DSNet: A Dynamic Squeeze Network for Real-time Weld Seam Image Segmentation

Jia Chen, Congcong Wang, Fan Shi, Mounir Kaaniche, Meng Zhao, Yan Jing, Shengyong Chen

DSNet: A Dynamic Squeeze Network for Real-time Weld Seam Image Segmentation

Jia Chen [a,b], Congcong Wang [a,b,*], Fan Shi [a,b,*], Mounir Kaaniche[c,d], Meng Zhao [a,b], Yan Jing [e] and Shengyong Chen [a,b]

[a] School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
[b] Laboratory of Computer Vision and System of Ministry of Education, Tianjin University of Technology, Tianjin 300384, China
[c] Université Sorbonne Paris Nord, L2TI, UR 3043, F-93430, Villetaneuse, France
[d] Université Paris-Saclay, CentraleSupélec, CVN, INRIA, Gif-sur-Yvette, 91190, France
[e] UMA Intelligent Group, China

ARTICLE INFO

Keywords:
Robot welding
Structured light vision
Seam tracking
Image segmentation
Lightweight network

ABSTRACT

The image noise generated by the welding process, such as arc light, splash, and smoke, brings significant challenges for the laser vision sensor-based welding robot to locate the weld seam and accurately conduct automatic welding. Currently, deep learning-based approaches surpass traditional methods in flexibility and robustness. However, their significant computational cost leads to a mismatch with the real-time requirement of automated welding. In this paper, we propose an efficient hybrid architecture of Convolutional Neural Network (CNN) and transformer, referred to as Dynamic Squeeze Network (DSNet), for real-time weld seam segmentation. More precisely, a lightweight segmentation framework is developed to fully leverage the advantages of the transformer structure without significantly increasing computational overhead. In this respect, an efficient encoder, which aims to increase its features diversity, has been designed and resulted in substantial improvement of encoding performance. Moreover, we propose a plug-and-play lightweight attention module that generates more effective attention weights by exploiting statistical information of weld seam data and introducing linear priors. Extensive experiments on weld seam images using NVIDIA GTX 1050Ti show that our approach reduces the number of parameters by 54x, decreases computational complexity by 34x, and improves inference speed by 33x compared to the baseline method TransUNet. DSNet achieves superior accuracy (78.01% IoU, 87.64% Dice) and speed performance (100 FPS) with lower model complexity and computational burden than most state-of-the-art methods. The code is available at https://github.com/hackerschen/DSNet.

1. Introduction

Welding is a crucial manufacturing process that connects multiple workpieces into a single unit and has a wide range of applications in various industrial domains such as automobile manufacturing, shipbuilding, and aerospace [1]. With the rapid development of the economy in recent years, higher requirements have been put forward for high-quality and efficient welding [2]. Limited by the technical level, harsh welding environments, and other objective and subjective factors [3], traditional manual welding methods can no longer meet the needs of the modern manufacturing industry. Nowadays, driven by the rapid development of robotics and artificial intelligence techniques, intelligent manufacturing is becoming a vital innovation direction in the industrial field. Under the guidance of this trend, intelligent welding robots emerged as an ideal choice for enhancing the quality and efficiency of welding [4]. One of the main objective of intelligent welding robots is to achieve an accurate perception of the workpieces through the seam tracking system.

*Corresponding author
congcong_wang@yeah.net (Congcong Wang); shifan@email.tjut.edu.cn (Fan Shi)

In this context, the laser structured light vision based system is widely adopted due to its good anti-interference performance and high measurement accuracy [5, 6, 7]. However, in current welding processes, a large number of welding arcs, smoke, and splashes can cause severe noise pollution to the laser stripe images captured by cameras, making extracting weld seam information more challenging. At the same time, due to the real-time requirement of welding tasks, the processing speed of the algorithm and the limited computing resources are also important criteria that should be considered. Therefore, there is an urgent need to develop a fast, high-precision, and lightweight image processing algorithm for weld seam segmentation and tracking.

Semantic segmentation, which is a fundamental challenge in computer vision, focuses on assigning a class label to every pixel within an image. The segmentation of laser stripes has been widely studied in the literature. Existing works can be divided into traditional- [8, 9] and deep learning-based methods [10, 11, 12].

Traditional image segmentation methods typically include two main steps. The first step is to filter the noise in the image using some popular techniques such as those based on morphological operations [8] and specific filters [13, 9]. The second step is to segment the target region from the image using different methods like thresholding, inter-class variance maximization, etc [14, 9, 15]. However, these methods are developed for specific noise and welding types and lack flexibility and accuracy, which limit their use in real application scenarios. Meanwhile, these methods usually rely on complex image processing operators, including critical parameters that need to be initialized through expert experience. This dependence leads to poor generalization ability of those traditional image segmentation methods [16]. Recent deep learning based methods can effectively solve the aforementioned drawbacks [17, 18] and significantly improve the accuracy of image segmentation. However, most of the existing methods [19, 16] use large models (e.g., UNet [20] and VggNet [21]) to obtain high accuracy, while ignoring the computation and latency aspects. In actual industrial environment, high accuracy and computational speed must be considered simultaneously. This brings difficulties to the practical application of the developed methods. Currently, there are relatively few works, described in Section 2, devoted to real-time segmentation of weld seam images. Therefore, the purpose of this paper is to implement an efficient laser stripe extraction algorithm that can significantly reduce the computational complexity and inference latency while maintaining high precision performance.

1.1. Motivation

Recent studies have focused on such specific problems for the practical application of image segmentation algorithms. However, most of the research [16, 22] mainly focus on the single metric of speed and assumes that such models should be low-latency and lightweight side by side when running in real-time on resource-constrained edge platforms. At the same time, Pan *et al.* [23] point out that the friendliness of the implementation is also essential for

real-world applications and broad deployment. Thus, the main objective of this work is not just to achieve high speed but to ensure a low latency, strong universality, lightweight, and highly accurate method for real-time weld segmentation.

More precisely, our approach relies on the autoencoder architecture widely used in semantic segmentation. While some recent approaches have focused on optimizing the computational efficiency of such architecture [24, 25], these approaches do not achieve a satisfactory trade-off between accuracy and speed. For this reason, we propose to reduce the high computational cost in the early stages of the encoder-decoder architecture by processing global information at a finer granularity. Indeed, by giving effective attention to important channels and regions through global information, we can achieve good results even using a plain CNN model-based encoder-decoder. In addition, although redundant feature maps in the network may contain a comprehensive understanding of the input data [26], too many redundant features also need to introduce more parameters and powerful computing resources, which are unacceptable for lightweight models. To this end, we design a new encoder that uses fewer parameters and less computation complexity to generate relevant weld seam features. The encoder projects the inputs into the latent space through a Dynamic Multilayer Perceptron (MLP) and then Restore MLP to capture the significant features of each downsampling stage. Compared with the baseline model, the proposed encoder reduces the computational overhead by about 4x and increases the inference speed by about 2x. Finally, since the weld data has prominent strip characteristics, we incorporate a lightweight attention mechanism to enhance the feature representation of the model and improve the segmentation accuracy. The proposed method was evaluated on both weld seam and road crack datasets. In comparison with the similar transformer-based generic segmentation framework TransUNet [27], our approach not only achieves a reduction in parameter count by 54x and a decrease in computational complexity by 34x, but also exhibits an increase in inference speed by 3x. Moreover, compared to the developed lightweight models, the proposed one not only outperforms the existing sizeable neural network models in terms of accuracy but also achieves a higher number of Frame Per Second (FPS) with lower floating-point operations per second (FLOPs).

1.2. Contributions

The main contributions of this work are summarized in what follows.

- First, we propose a lightweight central aggregation segmentation framework that performs global fine-grained fusion and interaction of deep-level and high-level semantic information based on an encoder-decoder architecture.

- Second, we develop a Dynamic Squeeze (DS) strategy to guide lightweight models to ensure feature diversity, while using few parameters. Based on the DS strategy, we design the Dynamic Squeeze Encoder (DSE) that generates more efficient feature maps through simple spatial transformations.

- Third, we propose Dynamic Squeeze Attention (DSA), a lightweight attention module that generates spatial attention maps closer to the actual features in each channel. This is achieved by introducing linear priors to efficiently extract the spatial mapping relationships between pixels.

- Finally, based on the above modules, we propose the Dynamic Squeeze Network (DSNet), a new lightweight segmentation network architecture for extracting laser stripes. More specifically, DSNet achieves a Dice score of 87.64% on a publicly available welding seam dataset with 100 FPS using NVIDIA GTX 1050Ti.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 describes the proposed DSNet-based segmentation approach. Experimental results are shown in Section 4 and further discussions are provided in Section 5. Finally, some conclusions and perspectives are given in Section 6.

2. Related Work

Laser Stripe Extraction: The precise extraction of laser stripes is directly related to the accurate positioning of welding robots on workpieces and thus significantly impacts the final welding quality. Some researchers have used filter and threshold segmentation to extract laser stripes. Muhammad *et al.* [28] employed averaging, color processing, and blob analysis for laser stripe segmentation. Ye *et al.* [15] obtained an image binarization threshold using an open operation and OSTU algorithm. Fan *et al.* [29] proposed a seam feature point acquisition method based on efficient convolution operators (ECO) and particle filter (PF). Such methods are often designed based on specific noise types and have low generalization capabilities. Recently, CNN-based methods [30] have achieved significant success with notable improvements in usability and accuracy. An attention dense convolutional block [16] is proposed to extract and accumulate multi-scale features. A restoration and extraction network [31], based on a conditional generative adversarial network, is also developed to extract the seam feature points. While these methods yield high accuracy performance, they use large models and lead to an increase in computational delay. As a result, some recent works [32, 33] have been developed to further improve existing networks.

Semantic Segmentation: Encoder-decoder is one of the most popular architectures in semantic segmentation. UNet [20] has achieved a breakthrough using the encoder-decoder structure and shortcut connections. UNet uses a U-shaped encoder-decoder network and has greatly succeeded in medical image segmentation. Although the U-shaped structure is simple, it performs outstandingly on small sample datasets. Subsequently, many enhanced networks using the U-shaped structure have been proposed. This include DeepLab [34], UNeXt [35], etc. These methods have achieved good performance, but most of them have not considered the computational complexity factor, resulting in high inference delay.

Real-time Semantic Segmentation: The goal of real-time semantic segmentation is to achieve a trade-off between accuracy and speed, and so enable the algorithm to be practically implemented in real-world environments. The developed models have a mainstream architecture based on an encoder-decoder structure [36]. The encoder-decoder structure reduces computation by gradually decreasing image resolution, thus saving inference time. Based on this structure, Li *et al.* [36] proposed an attention fusion module to strengthen feature representations. Such attention mechanism [37] allows the model to focus on important parts in the features and ignore irrelevant parts, thereby achieving better performance. Moreover, the lightweight network can enhance its predictive accuracy through attention mechanisms. For instance, HMANet [38] takes advantage of the attention modules to comprehensively capture feature correlations from the perspective of space, channel, and category. CSART [39] effectively improves tracking accuracy without adding much calculation and memory by considering different types of self-attention in two dimensions. In addition to segmentation, attention mechanisms are also extensively applied in various other fields [40, 41, 42].

In recent years, transformer-based architectures have performed better than CNN in various computer vision tasks. Some studies have employed the global information modeling feature of transformers to obtain better performance in real-time segmentation. Among them, one can mention the Simple and Efficient segmentation framework with Transformers (SegFormer) [43], the Adaptive Frequency Transformer (AFFormer) [44], and the Squeeze-enhanced Axial Transformer (SeaFormer) [45]. In addition to the above networks, which directly apply popular transformer blocks, other network design strategies have been developed by investigating and enhancing some critical components of transformer architectures. For instance, Melas-Kyriazi *et al.* [46] proposed to replace the attention layer with a Feed-Forward structure applied over the patch dimension. ConvMixer [47] is a new class of models that mixes the spatial and channel locations within a simple patch embedding stem. ConvNeXt [48] is a simple and efficient family of pure ConvNet models that competes favorably with transformers in terms of accuracy and robustness.

3. Proposed Dynamic Squeeze Network

In this section, we first present the overall architecture of the proposed DSNet architecture and then describe its main blocks.

3.1. Network Architecture

The proposed DSNet architecture is illustrated in Fig. 1. It is composed of three main modules referred to as Encoder, Central Aggregation, and Decoder. These modules and their main components will be described in what follows.

Encoder: Firstly, we extract high-level semantic information through the encoder while passing rich details to the decoder through residual mapping. A Dynamic Squeeze (DS) strategy is introduced in the encoder module to obtain

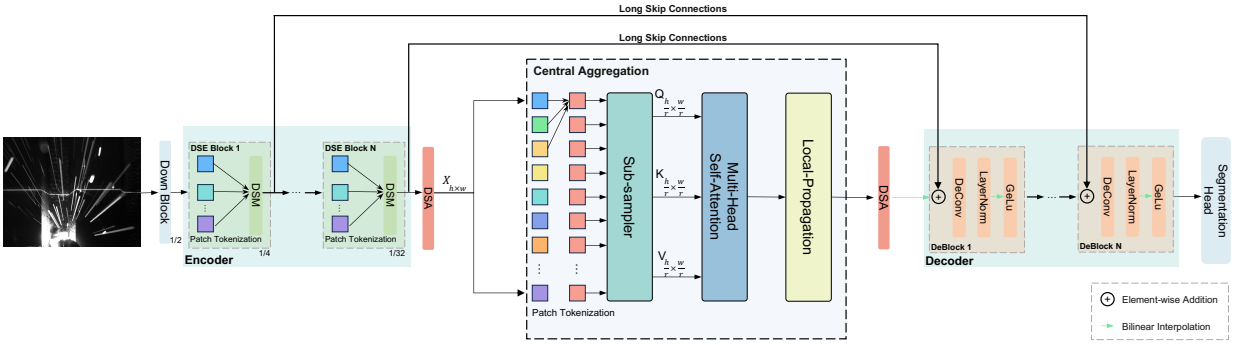


Figure 1. The overall architecture of DSNet. It consists of three main modules: Encoder, Central Aggregation, and Decoder. The hierarchical encoder is used for the progressive extraction of both low-level and high-level semantic features, and the encoded data is subject to adaptive attention through the Dynamic Squeeze Attention (DSA) module. The central aggregation is employed to fuse high-level semantic features and model global information. The lightweight decoder is used for predictive semantic segmentation encoding. The encoder and decoder preserve information details through long skip connections.

It is worth noting that the encoder and decoder are symmetric four-layer structures (i.e., $N = 4$).

more diverse feature maps through a simple and optimized feature extraction structure. The encoder is composed of lightweight Dynamic Squeeze Encoder (DSE) modules stacked together. The stacking level N is set to 4. Each DSE module consists of patch tokenization and Dynamic Squeeze MLP (DSM), which are illustrated in Figure 2 and described in Section 3.2. We employ long skip connections to reconstruct information loss caused by the downsampling process, thereby preserving more abundant information details. At the end of the encoder, a Dynamic Squeeze Attention (DSA) module, presented in Section 3.3, is used for more profound feature enhancement.

Central Aggregation: Secondly, to benefit from the lower spatial resolution size of the image and the good high-level semantic information, we can model global information by introducing operations with higher computational complexity and combining it with the local contextual features extracted by CNN. Through such a central aggregation approach, the performance of the model is greatly improved with a low overall computational complexity. More precisely, DSNet performs global context modeling via Multi-Head Self-Attention. In this respect, we adopt Global Sparse Attention module, which is a lightweight self-attention model referred to as EdgeViTs [23]. It uses a sub-sampler to shorten the length of the Query, Keys, and Values (denoted in Fig. 1 by Q, K, and V, respectively). Simultaneously, local propagation propagates the global contextual information encoded in the delegate tokens to their neighboring tokens.

Moreover, it should be noted that the computational complexity of the self-attention module is $O(N_p^2 d)$ where N_p denotes the number of patches and d represents the channel dimension. In order to reduce the computational effort without decreasing the accuracy, we add a stride of 2 for the patch module to shorten the sequence length and consider a deep convolution of size 3×3 for encoding and aggregating positional information. Thus, due to the excellent feature diversity, we can significantly reduce the number of channels and obtain a lightweight self-attention module.

Decoder: After applying the DSA mechanism to enhance the features from the central aggregation module, the decoder block is used while adopting the symmetric design of an encoder. To this end, we replace the DSE module with a deconvolution (DeConv) operator of size 3×3, followed by layer normalization and GELU activation function. The final output is obtained through Segmentation Head, which uses 1×1 convolution.

Comparison to existing designs. Our model shares a similar objective with recent works such as TopFormer [49] and SeaFormer, which aim to reduce the heavy computational cost associated with the self-attention mechanism. However, there are fundamental differences in the core architecture of these models. TopFormer and SeaFormer rapidly decrease the input resolution of images through swift downsampling, which will significantly reduce the computational burden of the self-attention module. While effective, this approach can easily overlook boundary details and appearances surrounding objects. Moreover, both SeaFormer and Afformer employ high-ratio upsampling to produce the final prediction masks, which may result in the loss of fine details and blurred edges. These compromising choices are made to reduce computational complexity. In this paper, we rethink the architectural approach of CNNs and transformers. More precisely, We propose to fully exploit the advantages of deep feature maps with smaller resolutions and rich semantic features through central aggregation. This approach not only substantially reduces the computational cost of the self-attention mechanism but also fully incorporates its global information modeling capabilities. Furthermore, our efficiently designed lightweight encoder and decoder modules retain the capability of the original encoder-decoder architecture to effectively handle information details. Finally, while BiSeNetV2 [24] reduces computational latency through a multi-branch approach, it also increases the computational complexity of the model. As shown in our experiments (Tables 2 and 3), our design achieves a better balance between model performance and computational costs such as latency, computational complexity, and model size.

In what follows, we will further describe the main blocks of the proposed DSNet, including the DSE and DSA modules, and the loss function used for training the overall architecture. Table 1 describes the structure of our DSNet architecture.

3.2. Dynamic Squeeze Encoder

The Dynamic Squeeze Encoder (DSE), illustrated in Fig. 2, consists of two main components: patch tokenization and Dynamic Squeeze MLP. Firstly, we use patch tokenization to project the image into the abstract image token to reduce the spatial resolution. Compared with the traditional convolution and pooling-based downsampling method, patch tokenization uses a single step to complete the downsampling and channel conversion, which is lighter and dramatically reduces the loss of information. Subsequently, inspired by [50], we propose to dynamically generate the positional encoding of the token while exploiting its local neighborhoods. Finally, compared to the MLP mixer [51], we rethink the hidden dimensions of the Feed-Forward structure and propose the Dynamic Squeeze MLP (DSM).

Table 1
Description of our DSNet architecture. Note that the Conv2d shown in the table refers to Conv-LN-GELU. KSize means kernel size. S, C, and T denote the stride, input channels, and a hyperparameter of the DSE module, respectively.

Input	Operator	T	C	KSize	S
512×512	Conv2d	–	3	2×2	2
256×256	DSE	4	64	2×2	2
128×128	DSE	2	96	2×2	2
64×64	DSE	4	128	2×2	2
32×32	DSE	4	160	2×2	2
16×16	DSA	–	320	–	–
16×16	Self-Attention	–	320	–	–
16×16	DSA	–	320	–	–
16×16	DeConv	–	320	3×3	–
32×32	DeConv	–	160	3×3	–
64×64	DeConv	–	128	3×3	–
128×128	DeConv	–	96	3×3	–
256×256	Transpose Convolution	–	64	2×2	2
512×512	Conv2d	–	8	1×1	–

Indeed, MobileNetV2 [52] uses the channel expansion structure to expand the hidden space dimension and avoid information loss. The feed-forward structure of the vision transformer [53] also follows this idea to design the hidden dimension. However, we found that this strategy is not always relevant. Such a design does not consider the hierarchical relationship of the network. The abstract expression capabilities of low-level and high-level semantic information are different. Moreover, adopting the same dimension expansion strategy in the higher layers of the network as in the lower layers may result in feature redundancy. For this reason, we propose the Dynamic Squeeze strategy to determine the hidden dimensions of the DSM, which bring lighter and better performance for DSNet.

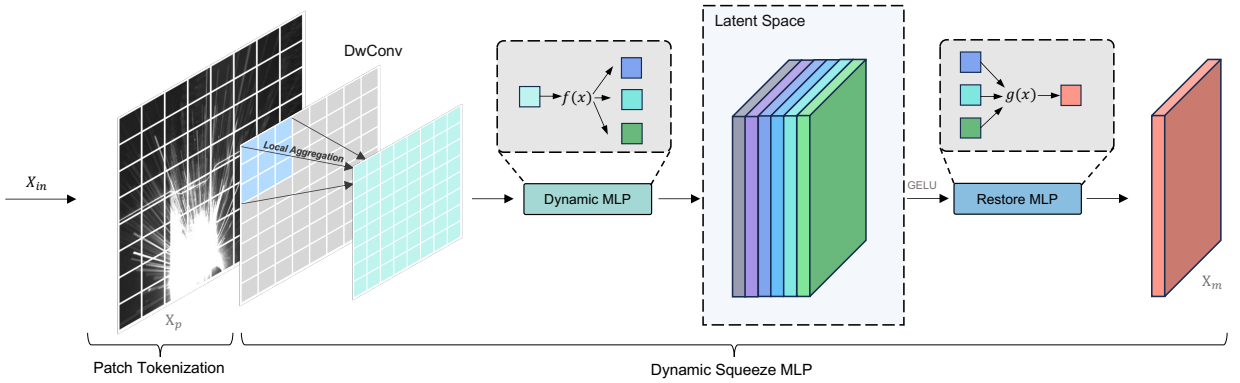


Figure 2. Illustration of DSE. DSE mainly consists of patch tokenization and Dynamic Squeeze MLP (DSM). The patch tokenization aggregates local information of neighboring tokens into the target token. The DSM is employed for dynamic feature selection. DwConv denotes depth-wise convolution, and GELU denotes Gaussian Error Linear Units.

$f(x)$ and $g(x)$ are nonlinear transformation functions, which are implemented here by Multilayer Perceptron.

As shown in Fig. 2, DSE is divided into two main modules. The first is the patch tokenization module for downsampling, and the second is the DSM module for extracting deep semantic information.

Patch Tokenization: It is performed by applying, to a given input feature map $\mathbf{X}_{in}$, a convolution with a kernel size of $K \times K$ and a stride of s along with Layer Normalization (LN) and GELU. This step is similar to the non-overlapping patch embedding in the vision transformer. Instead of using Batch Normalization and RELU, our design considers Layer Normalization and GELU, which have proven to be more stable and efficient in some recent architectures. Thus, the output feature maps $\mathbf{X}_p$ (i.e., the output of the patch) are obtained as follows:

$$\mathbf{X}_p = LN(GELU(Conv(\mathbf{X}_{in}))). \tag{1}$$

In order to ensure a lightweight structure design, the block of the first DSE module is composed of two patches with a stride of 2 to achieve a downsampling by factor 4 and quickly reduce the overall computational complexity. The remaining three DSE modules adopt a downsampling by factor 2.

Dynamic Squeeze MLP: As shown in Fig. 2, we pass the obtained feature maps $\mathbf{X}_p$ through a 3×3 depthwise convolution ($DwConv$) to perform position encoding and integrate information from locally neighboring tokens (i.e., local aggregation). The resulting features are then fed into a Dynamic MLP. Dynamic MLP uses the DS strategy (described later) to control the hidden dimensions of the MLP unit and adopt appropriate channel strategies according to different depth stages of the model. After that, the dimension is recovered by GELU and Restore MLP modules. Thus, the generated feature map $\mathbf{X}_m$ (i.e., the output of the DSM block) is given by

$$\mathbf{X}_m = MLP_r(GELU(MLP_d(DwConv(\mathbf{X}_p)))), \tag{2}$$

where r and d refer to the dimension restoration and the dimension transformation, respectively.

It should be noted here that the hidden layer dimensions were set to 4, 2, 1/4 and 1/2 times of the input dimension in the DSM modules of the first, second, third and fourth stages of the encoder, respectively. This allows to produce different characteristics of the features in different stages and increase their diversity.

Dynamic Squeeze strategy: Some of the latest existing works, such as ViT and ConvNeXt, adopt a design similar to the channel expansion strategy in the linear bottleneck of MobileNetV2 for the design of the hidden dimension of the Feed-Forward structure by setting the hidden dimension to four times of the input dimension. However, we found in our experiments that such a design is not always practical for encoder-decoder architecture, and it may affect the model accuracy of lightweight semantic segmentation models, especially when channel extensions are used in the deeper phases of the model, where the accuracy degradation is particularly noticeable. This may be related to the relationship between the model architecture level and the number of channels. Indeed, shallow networks with small receptive fields are mainly used for extracting low-level semantic information, and more channels are needed to encode and generate richer information, which coincides with the strategy of channel expansion. However, the high-level network has large

receptive fields and focuses more on processing global and high-level semantic information. Moreover, increasing the number of channels may cause information redundancy. Although redundancy in feature maps may help to train a deep neural network successfully, it may cause severe performance drop for lightweight models with few parameters and weak information processing capability.

To further check our conjecture, we refer to the method in ConvNeXt V2 [54] to evaluate feature diversity. Given a feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, the feature diversity can be computed as follows:

$$D_v = \frac{1}{C^2}\sum_{i=1}^{C}\sum_{j=1}^{C}\frac{1-cos(\mathbf{X}_i,\mathbf{X}_j)}{2}, \tag{3}$$

where $\mathbf{X}_i \in \mathbb{R}^{H\times W}$ is the feature vector of the i-th channel. Note that a higher value D_v indicates more feature diversity, while a lower value indicates feature redundancy.

In this respect, we select all the validation images of the experimental weld seam dataset and extract high-dimensional features from the encoder of different models, including those using the encoder of UNet and other strategies. Then, we average the values computed by each layer. The results are plotted in Fig. 3. In this figure, the baseline refers to the DSNet that replaces the DSE module with the encoder of UNet and removes the DSA module. Div, Multi, and DS denote the zooming in and out of the hidden dimensions in the DSE module and using the DS strategy. Thus, a large amount of redundancy is obtained with the encoder of UNet, which is consistent with our conjecture. This analysis also shows that tested strategies have different impacts on the feature maps diversity and model performance. Most importantly, the DS strategies not only effectively ensure features diversity but also improve accuracy performance and reduce computational cost.

Based on these observations, we propose the Dynamic Squeeze (DS) design strategy. Firstly, the feature expansion strategy is adopted in the shallow stage of the model. Thus, the dimension of the hidden layer is set to be multiple of the input dimension to capture more feature information. Secondly, in the deep stage of the model, as the number of channels increases, the input dimension is set to be multiple of the hidden dimension to retain truly effective information while reducing computational complexity to obtain an efficient and lightweight network.

The encoder, typically positioned in the early stages of the model, is responsible for extracting key features from raw data. However, large spatial resolutions can rapidly increase computational overhead if a complex encoder is used. Therefore, an efficient encoder is crucial for enhancing the efficiency of DSNet. As demonstrated in Table 4, the use of DSE not only achieves better accuracy but also reduces the computational complexity by approximately 4x and increased the inference speed by nearly 2x, confirming the effectiveness of the DSE module. However, due to the few number of parameters in lightweight encoders, this may limit their ability to capture complex data features.

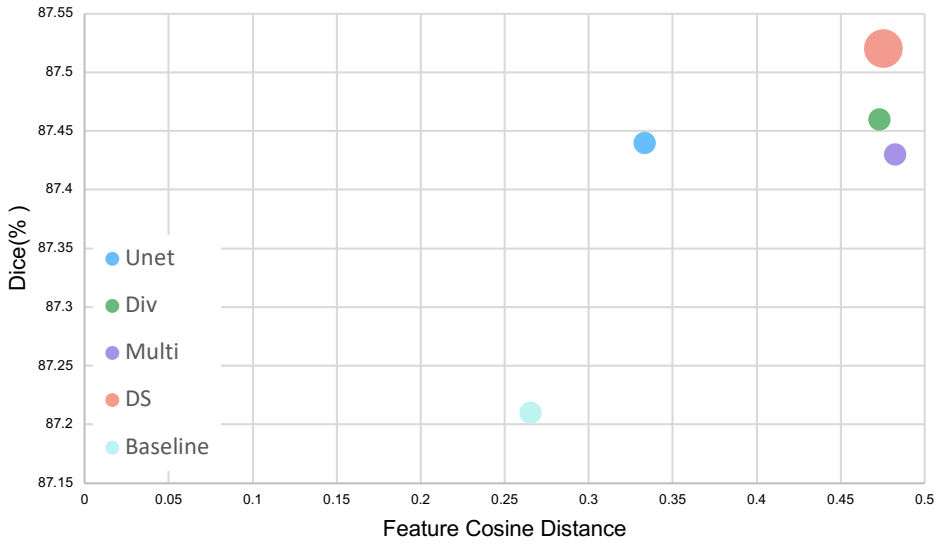


Figure 3. Feature diversity analysis. This is achieved by computing the cosine distance of the first four layers between UNet and DSNet (without the DSA modules) by using different strategies and averaging the results. We also show the baseline model, Div, Multi, and DS strategy to ensure fairness.

3.3. Dynamic Squeeze Attention

To generate attention weights, lightweight attention mechanism [37] often uses global pooling to extract statistical information and reduce computational complexity. However, this approach may lead to information loss. For example, using a single value to represent the entire feature map or channel can be easily affected by extreme values in the data distribution, which may lead to a drop of the model performance. The commonly used self-attention mechanism will greatly increase the computational complexity when performed on the original data. To this end, we adopt adaptive pooling to control the computational complexity while retaining more statistical information. In addition, since laser stripe images have obvious straight-line features, we introduce linear priors to guide DSA in generating more effective attention weights, thereby enabling the model to generate more effective features. In this respect, we add strip convolution as a complement to grid convolutions in attention to help extract strip-like features [55]. Moreover, the design of DSA also follows the idea of our DS strategy, which aims to control the hidden dimensions.

As shown in Fig. 4, the DSE output features $\mathbf{X}_m$ will serve as input for the attention module, composed of different steps. More specifically, we first perform adaptive pooling (denoted by $F_{AdaPool}$ function) on $\mathbf{X}_m$ to convert it to an $M \times M$ window size in order to control the computational complexity. Different from the global pooling method, we use more values to represent the features, which can extract more fine-grained and accurate statistical information. Then, we pass the image through a 3×1 convolution with the number of output channels that is half the number of input channels. Subsequently, we restore the channel number of the input to its original size by running it through a 1×3 convolution. After that, we resize the image to its original size by a bilinear interpolation (BI) algorithm and

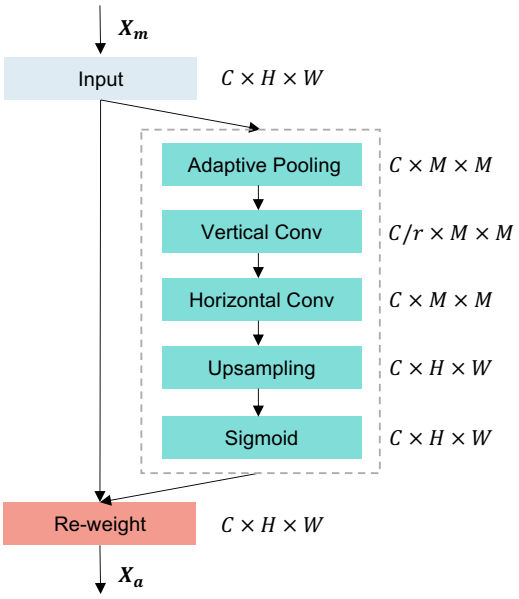


Figure 4. Diagram of the Dynamic Squeeze Attention (DSA) module. DSA controls computational complexity by using Adaptive Pooling and then re-scaling the generated attention weights back to the original size via upsampling. H and W denote the input image size, M denotes the window size, and c denotes the base number of the feature map channels, while k is the scaling factor.

generate the ultimate attention weight α through a Sigmoid function σ. Thus, the attention weight α is given by:

$$\alpha = \sigma(BI(Hconv_r(Vconv_d(F_{AdaPool}(\mathbf{X}_m))))), \tag{4}$$

where $Vconv_d$ denotes 3×1 vertical convolution with LN, and $Hconv_r$ denotes 1×3 horizontal convolution with GELU. Let us recall that the indexes d and r have been defined in (2).

Finally, the attention weight α is multiplied with the input features $\mathbf{X}_m$ to produce the output features $\mathbf{X}_a$:

$$\mathbf{X}_a = \mathbf{X}_m \cdot \alpha. \tag{5}$$

3.4. DSNet Loss Function

For the image segmentation network, the training direction of the model is to minimize the gap between the predicted result and the ground truth. In this regard, the design of the loss function plays a crucial role. Due to the small proportion of laser stripe pixels in the seam image, there is a serious imbalance between positive and negative samples. The Dice loss $\mathcal{L}_{dice}$ is firstly adopted to alleviate this problem. It is defined as follows:

$$\mathcal{L}_{dice}(\hat{y}, y) = 1 - \frac{2\,|\hat{y} \cap y|}{|\hat{y}| + |y|} \tag{6}$$

where y is the ground truth, and $\hat{y}$ is the prediction result.

In addition, to make the training process more stable, we introduce binary cross entropy loss $\mathcal{L}_{bce}$ given by

$$\mathcal{L}_{bce}(\hat{y}, y) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{7}$$

Therefore, we propose to define the DSNet loss function $\mathcal{L}$ as the arithmetic mean of the two previous expressions, i.e.

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}\Big(\mathcal{L}_{bce}(\hat{y}, y) + \mathcal{L}_{dice}(\hat{y}, y)\Big) \tag{8}$$

4. Experiments and Results

In this section, we first introduce the experimental platform, dataset, and implementation details. Then, we compare the proposed architecture in terms of accuracy and inference speed with state-of-the-art (SOTA) methods. Finally, an ablation study is conducted to illustrate the effects of the different modules in the proposed architecture.

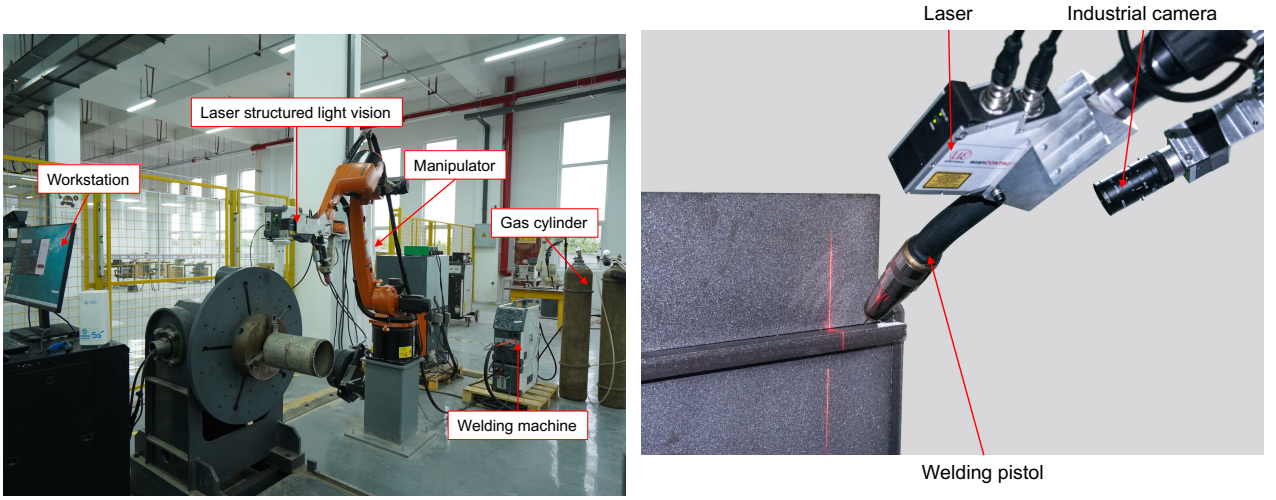
4.1. Experimental Platform

In order to validate the effectiveness of the proposed method, we built a primary welding experimental platform, as shown in Fig. 5a. The experimental platform is mainly divided into three parts: the robot system, the visual system, and the visual workstation. The robot system is composed of a gas cylinder, a welding machine, and a manipulator. The visual system consists of a laser for generating laser stripes and an industrial camera for image acquisition (as shown in Fig. 5b). The workstation uses NVIDIA GeForce GTX 1050Ti as the processing unit, and the entire algorithm test process is performed on the workstation. Our method is used to extract weld seams from acquired noisy images.

4.2. Dataset and implementation details

Dataset: We construct a new seam images dataset for laser stripe extraction based on a public seam image dataset[1]. The dataset consists of 193 images, composed of 48 images from the public dataset with a resolution of 1280×1024 and 145 images from our experiments with a resolution of 630×414. We randomly divided both sets (i.e., our experimental data and the public one) into training and test sets with an 8:2 ratio and finally merged them. In addition, due to the lack of accurate annotation data in the original public dataset, we re-annotated the data using LabelMe[2] and labeled it into two categories for ground truth. As shown in Fig. 6, the generated weld seam dataset includes weld seam data in multiple scenarios and different working conditions. To ensure the practical use of our algorithm, we collected images from different materials and devices during the welding process. These heterogeneous data introduce more realistic and diverse types of weld noise. For instance, Figs. 6 (a) and (b) exhibit more linear noise, while Figs. 6 (c) and (d)

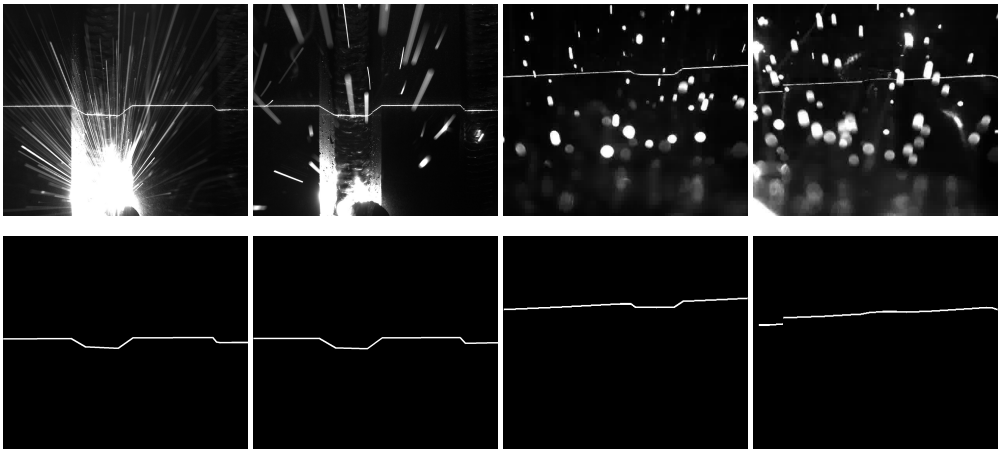
[1] https://aistudio.baidu.com/aistudio/datasetdetail/106021/
[2] https://github.com/wkentaro/labelme



(a) Structure of the experiment platform (b) The laser structured light vision

Figure 5. Illustration of the experimental platform.

display more point noise, posing higher demands on the adaptability and robustness of the algorithm. Furthermore, the dataset includes continuous and discontinuous weld seams to simulate the variability of real-world welding processes. Finally, we conducted more complementary experiments on a road crack dataset, composed of 151 images, with similar characteristics to the weld seam dataset to validate the robustness and generalizability of our method. This dataset is also split into two subsets using ratio of 8:2, which means that 120 images are used for training and 31 images are used for test purposes.



(a) (b) (c) (d)

Figure 6. Examples of welding seam images and their corresponding ground truth. (a)-(b) are from the public welding seam dataset, and (c)-(d) are from our experimental data.

Data augmentation: To improve the generalization ability and robustness of the model and reduce the risk of overfitting, we use offline data augmentation in training the weld dataset to transform and expand the training dataset. By random rotation, flip, and other transformations, we simulate samples under different viewpoints and scenarios to further expand the diversity of the training set. Eventually, the size of the whole training dataset is enlarged from 154 to 616. Through this strategy, we expect to learn a more robust and reliable DSNet model.

Training settings: To avoid potentially unfair comparison problems caused by different training sets, we adopt the control variable approach to train models. Additionally, when it comes to algorithms with pre-trained models, we loaded these models to enhance the accuracy and reliability of comparisons. Each model is trained using 600 epochs. The ADAM optimizer is employed with a learning rate equal to 2×10^{-3} while applying a decay of 10^{-4} and a batch size equal to 8. Moreover, a CosineAnnealingLR scheduler is used with a maximum number of iterations equal to 600 and a minimum learning rate equal to 10^{-5}. For the DSA module, we set $\boldsymbol{M}$ to 8 unless stated otherwise. Our model is implemented by PyTorch, and all experiments are conducted under the CUDA 11.4 and cuDNN 8.2.4 environments on the NVIDIA GeForce GTX 1050ti with 4GB of memory.

Inference settings: For fair and speed comparison, we adopt the same testing methods as performed in previous works [24]. More specifically, for the dataset images, we first resize the input image to a resolution of 512×512 for inference and then resize it to the original size. It is worth noting that the batch size is set to 2 to meet the requirements of certain comparison models while ensuring fairness. The cost of these steps is computed as the inference time. The latter is measured by using a single card and repeating the process for 5000 iterations to avoid the fluctuation in inference time caused by other factors.

4.3. Evaluation metrics

First, we employ the standard Dice coefficient and IoU, which are widely used to evaluate the accuracy of semantic segmentation. Let us recall that these metrics are computed as follows:

$$Dice = \frac{2\,|\hat{y} \cap y|}{|\hat{y}| + |y|} \tag{9}$$

$$IoU = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|} \tag{10}$$

where y is the ground truth and $\hat{y}$ is the prediction result.

Moreover, we consider the Giga-FLOPs (GFLOPs) metric for computing complexity comparison and the Frame Per Second (FPS) metric for inference speed comparison. These metrics are often used to evaluate real-time segmentation algorithms. In the following tables, the best evaluation results are marked in bold, while the second ones are underlined.

Table 2
Performance comparison with different non-real-time and real-time methods for the weld seam images dataset.

Model	venue	Backbone	GFLOPs	IoU(%)	Dice(%)	FPS	Params (M)
UNet[20]	2015	–	160.74	77.69	87.44	5	7.26
DeepLabV3+[34]	2018	MobileNetV2	6.16	77.15	87.1	39	4.38
DeepLabV3+[34]	2018	EfficientNet	2.34	76.84	86.9	23	0.94
TransUNet[27]	2021	–	126.25	77.51	87.32	3	93.23
Segdeformer[56]	2022	–	15.74	76.39	86.57	2	3.125
Swin-Unet[57]	2023	–	8.69	75.06	85.75	29	41.34
ESPNetV2[58]	2019	–	**0.82**	72.17	83.82	62	**0.33**
SFNet[36]	2020	ResNet18	30.47	77.71	87.46	21	12.87
BiSeNetV2[24]	2021	–	12.33	66.3	79.73	49	3.34
DDRNet[59]	2021	–	17.97	70.55	82.73	33	20.15
Segformer[43]	2021	–	6.76	76.07	86.4	24	3.71
UNeXt[35]	2022	–	2.29	76.31	86.56	82	1.47
TopFormer[49]	2022	–	1.68	71.18	83.16	65	5.02
Ffnet[60]	2022	–	15.63	71.86	83.62	40	25.28
Ppliteseg[61]	2022	–	31.54	71.36	83.29	47	21.58
Afformer[44]	2023	–	3.27	71.16	83.15	41	2.35
SeaFormer[45]	2023	–	1.75	70.65	82.8	29	8.50
DSNet	–	–	3.63	**78.01**	**87.64**	**100**	1.70

4.4. Performance comparison

Results on weld seam images dataset: The performance of the proposed DSNet approach is compared to several state-of-the-art methods using the aforementioned training and inference settings. The first group consists of non-real-time segmentation methods (given at the upper part of Table 2), and the second group consists of real-time semantic segmentation algorithms (given at the lower part of Table 2). Table 2 presents the model reference, backbone, IoU, Dice, GFLOPs, and FPS. As shown in Table 2, the proposed method yields a better trade-off between accuracy and inference speed than non-real and real-time segmentation methods, reaching a Dice coefficient of 87.64% with 100 FPS. Moreover, the GFLOPs and total number of parameters indicate that our method achieves comparable or better results than most state-of-the-art methods in model complexity and computational burden. Fig. 7 shows the results for the different methods to further demonstrate the trade-off between accuracy and inference speed. In the comparison of real-time semantic segmentation, CNN-based methods generally achieve better results than transformer-based methods. This behaviour is due to the number of samples. In the case of limited computational resources and limited data, the redundant attention mechanism backbone network of the transformer makes the learning of rich features more challenging, which could explain the poor performance of transformer-based networks. Our approach better integrates the advantages of both CNN and transformer and still achieves remarkable results with small sample datasets. Compared to the recent efficient method UNeXt [35], the proposed DSNet architecture achieves better accuracy and speed performance with similar GFLOPs and number of parameters.

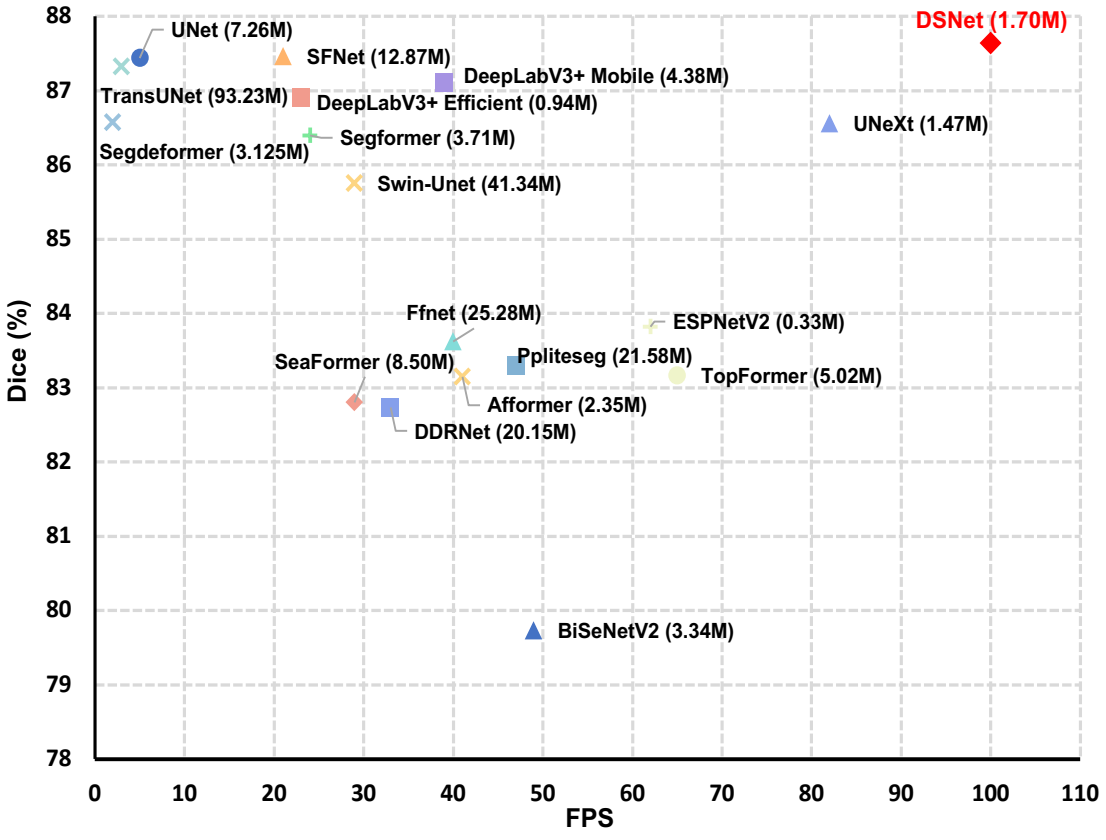


Figure 7. Speed-Accuracy trade-off comparison for the weld seam images dataset. Our methods are marked in red. The testing device is NVIDIA GTX 1050Ti. The experimental results demonstrate that our approaches achieve a state-of-the-art trade-off between accuracy and speed.

Results on road crack dataset: In addition, our experiments are conducted on the road crack dataset. Table 3 shows the comparison results for the different methods. Again, DSNet outperforms the real-time as well as the recent non-real-time segmentation methods, while achieving a Dice coefficient of 71.81%. This further demonstrates the effectiveness and generalization ability of the proposed DSNet model.

4.5. Ablation study

Ablation experiments are conducted to illustrate the contribution of the different modules to the overall DSNet architecture performance. The baseline model adopts DSNet while replacing the DSE module with the encoder of the UNet architecture and removing the DSA module. The specific experiment results are shown in Table 4. On the one hand, it can be seen from Table 4 that adding only the DSA module to the baseline model achieves an improvement of 0.14% in terms of accuracy. However, there is no significant change in FLOPs and FPS. On the other hand, by replacing only the UNet encoder with the DSE module (in the baseline model), an improvement of 0.26% in terms of accuracy is achieved while resulting in a more significant gain in terms of FPS (about 138%) and FLOPs (16.60 GFLOPs vs.

Table 3
Performance Comparison with different non-real-time and real-time methods for the road crack dataset.

Model	venue	Backbone	GFLOPs	IoU(%)	Dice(%)	FPS	Params (M)
UNet[20]	2015	–	160.74	**57.03**	**72.5**	5	7.26
DeepLabV3+[34]	2018	MobileNetV2	6.16	54.81	70.72	39	4.38
DeepLabV3+[34]	2018	EfficientNet	2.34	53.64	69.71	23	0.94
TransUNet[27]	2021	–	126.25	56.05	71.73	3	93.23
Segdeformer[56]	2022	–	15.74	55.15	70.52	2	3.125
Swin-Unet[57]	2023	–	8.69	52.85	69.02	29	41.34
ESPNetV2[58]	2019	–	**0.82**	50.52	67.01	62	**0.33**
SFNet[36]	2020	ResNet18	30.47	55.82	71.54	21	12.87
BiSeNetV2[24]	2021	–	12.33	45.49	62.33	49	3.34
DDRNet[59]	2021	–	17.97	50.29	66.80	33	20.15
Segformer[43]	2021	–	6.76	48.26	64.95	24	3.71
UNeXt[35]	2022	–	2.29	55.37	71.13	82	1.47
TopFormer[49]	2022	–	1.68	48.92	65.56	65	5.02
Ffnet[60]	2022	–	15.63	54.4	70.36	40	25.28
Ppliteseg[61]	2022	–	31.54	52.5	68.76	47	21.58
Afformer[44]	2023	–	3.27	51.06	67.52	41	2.35
SeaFormer[45]	2023	–	1.75	51.19	67.58	29	8.50
DSNet	–	–	3.63	56.13	71.81	**100**	1.70

3.81 GFLOPs). By integrating DSE and DSA modules, a high number of FPS is still obtained while reaching a Dice gain of 0.32%. Finally, by adding the DS strategy, the overall DSNet model achieves an improvement of 0.47% in Dice, an increase of 150% in FPS, and a decrease of approximately 5 times in FLOPs (16.60 GFLOPs vs. 3.63 GFLOPs), compared to the baseline model.

We also used Grad-CAM [62] to visualize the feature maps generated by different attention models. Grad-CAM highlights the key areas that contribute to the prediction of the model while processing input images. This not only aids in understanding the decision-making process of the model but also enhances the interpretability of the attention mechanism. The obtained feature maps, illustrated in Fig. 8, show that our DSA can better localize the objects of interest than other attention methods. Table 5 also depicts the impact of using different attention methods on the performance of DSNet. Compared to the current state-of-the-art plug-and-play attention mechanisms, DSA is more capable of capturing the data features of weld seams without significantly increasing computational costs and inference latency. Thus, it can be observed that our DSA module achieves the highest gain in terms of Dice and IoU compared to the remaining attention models.

5. Discussion

One of the most significant challenges in real-time weld seam segmentation is finding a balance between high processing speed and high accuracy. This challenge stems from the conflicting nature of two core requirements. On

Table 4
Ablation study for our proposed modules on the validation set of the weld seam images dataset. The baseline model is DSNet without the different proposed modules. The last row is the whole model.

Model	DSE	DSA	DS	GFLOPs	IoU(%)	Dice(%)	FPS
Baseline				16.607	77.26	87.17	40
DSNet		✓		16.608	77.48	87.31	39
DSNet	✓			3.811	77.69	87.43	95
DSNet	✓		✓	**3.636**	77.83	87.52	**102**
DSNet	✓	✓		3.812	77.78	87.49	93
DSNet	✓	✓	✓	3.637	**78.01**	**87.64**	100

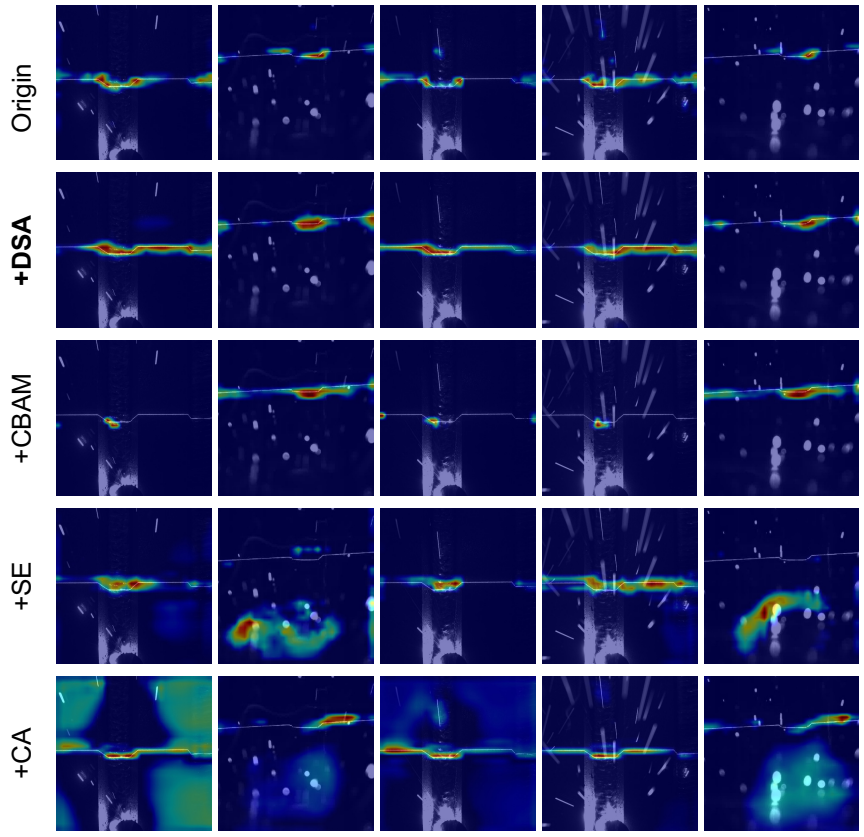


Figure 8. Visualization of feature maps before and after adding different attention models. The Origin model is the DSNet before adding a given attention module. Note that CBAM, SE and CA refer to the attention models proposed in [37], [63], and [64], respectively.

the one hand, segmentation algorithms must process input image data swiftly to achieve real-time processing. On the other hand, high accuracy is crucial to ensure the quality of weld seam segmentation and the reliability of subsequent welding operations. However, enhancing processing speed often comes with reduced model complexity, which can negatively impact segmentation accuracy. To address this dilemma, we propose DSNet, a lightweight method for weld seam segmentation. Tables 2 and 3 show a comparison of DSNet with the SOTA segmentation methods on weld seam and road crack datasets. The quantitative results in Table 2 demonstrate that the proposed method substantially

Table 5
Comparison of different attention modules. The Origin model is the DSNet before adding a given attention module.

Model	venue	GFLOPs	IoU(%)	Dice(%)	FPS
Origin	–	3.636	77.83	87.52	102
+DSA	–	3.637	**78.01**	**87.64**	**100**
+CBAM[37]	2018	**3.636**	77.59	87.38	94
+SE[63]	2018	**3.636**	77.82	87.52	98
+ECA[65]	2020	3.811	77.93	87.60	96
+CA[64]	2021	3.637	77.88	87.56	93
+FcaNet[66]	2021	4.048	77.78	87.49	88
+EMA[67]	2023	3.816	77.90	87.58	89
+OrthoNets[68]	2023	4.048	77.86	87.54	89

outperforms the state-of-the-art in real-time semantic segmentation in terms of both processing speed and accuracy, which confirms the effectiveness of the proposed model.

Figure. 6 displays the weld seam data used in this study, which includes two types of weld seams characterized by different environments, materials, and lighting conditions. The experimental results in Tables 2 and 3 reveal that, with adequate training, deep learning-based methods exhibit good robustness to different types of weld seam data. This robustness is inherently linked to the data-driven nature of deep learning models. Such models can automatically learn and extract complex features from large datasets, reducing reliance on expert knowledge and manual feature engineering. However, this characteristic also introduces a significant limitation: the models often fail to maintain high accuracy on new data types not seen during training. For practical industrial applications of the model, scene variation is an inevitable challenge. This issue might be addressed by providing a more diverse dataset or through domain adaptation. Furthermore, the computational resources required for the proposed method could be further reduced by using deep learning deployment tools such as TensorRT.

Although the proposed method has achieved promising results, there are still some limitations that require further research in future work. Firstly, the dimensionality transformation in the proposed encoder module is fixed and does not automatically adjust according to the noise intensity in weld images. This limitation hinders further optimization of computational overhead for images with slightly less noise. Secondly, to demonstrate the superiority of our proposed framework, only standard self-attention mechanisms were considered, while some recently advanced algorithms have begun to adopt more efficient self-attention designs to optimize computational efficiency. One potential solution to address the first issue could be to draw inspiration from dynamic convolution techniques [69], which dynamically determine encoder hyperparameters based on the input.

6. Conclusion and perspectives

The main research direction of this paper is the accurate and fast extraction of laser stripes in weld seam images affected by complex visual disturbances. Traditional image processing approaches rely on complex methodologies and result in poor robustness and flexibility. In contrast, current deep learning-based methods often employ deeper or more complex network architectures to enhance accuracy. However, the resulting substantial computational costs limit their applicability in real-world industrial scenarios. Inspired by the design of modern network and transformer structures, we propose an efficient laser stripe extraction method for the real-time seam tracking system of intelligent welding robots. This paper conducts extensive experiments on weld seam and road crack datasets to validate the effectiveness of the proposed DSNet. DSNet not only achieves approximately a threefold increase in processing speed (100 FPS) compared to the best-performing real-time segmentation frameworks but also reaches an accuracy comparable to larger models, which allows it to meet the requirements for real-time weld tracking. The main contributions of this paper can be summarized as follows:

- An end-to-end lightweight segmentation network is proposed for efficient weld feature extraction under limited computational resources.
- An efficient feature encoder has been designed. It significantly enhances encoding efficiency by increasing feature diversity within the encoder. This approach not only reduces computational complexity but also improves segmentation performance.
- A lightweight attention module is proposed by introducing linear priors to efficiently extract the spatial mapping relationships between pixels.

Currently, we have only achieved excellent results on small sample datasets. In the future, we will strive to integrate the proposed method into intelligent welding robots to improve their real-time performance and robustness. Moreover, it would be interesting to explore and extend the proposed approach to perform more tasks and handle other types of image datasets.

7. Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (62102285, 62020106004, 92048301, 62176183). The authors would like to thank Dr. Yao Zhang from the Tianjin University of Technology for valuable discussion.

References

[1] Baicun Wang, S Jack Hu, Lei Sun, and Theodor Freiheit. Intelligent welding system technologies: State-of-the-art review and perspectives. *Journal of Manufacturing Systems*, 56:373–391, 2020.

[2] Chao Liu, Hui Wang, Yu Huang, Youmin Rong, Jie Meng, Gen Li, and Guojun Zhang. Welding seam recognition and tracking for a novel mobile welding robot based on multi-layer sensing strategy. *Measurement Science and Technology*, 33(5):055109, 2022.

[3] Yanbiao Zou and Guohao Zeng. Light-weight segmentation network based on solov2 for weld seam feature extraction. *Measurement*, page 112492, 2023.

[4] Lei Yang, Junfeng Fan, Yanhong Liu, En Li, Jinzhu Peng, and Zize Liang. Automatic detection and location of weld beads with deep convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2020.

[5] Junfeng Fan, Fengshui Jing, Lei Yang, Long Teng, and Min Tan. A precise initial weld point guiding method of micro-gap weld based on structured light vision sensor. *IEEE Sensors Journal*, 19(1):322–331, 2018.

[6] Lei Yang, En Li, Teng Long, Junfeng Fan, and Zize Liang. A high-speed seam extraction method based on the novel structured-light sensor for arc welding robot: A review. *IEEE Sensors Journal*, 18(21):8631–8641, 2018.

[7] Nianfeng Wang, Kaifan Zhong, Xiaodong Shi, and Xianmin Zhang. A robust weld seam recognition method under heavy noise based on structured-light vision. *Robotics and Computer-Integrated Manufacturing*, 61:101821, 2020.

[8] Yanbiao Zou and Tao Chen. Laser vision seam tracking system based on image processing and continuous convolution operator tracker. *Optics and Lasers in Engineering*, 105:141–149, 2018.

[9] WP Gu, ZY Xiong, and W Wan. Autonomous seam acquisition and tracking system for multi-pass welding based on vision sensor. *The international journal of advanced manufacturing technology*, 69:451–460, 2013.

[10] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82:102616, 2022.

[11] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021.

[12] Jiangpeng Zheng, Fan Shi, Meng Zhao, Chen Jia, and Congcong Wang. Learning intra-inter-modality complementary for brain tumor segmentation. *Multimedia Systems*, 29(6):3771–3780, 2023.

[13] Liangyu Li, Lingjian Fu, Xin Zhou, and Xiang Li. Image processing of seam tracking system using laser vision. *Robotic welding, intelligence and automation*, pages 319–324, 2007.

[14] Yong-Hua Shi, Guo-Rong Wang, and Guo-Jin Li. Adaptive robotic welding system using laser vision sensing for underwater engineering. In *2007 IEEE International Conference on Control and Automation*, pages 1213–1218. IEEE, 2007.

[15] Hanmin Ye, Yingzhi Liu, and Wenjie Liu. Weld seam tracking based on laser imaging binary image preprocessing. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 756–760. IEEE, 2021.

[16] Lei Yang, Junfeng Fan, Benyan Huo, En Li, and Yanhong Liu. Image denoising of seam images with deep learning for laser vision seam tracking. *IEEE Sensors Journal*, 22(6):6098–6107, 2022.

[17] Jianhua Zhang, Jingbo Chen, Shengyong Chen, Zhenhua Wang, and Jianwei Zhang. Detection and segmentation of unlearned objects in unknown environment. *IEEE Transactions on Industrial Informatics*, 17(9):6211–6220, 2020.

[18] Arunabha M Roy, Jayabrata Bhaduri, Teerath Kumar, and Kislay Raj. Wildect-yolo: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics*, 75:101919, 2023.

[19] Kaixuan Wu, Tianqi Wang, Junjie He, Yang Liu, and Zhenwei Jia. Autonomous seam recognition and feature extraction for multi-pass welding based on laser stripe edge guidance network. *The International Journal of Advanced Manufacturing Technology*, 111(9-10):2719–2731, 2020.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Qi Wang, Jingwu Mei, Wuming Jiang, and Hegui Zhu. Shdm-net: Heat map detail guidance with image matting for industrial weld semantic segmentation network. *Engineering Applications of Artificial Intelligence*, 126:106946, 2023.

[23] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. EdgeViTs: Competing light-weight cnns on mobile devices with vision transformers. In *Proceedings of the European conference on computer vision (ECCV)*, pages 294–311, 2022.

[24] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:3051–3068, 2021.

[25] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021.

[26] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.

[27] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[28] Jawad Muhammad, Halis Altun, and Essam Abo-Serie. A robust butt welding seam finding technique for intelligent robotic welding system using active laser vision. *The International Journal of Advanced Manufacturing Technology*, 94:13–29, 2018.

[29] Junfeng Fan, Sai Deng, Yunkai Ma, Chao Zhou, Fengshui Jing, and Min Tan. Seam feature point acquisition based on efficient convolution operator and particle filter in gmaw. *IEEE Transactions on Industrial Informatics*, 17(2):1220–1230, 2020.

[30] Aditya Singh, Kislay Raj, Teerath Kumar, Swapnil Verma, and Arunabha M Roy. Deep learning-based cost-effective and responsive robot for autism treatment. *Drones*, 7(2):81, 2023.

[31] Chenfan Liu, Junqi Shen, Shengsun Hu, Dingyong Wu, Chao Zhang, and Hui Yang. Seam tracking system based on laser vision and cgan for robotic multi-layer and multi-pass mag welding. *Engineering Applications of Artificial Intelligence*, 116:105377, 2022.

[32] Zijian Wu, Peng Gao, Jing Han, Lianfa Bai, Jun Lu, and Zhuang Zhao. Real-time segmentation network for accurate weld detection in large weldments. *Engineering Applications of Artificial Intelligence*, 117:105008, 2023.

[33] Yunkai Ma, Junfeng Fan, Huizhen Yang, Hongliang Wang, Shiyu Xing, Fengshui Jing, and Min Tan. An efficient and robust complex weld seam feature point extraction method for seam tracking and posture adjustment. *IEEE Transactions on Industrial Informatics*, 2023.

[34] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[35] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: MLP-based rapid medical image segmentation network. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 23–33. Springer, 2022.

[36] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 775–793, 2020.

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[38] Ruigang Niu, Xian Sun, Yu Tian, Wenhui Diao, Kaiqiang Chen, and Kun Fu. Hybrid multiple attention network for semantic segmentation in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.

[39] Dawei Zhang, Zhonglong Zheng, Minglu Li, and Rixian Liu. Csart: Channel and spatial attention-guided residual learning for real-time object tracking. *Neurocomputing*, 436:260–272, 2021.

[40] Pengcheng Bian, Zhonglong Zheng, and Dawei Zhang. Light-weight multi-channel aggregation network for image super-resolution. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*, pages 287–297. Springer, 2021.

[41] Xiaoqiang Jin, Dawei Zhang, Qiner Wu, Xin Xiao, Pengsen Zhao, and Zhonglong Zheng. Improved siamcar with ranking-based pruning and optimization for efficient uav tracking. *Image and Vision Computing*, 141:104886, 2024.

[42] Arunabha M Roy and Jayabrata Bhaduri. Densesph-yolov5: An automated damage detection model based on densenet and swin-transformer prediction head-enabled yolov5 with attention mechanism. *Advanced Engineering Informatics*, 56:102007, 2023.

[43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[44] Dong Bo, Wang Pichao, and Fan Wang. Afformer: Head-free lightweight semantic segmentation with linear transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1–9, 2023.

[45] Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *Ninth International Conference on Learning Representation (ICLR)*, pages 1–19. The International Conference on Learning Representations (ICLR), 2023.

[46] Luke Melas-Kyriazi. Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet. *arXiv preprint arXiv:2105.02723*, 2021.

[47] Asher Trockman and J Zico Kolter. Patches are all you need? *Transactions on Machine Learning Research*, pages 1–19, 2023.

[48] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[49] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022.

[50] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations (ICLR)*, pages 1–19, 2023.

[51] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[52] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[53] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, pages 1–21, 2021.

[54] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023.

[55] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[56] Bowen Shi, Dongsheng Jiang, Xiaopeng Zhang, Han Li, Wenrui Dai, Junni Zou, Hongkai Xiong, and Qi Tian. A transformer-based decoder for semantic segmentation with multi-level context mining. In *European Conference on Computer Vision*, pages 624–639. Springer, 2022.

[57] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 205–218, 2023.

[58] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9190–9200, 2019.

[59] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*, 2021.

[60] Dushyant Mehta, Andrii Skliar, Haitam Ben Yahia, Shubhankar Borse, Fatih Porikli, Amirhossein Habibian, and Tijmen Blankevoort. Simple and efficient architectures for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2628–2636, 2022.

[61] Juncai Peng, Yi Liu, Shiyu Tang, Yuying Hao, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Zhiliang Yu, Yuning Du, et al. PP-liteseg: A superior real-time semantic segmentation model. *arXiv preprint arXiv:2204.02681*, 2022.

[62] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[63] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[64] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[65] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[66] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[67] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[68] Hadi Salman, Caleb Parks, Matthew Swan, and John Gauch. Orthonets: Orthogonal channel attention networks. In *2023 IEEE International Conference on Big Data (BigData)*, pages 829–837. IEEE, 2023.

[69] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6647–6656, 2021.