep learning," *Automation in Construction*, vol. 143, pp. 104582, 2022. https://doi.org/10.1016/j.autcon.2022.104582

[4] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau *et al.,* "Geometry-aware distillation for indoor semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 2869–2878, 2019. https://doi.org/10.1109/cvpr.2019.00298

[5] N. Banafian, R. Fesharakifard and M. B. Menhaj, "Precise seam tracking in robotic welding by an improved image processing approach," *The International Journal of Advanced Manufacturing Technology*, vol. 114, no. 1–2, pp. 251–270, 2021. https://doi.org/10.1007/s00170-021-06782-4

[6] W. J. Shao, Y. Huang and Y. Zhang, "A novel weld seam detection method for space weld seam of narrow butt joint in laser welding," *Optics & Laser Technology*, vol. 99, pp. 39–51, 2018. https://doi.org/10.1016/j.optlastec.2017.09.037

[7] Y. Z. Tian, H. F. Liu, L. Li, W. B. Wang, J. C. Feng *et al.,* "Robust identification of weld seam based on region of interest operation," *Advances in Manufacturing*, vol. 8, pp. 473–485, 2020.

[8] Y. Qiao, L. Jiao, S. Yang and B. Hou, "A novel segmentation based depth map up-sampling," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 1–14, 2018. https://doi.org/10.1109/tmm.2018.2845699

[9] X. Wang, R. Zhang, T. Kong, L. Li and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17721–17732, 2020.

[10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, USA, pp. 1290–1299, 2022. https://doi.org/10.1109/cvpr52688.2022.00135

[11] J. Wang, C. Mu, S. Mu, R. Zhu and H. Yu, "Welding seam detection and location: Deep learning network-based approach," *International Journal of Pressure Vessels and Piping*, vol. 202, pp. 104893, 2023. https://doi.org/10.1016/j.ijpvp.2023.104893

[12] W. Shao, X. Liu and Z. Wu, "A robust weld seam detection method based on particle filter for laser welding by using a passive vision sensor," *The International Journal of Advanced Manufacturing Technology*, vol. 104, pp. 2971–2980, 2019. https://doi.org/10.1007/s00170-019-04029-x

[13] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 3431–3440, 2015. https://doi.org/10.1109/cvpr.2015.7298965

[14] Y. Xu, G. Fang, S. Chen, J. J. Zou and Z. Ye, "Real-time image processing for vision-based weld seam tracking in robotic GMAW," *The International Journal of Advanced Manufacturing Technology*, vol. 73, no. 9–12, pp. 1413–1425, 2014. https://doi.org/10.1007/s00170-014-5925-1

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. https://doi.org/10.1109/tsmc.1979.4310076

[16] Q. -Q. Wu, J. -P. Lee, M. -H. Park, B. -J. Jin, D. -H. Kim *et al.,* "A study on the modified Hough algorithm for image processing in weld seam tracking," *Journal of Mechanical Science and Technology*, vol. 29, pp. 4859–4865, 2015. https://doi.org/10.1007/s12206-015-1033-x

[17] F. Shi, T. Lin and S. Chen, "Efficient weld seam detection for robotic welding based on local image processing," *Industrial Robot: An International Journal*, vol. 36, no. 3, pp. 277–283, 2009. https://doi.org/10.1108/01439910910950559

[18] Z. Ye, G. Fang, S. Chen and M. Dinham, "A robust algorithm for weld seam extraction based on prior knowledge of weld seam," *Sensor Review*, vol. 33, no. 2, pp. 125–133, 2013. https://doi.org/10.1108/02602281311299662

[19] Y. P. Huang and K. Bhalla, "Automatic generation of laser cutting paths in defective TFT-LCD panel images by using neutrosophic canny segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–16, 2022.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[21] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th Int. Conf.*, Munich, Germany, pp. 234–241, 2015. https://doi.org/10.1007/978-3-319-24574-4_28

[22] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *DLMIA and ML-CDS*, Granada, Spain, pp. 3–11, 2018. https://doi.org/10.1007/978-3-030-00889-5_1

[23] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang *et al.,* "UNet 3+: A full-scale connected UNet for medical image segmentation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 1055–1059, 2020. https://doi.org/10.1109/icassp40776.2020.9053405

[24] A. Paszke, A. Chaurasia, S. Kim and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint arXiv:1606.02147, 2016.

[25] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. https://doi.org/10.1109/tpami.2016.2644615

[26] C. Peng, X. Zhang, G. Yu, G. Luo and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 4353–4361, 2017. https://doi.org/10.1109/cvpr.2017.189

[27] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2881–2890, 2017. https://doi.org/10.1109/cvpr.2017.660

[28] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 801–818, 2018. https://doi.org/10.1007/978-3-030-01234-2_49

[29] L. C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[30] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang *et al.,* "Understanding convolution for semantic segmentation," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, pp. 1451–1460, 2018. https://doi.org/10.1109/wacv.2018.00163

[31] T. Xiao, Y. Liu, B. Zhou, Y. Jiang and J. Sun, "Unified perceptual parsing for scene understanding," in *Pro. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 418–434, 2018. https://doi.org/10.1007/978-3-030-01228-1_26

[32] R. B. Rusu, N. Blodow, Z. C. Marton and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Nice, France, pp. 3384–3391, 2008. https://doi.org/10.1109/IROS.2008.4650967

[33] R. B. Rusu, N. Blodow and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *IEEE Int. Conf. on Robotics and Automation*, Kobe, Japan, pp. 3212–3217, 2009. https://doi.org/10.1109/robot.2009.5152473