Wang *et al*. *Intell. Robot.* **2025**, *5*(1), 143-80
DOI: 10.20517/ir.2025.09

Intelligence & Robotics

Review

Open Access

An overview of industrial image segmentation using deep learning models

Guina Wang, Zhen Li, Guirong Weng, Yiyang Chen

School of Mechanical and Electric Engineering, Soochow University, Suzhou 215137, Jiangsu, China.

Correspondence to: Dr. Yiyang Chen, School of Mechanical and Electric Engineering, Soochow University, No. 8, Jixue Road, Suzhou 215137, Jiangsu, China. E-mail: yychen90@suda.edu.cn

How to cite this article: Wang, G.; Li, Z.; Weng, G.; Chen, Y. An overview of industrial image segmentation using deep learning models. *Intell. Robot.* **2025**, *5*(1), 143-80. http://dx.doi.org/10.20517/ir.2025.09

Received: 30 Dec 2024 **First Decision:** 22 Jan 2025 **Revised:** 24 Jan 2025 **Accepted:** 8 Feb 2025 **Published:** 24 Feb 2025

Academic Editor: Zuojin Li **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Image segmentation plays a vital role in artificial intelligence and computer vision with major applications such as industrial picking, defect detection, scene understanding and video surveillance. As parallel computing technologies develop, numerous deep learning (DL)-based segmentation algorithms have demonstrated practical performance with increased efficiency and accuracy. With the concept of DL image segmentation, a comprehensive review on recent literature is introduced in detail, including traditional image segmentation algorithms, DL schemes and the fusion of the former two algorithms. The seminal efforts of DL in image segmentation are elaborated in accordance with the quantity and quality of annotated labels, covering supervised, weakly-supervised, and unsupervised frameworks. Numerous methods on industrial benchmark datasets are compared and analyzed in standard evaluation indicators. Finally, the challenges and opportunities of DL image segmentation are discussed for further research.

Keywords: Image segmentation, deep learning, neural network

1. INTRODUCTION

As an important component of artificial intelligence, computer vision represents the science and technology on how to make machines perceive and make decisions from images. Image segmentation is one of the most fundamental tasks in computer vision. In general, it refers to the task of dividing an image or a frame in a video into several target regions according to the objects in the image or frame on demand. This task serves as a key component of many perception systems and plays a central role in a wide range of applications, such as industrial picking[1], autonomous driving[2], defect detection[3,4], virtual reality interaction[5], *etc.* The manufacturing of intelligent medical equipment, loading tools and identification systems is pivotal to promoting the digital upgrade and transformation in industry. Therefore, image segmentation is worthy of extensive research.

Image segmentation relies on prior knowledge and abstract representation of target and background to determine the target from the background or other erroneous objects. It is an important component of image understanding, which aims to separate targets and backgrounds for target recognition, and other subsequent processing. Accurate positioning will directly affect subsequent processing results. In order to achieve this goal, extensive and effective technologies have been developed over the years, including technologies based on traditional digital image processing methods such as threshold segmentation[6], edge detection[7] and region growing[8], active contour models (ACMs)[9], and recently utilized convolutional neural networks (CNNs)[10,11], generative adversarial networks (GANs)[12] or Transformer[13]. Although these methods have made significant progress, image segmentation still faces challenges in accurately depicting complex object boundaries, optimizing network structures, and processing multi-domain data.

Traditional approaches of image segmentation are mainly based on mathematical and image processing algorithms, dividing pixels of an image based on elementary image features. Traditional image segmentation divides images into several non-overlapping regions based on features such as grayscale, color, texture, and shape. These segmented targets make these features show similarities within the same region, while showing significant differences between different regions. Traditional image segmentation mechanism is widespread applied in textile, infrared, remote sensing, defect detection fields, *etc.* Common segmentation schemes include threshold method, edge detection method, clustering algorithm, and graph theory, and more, which achieve automatic extraction and recognition of interested target areas in images by different mathematical modeling. These methods are able to produce acceptable segmentation results with simple model implementation and low computational complexity.

The rapid development of parallel computing technologies such as deep learning (DL) and compute unified device architecture (CUDA) has directly propelled computer vision and image processing into a new technological era. DL has been extensively utilized in intelligent transportation navigation, defect detection, human behavior recognition[14], *etc.* As one of the fundamental problems in computer vision, image segmentation technologies are widely applied in fields such as robot control, defect detection, autonomous driving, assisted medical care, *etc.* DL-based image segmentation divides the image into several specific and unique regions based on features such as grayscale, color, spatial texture, and geometric shape, and extracts the target of interest. Unlike traditional image segmentation techniques that rely on manually designed features, DL-based image segmentation automatically learns and extracts complex features of images, thereby achieving more sophisticated image segmentation tasks.

DL-based image segmentation can be divided into three categories: semantic segmentation, instance segmentation, and panoptic segmentation. Semantic segmentation refers to dividing each pixel in an image into one of predefined categories, such as vehicles, pedestrians, insulators, *etc.* They are extensively applied in fields such as scene understanding and defect detection. However, the semantic segmentation framework is unable to distinguish different instances within the same category. Instance segmentation is more refined, aiming to detect all target instances in an image. It not only needs to distinguish diverse semantic categories, but also requires to differentiate distinct targets in each category. Panoptic segmentation combines semantic and instance segmentation to predict the semantic category of each pixel in the scene image and assign instance identification numbers to pixels belonging to instance targets. Panoptic segmentation allows for advisable visualization of different scene components, including detection, localization, and classification of various scene parts. The development of image segmentation approaches with DL is rapid, while they still face the following challenges:

- The performance of DL models is dependent on dataset quality and scale. Annotation cost is expensive, and annotation quality is difficult to guarantee without noise. The professional and accurate requirements for industrial data annotation are higher.
- The problem of multiple poses or perspectives of objects curtails the diverse practical application scenarios of DL image segmentation.
- The DL-based model structure is complicated, with numerous parameters, and requires a large amount of computing resources to train, including hardware devices such as CPU, GPU, TPU, *etc*., which reduces the possibility of real-time and flexible application.

In response to the above issues, scholars in this image segmentation field have explored various methods to address challenges such as multi-source domain data, multimodal scenes, and few samples. Pei *et al*. introduced multi-level adversarial learning and multi-model consistency loss into multi-modality cardiac segmentation and cross-modality liver segmentation[15]. This framework converts multi-source data features to the target domain and realizes multi-source unsupervised domain adaptation. In accordance with the spatial correlation between samples and intradomain sample expansion, a hyperspectral image (HSI) classification scheme with weak supervision is designed for few-sample training, which enhances classification stability and accuracy of HSIs[16]. Convolutional filters with dynamic parameters are utilized for various samples and learnable descriptive convolution adaptively learns fine texture features in lightweight dynamic convolution network (LDCNet)[17], decreasing computation complexity and improving real-time feasibility.

In accordance with DL, image segmentation that utilizes traditional image processing techniques to enhance segmentation performance and efficiency through network models and learning strategies has gradually emerged in public vision. DeepSnake[18] combines snake algorithm[19] in traditional ACMs, using a gradual adjustment strategy based on contours. DeepSnake gradually optimizes the initial contour to the boundary of targets, achieving excellent performance while still maintaining high real-time performance. However, DeepSnake propagates features of locally adjacent contour vertices to refine contour vertices instead of pixel-wise prediction, which is able to correct significant prediction errors. Yin *et al*. introduced the level set method (LSM) into neural networks, using the level set as input and constructing relevant loss functions to limit predicted level sets, such as using length terms to smooth boundaries[20].

With the advancement of DL technology, remarkable progress has been made in image segmentation research, which is widely exploited in fields such as defect detection, industrial quality inspection, bin-picking and map production. This survey sorts out the key nodes in the development process of this field and introduces a series of important algorithms and architectures focusing on the importance of data state for the development of DL. The practical applications of DL are widely explored in various fields such as defect detection and industrial picking. A forward-looking discussion is conducted on the future development trends of DL technology including the potential challenges and the opportunities contained in the future development process.

On the basis of previous work, this article summarizes the research results of traditional image segmentation, and systematically elaborates on DL-based image segmentation approaches. This survey mainly works on recent processes in three branches of industrial image segmentation according to the annotation level of datasets, namely supervised, weakly-supervised, and unsupervised image segmentation. Combined with traditional segmentation approaches, there are numerous papers in the DL-based image segmentation field. Influential works published in reputable conferences and journals are provided. Then, commonly used segmentation datasets, related evaluation metrics, and performance analysis are elucidated. Finally, there are a host of challenges and development directions in real-time and functional DL image segmentation, hoping to provide reference and assistance for more flexible and efficient DL image segmentation.

2. TRADITIONAL IMAGE SEGMENTATION

Traditional segmentation approaches commonly rely on topology, mathematical modeling, image processing, and so on to segment images, mainly containing threshold-, region-, edge-, clustering-, energy functional-, graph theory-, and wavelet-based approaches.

2.1 Threshold method

The basic notion of the threshold method is that various targets have diverse features such as color, grayscale, contour, *etc.* On account of small differences between features, specific thresholds are selected to classify pixels into appropriate categories according to comparison results, thereby distinguishing target objects from the background and achieving fast image segmentation. The most crucial step of the threshold method is to solve for the optimal threshold in the light of a certain criterion function. Given an original image $I(x, y)$, the feature value T is able to divide the original image into foreground and background with the segmented image obtained as

$$f(x, y) = \begin{cases} t_0, I(x, y) < T \\ t_1, I(x, y) \geq T. \end{cases} \tag{1}$$

Moreover, the Otsu algorithm, also known as the maximum inter-class variance algorithm, segments an image based on the probability density of the mean and variance, without assuming a specific probability density function, which significantly enhances the algorithm's computational speed. Many researches are conducted based on Otsu method. Bhandari *et al.* proposed an energy-based three-dimensional (3D) Otsu algorithm with contextual information, which combines with the pixel intensity values, same properties of histogram and spatial information[21]. This 3D Otsu algorithm yields desirable segmentation results in terms of daily color images. Ma *et al.* introduced a reverse learning strategy and an adaptive weighting strategy into the whale optimization algorithm using Otsu method as objective function, presenting efficient convergence accuracy on benchmark images[22].

Threshold segmentation approaches[23] are suitable for images, where the target and background occupy distinct grayscale ranges. However, uneven lighting and noise interference still need to be addressed. Uneven lighting causes the target peak in the histogram to mix with the background peak, thereby reducing the effectiveness of threshold methods. Noise has an impact on the entire process of image processing, causing the disappearance of separable grayscale peaks and a large number of misclassifications for segmentation.

2.2 Region-based methods

The region-based image segmentation strategy utilizes the spatial properties of images and similarity criteria of pixels in the same category, such as grayscale similarity, texture similarity, color similarity, *etc.*, to segment out the target region. It mainly includes three methods: watershed method, region division and merging method, and region growing method.

The watershed method[24] is a mathematical morphology-based segmentation approach with topology theory. Its basic idea is to model an image as a topographical terrain in geodesy. The grayscale value of each pixel in the image corresponds to the elevation of that point. Each local minimum and its affected area are called a watershed and its boundary forms a watershed. With the advantages of fast calculation speed, accurate positioning, and sensitivity to weak edge changes, this improved watershed algorithm presented by Tian *et al.* is combined with pseudo-color image transformation, principal component analysis (PCA) and band ratio methods to detect fruit surface defects unscathed[25]. However, for images with noise, low contrast, and bright edges such as bubble images, the interference causes the offset of segmentation contours. Therefore, Peng *et al.* proposed a watershed segmentation scheme with optimum labeling and edge constraints, employed to fuse foreground labels and correct segmentation results[26].

The region division and merging approach[27] is to divide an image into disjoint sub-regions, and then split or merge these subregions in accordance with relevant criteria. This method is proper for both grayscale and texture image segmentation with no need to specify seed points in advance. However, the splitting and merging algorithm is likely to damage boundaries of segmented regions. The region growing algorithm starts with a set of seed pixels representing various growth regions and then continuously merges eligible pixels in the neighborhood of seed pixels into the growing region represented by seed pixels. Meanwhile, it utilizes the newly added pixels as new seed pixels until no new pixels meet the growing criteria. Noise and uneven intensity still incur voids and over-segmentation, and this region growing method performs inadvisably due to low contrast regions in an image, although region growth algorithms are extensively applied in water quality analysis[28], visible point cloud segmentation[29], *etc*.

2.3 Edge-based methods

Edge refers to the collection of continuous pixels on the boundary lines of two different regions in an image, reflecting the discontinuity of local features such as grayscale and color. The edge-based segmentation method is rooted in the observation that the grayscale values of edges exhibit step or rooftop changes for image segmentation. There is a significant difference in the grayscale pixel values on both sides of the step edge, whereas the rooftop edge is located at the turning point where grayscale values rise or fall. Hence, the extremum of the first derivative and the zero crossing of the second derivative are adopted to decide edges, which is achieved by convolutional operators and more. Common edge segmentation operators include Sobel[30], Prewitt[31] operators, *etc*. These approaches convert images into grayscale forms and then calculate gradients or differences between pixel values to seek edges.

Traditional edge detection algorithms for segmentation usually require threshold processing on gradients or differences to distinguish between edges and noise. Therefore, threshold selection and noise have a serious influence on segmentation accuracy of these algorithms. Researchers have also conducted some research renewal for these issues and applications. Ghodrati *et al*. combined image edge detection solutions with different resolutions to measure material surface roughness, which provided a practical resolution for evaluating polymeric material roughness online[32]. An optimized edge detection method was designed by Xu *et al*. for detection of fissures in infrared images at diverse times based on the mathematical morphology, Canny and Laplacian of Gaussian operators[33]. Lu *et al*. introduced the local maximum inter-class variance calculation into Canny edge-detection framework to select threshold effectively for automatic hollowing checks in thermal images[34]. These improvements increase the practicality and flexibility of traditional edge detection algorithms.

2.4 Clustering-based methods

Clustering algorithm is an unsupervised machine learning method that discovers common cluster classes without labels. The essence of applying clustering to the field of image segmentation is to iteratively cluster pixels with similar features into the same region. The clustering technology is updated and employed in various image segmentation tasks. The spatial-temporal separation-based clustering segmentation (STSCS) algorithm[35] enhances pulsed infrared thermography images to detect sub-surface defects as small as 101 μm in diameter, using wavelet decomposition, and K-means clustering for effective segmentation. The global features with proposed self-representation coefficients modeled in fuzzy clustering are combined with fuzzy membership and neighbor information to resist intensity inhomogeneity and complex background in infrared images, which is proposed by Chen *et al*.[36]. Common clustering algorithms are summarized into three categories: distance-, density-, and connectivity-based algorithms.

The K-means algorithm uses distance as a measure of data similarity, meaning that the smaller the distance, the higher the likelihood of belonging to the same cluster. In K-means algorithm, the value of different clusters k needs to be selected according to different application scenarios. Basically, consistent with K-means algorithm, Achanta *et al*. suggested a simple linear iterative clustering (SLIC), which generated compact and neat superpixels with fewer hyperparameters, and had certain advantages in running speed, superpixel compactness, and wheel profile preservation[37]. Later, an image segmentation approach, combining SLIC with an automatically adjustable fuzzy c-means (FCM) algorithm, was presented[38] based on Gaussian radial basis function kernels. This algorithm is utilized for image segmentation and DL networks, with strong robustness and improved operational efficiency, although it is unable to obtain weak boundary information well, causing a lack of segmentation efficiency.

The mean shift algorithm originally proposed by Fukunaga[39] is a typical density-based non-parametric estimation. The mean shift algorithm achieves image segmentation by clustering pixels with the same mode points into the same region. When the number of cluster centers is unknown, it is capable of adapting clusters of any shape and has robustness to initialization and insensitivity to noise. Howbeit, the selection of its bandwidth parameter h significantly influences final segmentation. If h is too small, the convergence speed is slow. When h is set too large, the clustering effect is not inadvisable. Owing to overlapping organs and noise, Ranjbarzadeh[40] switched to utilize a mean-shift algorithm to enhance organ edges and used FCM and Kirsch filter for segmentation of liver and tumor, while its running time and complexity are high.

Density-based spatial clustering of applications with noise (DBSCAN), as a spatial clustering algorithm stemming from connectivity and density function, clusters grayscale data and finally performs coloring. Since multiple pixels in an image can be clustered together to form a larger object, each category determined in DBSCAN is mainly determined by the aggregation degree of sample distributions. Qiu *et al.* presented an improved DBSCAN with local window and multi-scale sliding windows extracting candidate objects quickly to realize adaptive threshold segmentation for infrared small target detection[41]. However, if the processed image has massive clusters or complicated background colors, the segmentation effect is undesirable and the computation time is high.

2.5 Energy functional based methods

The schemes based on energy functional mainly refer to ACMs. They employ continuous curves to express target edges and define energy functions so that their independent variables include edge curves. Accordingly, the segmentation process is converted into tackling the minimum value of energy functional. According to different forms of curve expression in these models, ACMs are divided into two categories: parametric ACMs[42] and geometric ACMs[43,44]. Parametric ACMs are based on Lagrange framework, which directly express curves in the parameterized form. The curve motion process of geometric ACMs is rooted in geometric measurement parameters of curves rather than expression parameters of contours.

The most representative model of parametric ACMs is Snake model proposed by Kasset *et al.*[19], which has been successfully employed in the early field of biological image segmentation, while its segmentation results are enormously affected by initial contour setting and laborious to cope with changes in curve topology. In addition, its energy functional only depends on choosing contour parameters, which can fall into local extremum and cannot segment concave object boundaries. In view of this situation, Xu *et al.* introduced a gradient vector flow (GVF) external force diffusing grayscale gradient vectors or binary edge maps with gradient amplitudes inversely proportional to distances from targets[45].

Compared to Snake model, initial curve position sensitivity of GVF model has been decreased, and it has the corresponding concave convergence ability[46]. However, noise robustness is reduced and its weak edge protection is inadvisable. For segmentation regarding images with weak edges and noise, an effective ACM using the adaptive radius of truncation equation and denoised variation-based term[47] is exploited to segment images with severe noise and uneven grayscale such as remote sensing images and brain magnetic resonance images in 2022. Later, a denoised ACM with optimized bias correction (BC) and a local pre-fitting equation[48] is designed to implement weak boundary image segmentation and obtain robustness to heavy noise. For automatic non-destructive inspection regarding printed circuit boards (PCB), layer identification proposed by Yun *et al.* automatically and unsupervisedly determines which slices of a 3D X-ray CT stack correspond to specific layers of a physical PCB, validated on a 4-layer PCB, and post level set-based segmentation effectively addresses grayscale inhomogeneity[49].

Geometric ACMs are able to handle changes in topological structure compared with parametric ACMs. This LSM[50] is introduced to transform curve evolution into level set solving problems which utilizes higher-order level set functions to describe contours. The geodesic active contour (GAC)[51] is the earliest edge-based geometric ACM, exploiting image gradient information to drive evolution curves closer to object borders and integrating geodesic construction concepts in Riemannian geometry, allowing contours to exactly detect fuzzy boundaries. However, this method has sensitivity to position selection of initial curves.

With further development, frameworks based on regional features have gradually emerged. On the basis of Mumford-Shah (MS) model[52], Chan-Vese (CV) model[53] replaces original smooth fitting equation with a simplified piecewise constant reckoning function. Using global information of images, it simplifies computational complexity, whereas failing in segmenting images with uneven grayscale. Given a pixel point x, the following CV energy function is written as

$$E^{CV}(C,c_1,c_2)=\lambda_1\int_{outside(C)}|I(x)-c_1|^2dx+\lambda_2\int_{inside(C)}|I(x)-c_2|^2dx+\nu\cdot|C|, \tag{2}$$

where c_1 and c_2 represent the average grayscale values outside and inside the curve, respectively. λ_1, λ_2, and ν are constants, and $|C|$ denotes the length of the curve. The first two terms in Equation (2) are data-driven terms that cause the curve to evolve towards target boundaries, while the length term acts as a smoothing curve.

Aimed at uneven intensity, numerous models[54,55] based on local image information have emerged, such as local binary fitting (LBF)[56] and local Kullback–Leibler divergence (LKLD)[57] models, *etc.* Taking minimizing LBF energy functional as an example, the curve C is expressed with the level set function ϕ, building its following remodeled energy equations as

$$E^{LBF}(\phi,f_1,f_2)=\sum_{i=1}^{2}w_i\int_{\Omega}\left[\int K_\sigma(x-y)|I(y)-f_i(x)|^2M_i(\phi)dy\right]dx+\zeta P(\phi)+\xi\mathfrak{L}(\phi), \tag{3}$$

where f_1 and f_2 respectively indicate mean values of nearby small fields within and outside the curve, K_σ represents a Gaussian kernel with standard deviation σ, and $w_i(i=1,2)$ are fixed weight parameters. Besides, $M_1(\phi(x))=H_\epsilon(\phi(x))$ denotes the approximated Heaviside function defined as

$$H_\epsilon(x)=\frac{1}{2}(1+\frac{2}{\pi}arctan(\frac{x}{\epsilon})). \tag{4}$$

$M_2(\phi)=1-H_\epsilon(\phi)$ explains that the pixel point is located outside the curve; ζ and ξ are constant values regarding the distance regularization expression $P(\phi)$, and the length restriction equation $\mathfrak{L}(\phi)$, respectively, and they are written as

$$P(\phi)=\int_{\Omega}\frac{1}{2}(|\nabla\phi|-1)^2dx, \tag{5}$$

$$\mathfrak{L}(\phi)=\int_{\Omega}\delta(\phi)|\nabla\phi|dx, \tag{6}$$

where $\delta_\epsilon(x)$ is Dirac function and its definition is

$$\delta_\epsilon(x)=H_\epsilon^{'}=\frac{\epsilon}{\pi(\epsilon^2+x^2)}. \tag{7}$$

Due to overlapping distribution of grayscale from bias field, this BC model[58] and its correspondingly developed ACMs[9,59] utilize the bias assumption regarding image multiplication model to derive the local intensity clustering properties of images and model energy functionals. Given that there are extensive convolution calculations in iterative evolution, significant operational consumption affects practicality.

2.6 Graph theory-based methods

The image segmentation algorithm based on graph theory transforms image segmentation into graph partitioning and seeks an optimal solution. Through different graph partitioning criteria and weight calculation expressions, extensive graph-based algorithms have been derived, such as GrabCut[60], and Random Walk. Felzenszwalb *et al.* introduced a minimum spanning tree strategy based on graph theory, clustering vertices in the graph via differences within and between regions to achieve segmentation[61]. This method possesses high running rate since its time complexity is $O(NlogN)$. The purpose of these algorithms is to maximize internal similarity of partitioned subgraphs and minimize the similarity between subgraphs.

Image segmentation is associated with the minimum cut problem of a graph according to graph theory. An image is mapped into a weighted undirected graph $G =< V, E >$, where each node $N \in V$ corresponds to each pixel in an image, and each edge belonging to E connects a pair of adjacent pixels. Values of edge weight represent non-negative similarity in gray intensity, color, or texture between adjacent pixels. Then, a cut s for an image is a cropping of an image, where each segmented region $C \in S$ corresponds to a subgraph in this image. In segmentation with graph theory, each pixel in an image is considered as a node in the graph, and edges between nodes denote their similarity. The optimal segmentation principle is to ensure that these partitioned subgraphs retain maximum similarity internally and minimum similarity between subgraphs.

The essence of graph-based segmentation is to remove specific edges and partition a graph into several subgraphs. In line with various similarity measurement approaches, diverse types of graphs are constructed, such as K-nearest neighbor graphs, fully connected graphs, *etc.* Subsequently, graph-based algorithms, such as minimum cut algorithm, spectral clustering, *etc.*, are employed to segment an image into multiple regions. The graph-based framework for segmentation is capable of handling complex scenes such as multiple objects overlapping together, without defining object numbers to be segmented in advance. Nevertheless, its deficiency is that it is not sensitive to noise and edges in an image, which induces over-segmentation or under-segmentation.

2.7 Wavelet-based methods

This wavelet theory was proposed by Morlet in 1980[62], effectively analyzing and processing signal mutations and non-stationary properties by introducing scaling and translation factors. Wavelet transform (WT) is a commonly applied mathematical tool. Unlike Fourier transform, which only obtains global spectral information, WT is equipped with high localization properties in both time and frequency domains and is able to unify time and frequency domains to analyze various signals. It possesses multi-scale characteristics and analyzes signals at different scales, and it has accordingly been applied to complete image segmentation. On the basis of multi-scale features from WT, Bi *et al.* designed a 3D discrete WT and subsequently adopted Markov random field (MRF) to realize image segmentation for polarimetric synthetic aperture radar (SAR)[63].

Since binary WT is equipped with the ability to detect local mutations in binary functions, it is applied as an image edge detection strategy. Image edges that appear at local grayscale discontinuities correspond to modulus maximum points of the dyadic WT. By detecting this maximum point of WT modulus, it is determined that image edges are located at various scales, and each scale's WT is capable of providing certain edge information. Therefore, multi-scale edge detection can be performed to obtain ideal image edges for image segmentation. With WT to attain detailed information from diagonal, vertical and horizontal directions, Gao *et al.* presented an image decomposition strategy to automatically decide segmentation thresholds[64]. This approach produces better conspicuous advantages to segment medical images than a maximum variance and histogram valley threshold segmentation.

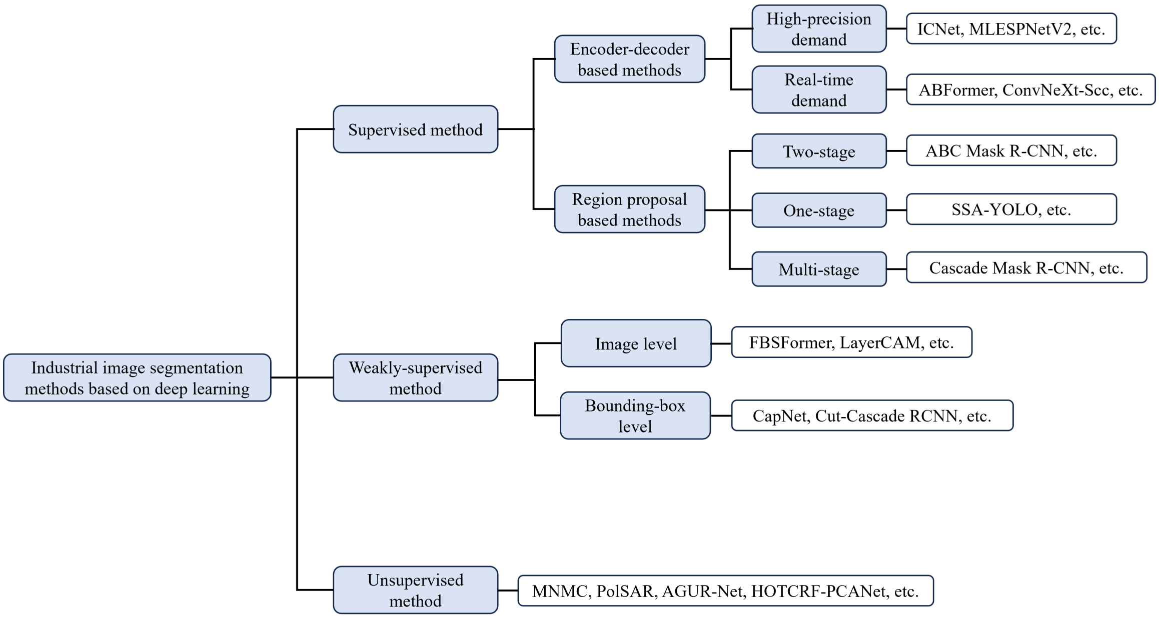
Industrial image segmentation methods based on deep learning
Supervised method
Encoder-decoder based methods
High-precision demand
ICNet, MLESPNetV2, etc.
Real-time demand
ABFormer, ConvNeXt-Scc, etc.
Region proposal based methods
Two-stage
ABC Mask R-CNN, etc.
One-stage
SSA-YOLO, etc.
Multi-stage
Cascade Mask R-CNN, etc.
Weakly-supervised method
Image level
FBSFormer, LayerCAM, etc.
Bounding-box level
CapNet, Cut-Cascade RCNN, etc.
Unsupervised method
MNMC, PolSAR, AGUR-Net, HOTCRF-PCANet, etc.

Figure 1. A framework of DL algorithms applied in industry. DL: Deep learning.

3. DL-BASED IMAGE SEGMENTATION APPROACHES

Traditional image segmentation methods have been effective in certain contexts; however, they often struggle with complex images and require extensive manual tuning. With the advent of DL, a paradigm shift has occurred in image segmentation. DL models have demonstrated remarkable capabilities in automatically learning hierarchical features from raw pixel data. This transition has enhanced the accuracy and efficiency of image segmentation tasks and handled diverse and intricate datasets, paving the way for advancements in various applications, including defect detection, autonomous driving, and path planning.

This section summarizes numerous DL-based segmentation approaches, including supervised models, weakly-supervised approaches and unsupervised schemes according to label completeness and accuracy of training data. Moreover, with the ability of traditional methods to extract low-level feature information such as texture, DL frameworks combined with traditional image segmentation are covered. According to data annotation level, a summarization of supervised, weakly-supervised and unsupervised models is provided in Table 1 and Figure 1.

3.1 Supervised deep-learning approaches

Modern supervised DL models learn from input images and their corresponding labeled masks shown in Figure 2 and models recognize new data and classify or predict it by learning these label masks. For image segmentation tasks, training data typically includes N input images and corresponding pixel-level labels, meaning that each pixel corresponds to a label. The standard evaluation manner for supervised DL with N inputs and N outputs is called loss minimization equation as

$$\widetilde{\Psi} \in \arg\min \frac{1}{N} \sum_{i} L(z(x_i), f(x_i)), \tag{8}$$

where L represents an error formulation measuring this estimated $f(x_i)$ against image segmentation outcomes related prior $z(x_i)$. In the light of Equation (8), the quality and quantity of label data have a serious impact on supervised DL models, which is time-consuming and labor-intensive prior knowledge for image segmentation.

Table 1. Summary of essential features for reviewed DL-based methods

Year	Method	Core architecture	Technical feature	Learning goal
2020	DeepSnake[18]	CenterNet GCN	Circular convolution	Supervided learning
	FCN-SFW[65]	FCN	Structured forests with WT	Supervided learning
	BlendMask[66]	ResNet FPN	Attention map learning	Supervided learning
	SOLO[67]	FCN	SOLO learning	Supervided learning
	PolarMask[68]	ResNet-50 FPN	Mask assembling Polar centerness Distance regression	Supervided learning
	CondInst[69]	-	Conditional convolutions	Supervided learning
	ODC[70]	ResNet-50	Unsupervised fine-tuning Joint clustering and feature Learning	Traditional method and DL
2021	HRNet[71]	HRNetv2-W48	Parallel multi-resolution convolution Repeated multi-resolution fusion	Supervided learning
	LayerCAM[72]	VGG-16	Weakly-supervised object localization	Weakly-supervised segmentation
	HOTCRF-PCANet[73]	Principal component analysis network	Unsupervised segmentation CRFs	Unsupervised segmentation
	Feng *et al.*[74]	VGG16	Weakly-supervised Graph cut	Traditional method + DL
2022	MDOAU-Net	U-Net	Multi-scaled feature-fusion Offset convolution	Supervided learning
	Wu *et al.*[75]	Siamese network	Max-pooling-based nonlocal attention Autofocus subwindow LSE Pooling	Weakly-supervised segmentation
	FreeSOLO[76]	SOLO	Free mask Self-supervise SOLO	Unsupervised segmentation
	Sledge *et al.*[77]	VGG13	weak saliency maps MB-CEDN Segmentation Postprocessing	Unsupervised segmentation
	Chen *et al.*[78]	FCN	Superpixel-guided unsupervised learning	Unsupervised segmentation
	E2EC[79]	DLA-34	Learnable contour initialization Multi-direction alignment	Traditional method and DL
2023	HSIs[16]	Continuous side window filter PSPNet HRNet	Cross-domain classification SE module -	Supervided learning Supervided learning Supervided learning
	ABFormer[80]	Transformer	Attention mechanism Split-attention boundary-aware fusion	Supervided learning
	ABC Mask R-CNN[81]	ResNet FPN	Attention block and context block Balanced block	Supervided learning
	Diaz *et al.*[82]	Resnet50 FPN	Depthwise separable convolution Transfer learning	Supervided learning
	CCRCNN[83]	ResNet101 FPN	Adaptive threshold segmentation	Weakly-supervised segmentation
	Wang *et al.*[84]	-	Self-attention	Unsupervised segmentation
	Wei *et al.*[85]	GAN	Structural similarity algorithm Multi-stage training strategy	Unsupervised segmentation
	Midwinter *et al.*[86]	CNN	Structural similarity	Unsupervised segmentation
	S^{3}Net[87]	Spectral–spatial Siamese network	Self-supervised learning Superpixel segmentation Transfer learning	Unsupervised segmentation
	Ma *et al.*[88]	CNN Channel attention	Image style transfer Multitask learning One-shot unsupervised domain adaptation	Unsupervised segmentation
	Wang *et al.*[89]	CNN	Unsupervised segmentation	Unsupervised segmentation
	Nouri *et al.*[90]	CNN	ACM	Traditional method and DL
	WDLS Qi *et al.*[91]	VGG16	Weakly-supervised learning ACM	Traditional method and DL
2024	Mobile-Deeplab[92]	Mobilenetv2 Deeplabv3+	Depthwise separable convolution Self-attention block	Supervided learning
	PCB-DeepLabV3[93]	Deeplabv3+	Attentional multi-scale two-space pyramid pooling network	Supervided learning
	ConvNeXt-SCC[94]	ConvNext	CBAM attention mechanism PCA	Supervided learning
	Deeplab-YOLO[95]	Deeplabv3+ YOLOv5	Depth separable convolution Convolutional block attention module	Supervided learning
	Shi *et al.*[96]	DeepLabv3+	ECA module and lightweight attention module Shallow feature fusion Dynamic threshold strategy	Weakly-supervised segmentation
	He *et al.*[97]	GAN U-GAT-IT	Adaptive threshold segmentation CAM quadratic optimization	Weakly-supervised segmentation
	FBSFormer[98]	ViT	Foreground–background separation module Attention-map refinement module	Weakly-supervised segmentation
	CapNet[99]	YOLOv8	Lightweight MirrorFill algorithm	Weakly-supervised segmentation
	OBBInst[100]	ResNet FPN	Oriented bounding box supervised learning	Weakly-supervised segmentation
	Jiang *et al.*[101]	-	Weakly-supervised segmentation mix Muti-instance loss Semi-supervised learning	Weakly-supervised segmentation
	Li *et al.*[102]	ResNet-18	Mutual information maximization	Unsupervised segmentation
	MNMC[103]	-	Mutual constraint	Unsupervised segmentation
	Goyal *et al.*[104]	Xception	Self-supervised learning Contrastive learning encoder	Unsupervised segmentation
	Wang *et al.*[105]	ResNet Convolutional long short-term memory	Semi-supervised classiffcation SimMatch Label information dissemination	Unsupervised segmentation
	S2VNet[106]	-	K-Means cross-attention Unified segmentation learning	Traditional method and DL

DL: Deep learning; FCN: fully convolutional network; WT: wavelet transform; FPN: feature pyramid network; SOLO: segmenting objects by locations; ODC: online deep clustering; CRFs: conditional random fields; HSIs: hyperspectral images; GAN: generative adversarial network; CNN: convolutional neural network; ACM: active contour model; PCA: principal component analysis.

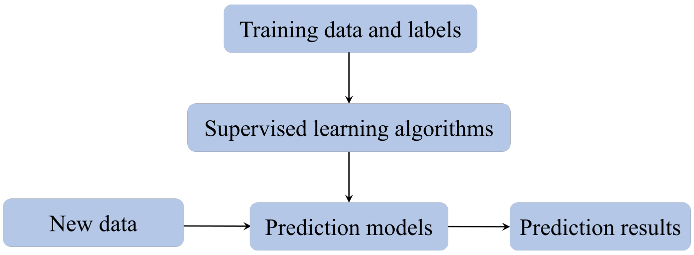

Figure 2. Supervised DL. DL: Deep learning.

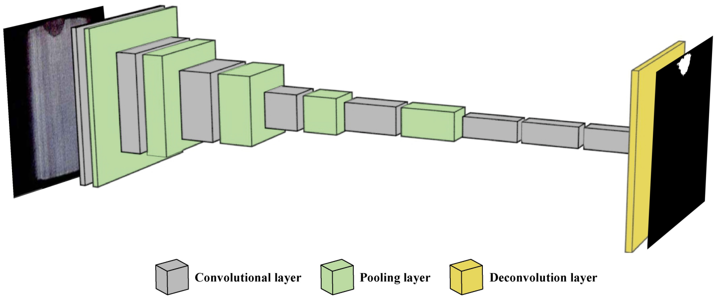

Figure 3. The network structure of FCN. FCN: Fully convolutional network.

3.1.1 Encoder-decoder-based models

Convolutional networks have been effectively applied in computer vision tasks, especially when initially used for object recognition and classification such as LeNet[107]. These images are processed by the input layer, hidden layer and output layer step by step, and the conception has also been applied in most subsequent computer vision tasks. Moreover, hidden layers comprise convolution layers, fully connected layers, and pooling layers, and CNNs tend to become an influential framework of image segmentation due to their translation invariance, parameter sharing, and sparse connections with improvement of computing resource level.

Different from treating semantic segmentation as a region classification, fully convolutional networks (FCNs) from Long *et al.* consider it as a pixel-level classification problem, and only convolutional layers are included in FCN[108]. As depicted in Figure 3, FCNs are encoder-decoder-based models, and their encoder is designed for downsampling to extract high-level semantic features, and then deconvolution is used for upsampling, enabling a segmentation map to be the same size as its input image. Combining FCN and structured forests with wavelet transform (SFW), FCN-SFW[65] is a fusion segmentation algorithm to effectively detect tiny cracks in steel beams, achieving superior segmentation performance against existing methods. The result resolution from FCN is too low due to 16 times upsampling rate, and its segmentation boundaries between objects tend to appear incorrectly. Without considering the correlation between pixels, spatial information is prone to be lost in FCN.

To address above two issues, DeepLab V1[109] introduces dilated convolution that enlarges dilation in the convolutional kernel to capture a larger range of contextual information without increasing computational complexity, improving CNN receptive field while ensuring parameter quantity. To effectively learn relationships between pixels, DeepLab V1 only improves accuracy compared to FCN, whereas the overall accuracy is still not high. Both SegNet[110] and U-Net[111] utilize a similar encoder-decoder structure as FCNs, and fine-grained upsampling to recover to original maps. In SegNet, unpooling in the decoder to upsample feature maps, which reduces the number of parameters and computations compared to deconvolution in FCN. As shown in Figure 4, U-Net improves a long skip connection, different from DenseNet[112], to connect feature maps at different levels and enable the model to simultaneously focus on low-level and high-level semantic features. Hence, U-Net is used to obtain preliminary insulator masks by training insulator images[4]. Nevertheless,

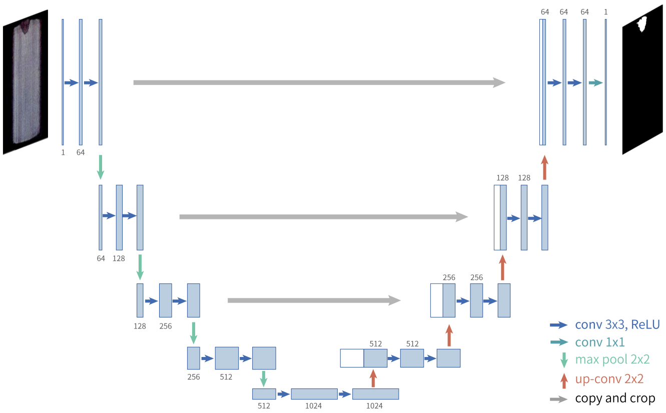

Figure 4. The network structure of U-Net.

the skip connection imposes unnecessary restrictive fusion schemes, forcing fusion only on feature maps of the same proportion in the encoder and decoder. Li *et al.* improved U-Net combining VGG16 and hybrid attention module to achieve high-precision segmentation of PCB welding defects and meet real-time detection requirements[4]. In accordance with U-Net, a multi-scale attention DL model called MDOAU-Net is designed by Wang *et al.* with extended convolution and offset convolution for monitoring aquaculture rafts with SAR image segmentation[113].

Moreover, DeepLab V2[114], DeepLab V3[115], DeepLab V3+[116], and PSPNet[117] adopt a spatial pyramid pooling (SPP) structure with granularly multi-receptive field learning the relationships between pixels, which extracts feature maps and convert them into fixed-size feature vectors for subsequent classification and regression tasks. SPP makes the model handle input images of different sizes and improves robustness to changes in object position and size by pooling feature maps at different scales. Inspired by PSPNet, SE-PSPNet[118] accurately segments defects in 3D braided composites, enhancing small object detection and addressing class imbalance for improved performance. Due to multi-scale input, RefineNet[119] can obtain global contextual information of small-sized targets. Although these models enhance their ability to learn relationships between pixels, their parameter quantity, and inference speed cannot meet real-time demands. Due to actually high real-time requirement, Mobile-Deeplab[92] is a lightweight pixel segmentation-based fabric defect detection strategy with 87.11 frames per second (FPS) on a 256 × 256 size image. The improved PCB-DeepLabV3 model[93] combined with Mobilenetv2 and multi-scale attention pooling achieves high-precision intelligent segmentation of voids inside chip solder joints, improving industrial detection efficiency.

Instead of upsampling to restore resolution in the decoder, HRNet[71] parallelly connects high-resolution to low-resolution convolutional streams and exchanges information across resolutions with multi-scale feature fusion to remain high-resolution expression. Wang *et al.* achieved a mean intersection to union ratio of 70.60% with HRNet by denoising Sentinel-1 SAR radar images and enhancing them with deep neural networks in industrial applications[120]. Besides, some DL-based models efficiently learn contextual connections between pixels from a spatial perspective. DANet[121] simultaneously introduces spatial attention and channel attention to capture contextual information between any two positions and channel dimensions. Given that Transformer is able to capture global information of input data effectively with multiple Transformer blocks in Figure 5, ABFormer[80] improves the accuracy of defect semantic segmentation and solves the problems of intra-class differences and inter-class ambiguity in industrial products by introducing boundary perception modules and attention mechanisms. Due to these drawbacks, it is essential to utilize image features from multiple scales containing features from shallow and deep layers in the backbone network for accuracy.

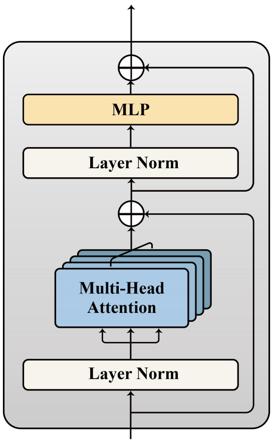

Figure 5. The standard network block of Transformer.

For real-time industrial demand, professional lightweight architecture is essential for practical application efficiency. ERFNet[122], ESPNetV2[123], and other lightweight models adopt an asymmetric encoder-decoder structure, removing low-level feature extraction results via exchanging accuracy for speed. With ESPNetV2, a high-precision surface crack detection architecture for steel rolling equipment is presented by Peng *et al.*, which utilizes coordinate attention-deep convolution generative adversarial networks (CA-DCGANs) for data augmentation, MLESPNetV2 for accurate detection, and MLESP-GAN for semi-supervised learning, enhancing real-time performance significantly[124]. Shallow and deep features respectively contain more spatial information and contextual information for images and they are indispensable in DL models. Therefore, some researchers have utilized a lightweight network to achieve multi-scale input while avoiding large-sized convolutional kernels similar to approaches in RefineNet[119]. These typical models include ICNet[125], *etc.*, while their segmentation accuracy is acceptable with reluctantly real-time speed.

Moreover, the feature extraction network should be as lightweight as possible for less occupied computing resources. In glaucoma diagnosis, Mallick *et al.* presented a ConvNext encoder and decoder with skip connections realizing light weight, and a novel attention gate module was designed for smoothing[126]. Besides, this method improves its loss function for accurate model training. ConvNeXt-SCC[94] visualizes tire crown defects and enhances defect detection and localization performance in tire production despite challenges such as limited datasets and small defect sizes by a modified convolutional block attention module and ConvNext.

3.1.2 Region proposal based methods

Region proposal based methods are generally identified as the combination of semantic segmentation and object detection tasks, for which they are known as a two-stage method. In accordance with the series sequence of object detection and semantic segmentation, two-stage approaches are further divided into top-down and bottom-up schemes. The top-down DL segmentation technique adopts the notion of detection before segmentation, predicting a bounding box for each object first and then generating an instance mask within each bounding box. By contrast, the bottom-up approach utilizes a semantic segmentation model for pixel embedding projection, and then uses post-processing to project pixels onto each instance, i.e., completes segmentation before detection.

In top-down DL image segmentation, Mask Region-based CNN (R-CNN) proposed by He *et al.* adopts ResNet as its backbone, adds a branch to predict segmentation masks in each region of interest (RoI), constructs a feature pyramid network (FPN) to process features at different scales, and utilizes non-maximum suppression (NMS) to remove masks with excessive intersections[127]. Since Mask R-CNN outperforms previous benchmarks in standard datasets such as COCO, an intelligent welding quality detection model[81] called ABC Mask R-CNN using enhanced Mask R-CNN with attention mechanisms, achieves 98.20% accuracy, improves efficiency over traditional expert-based methods, and supports intelligent maintenance in metro train body manufacturing. However, this mask R-CNN evaluation function only scores bounding boxes outputted by its target detector instead of mask prediction, disrupting segmentation results.

Mask Scoring R-CNN[128] adds MaskIoU head to learn intersection over union (IoU) scoring templates, and obtains more accurate masks. BlendMask[66] locks in the approximate range of instance pixels by object detection, and within this range, mask refinement is applied to the instance pixels to ensure the quality of the instance mask. Later, Geng *et al.* combined BlendMask with top-down and bottom-up structures to detect tunnel defects[129]. Although bottom-up methods such as deep metric learning (DML)[130] conform to human intuition, they generally depend on post-processing methods such as performance, causing over- or under-segmentation. Hence, DL-based bottom-up image segmentation has generally been less mainstream than top-down methods.

If the order of detection and segmentation is only distinguished, it is difficult to bring performance gains to two-stage image segmentation schemes. Therefore, the multi-stage approach has emerged, where segmentation and detection no longer have a sequential order. In the typical multi-stage Cascade Mask R-CNN[131], each cascade has an independent RPN to produce proposed boxes, and each detector filters bounding boxes given by its previous detector for boxes with high confidence so that this more accurate box proposal is used for detection and segmentation. By cascade, it enhances accuracy for segmentation with inconspicuously increased computational burden. In accordance with Cascade Mask R-CNN, Diaz *et al.* presented a fast wind turbine defect detection model with depthwise separable convolutions, enhanced by image augmentation and transfer learning, with a superior performance of 82.42% mAP[82].

Although two-stage and multi-stage DL approaches bring object instances of different scales into a standard scale and improve segmentation accuracy, they lack sufficient representational power and low efficiency. As demonstrated in the experimental literature[132], abundant training resources are needed in both two-stage and multi-stage DL schemes. Therefore, the one-stage segmentation model has emerged that consolidates detection and segmentation into a single network. Based on the presence or absence of a preset fixed-size object detection bounding box, the one-stage DL model is divided into two categories, i.e., anchor-based or anchor-free schemes. Anchor-based image segmentation mainly develops from you only look once (YOLO)[133] series. Deeplab-YOLO[95] is a lightweight hot spot defect detection model in infrared photovoltaic (PV) panels with MobileNetV3 to replace the YOLOv5 backbone, and MobileNetV2 was introduced into Deeplabv3+ for segmentation. SSA-YOLO[134] is an enhanced YOLO by a convolution squeeze-and-excitation module, Conv2d-BatchNorm-SiLU with Swin transformer, and adaptive spatial feature fusion module, thereby advancing quality control in steel production.

The anchor-free segmentation detector is mainly built on fully convolutional one-stage object detectors (FCOS)[135], including models that classify masks only by location and use contour regression to segment labels. Explicit shape encoding-based segmentation (ESE-Seg)[136] and polar coordinate-based instance segmentation (PolarMask)[68] predict contour regression to output instance masks. They typically use 20 to 40 coefficients to parameterize a mask contour, have fast inferring speed and are easy to optimize. However, they cannot accurately depict masks and objects with holes in their centers. Segmenting objects by locations (SOLO)[67] and conditional convolutions for instance segmentation (CondInst)[69] define instance categories by position and size. SOLO sets instance segmentation as a classification problem and removes any regression-dependent problem, making SOLO naturally independent of object detection. CondInst is prone to being affected by imprecise contour regression methods. By contrast, the performance of anchor-based segmentation is affected by detection, and an effective anchor is relatively time-consuming to obtain. Moreover, these anchor-free segmentation models serve as a comparative model to evaluate the performance of designed models for industrial

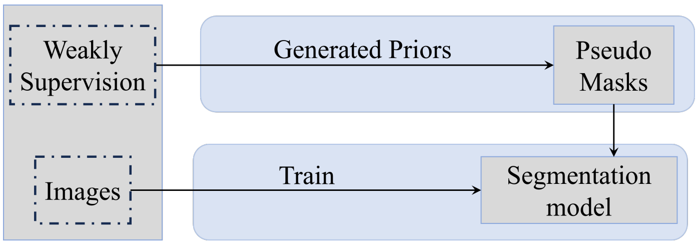

Figure 6. Weakly-supervised DL. DL: Deep learning.

applications due to their fast inferring speed.

3.2 Weakly-supervised image segmentation

The generalization ability of DL segmentation models relies on large-scale and high-quality pixel-level annotated data, while image segmentation annotation is a costly and time-consuming process for manual inefficiency. For new tasks or scenarios that require rapid application, data scarcity is even more severe. Therefore, the high cost of data annotation reduces the model's applicability in new tasks and scalability, thereby hindering the practical application of DL-based image segmentation models in industry, such as defect detection, *etc.* To alleviate the pressure of data annotation for actual industrial applications, a large amount of work has been done on semi-supervised and weakly-supervised image segmentation. In semi-supervised and weakly-supervised learning, partial image pixels are labeled while others are unlabeled. Displayed in Figure 6, algorithms utilize labeled pixel labels to learn prior information maps for generated Pseudo masks and then apply knowledge to unlabeled pixels for category and target location.

3.2.1 Image-level weakly-supervised image segmentation

Image-level supervision refers to using image-level labels such as category labels of objects contained in each image to train a model, which ignores the object position and size in an image and poses obstacles to learning object boundaries. Compared to pixel-level annotation, image-level annotation greatly reduces the manpower, time, and capital costs of data annotation. Mayr *et al.* designed L_p normalization with ResNet50 for defect detection in the solar cell by image-level supervision[137]. Shi *et al.* improved defect detection in diverse scenarios by a semi-supervised semantic segmentation with more reliable pseudo labels from dynamic threshold strategy[96]. However, due to lacking the position, shape, and contour information of targets in an image, training segmentation networks through image level annotation is a highly challenging problem, to which the key lies in how to convert image category information into corresponding target location information.

The segmentation process of most existing image-level semantic models mainly includes three stages. Object features are firstly preliminarily learned with coarse-grained annotation information. Since the class activation map (CAM)[138] cannot cover the entire corresponding target, different CAMs need integration to obtain a more complete target activation region. CAM-based models such as gradient-weighted class activation mapping (Grad-CAM)[139] enable classification-trained CNNs to learn object localization without the need for any bounding boxes. LayerCAM[72] combines activation maps from different layers to simultaneously obtain fine-grained object details and accurate spatial localization, achieving a segmentation effect of 27.26% mean IoU (mIoU) on the industrial product defect dataset Deutsche Arbeitsgemeinschaft fuer Muster-erkennung (DAGM)-2007. Wu *et al.* introduced a weakly-supervised defect segmentation algorithm with image-level labels and CAMs, enhancing performance with a Siamese network and improved modules[75]. He *et al.* proposed a weakly-supervised pavement crack segmentation method using CAM based on a GAN (U-GAT-IT) model to generate and refine CAMs, effectively bridging the gap between unsupervised and fully supervised approaches[97]. A foreground-background separation transformer (FBSFormer)[98], a weakly-supervised pixel-level defect detection method, enhances CAMs through a foreground-background separation module and attention-map refinement, achieving superior performance on industrial defect datasets.

Introducing the prompt paradigm from natural language processing into computer vision for segmentation, the segment anything model (SAM) is applied into semi-/weakly-supervised models to obtain exact segmentation outputs. SAM is Meta's open source generic model for image segmentation tasks. It utilizes a combination of CNNs and Transformer architectures to process images in a hierarchical and multi-scale manner, and prompt engineering idea is introduced to realize prompt segmentation based on point, box, mask and even free-form text. In addition, SAM employs a large dataset (SA-1B) containing at least 1 billion masks and 11 million images for model pre-training, enabling powerful generalization capabilities.

CS-WSCDNet[140] adopts CAM with localization capability and SAM to localize changed regions of image pairs and generate pixel-level pseudo labels to train a model for high-resolution scenes. The self-attention and model global correlation are integrated for optimized initial CAMs, a Siamese structure and generic SAM for a joint optimization strategy are used to produce high-precision target boundaries in remote sensing images at different scales[141]. In PV power generation, Yang *et al.* strategically combined both SAM and CAM to refine efficient pseudo-labels, and a boundary-aware loss function was employed to manage error features derived from generated pseudo-labels[142].

3.2.2 Bounding-box weakly-supervised image segmentation

Weakly-supervised schemes based on bounding boxes adopt bounding boxes surrounding an object to train their segmentation models, allowing the model to generate segmentation results based on produced bounding box range. Bounding boxes provide category labels for objects, and include the quantity and rough location information of objects. The bounding box-based segmentation methods utilize the position and size information of object bounding boxes instead of pixel-level annotation information inside the object, making it a more powerful supervision signal than image-level annotation. Thereupon, such methods tend to achieve better performance than image-level annotation methods. Because all regions outside bounding boxes are divided into background, and both foreground and background regions exist inside the minimum bounding box, it is essential for bounding box weakly supervision to accurately distinguish foreground objects and background regions decided by bounding boxes.

In 2015, Dai *et al.* proposed BoxSup, the first segmentation network based on bounding box annotation[143]. BoxSup uses border annotation and multi-scale combinatorial grouping (MCG)[144] to generate rough segmentation outputs of target classes, and uses them as supervised information to train parameters in VGG. Then, based on the current network, the predicted target segmentation results are used as supervision to continue training a segmentation network, iteratively optimizing segmentation results as supervisory information and updating network parameters. Due to category information annotated with borders and foreground segmentation results from MCG, BoxSup achieves good segmentation accuracy.

In Box2seg[145], pixel-level cross entropy loss is optimized by predicting attention maps of each object type, and the filling rate of each object kind in the minimum bounding box is calculated to constrain the generation of attention maps, thereby making the attention maps better focus on the foreground region and reducing incorrect gradient propagation. Zhang *et al.* trained Cut-Cascade R-CNN with simple bounding box annotations, obtaining high accuracy on welding defect detection in radiographic images[83]. CapNet[99] is a novel insulator self-blast defect detection network using normal samples and bounding box annotations, featuring a memory mechanism, polar alignment loss, and MirrorFill algorithm to enhance defect detection without additional manual effort. These schemes guide the segmentation process by calculating the fill rate of each foreground object in the minimum bounding box, improving the accuracy of pixel-level pseudo labels.

Some weakly-supervised schemes leverage various bounding boxes to guide network training for effective segmentation results and obtain object instances. In OBBInst[100], oriented bounding box annotations with compact object expression have a lower annotation burden than horizontal bounding box annotations, which

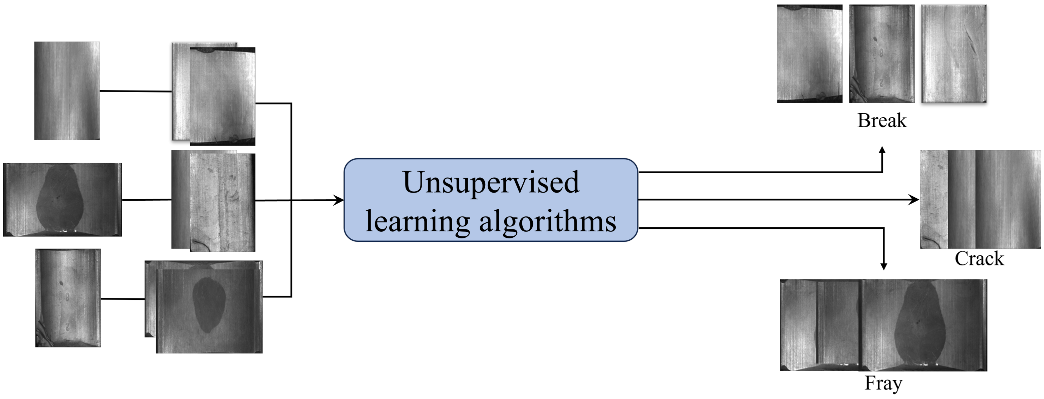

Figure 7. Unsupervised DL. DL: Deep learning.

are combined with corresponding loss function for oriented projection. Besides, OBBInst introduces Canny edge prior into DL to identify more precise edges for instance segmentation. Moreover, Jiang *et al.* presented single object box-supervised segmentation for insulator images using horizontal bounding boxes, generating segmentation maps to synthesize high-quality images for weakly-supervised segmentation mix and providing pseudo-labels for weakly-supervised instance segmentation[101].

Compared with weakly-supervised semantic segmentation, the supervised signal used in weakly-supervised instance segmentation is directly related to target instances. There is relatively less general prior information that can be learned and utilized compared to semantics, making it more sensitive to changes in the shape, size, quantity, and other aspects of the target instance. However, coarse-grained labels in weakly-supervised image segmentation models have inaccuracies, missing labels, or mutual interference between labels, resulting in lower accuracy and stability of the training model. Therefore, further research and improvement are needed to enhance the accuracy and robustness of weakly-supervised semantic segmentation with more accurate bounding box prior and labels.

3.3 Unsupervised DL-based image segmentation

The combination of unsupervised learning and DL enables DL models to discover hidden structures in data, improve their generalization performance and better deal with structurally complex image segmentation problems. By learning the hidden patterns and structures of data on unlabeled data for same semantic feature descriptions, unsupervised DL is not dependent on labels to address problems of difficult data annotation and extensive unlabeled images in Figure 7.

3.3.1 Pixel-level feature relation as prior information

Traditional unsupervised segmentation schemes are mainly realized through artificially designed image features and low-level image features, such as Gaussian mixture models, spectral clustering and K-means clustering, which are adopted to divide images into multiple regions with high self-similarities or discrepancies. In view of the abstract and high-level feature representation capacity of DL, unsupervised feature representation learning combined with DL develops rapidly. Unsupervised intensive feature representation learning, as indicated in SimCLR[146] and MoCo[147], substantially promotes unsupervised DL image segmentation strategies. In unsupervised image segmentation, it is essential for an effective network to learn a dense feature map for a given image without any annotation. An efficient network analyzes and models the similarity between pixels based on features and these learning feature maps represent pixels in the same semantic region (object/stuff) with similar feature descriptions, while pixels from different semantic regions are represented with various feature representations. Since the network learning process is not guided by labeled data, unsupervised segmentation schemes are constructed by extracting pixel-level features and grouping pixels into different semantic groups according to a variety of prior rules independent of training data.

Kanezaki[148] used a classical machine learning unsupervised segmentation algorithm for input image pre-classification to assign identical semantic labels to small areas where the semantic information is clearly identical, and CNN was adopted to process fine-grained pre-classification results from machine learning and merge these small semantic pixel regions for expected segmentation results step by step in iteration. Wang *et al.* designed an unsupervised CNN model with feature learning and self-attention modules to capture features and this proposed model clustered labels according to feature similarity to segment surface defects from components fabricated by selective laser melting with various materials[84]. A multi-stage unsupervised framework was proposed by Wei *et al.* based on DCGAN, and completed defect detection by comparing differences between test and reconstructed images from minimal normal samples[85]. In these models, superpixel segmentation and image reconstruction are combined into a unified network model and similarities or discrepancies between superpixel feature blocks are utilized to upgrade network parameters. Pixels extracted from randomly initialized CNNs are grouped on feature maps to minimize the distance between similar pixels while maximizing the distance between different features to maintain feature diversity.

There are proposed models to learn feature representation of repeated semantic objects in the image via maximizing mutual information between pixel features from two views of the same input image[149]. Information maximization and adversarial regularization segmentation (InMARS)[150] clusters generated superpixel regions with mutual information for maximizing the dependency between output and input, and then an adversarial regularization is combined with data augmentation strategy for training and optimization. Li *et al.* designed a PCA stretch alignment operation and loss function for fusing infrared and visible image features, and utilized mutual information maximization between visible feature maps and infrared feature maps to obtain these interimage relations among pixels, which is unified with loss functions to optimize computing and fusion representation[102]. A mixed noise-guided mutual constraint (MNMC)[103], an unsupervised anomaly model, utilizes mixed noise generation and mutual constraints to enhance feature learning, effectively addressing challenges in anomaly detection, and demonstrating competitive performance on MVTec dataset. Midwinter *et al.* proposed an unsupervised semantic segmentation method leveraging pose information to enhance visual inspection, validated through a spalling quantification task[86].

Furthermore, other models generate dense self-supervised information including cross-view consistency, cross-pixel similarity and cross-image relation according to heuristic priors to complete unsupervised image segmentation tasks. Despite different views, the same target object is considered to be consistent, which is so-called cross-view consistency. SimSiam[151] is a simple Siamese network model, where neither large batches nor negative sample pairs and momentum encoders are needed. It employs cross-view consistency between two randomly enhanced images by two decoder branches in the same input image. A self-supervised learning approach[104] for hotspot detection in thermal images, utilizing a SimSiam-based ensemble classifier, achieves high accuracy and precise hotspot isolation, addressing industrial safety concerns.

Cross-pixel similarity is that the same semantic region in an image contains pixels with highly similar cues such as color, texture or brightness. FreeSOLO[76] is an unsupervised framework to study instance segmentation combining segmentation and dense self-supervised learning for generated class-agnostic masks via cross-view consistency. Pixels from the same category target across images possess semantic relations. Due to the lack of supervised information, extracting semantically similar pixel features across images is challenging. Zhang *et al.* introduced an implicitly cross-image relation unsupervised segmentation scheme with pixel-wise contrastive learning[152]. This model assigns pseudo labels to all pixels in training images by clustering image features and utilizes these pseudo labels to choose appropriate positive or negative pairs with contrastive learning.

It is a preliminary stage for unsupervised DL to explore how to produce intensive representation information.

Different from pixel-level feature learning, dense representation requires regional priors to extract the relationship between pixels, which is used to determine whether these pixels are subordinate to the same semantic space. However, prior rules are generally unable to handle complex scenes. Objects of different categories are likely to possess similar appearances, while it is considerably possible for objects of the same category to have significant intra-class appearance variations, causing algorithms to be unable to effectively distinguish the differences in semantic features for each pixel in the feature space. Therefore, it is worth further exploration on how to introduce more accurate region priors and refine them for a more efficient segmentation learning process.

3.3.2 Intermediate features as additional clues

Instead of directly processing pixel-level features, a portion of unsupervised learning approaches adopt intermediate features as additional clues to guide semantic feature grouping, including saliency maps, superpixels, and shape priors. Saliency is a pattern of image partitioning, while saliency maps represent features on the uniqueness of each pixel. Saliency maps are extracted to simplify or change the representation of general images into a style that is appropriate for DL strategies to discover and analyze the salient and common foreground objects with the same semantic class. Superpixels are generated by directly clustering spatially connected groups of pixels instead of pixels, increasing the reliability of statistics when increasing batch size for training, and decreasing operation cost. Different from texture and color, the shape prior representation is more applicable to application scenarios with limited data scale and large shape changes but insignificant color features such as organ or tumor region extraction.

Utilizing features from a set of pixel groups, namely superpixels, transforms unsupervised image segmentation from pixel-level classification into unsupervised sub-region classification. MaskContrast[153] employs two additional unsupervised saliency models to generate pseudo labels for foreground objects in each image, and trains its final segmentation model a two-step bootstrap mechanism. Subsequently, collected foreground masks are clustered to form semantic concepts via contrastive learning. An unsupervised PolSAR image classification segmentation framework[105] utilizes high-confidence superpixel pseudo-labels and semi-supervised methods, achieving significant accuracy improvements over existing methods, as demonstrated on three real PolSAR datasets. A saliency-based multitarget detection and segmentation framework[77] utilizes a multibranch convolutional encoder-decoder network for circular-scan synthetic aperture sonar imagery. Although saliency maps are obtained to reduce the complexity of pixel-level feature expression, it is difficult to capture effective semantic information without the guidance of common semantic category information.

It is inadvisable for DL schemes to merely use low-level image information; the segmentation accuracy at object edge details is limited. Superpixels not only make for retaining the object edge information, but also represent image features instead of a large number of pixels, hence decreasing the noise interference for model training. Chen *et al.* utilized superpixel segmentation results to supervise the gradient descent direction for better FCN training, which realizes pixel-level segmentation for remote sensing images with various scales[78]. S^3Net[87] is a deep Siamese network guided by superpixels and inspired by transfer learning, and it generates pseudocolor maps to obtain homogeneous superpixel objects for desirable boundary adherence and multi-scale object-level difference features. Methods based on superpixels use a small region as a computing unit to independently extract features and determine saliency values, while it is possible for superpixels to generate salient object feature maps with fuzzy boundaries, incurring training instability.

Explicit shape priors are conducive to mitigating the dependence on the color features and learnable prototype. SegSort[154] is a pixel-level end-to-end unsupervised model with spherical K-means clustering to obtain pseudo segments and create segments aligned with semantic boundaries, which is crucial for subsequent clustering, prototype extraction and segmentation. For ultrasound vessel segmentation, an unsupervised and model-based domain adaptation segmentation with two-stage training strategy was suggested to achieve in-vivo data segmentation and the prior information from the elliptical shape of segmentation masks was applied to recognize uncontemplated outputs[155]. In rail surface defect segmentation, Ma *et al.* used prior information on content features and style features as input of a transfer module to obtain shape-consistent style features, and inputted constructed images using multitask learning to avoid distortion with SegFormer[156] as a feature extractor[88].

By implementing intermediate feature transformation, consistency information comparison and other schemes according to saliency maps, superpixels, and shape priors, the semantic discrimination of features is enhanced, promoting the separation of heterogeneous pixels. It is effective to adopt intermediate features as guidance to train segmentation for lower computing cost and complexity than pixel-level feature map classification. However, these methods assume that the objects in the image have sufficient saliency, which is not always true in some complex scenario-based situations and is likely to cause segmentation inaccuracy and instability.

3.3.3 Post-processing as a refinement

In the image segmentation process, an unsupervised model needs to model and analyze the similarity between pixels according to features such as color, texture and edge. Subsequently, this unsupervised model divides all pixels in an image into different groups, and different groups represent various object categories to complete segmentation tasks. However, since the network learning process is not guided by labeled data, segmentation results are prone to be undesirable. As unsupervised learning algorithms involve the processing and learning of unlabeled data, it is essential to be fully aware of the prior structure regarding the data and the model for expected convergence. Moreover, unsupervised learning algorithms usually require more domain knowledge to select and tune models, and further post-processing or manual intervention is hence needed to improve segmentation accuracy and efficiency.

Upon careful examination of the predicted mask, small prediction edges or regions with incorrectly predicted pixels are segmented out. Therefore, a step to finer adaptation and segmentation boundaries is essential to be added into unsupervised learning. An unsupervised recurrent scheme for DL and its corresponding loss function were suggested by Wang *et al.* to generate optimized pixel-level feature expression and semantic labels[89]. An attention-gate-based U-shaped reconstruction network (AGUR-Net)[157] achieves 59.38% precision on YDFID-1 dataset for unsupervised defect detection in color-patterned fabrics, and a dual-threshold segmentation post-processing optimizes detection results. In this recurrent framework, an over-segmentation method was implemented into model training, and the post-processing steps regarding clustering, threshold limitations, *etc.* were applied to coalesce over-segmented model outputs into the expected clustering number, further improving segmentation effectiveness.

Moreover, conditional random fields (CRFs) tend to be employed as a post-processing tool to improve the algorithm performance in that they are able to consider the feature information of the entire sequence and better capture the dependencies between sequences. HOTCRF-PCANet[73], combining high-order triplet CRFs and unsupervised principal component analysis network (PCANet), was proposed for effective segmentation of nonstationary SAR images, enhancing label consistency and feature representation. GAN continuously learns from the generative network and discriminative network through game theory in the framework to achieve better distribution convergence, where the one forges while the other performs detection, allowing models to train each other and achieve equilibrium. Although GAN achieves high performance in the game shown in Figure 8, the relationships between pixels are easily ignored, resulting in feature loss and segmentation results with discontinuity or significant deformation. Therefore, CRFs serve as a post-processing approach to capture entire sequence features and better-structured segmentation.

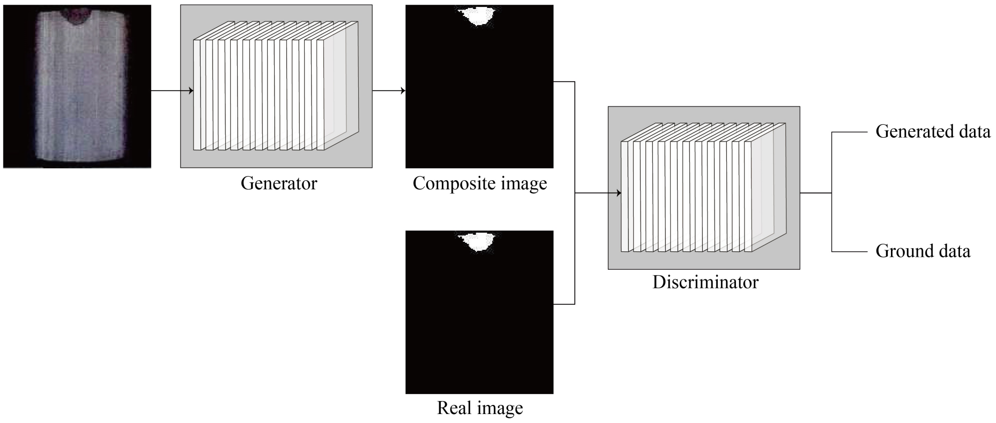

Figure 8. The model structure of GAN. GAN: Generative adversarial network.

3.4 DL combined with traditional image segmentation models

Traditional image segmentation generally utilized low-level feature information such as texture, color, *etc.*, while DL is equipped with a powerful expression in high-level semantic feature extraction. However, it is inevitable for DL to break shallow details in the image with feature extraction due to pooling, *etc.*, which is unfavorable to the feature recognition of small objects[158]. In addition, DL has large requirements for device performance and hardware consumption, limiting the possibility of landing applications. Therefore, considering the complexity and cost of implementation, many effective image segmentation methods have emerged by combining traditional image segmentation algorithms with DL.

3.4.1 Frameworks with improved theory

Some frameworks improve the construction mechanism of DL model in accordance with traditional image segmentation conception. Some schemes generate a joint distribution of two adjacent pixels based on similar semantic information of the same object, and obtain the segmentation result according to the similarity distribution on the centroid of a series of clusters. The segmentation route of ACMs is applied into DL image segmentation, which includes determining an initial contour, calculating the offset of the control point on the contour, shrinking the contour curve according to this offset to approximate the target boundary, and then segmenting out contours. Moreover, there are models to blend the concept that pixels with similar features are iteratively fitted to a region from GrabCut into DL-based segmentation frameworks, such as DeepCut[159].

Unsupervised clustering is applied to aggregate similar pixel-level features into a certain number of clusters, and each cluster represents a kind of semantic object. DeepClustering[160] exploits a simple process to generate the semantic distribution. In DeepClustering, K-means algorithm is used to cluster the eigenvalues of the model output for generated classification pseudo-labels. Thereafter, artificial labeling of general classification tasks is replaced by pseudo-labels generated by K-means to train neural network classifiers. As one of the earliest self-supervised representation learning methods, this paradigm has not achieved considerably acceptable performance, while its ideas have inspired numerous later works. Online deep clustering (ODC)[70] further improves the updating scheme of cluster centers in each iteration. The original updating strategy of resetting cluster centers in iteration is improved into an iterative updating method, and the two adjacent cluster centers are updated using the optimal matching algorithm, alleviating convergence instability to a certain extent.

The segmentation train of thought regarding Snake model[19] is fused into DL for accurate contour segmentation. Within the bounding box provided by the object detector, DeepSnake first generates an initial contour, and then deforms the contour by predicting vertex offsets to accurately match the object shape, implying the possibility of applying structured contour feature learning to instance segmentation. It uses cyclic convolution to efficiently learn contour features, reducing reliance on precise bounding boxes and enhancing the ability to handle object localization errors. Compared with DeepSnake, an end-to-end contour-based method (E2EC)[79] firstly learns the initialization contour by performing backbone operations on the image to generate a heatmap for the location of instance target centers and regresses the initial bias based on the center point features. Then, the generated initialization contour is deformed through a global deformation module to obtain a coarse contour, and finally two deformations are implemented to obtain a fine contour. These models pay special attention to boundary detection and segmentation of complex objects. They effectively handle objects with irregular shapes and fuzzy boundaries, and understand the shape and structure of objects in different contexts, thereby improving segmentation accuracy. Although the specific implementation and optimization strategies differ, they focus on real-time performance, especially in industrial application scenarios that require rapid response.

3.4.2 Frameworks with designed loss functions

In DL, all algorithms rely on minimizing or maximizing a function, called loss function. The loss function is used to measure the error between predicted values and their corresponding true values. When the model is trained, the loss function is minimized or maximized to make the model reach the convergence state and update model parameters to reduce the error of model prediction values. Therefore, the impact of different loss functions on the model is crucial. With help of additional prior information, such as shape, region and edge, traditional image segmentation algorithms achieve promising performance.

LevelSet R-CNN[161] is a deep variational instance segmentation framework, combining the architecture of CV model and Mask R-CNN, and its loss functions are designed corresponding to CV model and Mask R-CNN. Mask R-CNN is intended to categorize all objects in an image, and then a RoI is obtained by convolution operation to initialize a truncated signed distance function (TSDF), and instance hyperparameters. TSDF, hyperparameters and feature tensors are inputted into a CV optimization model to output a target TSDF, and a mask is produced by applying Heaviside function to this TSDF. Accordingly, a loss function on the basis of final and initial TSDFs is established to calculate the error between ground truth and prediction for joint end-to-end model training. Later, Nouri *et al.* utilized an original image and its encoded image generated by local word directional pattern texture descriptor as inputs to CNN for parameter maps and GVF maps, and introduced the loss expression concerning predicted ACM parameter and GVF maps for more accurate and robust fuzzy boundary segmentation[90].

The loss function according to traditional image segmentation algorithms is defined to optimize segmentation results or constraint prediction contour length for less non-target prediction results. Due to the striking similarity between the softmax layer in those networks and these feature functions in MS model, the LSM is naturally incorporated into CNN. The weakly-supervised deep level set model[91] combines a multibranch structure with a self-supervised objective, and presents an adaptive algorithm with level set function as a self-supervised loss term, effectively segmenting aeroengine defects in videoscope images and achieving real-time processing on Turbo19 dataset. These designed loss functions can not only serve as a regularization term for supervised neural network segmentation, but also as a semi-supervised or unsupervised segmentation since it is difficult to obtain the dataset required for training supervised segmentation networks.

3.4.3 Frameworks with pre-processing and post-processing

DL is suitable for high-dimensional data processing to solve problems that traditional image algorithms find difficult to solve, such as diverse defect segmentation and detection. However, DL devices have high-performance requirements, high hardware consumption, and high model training costs. In addition, DL segmentation has a dependency on training data, and generally speaking, the larger the amount of data, the better its performance. Therefore, some researchers, starting from optimizing model training efficiency and smoothing segmentation boundaries, utilize traditional image segmentation as part of model pre-processing or post-processing. These approaches rely on traditional image processing in data pre-processing to greatly reduce the learning pressure of the model and noise interference.

Due to sample distribution imbalance and interference such as noise and lighting changes in image datasets, there are DL-based approaches to utilize traditional strategies such as pre-processing for noise reduction and smoothing. Lin *et al.* proposed ScribbleSup that iteratively trains segmentation networks and uses graph cuts to correct pixel labels[162]. ScribbleSup first divides the image into superpixel forms, and then corresponds scribble annotations to the superpixel annotations, executing the graph-cuts algorithm on superpixels. Feng *et al.* adopted the graph cut scheme to obtain more effective supervision information according to activation seeds from a classification network, and employed fully-connected CRF for segmentation smoothing[74].

Moreover, DL-based unsupervised algorithms generally require more domain knowledge to optimize models, and post-processing is hence adopted to improve segmentation efficiency. Raja *et al.* exploited a non-local mean filter to denoise in the pretreatment stage, and then Bayesian fuzzy clustering, deep autoencoder and softmax regression were jointly suggested to complete brain tumor segmentation and classification[163]. S2VNet[106] is a general purpose framework intended to address medical image segmentation problems, using clustering techniques to initialize cluster centers from prior slices and perform fast inference speed and less memory consumption with 2D networks compared to mainstream 3D solutions.

3.4.4 Discussion on traditional and DL-based segmentation

Traditional methods are relatively easy to implement and suitable for rapid development and prototype design. In industrial scenarios with small data volumes and distinct features, traditional methods can achieve good results. Due to their clear features, the results are usually easy to interpret, making it easier for engineers to understand and debug. However, traditional methods are sensitive to image noise and lighting changes, which may result in poor segmentation performance and affect the accuracy of industrial detection. In complex scenes, selecting appropriate features and parameters requires experience and is not suitable for diverse industrial images. When dealing with complex or diverse images, the performance of traditional methods often declines, making it difficult to adapt to different industrial applications.

On large-scale datasets, DL-based methods typically achieve higher segmentation accuracy and have better robustness to noise and lighting changes. These models are able to adapt to different industrial environments and are suitable for complex industrial inspection tasks. Additionally, they automatically learn optimal features, reducing manual intervention without manually selecting features. However, training DL models requires a significant amount of computing resources and time, which is not suitable for resource-limited industrial environments. The demand for annotated data is high, and performance may decrease when there is insufficient data, especially in specific industrial scenarios. The poor interpretability of DL models affects engineers' trust in detection results.

Traditional methods are suitable for scenarios that are simple, feature-rich, and require high real-time and interpretability, while DL methods perform well in applications with complex backgrounds, high-precision requirements, and large-scale datasets. It is crucial to choose appropriate segmentation strategies based on specific industrial needs, which may require comprehensive consideration of characteristics from practical applications and resource limitations. Therefore, designing pre-processing, post-processing or loss functions using traditional methods is an effective component for weakly-supervised or unsupervised segmentation, when training supervised segmentation networks is challenging for lack of uniform and efficient datasets.

3.5 Data analysis of industrial image segmentation models

According to statistical results through a Web of Science search from 2020 to 2024, DL-based industrial image segmentation models are summarized in Figure 9. Technique terms are set based on image annotation level, i.e., supervised image segmentation, weakly-supervised image segmentation and unsupervised image segmentation. Industrial application terms include six widely concerned image segmentation search fields: insulator defect, PCB, fabric defect, magnetic tile, rail defect, and steel defect. The type of these statistically

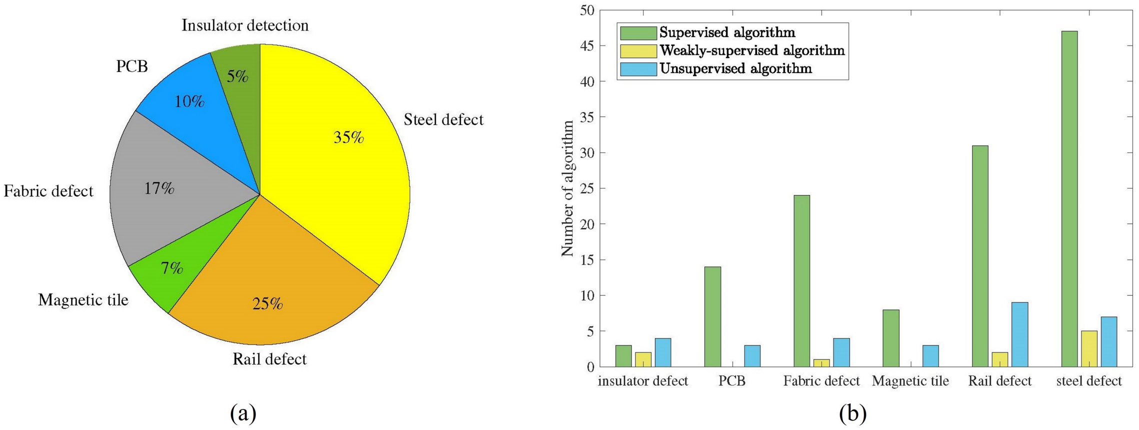

Figure 9. Statistical results for application of DL-based models in industry. DL: Deep learning.

analyzed papers is limited to research papers, not including conference papers, review articles, and doctoral and master dissertations.

The proportion of DL methods in different industrial applications is shown in Figure 9A. It is seen that the research papers on steel defect have the largest proportion 35 % and papers on rail defect have second largest proportion 25% in image segmentation. The application of defect detection accounts for the largest proportion 77% in Figure 9A, meaning that image segmentation in defect detection is considerably widespread. There are a small amount of DL methods in industrial applications of insulator detection and magnetic tile since the detection requirements for insulator and tile defects may be relatively small, compared to other types of defects such as PCB defects or fabric defects, resulting in insufficient investment in related research and development.

From Figure 9B, compared to supervised image segmentation, the application of weakly-supervised and unsupervised methods accounts for a smaller proportion in industry. Relatively mature supervised learning utilizes a large amount of labeled training data to achieve high-precision segmentation results, making it suitable for industrial applications that require high precision. The goal of supervised learning is clear, and the model accurately identifies and segments target objects by learning specific input-output relationships. Weakly-supervised learning reduces the cost and time of annotated data, while it requires designing effective strategies to utilize unlabeled data, which may face complex backgrounds and changing conditions in industrial environments, increasing implementation difficulty. Although unsupervised learning has advantages in data annotation cost, its ability to identify specific targets or details is weak, making it difficult to meet industrial requirements for high accuracy and reliability.

4. DL-BASED MODEL SEGMENTATION PERFORMANCE

In this section, the metrics and datasets for DL-based segmentation models are summarized, which are generally adopted to evaluate DL-based segmentation performance. The quality, scale, and diversity of datasets have a significant impact on the performance of image segmentation approaches. Therefore, the performance estimation on corresponding several segmentation benchmarks is also reported for DL-based segmentation models in this section.

4.1 Metrics for image segmentation models

Image segmentation model performance should be evaluated from multiple aspects, including accuracy, visual quality, inference speed, and memory storage requirements. Nevertheless, most researchers have focused on evaluation metrics to quantify model accuracy. The corresponding popular metrics are as follows:

The calculation of pixel accuracy (PA) is based on the number of pixels that accurately predict the category, and the accuracy ratio is obtained through ratio calculation. PA is obtained by dividing the number of correctly classified pixels by the total number of pixels. For $N + 1$ categories including the background and N foreground categories, PA is expressed as

$$\mathrm{PA} = \frac{\sum_{i=0}^{N} p_{ij}}{\sum_{i=0}^{N} \sum_{j=0}^{N} p_{ij}}, \tag{9}$$

where p_{ij} denotes the number of pixels belong to class i predicted as class j. However, the correspondence between pixels and classification is likely to be inaccurate, and the accuracy ratio is fed back through the mean PA (mPA), which is calculated for each category and then averaged over the totality number of categories as

$$\mathrm{mPA} = \frac{1}{N+1} \sum_{i=0}^{N} \frac{p_{ij}}{\sum_{j=0}^{N} p_{ij}}. \tag{10}$$

Dice coefficient (Dice) is mainly employed to calculate the similarity between different samples and is a similarity measurement function between two sets. The larger the coefficient, the better the segmentation effect. Dice is expressed as twice the overlap of predicted and ground-truth pixel maps divided by the total number of pixels, and its definition is

$$\mathrm{Dice} = \frac{2|A \cap B|}{|A| + |B|}, \tag{11}$$

where A is a predicted segmentation result and B represents ground truth.

IoU or Jaccard Index represents the number of overlapping pixels between the real label and the predicted image divided by the merged pixels between the real label and the predicted image. IoU ranges from 0 to 1, and is defined as

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|}. \tag{12}$$

Precision, Recall and F1 score can be specifically defined for each segmentation class or at the overall level. F1 score is computed by the harmonic mean of precision and recall rates. They are respectively defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{13}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{14}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{15}$$

4.2 Datasets

A suitable dataset ensures that models can learn correct and critical features, which directly affects their accuracy and efficiency. When selecting a dataset, it is necessary to make the selection based on specific application scenarios and algorithm requirements. The commonly used public datasets in image segmentation include NEU, Kolektor Surface Defect Dataset (KolektorSDD), *etc.* Some basic information is summarized in Table 2 respecting commonly applied datasets.

4.2.1 Dataset for daily scenarios

COCO dataset[164] created by Microsoft covers 80 categories of common objects and scenes, such as humans, animals, vehicles, furniture, *etc.* Each image is equipped with dense pixel-level segmentation annotations to identify boundaries of each object. COCO dataset consists of three subsets: training, validation, and test sets.

Table 2. Common industrial image segmentation datasets

Name	Objects	Links	Annotation
KolektorSDD	Electronic commutator	https://www.vicos.si/Downloads/KolektorSDD	Pixel level
NEU	Hot rolled steel belt	http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_dataset.html	Bounding box
RSDDs	Rail	http://icn.bjtu.edu.cn/Visint/resources/RSDDs.aspx	Pixel level
MVTec AD	Surface defect	https://www.mvtec.com/company/research/datasets/mvtec-ad	Pixel level
DAGM 2007	Optical texture	https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection	Label level
Aitex fabric image	Fabric defect	https://www.aitex.es/afid/	Pixel level
FUSAR-Ship	Ship	https://radars.ac.cn/web/data/getData?dataType=FUSAR	Bounding box
Magnetic tile dataset	Magnetic tile	https://github.com/abin24/Magnetic-tile-defect-datasets	Pixel level

The training set contains approximately 118,000 images, the validation set has approximately 5,000 images, and the test set includes approximately 20,000 images.

Pascal VOC dataset[165] comes from the Pascal VOC Challenge and is a popular benchmark dataset used to evaluate image segmentation algorithms. It contains images of 20 object categories such as humans, vehicles, animals, and 10 action categories. Moreover, pixel-level annotations are provided in Pascal VOC for each image. This dataset contains 11,540 images for object detection and classification recognition, and 6,929 images for image segmentation.

4.2.2 Dataset for surface defect detection

KolektorSDD includes defect images of electronic commutators. Specifically, there are minor damages or cracks on the plastic embedded surface of the electronic commutator. The image is collected under controlled conditions such as uniform lighting. Provided pixel-level annotation of defects for defect images. This dataset includes eight non-overlapping images collected from each surface of 50 defective electronic commutators, resulting in a total of 399 images, including 52 defect images and 347 defect-free images.

NEU surface defect database contains 1,800 grayscale images about surface defects in hot-rolled strip steel. Six defect categories including scale, patches, cracking, pitted surface, inclusion, and scratches are collected in NEU. Each category contains 300 images with a size of 200×200. The characteristics of the NEU dataset are significant differences in intra-class defects and certain similarities between inter-class defects, which pose difficulties and challenges for achieving accurate classification in defect detection.

MVTec Anomaly Detection (MVTec AD) dataset contains 5,354 high-resolution color images of 15 different objects and texture categories, where 3,629 images were used for training and validation, and 1,725 images were used for testing. Among 15 categories, five texture categories cover different types of regular textures such as carpets, grids and random textures such as leather, tiles, wood. The remaining ten object categories include rigid objects with a fixed appearance such as bottles, metal nuts and non-rigid and deformable objects such as cables, and some contain natural objects. All image resolutions are between 700×700 and $1,024 \times 1,024$ pixels, and pixel-level ground truth annotations are provided for all anomalies in MVTec AD.

RSDDs dataset is designed specifically for detecting surface defects in steel rails and includes two types of surface defect datasets: Type-I and Type-II. This Type-I dataset is collected from the fast lane and contains 67 images with $160 \times 1,000$ pixels. This Type-II dataset is collected from regular/heavy transport tracks and contains 128 images with $55 \times 1,250$ pixels. Each image in both datasets contains at least one defect, with a complex and noisy background. Defects in RSDDs include various common rail surface defects such as cracks, wear, grooves, and foreign object embeddings, which have been pixel level annotated by some professionals in rail surface detection.

4.2.3 Dataset for other industrial scenes

DAGM-2007 dataset is designed specifically for detecting miscellaneous defects on textured backgrounds in industrial optical inspection. It contains ten sub-datasets, with the first six being training datasets and the last four being testing datasets. The characteristic of DAGM 2007 is that it provides images in grayscale 8-bit PNG format. Each dataset contains 1,000 defect-free images and 150 defective images, where the "defect free" images display texture backgrounds without miscellaneous defects, while defective images have exactly one marked defect on the texture background.

The high-resolution ship dataset FUSAR Ship is a unified and standardized dataset for ship target recognition. This dataset slice contains 15 key ship targets with 98 subcategories and non-ship interference targets. A total of 16,144 slices have been accumulated, including 6,252 ships that match AIS information, 2,045 strong false alarms such as bright spots resembling ships, 1,461 bridges and coastlines, 1,010 coastal areas and islands, 1,967 complex sea waves and clutter, 1,785 ordinary sea surface images, and 1,624 land images. The annotation method for ship slices in the dataset is to expand 256 pixels outward with the center of the minimum outer circle of the target as the midpoint, and store them in a fixed size of 512×512 pixels.

Aitex fabric image dataset contains seven different fabrics with a total of 245 images, each with a resolution of $4,096 \times 256$. The fabrics in this dataset are mainly in plain colors. Among them, there are 140 flawless images, with 20 images for each type of fabric. There are a total of 105 defect images, including 12 common types of fabric defects in the textile industry. The large size of the image allows users to use different window sizes, thereby increasing the sample size. All defects are annotated at the pixel level, with white pixels representing the defect area and the remaining pixels being black.

Magnetic-tile defect dataset was automatically collected by the Chinese Academy of Sciences, and six common defects of the magnetic tile including crack, blowhole, break, fray, uneven and free images were marked by image acquisition and semantic segmentation. The dataset contains 1,344 images, each cropped to a RoI with different sizes. Suitable for studying the detection and segmentation techniques of small defects in complex backgrounds, especially those defects with similar colors, textures, and backgrounds.

4.3 Quantification performance for DL models

To facilitate the analysis of current research progress in image segmentation, the quantification performance of DL models is tabulated on current segmentation benchmarks in this section. Model segmentation accuracy such as PA and mIoU are generally utilized as standard metrics to evaluate model performance. Although most works report standard datasets and measure metrics to evaluate model performance, the downstream scenarios and implementation purposes of image segmentation frameworks are diverse, causing some differences between evaluation indicators reported by numerous research works. The following tables summarize the performance of several DL-based segmentation models on various datasets, where "-" indicates that the original work does not provide the corresponding value.

Combining with Table 3, MDOAU-Net[113] segments SAR images with fewer non-target or background pixels, and its overall accuracy is 0.906 in segmenting its test images, which is 4.1% over SegNet[110], 2.5% over U-Net[111], 1.8% over DeepLab V3+[116]. Shape-consistent one-shot unsupervised domain adaptation (SC-OSDA)[88] obtains 81.7% precision, outperforming SegFormer[156] over 0.7%, Mask R-CNN[127] over 10.8% with results verified using only RSDDs dataset while other models are trained with RSDDs dataset listed in Table 4. According to the literature[88], SC-OSDA generates finer segmentation with easily ignored appearance.

mIoU metric is the most extensive usage in industrial image segmentation. Table 5 summarizes the results of reviewed approaches on KolektorSDD datasets. As seen, FBSFormer[98] performs the best, with 0.787 in mIoU, which is 0.157 higher than U-Net[111]. In addition to mIoU, there are also precision, recall, and F1 score used

Table 3. Performance comparison of algorithms in SAR images

Time	Model	Feature	Accuracy	Parameters
2015	U-Net [111]	Skip connection	0.881	31.1M
2017	SegNet [110]	VGG16	0.865	29.4M
2018	DeepLab V3+ [116]	SPP + decoder-encoder	0.888	59.3M
2022	MDOAU-Net [113]	U-Net	0.906	4.1M

SAR: Synthetic aperture radar; SPP: spatial pyramid pooling.

Table 4. Performance comparison of algorithms in RSDDs dataset

Time	Models	Feature	Precision	Recall	IoU	F1
2015	U-Net [111]	Skip connection	0.869	0.903	0.856	0.886
2017	Mask R-CNN [127]	FCN	0.709	0.967	0.692	0.818
2017	PSPNet [117]	ResNet50	0.901	0.908	0.826	0.905
2018	DeepLab V3+ [116]	SPP	0.906	0.889	0.814	0.898
2021	SegFormer [156]	Transformer	0.810	-	-	0.833
2023	SC-OSDA [88]	-	0.817	0.834	0.748	0.825

IoU: Intersection over union; FCN: fully convolutional network; SPP: spatial pyramid pooling.

for steel defect detection. Table 6 gathers steel defect segmentation results on MVTec dataset, illustrating that ABFormer [80] obtains the mIoU value of 0.915, 0.962 Precision value, Recall value of 0.947 and F1 score of 0.954. Transformer-based ABFormer [80] greatly improves segmentation performance.

Giga floating-point operations per second (GFLOPs) and FPS are two common metrics for measuring computational complexity in DL. NEU dataset and DAGM-2007 dataset are chosen for testing image segmentation methods. As listed in Table 7, ABFormer [80] is the top one in terms of mIoU, precision, recall and F1 score, and the parameter number metric GFLOPs is the lowest, implying that ABFormer is significantly superior to all other networks in terms of computational complexity, inference accuracy, and number of parameters. In Table 8, FBSFormer [98] performs the best in DAGM-2007 dataset, with 79.1% in mIoU and scores 78.6 in terms of FPS.

Over the past few years, the performance of DL-based segmentation models has improved significantly, with relative improvements of 30%-40% on various evaluation metrics across different datasets. However, for diverse reasons, the performance evaluation of several models lacks reproducibility. They evaluate some performance metrics against non-standard databases, or report only a few metrics on the subsets of test sets from popular benchmarks. In addition, the source code for many model implementations is not provided or the model architecture is difficult to reproduce. However, with the increasing popularity of DL models, the development trend of DL-based image segmentation models is positive and inspiring.

5. CHALLENGES AND PROSPECTS

Image segmentation based on DL is an important research direction in the field of computer vision, which has enormous theoretical research value and a wide range of application prospects, while several challenges lie ahead. The rapid development of DL technology also brings new challenges and opportunities. Some promising and potential development directions and challenges are illustrated for advancing image segmentation algorithms.

Table 5. Performance comparison of algorithms in KolektorSDD dataset

Time	Model	Feature	mIoU
2015	U-Net [111]	Skip connection	0.630
2016	CAM [138]	ResNet50	0.703
2019	ESPNetV2 [123]	SPP	0.662
2021	LayerCAM [72]	ResNet50	0.705
2022	Wu *et al*. [75]	Weakly-supervised	0.69
2024	MLESPNetV2 [124]	GAN	0.754
2024	FBSFormer [98]	ViT	0.787

mIoU: Mean intersection over union; CAM: class activation map; SPP: spatial pyramid pooling; GAN: generative adversarial network.

Table 6. Performance comparison of algorithms in MVTec dataset

Time	Model	Feature	mIoU	Precision	Recall	F1
2015	U-Net [111]	Skip connection	0.691	0.903	0.710	0.779
2015	FCN [108]	-	0.895	0.948	0.938	0.943
2017	SegNet [110]	VGG16	0.671	0.798	0.725	0.756
2017	PSPNet [117]	ResNet50	0.901	0.951	0.942	0.946
2018	DeepLab V3+ [116]	SPP	0.611	0.831	0.627	0.683
2018	ICNet [125]	PSPNet	0.853	0.929	0.909	0.919
2019	DANet [121]	Dual attention network	0.900	0.951	0.941	0.946
2020	HRNet [71]	HRNetv2-W48	0.626	0.792	0.632	0.715
2022	OBFTNet [11]	VGG16	0.426	-	-	-
2022	OBFTNet [11]	ResNet50	0.522	-	-	-
2023	ABFormer [80]	Transformer	0.915	0.962	0.947	0.954
2024	Mobile-Deeplab [92]	Mobilenetv2	0.871	0.912	0.941	0.926

mIoU: Mean intersection over union; FCN: fully convolutional network; SPP: spatial pyramid pooling.

Table 7. Performance comparison of algorithms in NEU dataset

Time	Model	Feature	mIoU	Precision	Recall	F1	GFLOPs
2015	FCN [108]	-	0.798	0.922	0.851	0.885	30.17
2017	PSPNet [117]	ResNet50	0.824	0.903	0.898	0.900	27.27
2018	DeepLab V3+ [116]	SPP	0.804	0.902	0.875	0.887	41.15
2018	ICNet [125]	PSPNet	0.786	0.865	0.886	0.875	2.28
2019	DANet [121]	Dual attention network	0.828	0.907	0.899	0.903	30.37
2023	ABFormer [80]	Transformer	0.847	0.908	0.922	0.915	1.26

mIoU: Mean intersection over union; GFLOPs: giga floating-point operations per second; FCN: fully convolutional network; SPP: spatial pyramid pooling.

5.1 Challenges for DL models

Equipment dependency. DL segmentation models have a high dependence on hardware devices. Generally speaking, segmentation results will be more advisable when calculated on GPUs with fast speed and high accuracy. However, it is challenging for most researchers to obtain available hardware devices. Therefore, it is a thought-provoking question on how to break away from the dependence on hardware devices and enable segmentation models to achieve desirable experimental results on ordinary devices for the future development of real-time image segmentation. Besides, for low-storage hardware such as portable and mobile devices, it is difficult to carry out large-scale operations, and it is a major challenge to achieve high-precision computing results in a short period of time. Therefore, it is a major challenge for real-time DL-based image segmentation to reduce the computational load and storage requirements in applying segmentation networks to lightweight

Table 8. Performance comparison of algorithms in DAGM-2007 dataset

Time	Model	Feature	mIoU	FPS
2016	CAM [138]	ResNet50	0.676	66.32
2017	SegNet [110]	VGG16	0.220	17.92
2017	RefineNet [119]	-	0.329	31.05
2020	Grad-CAM [139]	CAM	0.350	60.97
2021	LayerCAM [72]	ResNet50	0.717	70.34
2024	FBSFormer [98]	ViT	0.791	76.8

FPS: Frames per second; CAM: class activation map.

devices such as mobile devices in the future.

Model performance. Compared to semantic features, it is more difficult for DL-based models to learn feature expressions of different individual instances within the same semantics. Given that the same semantic objects generally share similar semantic features, feature sharing reduces the dimensionality of the feature space. Despite sparse vectors representing specific objects, these features corresponding to instance objects are related to random differences in appearance, position and size of each individual, and these random differences incur various individual features. Therefore, if it is expected that models are able to learn general instance prior features, the difficulty is generally greater than that of semantic features. As the search dimension of instance features is larger than that of semantic features in the feature space, learning and optimizing instance feature expression is more challenging. How to enable DL models to fully utilize finite prior knowledge to learn instance features is crucial for efficient instance segmentation and panoptic segmentation in autonomous driving, defect detection, track monitoring, *etc*.

Data scarcity. Due to inconsistent constraints of real labels, sample distribution is considerably likely to be unbalanced, where the number of samples for some categories is too small to fully learn their feature representations. The sample imbalance results in undesirable model segmentation performance for these categories. Due to noise interference, there are issues such as abnormal pixel values, incomplete target regions, or inferior image quality, exerting a negative impact on segmentation results. For lack of constraints from real labels for weakly-supervised and unsupervised segmentation, pixels within the same target region is also likely to have different features, especially when segmenting image with unclear target boundaries. Moreover, due to the uncertainty of manually labeled labels and the diversity of application scenarios, it is challenging to form consistent evaluation criteria in image segmentation, which also brings certain predicaments to the training and optimization of models.

Application scenarios. Industrial scenes often have complex backgrounds and diverse textures, which may interfere with the performance of segmentation models. The size, shape, and arrangement of industrial objects vary, making segmentation tasks more complicated. The lighting conditions in industrial environments vary, affecting image quality and segmentation accuracy. In industrial applications, especially for specific categories of objects, a lack of sufficient labeled data is a difficult challenge to overcome for effective training. In many industrial applications, segmentation models require real-time processing capabilities to meet the needs of production lines. There are significant domain differences between natural images and industrial images, leading to poor performance of models trained in one domain in another domain.

5.2 Development trends for DL models

Adaptive learning model. Most existing DL-based segmentation models are trained and tested on existing datasets with fixed categories that do not exceed the known range of the model. However, real-world scenarios are open and the object categories are unfixed. Models will encounter constantly emerging new categories during their application process. In autonomous driving, unprecedented categories are likely to encounter, such as flood during driving. Therefore, whether algorithms continuously incorporate unseen scenarios and targets into the knowledge base and automatically adjust model parameters to maintain lifelong learning ability is a worthwhile research direction.

Cross-modality federated learning. At present, a small number of researchers are attempting to combine visual tasks with language processing. Cross modal federated learning includes two aspects. On the one hand, visual tasks are responsible for providing pixel-level clues, and then language models generate titles or other text. On the other hand, semantic models provide textual clues, and visual models complete segmentation tasks based on textual information, such as reference segmentation and reference video segmentation. The cross-modality joint learning will also become a major research trend in the future.

Real time segmentation centered on data. The DL scheme mainly focuses on the model as the center, with relatively fixed data, and then designs various models to improve prediction accuracy. Given possible various issues arising with data in practical applications, it is hence essential to show solicitude for high-quality annotated data in real-world scenarios to improve the quality and quantity of data. The existing real-time image segmentation is generally aimed to segment object classes with a large pixel proportion in the image, while there is little segmentation for classes with a small pixel proportion. Therefore, the image quality and quantity deserve attention for data for fine-grained small-sized object segmentation.

Domain adaptation technology can reduce the domain shift between natural images and industrial images, and enhance the generalization ability of the model. By simultaneously training multiple related tasks such as classification and segmentation, the robustness and accuracy of the model can be improved. Introducing attention mechanism can help the model focus on important features, reduce the interference of background noise, and improve segmentation accuracy. The use of lightweight network architectures such as MobileNet and EfficientNet, along with model compression techniques, enables real-time processing of segmentation models in industrial applications. These technological advancements make the application of DL image segmentation more effective and reliable in industrial scenarios.

Moreover, with the ability of traditional methods for shallow image feature extraction, it is advisable for DL segmentation strategies to adopt traditional algorithms for noise reduction and low-level feature enhancement on data. With these shallow visual features, which are similar to visual features acquired by human direct perception, it is conducive to improving the segmentation results regarding images with distinct imaging equipment and environments in various industry communities. Although these fusion approaches yield advisable segmentation results, it is essential for these algorithms to integrate weakly-supervised learning strategies and unsupervised DL strategies for better universality for industrial scenarios with few labels.

6. CONCLUSIONS

This work investigates the recent advancement in image segmentation approaches relying on DL in various complicated circumstances. The description on DL image segmentation tasks and recent developments in three branches of industrial image segmentation, categorized by the level of dataset annotation, i.e., supervised, weakly-supervised, and unsupervised image segmentation, are provided respecting designing network architecture, improving segmentation accuracy or speed, optimizing the trade-offs between speed and accuracy, and reducing the burden of manually annotated data. In conjunction with traditional segmentation methods, there is a wealth of literature in the field of DL-based image segmentation. The commonly applied benchmark metrics are subsequently illustrated, followed by benchmark datasets used to evaluate the accuracy and effectiveness of proposed models. Meanwhile, performance tables and figures are displayed to analyze quantitative performance of these approaches in industrial image segmentation. In conclusion, DL-based image segmentation faces both challenges and embraces bright prospects in flexible and practical applications.

DECLARATIONS

Authors' contributions

Writing - original draft preparation: Wang, G.; Li, Z.
Writing - reviewing and editing: Chen, Y.; Weng, G.
Conceptualization, methodology: Chen, Y.; Wang, G.; Li, Z.
Project administration: Chen, Y.; Wang, G.
Recourses: Chen, Y.; Weng, G.
Supervision: Weng, G.; Chen, Y.
Data curation: Chen, Y.; Li, Z.
Software: Li, Z.; Weng, G.
Investigation, visualization: Wang, G.

Availability of data and materials

Not applicable.

Financial support and sponsorship

This research paper was supported by the National Natural Science Foundation of China under Grant 62103293, the Natural Science Foundation of Jiangsu Province under Grant BK20210709, and the Entrepreneurship and Innovation Plan of Jiangsu Province under Grant JSSCBS20210641.

Conflicts of interest

Chen, Y. is a Junior Editorial Board Member of the journal *Intelligence & Robotics* and Guest Editor of the Special Issue of "Performance Evaluation and Optimization for Intelligent Systems". He is not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, or decision-making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Xu, Y.; Arai, S.; Liu, D.; Lin, F.; Kosuge, K. FPCC: fast point cloud clustering-based instance segmentation for industrial bin-picking. *Neurocomputing.* **2022**, *494*, 255-68. DOI
2. Wang, H.; Chen, Y.; Cai, Y.; et al. SFNet-N: an improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Trans. Intell. Transport. Syst.* **2022**, *23*, 21405-17. DOI
3. Wang, J.; Li, G.; Zhou, Y.; et al. A pixel-wise segmentation method for automatic X-ray image detection of chip packaging defects. *IEEE Trans. Compon. Packag. Manufact. Technol.* **2024**, *14*, 1520-7. DOI
4. Li, Z.; Liu, X. Soldering defect segmentation method for PCB on improved UNet. *Appl. Sci.* **2024**, *14*, 7370. DOI
5. Choi, S. H.; Park, K.; Roh, D. H.; et al. An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation. *Robotics Comput. Integr. Manuf.* **2022**, *73*, 102258. DOI
6. Guo, A.; Wang, Y.; Guo, L.; Zhang, R.; Yu, Y.; Gao, S. An adaptive position-guided gravitational search algorithm for function optimization and image threshold segmentation. *Eng. Appl. Artif. Intell.* **2023**, *121*, 106040. DOI
7. Qiao, Z.; Zhang, Q. Two-phase image segmentation by the Allen-Cahn Equation and a nonlocal edge detection operator. *NMTMA.* **2022**, *15*, 1147-72. DOI
8. Zerweck, L.; Wesarg, S.; Kohlhammer, J.; Köhm, M. Combining seeded region growing and k-nearest neighbours for the segmentation of routinely acquired spatio-temporal image data. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 2063-72. DOI

9. Weng, G.; Dong, B.; Lei, Y. A level set method based on additive bias correction for image segmentation. *Expert Syst. Appl.* **2021**, *185*, 115633. DOI
10. Byrne, N.; Clough, J. R.; Valverde, I.; Montana, G.; King, A. P. A persistent homology-based topological loss for CNN-based multiclass segmentation of CMR. *IEEE Trans. Med. Imaging.* **2023**, *42*, 3-14. DOI
11. Shan, D.; Zhang, Y.; Coleman, S.; Kerr, D.; Liu, S.; Hu, Z. Unseen-material few-shot defect segmentation with optimal bilateral feature transport network. *IEEE Trans. Ind. Inf.* **2023**, *19*, 8072-82. DOI
12. Sekar, A.; Perumal, V. SS-GAN based road surface crack region segmentation and forecasting. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108300. DOI
13. Li, J.; Chen, N.; Zhou, H.; et al. MCRformer: morphological constraint reticular transformer for 3D medical image segmentation. *Expert Syst. Appl.* **2023**, *232*, 120877. DOI
14. Chen, Y.; Li, Y.; Cheng, C.; Ying, H. Neural network based cognitive approaches from face perception with human performance benchmark. *Pattern Recogn. Lett.* **2024**, *184*, 155-61. DOI
15. Pei, C.; Wu, F.; Yang, M.; et al. Multi-source domain adaptation for medical image segmentation. *IEEE Trans. Med. Imaging* **2024**, *43*, 1640-51. DOI
16. Wang, X.; Liu, J.; Wang, W.; Chi, W.; Feng, R. Weakly supervised hyperspectral image classification with few samples based on intradomain sample expansion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5769-81. DOI
17. Yin, Y.; Luo, S.; Zhou, J.; Kang, L.; Chen, C. Y. LDCNet: lightweight dynamic convolution network for laparoscopic procedures image segmentation. *Neural Netw.* **2024**, *170*, 441-52. DOI
18. Peng, S.; Jiang, W.; Pi, H.; Li, X.; Bao, H.; Zhou, X. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, USA, Jun 13-19 2020; IEEE, 2020; pp. 8533–42. DOI
19. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: active contour models. *Int. J. Comput. Vision* **1988**, *1*, 321-31. DOI
20. Yin, P.; Xu, Y.; Zhu, J.; et al. Deep level set learning for optic disc and cup segmentation. *Neurocomputing* **2021**, *464*, 330-41. DOI
21. Bhandari, A. K.; Singh, A.; Kumar, I. V. Spatial context energy curve-based multilevel 3-D Otsu algorithm for image segmentation. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *51*, 2760-73. DOI
22. Ma, G.; Yue, X. An improved whale optimization algorithm based on multilevel threshold image segmentation using the Otsu method. *Eng. Appl. Artif. Intell.* **2022**, *113*, 104960. DOI
23. Yu, Y.; Bao, Y.; Wang, J.; et al. Crop row segmentation and detection in paddy fields based on treble-classification Otsu and double-dimensional clustering method. *Remote Sens.* **2021**, *13*, 901. DOI
24. Beucher, S. Use of watersheds in contour detection. 1979. https://cir.nii.ac.jp/crid/1572261550878454016. (accessed 2025-02-18)
25. Tian, X.; Liu, X.; He, X.; Zhang, C.; Li, J.; Huang, W. Detection of early bruises on apples using hyperspectral reflectance imaging coupled with optimal wavelengths selection and improved watershed segmentation algorithm. *J. Sci. Food Agric.* **2023**, *103*, 6689-705. DOI
26. Peng, C.; Liu, Y.; Gui, W.; Tang, Z.; Chen, Q. Bubble image segmentation based on a novel watershed algorithm with an optimized mark and edge constraint. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1-10. DOI
27. Gonzalez, R. C.; Wintz, P. Digital image processing, 2nd ed. Prentice Hall, 2002. https://dl.acm.org/doi/abs/10.5555/22881. (accessed 2025-02-18)
28. Kadapala, B. K. R.; K, A. H. Region-growing-based automatic localized adaptive thresholding algorithm for water extraction using sentinel-2 MSI imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*,1-8. DOI
29. Zhang, W.; Zhou, F.; Wang, L.; Sun, P. Region growing based on 2-D–3-D mutual projections for visible point cloud segmentation. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1-13. DOI
30. Lang, Y.; Zheng, D. An improved sobel edge detection operator. In *Proceedings of the 2016 6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016)*. Atlantis Press; 2016; pp. 590–3. DOI
31. Yang, L.; Wu, X.; Zhao, D.; Li, H.; Zhai, J. An improved prewitt algorithm for edge detection based on noised image. In *2011 4th International Congress on Image and Signal Processing*, Shanghai, China, Oct 15-17 2011; IEEE, 2011, pp. 1197–200. DOI
32. Ghodrati, S.; Mohseni, M.; Gorji Kandi, S. Application of image edge detection methods for precise estimation of the standard surface roughness parameters: polypropylene/ethylene-propylene-diene-monomer blend as a case study. *Measurement* **2019**, *138*, 80-90. DOI
33. Xu, D.; Zhao, Y.; Jiang, Y.; Zhang, C.; Sun, B.; He, X. Using improved edge detection method to detect mining-induced ground fissures identified by unmanned aerial vehicle remote sensing. *Remote Sens.* **2021**, *13*, 3652. DOI
34. Lu, Y.; Duanmu, L.; Zhai, Z.; Wang, Z. Application and improvement of Canny edge-detection algorithm for exterior wall hollowing detection using infrared thermal images. *Energy Build.* **2022**, *274*, 112421. DOI
35. Zhang, X.; Fang, T.; Saniie, J.; Bakhtiari, S.; Heifetz, A. Unsupervised learning-enabled pulsed infrared thermographic microscopy of subsurface defects in stainless steel. *Sci. Rep.* **2024**, *14*, 14865. DOI
36. Chen, Y.; Wang, Z.; Bai, X. Fuzzy sparse subspace clustering for infrared image segmentation. *IEEE Trans. Image. Process.* **2023**, *32*, 2132-46. DOI
37. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274-82. DOI
38. Kishorjit Singh, N.; Johny Singh, N.; Kanan Kumar, W. Image classification using SLIC superpixel and FAAGKFCM image segmentation. *IET. Image Process.* **2020**, *14*, 487-94. DOI
39. Fukunaga, K.; Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* **1975**, *21*, 32-40. DOI

40. Ranjbarzadeh, R.; Saadi, S. B. Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering. *Measurement* **2020**, *150*, 107086. DOI
41. Qiu, Z.; Ma, Y.; Fan, F.; Huang, J.; Wu, L.; Du, Y. Improved DBSCAN for infrared cluster small target detection. *IEEE Geosci. Remote Sensing Lett.* **2023**, *20*, 1-5. DOI
42. Tang, K.; Zhou, X. Evolution algorithm of parametric active contour model based on Gaussian smoothing filter. *Mach. Vision. Appl.* **2022**, *33*, 1336. DOI
43. Pramanik, S.; Banik, D.; Bhattacharjee, D.; Nasipuri, M.; Bhowmik, M. K.; Majumdar, G. Suspicious-region segmentation from breast thermogram using DLPE-based level set method. *IEEE Trans. Med. Imaging* **2019**, *38*, 572-84. DOI
44. Chen, Y.; Wu, L.; Wang, G.; He, H.; Weng, G.; Chen, H. An active contour model for image segmentation using morphology and nonlinear Poisson’s equation. *Optik* **2023**, *287*, 170997. DOI
45. Xu, C.; Prince, J. L. Gradient vector flow: a new external force for snakes. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, USA, Jun 17-19 1997; IEEE, 1997; pp. 66–71. DOI
46. Chen, Y.; Ge, P.; Wang, G.; Weng, G.; Chen, H. An overview of intelligent image segmentation using active contour models. *Intell. Robot.* **2023**, *3*, 23-55. DOI
47. Cai, Q.; Qian, Y.; Zhou, S.; et al. AVLSM: adaptive variational level set model for image segmentation in the presence of severe intensity inhomogeneity and high noise. *IEEE Trans. Image. Process.* **2022**, *31*, 43-57. DOI
48. Wang, G.; Li, Z.; Weng, G.; Chen, Y. An optimized denoised bias correction model with local pre-fitting function for weak boundary image segmentation. *Sign. Process.* **2024**, *220*, 109448. DOI
49. Yun, X.; Zhang, X.; Wang, Y.; et al. Automated layer identification and segmentation of x-ray computer tomography imaged PCBs. *X-Ray Spectrom.* **2024**, *53*, 315-25. DOI
50. Osher, S.; Sethian, J. A. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.* **1988**, *79*, 12-49. DOI
51. Caselles V, Kimmel R, Sapiro G. Geodesic active contours. *Int. J. Comput. Vis.* **1997**, *22*, 61-79. DOI
52. Mumford D, Shah J. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure. Appl. Math.* **1989**, *42*, 577-685. DOI
53. Chan TF, Vese LA. Active contours without edges. *IEEE Trans. Image. Process.* **2001**, *10*, 266-77. DOI
54. Ma P, Yuan H, Chen Y, Chen H, Weng G, Liu Y. A Laplace operator-based active contour model with improved image edge detection performance. *Digit. Signal. Process.* **2024**, *151*, 104550. DOI
55. Ge P, Chen Y, Wang G, Weng G. An active contour model based on Jeffreys divergence and clustering technology for image segmentation. *J. Vis. Commun. Image. R.* **2024**, *99*, 104069. DOI
56. Peng, Y.; Liu, F.; Liu, S. A normalized local binary fitting model for image segmentation. In *2012 Fourth International Conference on Intelligent Networking and Collaborative Systems*, Bucharest, Romania, Sep 19-21, 2012; IEEE, 2012; pp. 77–80, DOI
57. Yang C, Weng G, Chen Y. Active contour model based on local Kullback–Leibler divergence for fast image segmentation. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106472. DOI
58. Li C, Huang R, Ding Z, Gatenby JC, Metaxas DN, Gore JC. A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. *IEEE Trans. Image. Process.* **2011**, *20*, 2007-16. DOI
59. Wang G, Zhang F, Chen Y, Weng G, Chen H. An active contour model based on local pre-piecewise fitting bias corrections for fast and accurate segmentation. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1-13. DOI
60. Rother C, Kolmogorov V, Blake A. “GrabCut”: interactive foreground extraction using iterated graph cuts. *ACM. Trans. Graph.* **2004**, *23*, 309-14. DOI
61. Felzenszwalb PF, Huttenlocher DP. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167-81. DOI
62. Morlet, J. Sampling theory and wave propagation. In *Issues in acoustic Signal - image processing and recognition*. Springer, 1983, pp. 233–61. DOI
63. Bi H, Xu L, Cao X, Xue Y, Xu Z. Polarimetric SAR image semantic segmentation with 3D discrete wavelet transform and markov random field. *IEEE Trans. Image. Process.* **2020**, *29*, 6601-14. DOI
64. Gao J, Wang B, Wang Z, Wang Y, Kong F. A wavelet transform-based image segmentation method. *Optik* **2020**, *208*, 164123. DOI
65. Wang S, Pan Y, Chen M, Zhang Y, Wu X. FCN-SFW: steel structure crack segmentation using a fully convolutional network and structured forests. *IEEE Access.* **2020**, *8*, 214358-73. DOI
66. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. Blendmask: top-down meets bottom-up for instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, Jun 13-19, 2020; IEEE, 2020; pp. 8573–81. DOI
67. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: segmenting objects by locations. In *ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020; pp. 649-65. DOI
68. Xie, E.; Sun, P.; Song, X.; et al. Polarmask: single shot instance segmentation with polar representation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, Jun 13-19, 2020; IEEE, 2020; pp. 12190-9. DOI
69. Tian, Z.; Shen, C.; Chen, H. Conditional convolutions for instance segmentation. In *Computer Vision - ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020; pp. 282–98. DOI
70. Zhan, X.; Xie, J.; Liu, Z, Ong, Y. S.; Loy, C. C. Online deep clustering for unsupervised representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, Jun 13-19, 2020; IEEE, 2020; pp. 6688–97. DOI
71. Wang, J.; Sun, K.; Cheng, T.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern. Anal.*

Mach. Intell. **2021**, *43*, 3349-64. DOI

72. Jiang, P. T.; Zhang, C. B.; Hou, Q.; Cheng, M. M.; Wei, Y. LayerCAM: exploring hierarchical class activation maps for localization. *IEEE Trans. Image. Process.* **2021**, *30*, 5875-88. DOI
73. Zhang, P.; Boudaren, M. E. Y.; Jiang, Y.; et al. High-order triplet CRF-PCANet for unsupervised segmentation of nonstationary SAR image. *IEEE Trans. Geosci. Remote. Sensing.* **2021**, *59*, 8433-54. DOI
74. Feng, J.; Wang, X.; Liu, W. Deep graph cut network for weakly-supervised semantic segmentation. *Sci. China Inf. Sci.* **2021**, *64*, 3065. DOI
75. Wu, X.; Wang, T.; Li, Y.; Li, P.; Liu, Y. A CAM-based weakly supervised method for surface defect inspection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1-10. DOI
76. Wang, X.; Yu, Z.; De Mello, S.; et al. FreeSOLO: learning to segment objects without annotations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, Jun 18-24, 2022; IEEE, 2022; pp. 14176–86. DOI
77. Sledge, I. J.; Emigh, M. S.; King, J. L.; Woods, D. L.; Cobb, J. T.; Principe, J. C. Target detection and segmentation in circular-scan synthetic aperture sonar images using semisupervised convolutional encoder–decoders. *IEEE J. Oceanic. Eng.* **2022**, *47*, 1099-128. DOI
78. Chen, G.; He, C.; Wang, T.; Zhu, K.; Liao, P.; Zhang, X. A superpixel-guided unsupervised fast semantic segmentation method of remote sensing images. *IEEE Geosci. Remote. Sensing. Lett.* **2022**, *19*, 1-5. DOI
79. Zhang, T.; Wei, S.; Ji, S. E2EC: an end-to-end contour-based method for high-quality high-speed instance segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, Jun 18-24, 2022; IEEE, 2022. pp. 4443–52. DOI
80. Yeung, C.; Lam, K. Attentive boundary-aware fusion for defect semantic segmentation using transformer. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1-13. DOI
81. He, D.; Ma, R.; Jin, Z.; et al. Welding quality detection of metro train body based on ABC mask R-CNN. *Measurement* **2023**, *216*, 112969. DOI
82. Diaz, P. M.; Tittus, P. Fast detection of wind turbine blade damage using Cascade Mask R-DSCNN-aided drone inspection analysis. *SIViP* **2023**, *17*, 2333-41. DOI
83. Zhang, B.; Wang, X.; Cui, J.; et al. Welding defects classification by weakly supervised semantic segmentation. *NDT. E. Int.* **2023**, *138*, 102899. DOI
84. Wang, R.; Cheung, C. F.; Wang, C. Unsupervised defect segmentation in selective laser melting. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1-10. DOI
85. Wei, C.; Liang, J.; Liu, H.; Hou, Z.; Huan, Z. Multi-stage unsupervised fabric defect detection based on DCGAN. *Vis. Comput.* **2023**, *39*, 6655-71. DOI
86. Midwinter, M.; Al-Sabbag, Z. A.; Yeum, C. M. Unsupervised defect segmentation with pose priors. *Computer. Aided. Civil. Eng.* **2023**, *38*, 2455-71. DOI
87. Zhan, T.; Gong, M.; Jiang, X.; Zhang, E. S^{3}Net: superpixel-guided self-supervised learning network for multitemporal image change detection. *IEEE Geosci. Remote. Sensing. Lett.* **2023**, *20*, 1-5. DOI
88. Ma, S.; Song, K.; Niu, M.; Tian, H.; Wang, Y.; Yan, Y. Shape-consistent one-shot unsupervised domain adaptation for rail surface defect segmentation. *IEEE Trans. Ind. Inf.* **2023**, *19*, 9667-79. DOI
89. Wang, H.; Dalton, L.; Guo, R.; Mcclure, J.; Crandall, D.; Chen, C. Application of unsupervised deep learning to image segmentation and in-situ contact angle measurements in a CO2-water-rock system. *Adv. Water. Resour.* **2023**, *173*, 104385. DOI
90. Nouri, M.; Baleghi, Y. An active contour model reinforced by convolutional neural network and texture description. *Neurocomputing* **2023**, *528*, 125-35. DOI
91. Qi, H.; Cheng, L.; Kong, X.; Zhang, J.; Gu, J. WDLS: deep level set learning for weakly supervised aeroengine defect segmentation. *IEEE Trans. Ind. Inform.* **2024**, *20*, 303-13. DOI
92. Bai, Z.; Jing, J. Mobile-Deeplab: a lightweight pixel segmentation-based method for fabric defect detection. *J. Intell. Manuf.* **2024**, *35*, 3315-30. DOI
93. Kong, D.; Hu, X.; Gong, Z.; Zhang, D. Segmentation of void defects in X-ray images of chip solder joints based on PCB-DeepLabV3 algorithm. *Sci. Rep.* **2024**, *14*, 11925. DOI
94. Zhou, Y.; Zhang, J.; Ni, P.; Cao, Q.; Hu, J. A customised ConvNeXt-SCC network: integrating improved principal component analysis with ConvNeXt to enhance tire crown defect detection. *Nondestruct. Test. Eva.* **2024**, 1-29. DOI
95. Lei, Y.; Wang, X.; An, A.; Guan, H. Deeplab-YOLO: a method for detecting hot-spot defects in infrared image PV panels by combining segmentation and detection. *J. Real. Time. Image. Proc.* **2024**, *21*, 1415. DOI
96. Shi, C.; Wang, K.; Zhang, G.; Li, Z.; Zhu, C. Efficient and accurate semi-supervised semantic segmentation for industrial surface defects. *Sci. Rep.* **2024**, *14*, 21874. DOI
97. He, T.; Li, H.; Qian, Z.; Niu, C.; Huang, R. Research on weakly supervised pavement crack segmentation based on defect location by generative adversarial network and target re-optimization. *Construct. Build. Mater.* **2024**, *411*, 134668. DOI
98. Jiang, X.; Feng, J.; Yan, F.; et al. Foreground–background separation transformer for weakly supervised surface defect detection. *J. Intell. Manuf.* **2024**. DOI
99. Jiang, D.; Cao, Y.; Yang, Q. CapNet: learning insulator self-blast from bounding box. *IEEE. Trans. Instrum. Meas.* **2024**, *73*, 1-10. DOI
100. Cao, X.; Zou, H.; Li, J.; Ying, X.; He, S. OBBInst: remote sensing instance segmentation with oriented bounding box supervision. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103717. DOI
101. Jiang, D.; Cao, Y.; Yang, Q. Weakly-supervised learning based automatic augmentation of aerial insulator images. *Expert. Syst. Appl.*

2024, *242*, 122739. DOI

102. Li, Y.; Fang, A.; Guo, Y.; Wang, X. Image fusion via mutual information maximization for semantic segmentation in autonomous vehicles. *IEEE Trans. Ind. Inf.* **2024**, *20*, 5838-48. DOI
103. Zhao, Q.; Wang, Y.; Lin, Y.; et al. Mixed noise-guided mutual constraint framework for unsupervised anomaly detection in smart industries. *Comput. Commun.* **2024**, *216*, 45-53. DOI
104. Goyal, S.; Rajapakse, J. C. Self-supervised learning for hotspot detection and isolation from thermal images. *Expert. Syst. Appl.* **2024**, *237*, 121566. DOI
105. Wang, L.; Peng, L.; Gui, R.; Hong, H.; Zhu, S. Unsupervised PolSAR image classification based on superpixel pseudo-labels and a similarity-matching network. *Remote Sens.* **2024**, *16*, 4119. DOI
106. Ding, Y.; Li, L.; Wang, W.; Yang, Y. Clustering propagation for universal medical image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA; June 16-22, 2024; IEEE, 2024; pp. 3357–69. DOI
107. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE.* **1998**, *86*, 2278-324. DOI
108. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015; pp. 3431–40. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html. (accessed 2025-02-19)
109. Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *International Conference on Learning Representations*, San Diego, United States, May 2015. https://inria.hal.science/hal-01263610. (accessed 2025-02-19)
110. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2017**, *39*, 2481-95. DOI
111. Ronneberger, O.; Fischer, P.; Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention - MICCAI 2015: 18th international conference*, Munich, Germany, Oct 5-9, 2015; Springer, Cham, 2015; pp. 234–41. DOI
112. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–8. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf. (accessed 2025-02-19)
113. Wang J, Fan J, Wang J. MDOAU-Net: a lightweight and robust deep learning model for SAR image segmentation in aquaculture raft monitoring. *IEEE Geosci. Remote. Sensing. Lett.* **2022**, *19*, 1-5. DOI
114. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2018**, *40*, 834-48. DOI
115. Chen, L. C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* 2017; arXiv:1706.05587. Available from: https://arxiv.org/abs/1706.05587. [Last accessed on 19 Feb 2025]
116. Chen, L. C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–18. https://openaccess.thecvf.com/content_ECCV_2018/papers/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.pdf. (accessed 2025-02-19)
117. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–90. https://openaccess.thecvf.com/content_cvpr_2017/papers/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.pdf. (accessed 2025-02-19)
118. Guo Y, Xiao Z, Geng L. Defect detection of 3D braided composites based on semantic segmentation. *J. Text. Inst.* **2023**, *114*, 574-83. DOI
119. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–34. https://openaccess.thecvf.com/content_cvpr_2017/papers/Lin_RefineNet_Multi-Path_Refinement_CVPR_2017_paper.pdf. (accessed 2025-02-19)
120. Erten H, Bostanci E, Acici K, Guzel MS, Asuroglu T, Aydin A. Semantic segmentation with high-resolution sentinel-1 SAR data. *Appl. Sci.* **2023**, *13*, 6025. DOI
121. Fu, J.; Liu, J.; Tian, H.; et al. Dual attention network for scene segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, Jun 15-20, 2019; IEEE, 2019; pp. 3146–54. DOI
122. Romera E, Alvarez JM, Bergasa LM, Arroyo R. ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation. *IEEE Trans. Intell. Trans. Syst.* **2018**, *19*, 263-72. DOI
123. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. Espnetv2: a light-weight, power efficient, and general purpose convolutional neural network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beache, USA; Jun 15-20, 2019; IEEE, 2019. pp. 9190–200. DOI
124. Peng Y, Wang C, Hao Y, et al. High-precision surface crack detection for rolling steel production equipment in ICPS. *IEEE Int. Things. J.* **2024**, *11*, 4586-99. DOI
125. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 405–20. https://openaccess.thecvf.com/content_ECCV_2018/papers/Hengshuang_Zhao_ICNet_for_Real-Time_ECCV_2018_paper.pdf. (accessed 2025-02-19)
126. Mallick S, Paul J, Sil J. Response fusion attention U-ConvNext for accurate segmentation of optic disc and optic cup. *Neurocomputing* **2023**, *559*, 126798. DOI

127. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–69. https://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf. (accessed 2025-02-19)
128. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang X. Mask scoring R-CNN. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA; Jun 15-20, 2019; IEEE, 2019; pp. 6409–18. DOI
129. Geng P, Jia M, Ren X. Tunnel lining water leakage image Segmentation based on improved BlendMask. *Struct. Health Monit.* **2023**, *22*, 865-78. DOI
130. Fathi, A.; Wojna, Z.; Rathod, V.; et al. Semantic instance segmentation via deep metric learning. *arXiv* 2017, arXiv:1703.10277. Available from: https://arxiv.org/abs/1703.10277. [Last accessed on 19 Feb 2025]
131. Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2021**, *43*, 1483-98. DOI
132. Gu W, Bai S, Kong L. A review on 2D instance segmentation based on deep neural networks. *Image. Vis. Comput.* **2022**, *120*, 104401. DOI
133. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–88. https://www.academis.eu/machine_learning/_downloads/51a67e9194f116abefff5192f683e3d8/yolo.pdf. (accessed 2025-02-19)
134. Huang X, Zhu J, Huo Y. SSA-YOLO: an improved YOLO for hot-rolled strip steel surface defect detection. *IEEE Trans. Instrum Meas.* **2024**, *73*, 1-17. DOI
135. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct 27 - Nov 02, 2019; IEEE, 2019; pp. 9627–36. DOI
136. Xu, W.; Wang, H.; Qi, F.; Lu, C. Explicit shape encoding for real-time instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct 27 - Nov 02, 2019; IEEE, 2019; pp. 5168–77. DOI
137. Mayr, M.; Hoffmann, M.; Maier, A.; Christlein, V. Weakly supervised segmentation of cracks on solar cells using normalized Lp norm. In *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep 22-25, 2019; IEEE, 2019; pp. 1885–9. DOI
138. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–9. https://openaccess.thecvf.com/content_cvpr_2016/papers/Zhou_Learning_Deep_Features_CVPR_2016_paper.pdf. (accessed 2025-02-19)
139. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336-59. DOI
140. Wang L, Zhang M, Shi W. CS-WSCDNet: class activation mapping and segment anything model-based framework for weakly supervised change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1-12. DOI
141. Zhang J, Zhang Q, Gong Y, Zhang J, Chen L, Zeng D. Weakly supervised semantic segmentation with consistency-constrained multiclass attention for remote sensing scenes. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1-18. DOI
142. Yang R, He G, Yin R, et al. Weakly-semi supervised extraction of rooftop photovoltaics from high-resolution images based on segment anything model and class activation map. *Appl. Energy.* **2024**, *361*, 122964. DOI
143. Dai, J.; He, K.; Sun, J. Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1635–43. https://openaccess.thecvf.com/content_iccv_2015/papers/Dai_BoxSup_Exploiting_Bounding_ICCV_2015_paper.pdf. (accessed 2025-02-19)
144. Arbeláez, P.; Pont-Tuset, J.; Barron, J.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA; Jun 23-28, 2014; IEEE, 2014; pp. 328–35. DOI
145. Kulharia, V.; Chandra, S.; Agrawal, A.; Torr, P.; Tyagi, A. Box2seg: attention weighted loss and discriminative feature learning for weakly supervised segmentation. https://www.robots.ox.ac.uk/~tvg/publications/2020/box2seg.pdf. (accessed 2025-02-19)
146. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA; Jun 20-25, 2021; IEEE, 2021; pp. 2613–22. DOI
147. Zhang B, Xiao J, Wei Y, Sun M, Huang K. Reliability does matter: an end-to-end weakly supervised semantic segmentation approach. *AAAI.* **2020**, *34*, 12765-72. DOI
148. Kanezaki, A. Unsupervised image segmentation by backpropagation. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Calgary, Canada, Apr 15-20, 2018; IEEE, 2018; pp. 1543–7. DOI
149. Ouali, Y.; Hudelot, C.; Tami, M. Autoregressive unsupervised image segmentation. In *Computer Vision - ECCV 2020: 16th European Conference*, Glasgow, UK, Aug 23-28, 2020; Springer, 2020; pp. 142–58. DOI
150. Mirsadeghi SE, Royat A, Rezatofighi H. Unsupervised Image Segmentation by Mutual Information Maximization and Adversarial Regularization. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6931-8. DOI
151. Chen, X.; He, K. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, Jun 20-25, 2021; IEEE, 2021; pp. 15750–8. DOI
152. Zhang, F.; Torr, P.; Ranftl, R.; Richter, S. Looking beyond single images for contrastive semantic segmentation learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3285–97. https://proceedings.neurips.cc/paper_files/paper/2021/file/1a68e5f4ade56ed1d4bf273e55510750-Paper.pdf. (accessed 2025-02-19)
153. Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Van Gool, L. Unsupervised semantic segmentation by contrasting object mask proposals. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, Oct 10-17, 2021; IEEE, 2021; pp. 10052–62. DOI

154. Hwang, J. J.; Yu, S.; Shi, J.; et al. SegSort: segmentation by discriminative sorting of segments. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct 27 - Nov 02, 2019; IEEE, 2019; pp. 7334–44. DOI
155. van Knippenberg L, van Sloun RJ, Mischi M, de Ruijter J, Lopata R, Bouwman RA. Unsupervised domain adaptation method for segmenting cross-sectional CCA images. *Comput. Methods. Prog. Biomed.* **2022**, *225*, 107037. DOI
156. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P. Segformer: simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–90. https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf. (accessed 2025-02-19)
157. Zhang H, Wang S, Lu S, Yao L, Hu Y. Attention-gate-based U-shaped reconstruction network (AGUR-Net) for color-patterned fabric defect detection. *Text. Res. J.* **2023**, *93*, 3459-77. DOI
158. Ma P, He X, Chen Y, Liu Y. ISOD: improved small object detection based on extended scale feature pyramid network. *Vis. Comput.* **2025**, *41*, 465-79. DOI
159. Rajchl M, Lee MCH, Oktay O, et al. DeepCut: object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans. Med. Imaging.* **2017**, *36*, 674-83. DOI
160. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–49. https://openaccess.thecvf.com/content_ECCV_2018/papers/Mathilde_Caron_Deep_Clustering_for_ECCV_2018_paper.pdf. (accessed 2025-02-19)
161. Homayounfar, N.; Xiong, Y.; Liang, J.; Ma, W. C.; Urtasun, R. LevelSet R-CNN: a deep variational method for instance segmentation. In *Computer Vision - ECCV 2020: 16th European Conference*, Glasgow, UK, Aug 23–28, 2020; Springer, 2020; pp. 555–71. DOI
162. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–67. https://openaccess.thecvf.com/content_cvpr_2016/papers/Lin_ScribbleSup_Scribble-Supervised_Convolutional_CVPR_2016_paper.pdf. (accessed 2025-02-19)
163. Siva Raja P, rani AV. Brain tumor classification using a hybrid deep autoencoder with Bayesian fuzzy clustering-based segmentation approach. *Biocybern. Biomed. Eng.* **2020**, *40*, 440-53. DOI
164. Lin, T. Y.; Maire, M.; Belongie, S.; et al. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014: 13th European Conference*, Zurich, Switzerland, Sep 6-12, 2014; Springer, 2014; pp. 740–55. DOI
165. Mottaghi, R.; Chen, X.; Liu, X.; et al. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA; Jun 23-28, 2014; IEEE, 2014; pp. 891–8. DOI