SN Computer Science (2022) 3:499
https://doi.org/10.1007/s42979-022-01407-3

SURVEY ARTICLE

A Survey on Object Instance Segmentation

Rabi Sharma[1] · **Muhammad Saqib**[1] · **C. T. Lin**[1] · **Michael Blumenstein**[1]

Received: 19 May 2022 / Accepted: 7 September 2022 / Published online: 29 September 2022

Abstract
In recent years, instance segmentation has become a key research area in computer vision. This technology has been applied in varied applications such as robotics, healthcare and intelligent driving. Instance segmentation technology not only detects the location of the object but also marks edges for each single instance, which can solve both object detection and semantic segmentation concurrently. Our survey will give a detail introduction to the instance segmentation technology based on deep learning, reinforcement learning and transformers. Further, we will discuss about its development in this field along with the most common datasets used. We will also focus on different challenges and future development scope for instance segmentation. This technology will provide a strong reference for future researchers in our survey paper.

Keywords Instance segmentation · Deep learning · Reinforcement learning · Convolutional neural network · Transformers

Introduction

Segmenting a mask from an image/video channel is a challenging task that recently attracted significant attention from the computer vision community. Analyzing an instance mask is not only detecting and classifying the object in an image; rather, it is also a description of the exact individual object boundaries, labelings them according to their classes. So, it has become a key component in image analysis and visual understanding, which are both vital for various domains which include video surveillance, robotics, healthcare, video tracking, human–computer interaction, etc. On the other hand, video data are different from still image segmentation, where the target objects cannot be identified clearly due to different viewpoints. Additionally, camera motion obscures the targeted object in the video data.

✉ Rabi Sharma
rabi.sharma@student.uts.edu.au

Muhammad Saqib
muhammad.saqib@uts.edu.au

C. T. Lin
chin-teng.lin@uts.edu.au

Michael Blumenstein
michael.blumenstein@uts.edu.au

[1] Faculty of Engineering and Information Technology, School of Computer Science, University of Technology Sydney, Broadway, NSW 2007 Sydney, Australia

Recently, considerable work has been completed in distinct areas of computer vision areas comprising, among others such as object classification [110–112], object detection [15, 29, 113, 114], semantic segmentation [21, 23, 24] and instance segmentation [28, 54]. The aim of the object classification task is to identify the object class by classifying them from a set of input labels. The object detection task helps to detect the object according to their location and also classify them as per categories. The segmentation task has been split into two distinct areas, Semantic Segmentation and Instance segmentation. Semantic segmentation aims to label each pixel of an image as belonging to a corresponding class of what is being represented, but the interesting thing is that it does not differentiate instances. The aim of instance segmentation is to delineate all instances of each class along with its accurately noted location. There has been a significant exploration of the research on instance segmentation, but there are still serious challenges, such as segmenting smaller objects, image degradation, occlusions, inaccurate depth estimation, and handling aerial images, named only the major tasks. Researchers have used several approaches to generate a bounding box and then segment the instance mask. However, the most challenging task is to get the exact instance mask at a pixel level, along with their class names. The task of instance segmentation solves the object detection and semantic segmentation problems. In Fig. 1, we show the four basic tasks of computer vision such as object

Fig. 1 First row (Left to Right) (**a**) Classification (**c**) Semantic Segmentation, Second row (Left to Right) (**b**) Object Detection (**d**) Instance Segmentation

classification, object detection, semantic segmentation, and instance segmentation.

Lately, deep-learning-based methods have become very popular using Convolutional Neural Networks (CNNs) [1–5] introducing many instance segmentation frameworks. Modern proposal-based instance segmentation approaches depend heavily on the bounding box technique, which forms the basis for predicting instance masks. These bounding box techniques are time-consuming and difficult to train because they detect the object first before they segment them. The proposal-free instance segmentation approaches to overcome the limits of the bounding box to generate a pixel-level segmentation map across the image and localize object instances. This method is simpler to train and very efficient. The Deep Learning algorithms have various drawbacks, including but not only being time-consuming and presenting the need for large data for training and simply demanding more work.

Reinforcement Learning (RL) recently showed potentially promising outcomes for solving complex tasks [6, 7]. However, instance segmentation using reinforcement learning is an active research field that presents several challenges, such as the complexity of state and action space, large-scale segmentation, and occlusion. Little of this work has been explored in the domain of instance segmentation. Figure 2 shows the existing methods, for instance segmentation with the help of a flowchart based on deep learning, reinforcement learning and transformers. We have split instance segmentation into 3 parts,i.e., deep learning, reinforcement learning and transformers. First, we have deep learning techniques that consist of two types of methods, i.e., Proposal Based and Proposal-free. Proposal-based methods are based on the bounding box technique that detects the object with a bounding box and then segment them. On the other hand, proposal-free methods are based on clustering or grouping technique at the pixel level to generate object masks. Second, we have reinforcement learning techniques and third, are transformer techniques.

Another recent research area in the computer vision community shows promising performance, i.e., Transformers. Transformers shows great success in natural language processing (NLP); however, researcher focus on using transformer to address computer vision problem [115, 116]. Furthermore, Transformers have shown promising results on several computer vision tasks such as image recognition [117, 118], object detection [119, 120], segmentation [121] and some others use cases [122, 123].

Motivation and Contributions Recently, instance segmentation technology has witnessed exponential growth in the computer vision community based on different techniques i.e., deep learning, reinforcement learning, and transformers. This technology is now dominant in distinguished instance segmentation conferences and journals, and it is getting tough to keep pace with the current progress due to the rapid inflow of papers. As such, there are limited survey papers on instance technology such as [133, 134]. However, this survey paper focuses on different techniques along with their methods and datasets. To this end, we provide a holistic overview of the instance segmentation technology based on different techniques and methods. We hope this survey will provide

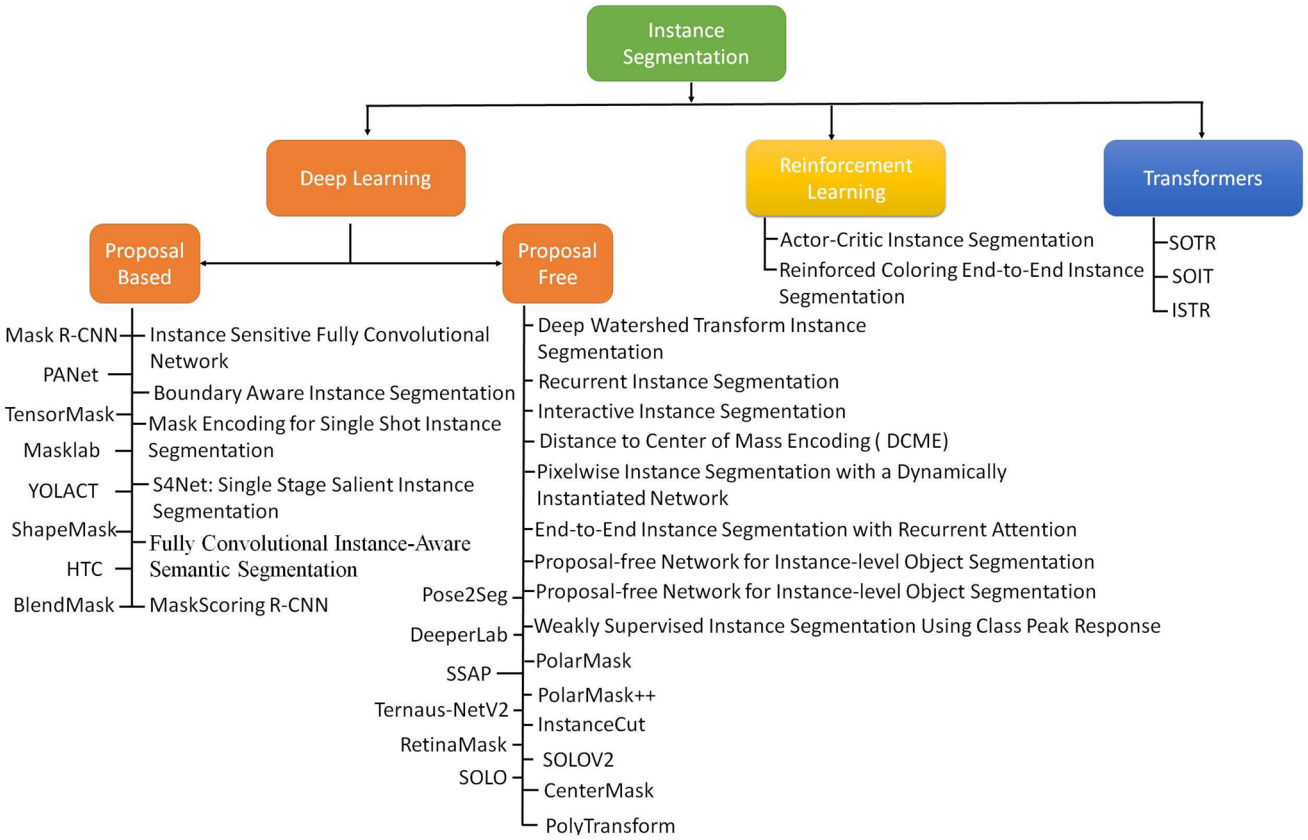

Fig. 2 Instance segmentation flowchart based on deep learning, reinforcement learning and transformers

a roadmap for the researcher to explore further. Our major contributions include:

- This is the first survey paper on instance segmentation that broadly covers the technology based on different techniques such as deep learning, reinforcement learning and transformers. Specifically, we present a comprehensive overview of more than 40 papers to cover the recent progress in detail.
- We provide complete coverage of this field by sorting the paper based on the techniques used. This survey paper represents the different applications using instance segmentation technology in Fig. 3.
- Finally, we provide a discussion about the key challenges, highlighting open problems and outlining the future scope.

Paper Organizations The rest of the survey paper is organized as follows. "Instance Segmentation Using Deep Learning" will introduce instance segmentation using deep learning algorithms in detail. "Instance Segmentation Using ReinforcementLearning" will describe this technology based on reinforcement learning. Section 4 will explain the current mainstream algorithms based on transformers. "Results" will compare the results based on different techniques and their datasets. "Datasets" will discuss the common dataset along with their images. "Discussion, Challenges and Future Scope" will discuss the challenges and future scope of instance segmentation technology. Finally, in "Conclusion", conclusions are given.

Instance Segmentation Using Deep Learning

Deep learning has explored several areas in computer vision extensively, leading to numerous modern tasks such as object classification [8–10], object detection [11–20], and semantic segmentation [21–24]. From the advances of deep learning in computer vision, it is encouraging that the capability of deep learning is growing rapidly towards instance segmentation which is the most difficult task in the vision community and the primary focus of this paper. We have split the techniques of instance segmentation into the two domains of a proposal-based and proposal-free approach.

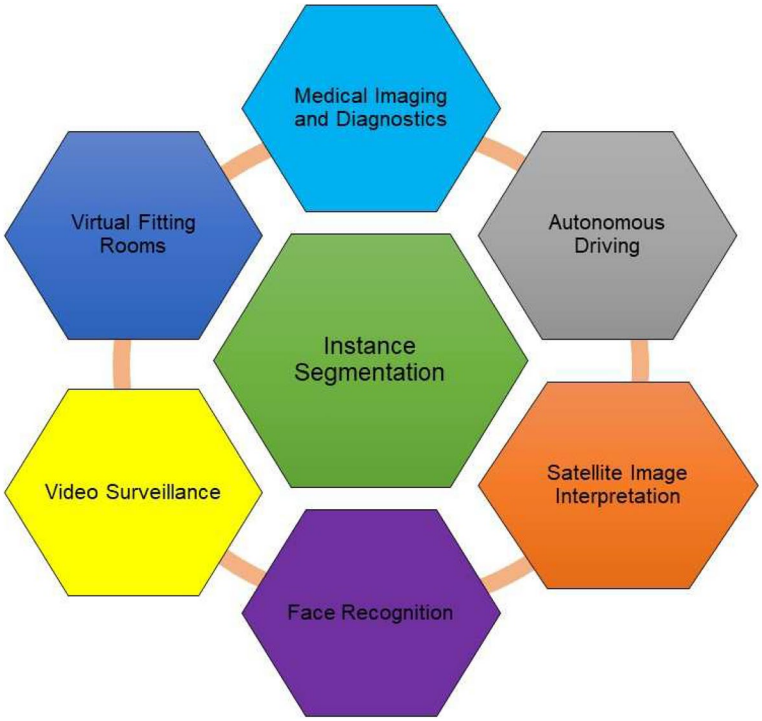

Fig. 3 A diverse set of applications using instance segmentation technology

Proposal-Based Approaches

The proposal-based approach is the baseline technique of instance segmentation generating a fine instance mask with the help of the bounding box, including the sub-tasks of object localization and classification. We use Mask R-CNN [28] framework to illustrate the proposal-based approach Fig. 4. Current prominent techniques selected from the many methods used for instance segmentation are discussed below.

Mask R-CNN

Mask R-CNN [28] is a simple and flexible framework driven by the existing method in [15], one that displays the two heads of classification and regression used for object detection using bounding box techniques. By adding another head on top of Faster R-CNN, such as a mask head that works in parallel with classification, regression and mask prediction from the Region of Interest(ROI) layer becomes possible.

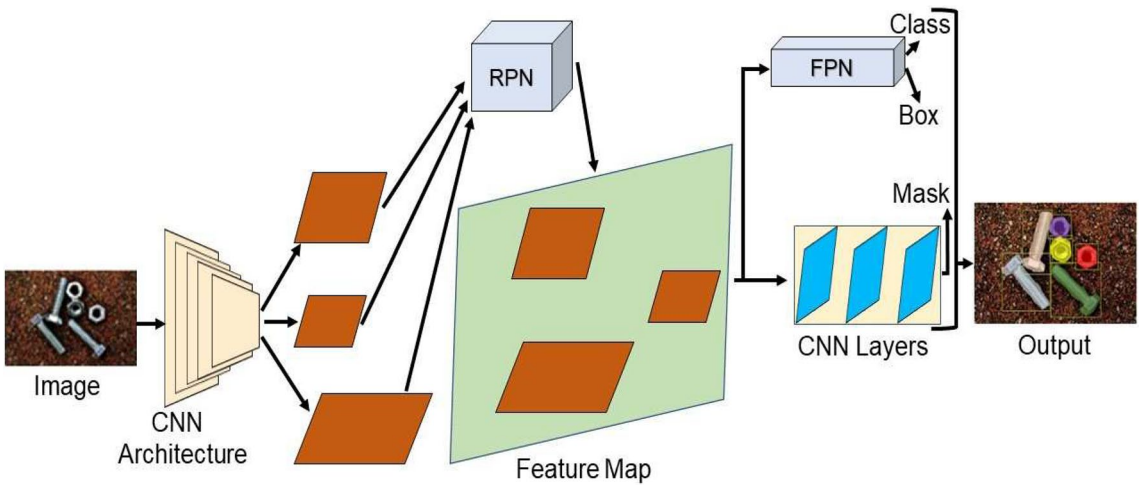

Fig. 4 Architecture of Mask R-CNN: In the first stage, an image is fed into CNN layers that generates feature maps, RPN helps to identify multiple objects present within a particular image. Later, the second stage consists of two parallel tasks, i.e FPN and Mask prediction, where the FPN generates the object class along with the bounding box and mask prediction outputs binary mask for individual object

ROIpool [29, 30] is replaced by ROIAlign, which does not use the quantization layer to smoothly transform features from ROIs into a fixed-size feature vector. The backbone architecture used in Mask R-CNN[28] are ResNet-101-C4, ResNet-101-FPN and ResNext-101-FPN using the publicly available datasets MS COCO [31] and cityscapes [32]. Furthermore, these datasets are also used for human-pose estimation through key-point techniques. Mask RCNN also achieves 1st place for COCO suite 2016 challenges for all three tasks of segmentation, object detection, and keypoint detection for human pose estimation.

Instance-Sensitive Fully Convolutional Networks (FCNs)

Although [26] uses FCNs that do not differentiate individual object instances of the same category as in the semantic segmentation process. In [25], the author uses FCNs instead of fully connected layers to segment object instances. Previously the FCN worked for semantic segmentation that involves only one object class and generates one score map, but the author proposes an instance sensitive score map where the results are in the form of pixel-wise classifiers that locate the positions of object instances. It consists of two FCN branches; it begins by using the features to generate instance-sensitive score maps, which, with the help of an assembling module, used these maps used to create object instances by employing a sliding window technique in each case to produce segmented instances. As a second step, the scored instances are calculated to make an object score map. As compared to Deepmask [27], FCNs minimize the number of parameters and the computational cost.

Fully Convolutional Instance-Aware Semantic Segmentation

This method has similar score maps in FCNs for semantic segmentation [26] and in an instance-sensitive fully convolutional network [25]. In [33], the author has implemented two position-sensitive inside/outside score maps using ROI employed to compute both object segmentation and detection cooperatively and concurrently. Taking the inside/outside score maps, one is designated for the instance mask and the other for category probability. Finally, these score maps and convolutional representations are fully allocated for both detection and segmentation tasks. Thus the performance has successfully achieved the paired results, i.e., accuracy and efficiency, as well as gaining the COCO 2016 segmentation competition.

Boundary-Aware Instance Segmentation

To avoid any inaccuracy/errors for the bounding box proposals during the object generation procedure, the author in [34], develops a method that computes the distance transform-based technique to predict instance segmentation beyond the limitations of anchor boxes. Furthermore, Object Mask Network (OMN), a novel architecture was designed for generating segment mask feed from the ROI warping features from each proposal, i.e., generating the bounding box proposal from Region Proposal Network (RPN) [15]. This helps to predict the mask beyond the scope of anchor boxes. Finally, OMN is merged with a multitasking network cascade framework to learn the output end-to-end. Pascal VOC 2012 [36] and Cityscapes datasets were used to evaluate both object proposal generation and instance segmentation.

S4Net: Single-Stage Salient-Instance Segmentation

Traditionally, foreground and background features are separated using several techniques such as local and global methods [95, 96], background priority [97], and GrabCut [98] which each help to perform the tasks of salient object detection and segmentation. Later a Convolutional Neural Network(CNN) plays an important role in both salient object detection [99–101], and semantic segmentation [102–104] along with common object detection [15, 105, 106] by learning from features extraction at different levels for the targeted categories. This approach achieves incredible results. However, a process is needed to implement the salient object-detecting technique for an instance segmentation framework, one which identifies the individual instance in a scene. An existing CNN-based technique such as ROIpooling [29], ROIWarp [35] and ROIAlign [28] to extract features from each bounding box is performed. In S4Net [37], the author introduced a new ROIMasking layer for feature extraction, which not only extracts information from inside the bounding box but also separates its surroundings information from the targeted object to enhance the segmentation task. To improve performance the salient instance discriminator increases the receptive field of the segment branch. The model is an end-to-end trainable approach that evaluates Pascal VOC 2012 datasets and works in real time.

PANet: Path Aggregation Network for Instance Segmentation

There is a need to improve the information flow from the lower layer to the topmost feature in Mask R-CNN [28], which suffers loss of localization information at high-level features due to the extensive path from lower-level to top-level features. PANet [38] is an extension of Mask R-CNN [28], which aims to boost the information flow through different levels of pool features and shorten the distance between lower-level and topmost-level features by creating a bottom-up augmentation technique. In addition, adaptive feature pooling has been used to retrieve the broken

information path between all feature levels and for each proposal. Finally, to enhance the quality of the segmented mask, an alternative branch captures each proposal from a different view with the help of fully connected layers. The method performs on publicly available datasets such as MS COCO, Cityscapes, and Mapillary Vista, achieving 1st in the COCO 2017 challenge as well as 2nd on the detection challenge.

Masklab : Instance Segmentation by Refining Object Detection with Semantic and Direction Features

There are improvements made possible by combining the two existing techniques such as detection based and segmentation based, to address the problem of instance segmentation. The author in [40] proposes Masklab, i.e., known as Mask Labeling built on top of Faster R-CNN [15] generate two outputs, i.e., semantic segmentation and instance center direction. The output of the predicted bounding box from [15] will give the object instances at a different scale to a canonical scale. Using the individual box prediction, foreground/background segmentation has been performed by combining the pair, semantic along with direction prediction. The task of semantic segmentation prediction, which encodes the pixel-level classification as well as background class, is implemented to distinguish among objects of different semantic categories. However, this method eliminates the duplicate background encoding in [33]. Furthermore, direction prediction is used to split object instances with identical semantic labels. After gathering the direction information using the assembling technique in [25, 33] the complicated template matching method is discarded [41]. Additionally, inspired by the current progress in segmentation and detection both together, the present technique incorporates atrous convolution [21] for denser features map extraction, hypercolumn features [42] for filtering the mask segmentation, multi-grid technique [43–45] for capturing unrelated context scales and a new Tensorflow operation technique [46], for deformable crop and resize, as motivated by the deformable pooling operation [44]. The model was evaluated on MSCOCO [31] and the results were like those obtained via more state-of-art models.

TensorMask: A Foundation for Dense Object Segmentation

Implementing the sliding-window technique for object detection has shown rapid advancement and popularity, where it generates a bounding box for object prediction in support of a dense and regular grid. In [47], TensorMask has been introduced to execute instance segmentation using the dense sliding window technique, which remains largely unexplored. It works with 4D tensors with shapes (H, W, V, U), where (H, W) represents the image position and (V, U) represents the mask location as sub-tensors. A pyramidal structure has been developed on the head of an indexed list of 4D tensors known as tensor bipyramid enabled in the TensorMask framework. They grow in opposite directions in a pyramid shape in (H, W) and (V, U) sub-tensors. These portions are combined into a network by following the training process of the RetinaNet [48] where the dense mask predictors extend the original dense anchor box predictors. Finally, TensorMask produces results related to the Mask R-CNN [28] counterpart. The author claims the results show the proposed framework can pave the way toward future work for dense sliding window instance segmentation.

ShapeMask: Learning to Segment Novel Objects by Refining Shape Priors

In [49], the author introduced Shapemask to improve the generalization for instance segmentation which needs many mask annotations. It can be very difficult and costly to segment objects and simultaneously learn from object shapes. Initially, the proposed technique uses Retinanet [50] for bounding box detection and refines it by approximating the detected object shape by collecting prior shapes. Later, it filters the coarse shape into an instance mask learned from Instance embeddings. The shape priors have a strong indication for object prediction and Instance Embeddings which gives the instance-level information. This outperforms the state-of-the-art learning categories while providing competitive performance in a fully supervised setting. This method is robust for inaccurate detections, which in turn decreases the system capacity and trains in limited data.

YOLACT: Real-Time Instance Segmentation

YOLACT is a real-time instance segmentation that is fast and simple and that uses a fully convolutional model. Using a single-stage object detector [51], not easy for a real-time scenario, the author proposed the You Look Only At Coefficient (YOLACT), which is split into two concurrent modules; the first task is the generation of a prototype masks branch which takes the entire image and using FCN creates a set of prototype mask. Secondly, the vector of mask coefficients has a prediction for each proposal which adds a new head in object detection that converts instance representations by the prototype space. Finally, both branches are combined linearly to construct a mask. The author has examined the outcome of developing a prototype mask and shows that the system learns to localize object instances automatically in a translational variant regardless of a fully convolutional network.

Hybrid Task Cascade for Instance Segmentation

A cascade combination of two powerful architectures is a useful approach for boosting performance in different tasks. One of the challenges is to implement cascade to instance segmentation. Combining cascade R-CNN [52] and Mask R-CNN [28] has achieved limited improvement. To explore this, the author proposed Hybrid Task Cascade (HTC) [53] split into two important aspects, (1) In place of performing cascading refinement tasks separately, it merges them for mutual multi-stage processing. (2) By adopting a fully convolutional branch to deliver spatial context, which helps to differentiate the solid foreground from the complex background. The author declares that the proposed method is capable of acquiring more selective features by integrating them gradually at each stage. Without fine-tuning, the HTC model gains 38.4% on mask AP along with 1.5% improvement on cascade mask R-CNN on MS COCO datasets. The system also obtains 48.6 mask AP for the test challenge and ranked 1st on COCO 2018 in object detection.

Mask Scoring R-CNN

A major job in the field of image segmentation is where deep neural networks are aware of their prediction quality. In instance segmentation, the estimation of confidence for classification of the instance is used for scoring the mask quality in major segmentation methods. Contrary to this technique, mask quality is measured by the Intersection over Union (IoU) between the instance mask and their ground truth. But in this method, they have not related correctly for calculating the classification score. The author in mask scoring R-CNN [54] has analyzed and addressed the problem without dropping the generality, employing in Mask R-CNN [28] that adds a new MaskIoU prediction head. This head acquires the MaskIoU associated with the mask score. The author declares that the proposed technique is simple to use and implement. By combining Mask R-CNN and MaskIoU prediction head that jointly fed the instance features and predicted mask, which helps to predict IoU among input and ground-truth masks. Moreover, the mask scoring technique computes the alignment fault within the mask quality and the mask score. Therefore, it will improve the performance of instance segmentation tasks that prioritize the prediction of a correct mask during COCO AP evaluation. With the extensive experiments conducted on [31], the proposed method increases steadily and noticeably on many models. It even beats the existing model in [28].

Blendmask: Top-Down Meets Bottom-Up for Instance Segmentation

Instance segmentation is a challenging problem in computer vision. Currently, FCIS [33] technique has come to prominence, though simpler and more effective than two-stage methods such as Mask R-CNN [28]. Several approaches lag behind the existing Mask R-CNN [28] approach due to the mask accuracy. But the models are not different in their computation complexity which presents a great opportunity for improvement. Blendmask [55] has been introduced because it achieves enhanced mask prediction by merging the instance-level information of the object along with semantic information at the lower-level fine-grained. The author contributes a blender module that encourages both top-down and bottom-up instance segmentation methods. The proposed module uses a few channels which can predict the positive-sensitive instance-level feature map and learns the attention maps from individual instances by implementing a single convolutional layer, thus making inference faster. Blendmask can easily integrate into the state-of-the-art single-stage detection models that outperform the existing benchmark model Mask R-CNN along with the same training time with shorter intervals. Furthermore, the framework can be used for a wide variety of instance prediction tasks.

Mask Encoding for Single-Shot Instance Segmentation

Until today, Instance Segmentation has been influenced by two-stage techniques such as Mask R-CNN [28]. By comparison, one-stage techniques cannot compete with the two-stage Mask RCNN on mask AP due to correctly representing the mask, which makes the development of a single-stage approach exceedingly difficult. The author proposes a single-shot approach named Mask Encoding for Single Shot Instance Segmentation (MEInst) [56]. Without predicting the 2D mask directly, the proposed approach distills data into a fixed-dimensional vector, which allows integration into one-stage detectors for the instance segmentation task generating effective results. This is seen as a simple and highly flexible single-stage instance segmentation technique that achieves a competitive performance and is easy to apply to other instance-level tasks.

Boundary-Preserving Mask R-CNN

To improve mask location and accuracy incredible efforts have been made in instance segmentation. Recently, Fully Convolutional Networks are being used for instance segmentation approach at the pixel-level classification that disregards the object edges and contours, which have generally led to rough and fuzzy mask outcomes, along with imprecise locations. The author proposes Boundary-preserving Mask R-CNN [57] to overcome these problems and obtain boundary information that increases accurate mask prediction. It contains a new head known as a boundary preserving mask which jointly learns the object boundary and masks with the

help of fusion blocks. Finally, the resulting mask is correctly aligned with object edges.The proposed method outperforms existing frameworks such as Mask R-CNN [28] in terms of accuracy and object location while using the precise existing ground truth.

Proposal-Free Approaches

The research aims to segment the instance mask in an image based on the proposal-free approach. The success of semantic segmentation [20, 22–24] resolves the problem of instance segmentation for this approach. The basic concept of this method is the application of grouping/clustering techniques at a pixel level to generate instances. We have used PolarMask [69] architecture for the proposal-free approach shown in Fig. 5. There are numerous algorithms in use, for instance segmentation that employs proposal-free techniques as described below.

PolarMask: Single-Shot Instance Segmentation with Polar Representation

Polarmask [69] is an anchor-free and single-shot instance segmentation technique that uses a Fully Convolutional Network and is easily implemented on existing object detection methods. The proposed method predicts the instance contour to address the problem of instance segmentation using two parallel tasks such as for example, center classification and dense distance regression in a polar coordinate. Furthermore, the author proposes two effective approaches, such as polar IoU loss and polar centeredness which improve optimization and centeredness [70] to boost performance.

PolarMask++: Enhanced Polar Representation for Single-Shot Instance Segmentation and Beyond

To overcome and reduce the complexity of the PolarMask [69], the author proposed PolarMask++ [107] as an anchor-free and single-shot pipeline that works for both instance segmentation and rotated object detection. The proposed methods have two parallel tasks, such as polar IoU loss and soft polar centerness for center classification and dense regression. Later, a Refined Feature Pyramid module has been proposed to increase the accuracy for features at different scales. The experiments of PolarMask++ has done on different datasets such as COCO [31], ICDAR2015 [108] and DSB2018 [109].

RetinaMask: Learning to Predict Masks Improves State-of-the-Art Single-Shot Detection for Free

Two-stage frameworks have developed toward single-shot detectors in terms of accuracy and speed trade-off. However, single-shot detectors are immeasurably popular in computer vision tasks. The proposed technique [74], consists of three ways to increase accuracy during training; One step could be by first adding a new branch such as mask prediction in RetinaNet [50] through training. Second, by tuning the parameter, an adaptive loss function has been adopted that improves robustness. Finally, adding more positive samples for training. This approach is known as RetinaMask, and the computational cost of the detection element is the same as Retinanet [50] with solid accuracy. Furthermore, group normalization will considerably improve the system performance of the RetinaMask-101 architecture.

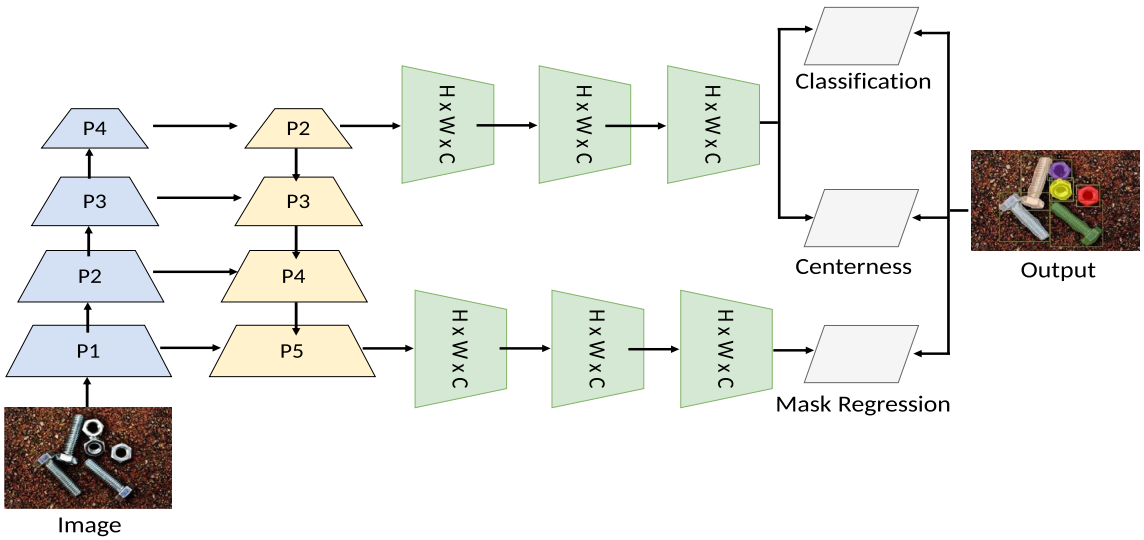

Fig. 5 Architecture of Polarmask: To extract features from different levels, the image pass into backbone and feature pyramid (P1 to P4 and P2 to P5). The middle part contains two heads, i.e., first head contains Object classification and Object center, and second head generates Mask regression. However, H, W and C represent height, width and channels of feature maps

Deep Watershed Transform for Instance Segmentation

To take up the challenges of contemporary methods for instance segmentation that uses composite pipelines such as Conditional Random Fields, Recurrent Neural Networks, Object Proposals, or Template Matching techniques, the author in [58] has introduced a novel tactic. The proposed approach is a simple and powerful end-to-end neural network that can tackle these problems. By combining the techniques from traditional watershed transformation and deep learning methods, it is indicated we can generate the energy level of an image that represents the object instance precisely in the energy basins. Finally, the Cut technique was implemented at a sole energy level to avoid over-segmentation. The model has achieved double performance over the existing Cityscapes [32] dataset.

InstanceCut: from Edges to Instances with MultiCut

InstanceCut [59] has been introduced for Instance-aware Semantic Segmentation tasks which, it is proposed, can be solved by designing a new paradigm that can assess the trade-off between advantages and disadvantages related to known techniques. The proposed technique is denoted by two outcomes: (1) an instance-agnostic semantic segmentation and (2) entire instance boundaries. The antecedent for semantic segmentation is calculated using a convolutional neural network which allows the derivation of the edge detection model. Finally, the author proposes a new multicut approach by merging the two processes to permit universal reasoning regarding the optimal separation of images into instances. A publicly available dataset has been used for evaluation.

Recurrent Instance Segmentation

In [60], the author understands the problem of detecting and delineate individual instances in an image for instance segmentation. But the existing approaches are based on a pipeline of components that are trained autonomously for each other, hence losing the chances of joint learning. The proposed approach contains an end-to-end method that masters to fragment instances sequentially. However, the implementation of a recurrent neural network will search the objects and perform segmentation concurrently. This recurrent neural network helps by adding a spatial memory that keeps track of the pixels as well as handles occlusion. A loss function has been designed for model training that keeps track of accuracy for instance segmentation problems. The proposed approach outperforms the existing state-of-the-art approach to multiple-person segmentation.

Iterative Instance Segmentation

Previous dependence on, Pixel-wise labeling has ignored the fundamental structure of annotation, leading to visually implausible outcomes. To improve the predicted quality, we initiate the challenging task of integrating the structure into the model. However, defining the structural procedure manually has certain limitations, which may make it unfeasible as well as present unmanageable inference. The proposed approach [61] used for instance segmentation task, which can learn the shape, nearest regions, and smooth region contours information without any prior evidence. The model outperforms the state-of-the-art on instance segmentation tasks by achieving good results.

Pixelwise Instance Segmentation with a Dynamically Instantiated Network

Recent advances in the field of computer vision tasks such as semantic segmentation and object detection indicate its increasing popularity. But there is no improvement for instance segmentation able to delineate individual objects with the same technique using the bounding box. In [62], the author introduces a method that can apply to instance segmentation by generating a semantic map that can assign an object class with an instance label at each pixel level. Predominantly, existing object detectors are used to segment as a replacement for anchor boxes. The proposed approach builds on a semantic segmentation module that feeds into an instance sub-network using information taken from the object detector following part of the dynamic network within the Conditional Random Field (CRF) is thereby able to differentiate diverse instances. It involves no postprocessing and contemplates the entire image rather than processing the individual proposals. Finally, the outcome of the end-to-end network is to dynamically yield the variable number of instances per image. So, also being applicable in related but dissimilar work, where multiple instances cannot depend on a single pixel. Furthermore, the proposed network shows improvement on semantic segmentation tasks, as well as in the task of instance segmentation at high AP thresholds.

End-to-End Instance Segmentation with Recurrent Attention

So far, Convolutional Neural Network (CNN) [3] has gained a lot of attention for solving the structured output task in semantic segmentation, but there is still the challenging task of separating individual objects in an image. Several applications such as autonomous driving, image captioning, and visual question answering are important, for instance segmentation tasks. Combining the large graphical model with the low-level counterpart technically helps to solve the

problem. So, the author proposes end-to-end instance segmentation with recurrent attention [63], which is based on a recurrent neural network utilizing visual attention and taking account of the human-like counting method to build instance segmentation. However, the network is trained jointly to yield a region of interest along with the segmentation of each object region.

SGN: Sequential Grouping Networks for Instance Segmentation

In [64], the author addresses the problem of instance segmentation using the grouping technique. Here, the task of instance segmentation breaks into several sub-tasks which are easily trackable. The grouping approach is used to employ a chain of Convolutional Neural Networks that can solve the sub-grouping problem, but the semantic complexity increases to compose objects from pixels. The initial neural network aims to predict horizontal and vertical breakpoints by grouping pixels for each image by row and column. Using these breakpoints, line segments are created. Utilizing the two-dimensional information, the next network groups the horizontal and vertical lines by connected components. Finally, the connected components of the second network are grouped to generate object instances. The proposed approach outperforms existing methods.

Proposal-Free Network for Instance-Level Object Segmentation

Existing methods are based on Region Proposal Methods (RPN), where extracting the object elements to produce results that generate accurate proposals is a tough task. But instance-level object segmentation remains unexplored. In [76], the author presents a proposal free network that gives an output from the instance number of different classes and retrieves the pixel information from the following, (1) with the help of bounding box coordinates for each pixel for instance locations, (2) the pixel level neural network depends on the confidence scores of different classes for each pixel. Finally, the output uses the clustering method for post-processing that can generate instance-level object segmentation results. The proposed approach can easily be trained in an end-to-end manner without using the proposal generation branch.

Distance to Center of Mass Encoding for Instance Segmentation

In [77], the author proposes a technique for instance segmentation, to encounter the problem of object detection, which uses object contours as an alternative to bounding boxes. However, the task is challenging because it requires the technique to recognize single pixel and their classes independently. Instance segmentation has led to object detection by delineating the object accurately to move towards object localization. Furthermore, the basic image processing techniques can be used to evaluate any partial occlusion from object contours. So, this work presents instance segmentation as an annotation problem that designs new methods capable of encrypting and decrypting the ground truths. The author proposes a mathematical notation that makes semantic models learn and generalize from them. Individual instances are denoted by the center of mass, and a set of vectors are used to locate them.

TernausNetV2: Fully Convolutional Network for Instance Segmentation

Instance segmentation techniques are complex, along with using two-stage methods with object proposal, Conditional Random Fields (CRF), Pattern Matching, or Recurrent Neural Networks (RNNs). In [78], a fully convolutional network that allows extracting objects from high satellite imagery at the instance level is proposed. The architecture of the encoder–decoder has been implemented along with skip connections which have made some modifications to better suit semantic segmentation and instance segmentation. Additionally, the encoder network has been generalized by pre-training the RGB images for the input parameters. It can be used to transfer learning from visual to the wider spectral range.

Weakly Supervised Instance Segmentation Using Class Peak Response

To explore the pixel-level mask for instance segmentation is a challenging problem. To address this problem, the author proposes a new technique [75] that enables classification networks by exploiting class peak response for extraction of the instance masks. The CNN classifier uses a fully convolutional network that can generate class response maps to obtain a classification confidence score at each location in an image. From the local maxima's observation, the peaks of the class response map reflect the strong visualization cues that reside inside each instance. From this inspiration, a new strategy has been implemented where a class response map is processed by merging the stimulated peaks and backpropagating the peaks to map towards the informative part of each instance, i.e., the object boundaries. This class response map is called a Peak Response Map (PRM) and is a superior instance-level representation that can be used to extract object mask from existing methods.

SSAP: Single-Shot Instance Segmentation with Affinity Pyramid

Currently, anchor-free instance segmentation has become most popular for its compact and well-organized neural network. This is due to its use of proposal-free approaches [21, 23, 67], which help to construct the instance-agnostic for semantic class labels and thereby group the pixels into different object instances while reflecting instance-aware features. Existing approaches use the two sub-tasks that split into two separate phases, utilizing the different modules that are sub-optimal. Exploiting the two sub-tasks benefits, which can further improve the execution of instance segmentation and employ multiple modules separately, might reduce the computational cost for applications. To address this problem, the author proposes in [66], a single-shot proposal-free method that involves a single pass for the prediction. The proposed method depends on the pixel-pair affinity pyramid that computes the probability that two pixels depend on the same instance in an ordered fashion. However, the affinity pyramid simultaneously can learn labeling for the semantic class along with mutual benefits. Finally, a cascaded graph partition module is proposed by integrating the affinity pyramid that can generate instances in sequence from coarse to fine. The existing graph partition techniques are time-consuming, but this proposed method achieves a speed of 5 × along with a 9% increase in precision on average.

DeeperLab: Single-Shot Image Parser

Image parsing is a challenging task in computer vision, one which is less explored due to the problem of efficiency in parsing the whole image. Image parsing is extremely difficult in two popular tasks such as semantic segmentation for stuff classes and instance segmentation for thing classes that both assign semantic and instance labels at a pixel level. Existing approaches parse the whole image to separate modules such as semantic and instance tasks that consist of different epochs for inference. In [80], the author proposes a simple single shot and a bottom-up approach called DeeperLab, which parses the entire image using a fully convolutional technique that works cooperatively for semantic and instance segmentation tasks in a single pass. The outcomes are fused into the final image parsing branch for fast processing. To improve the quantitative evaluation, the author proposes two metrics such as the Instance-Based Panoptic Quality (PQ) and the Region-Based Parsing Covering (PC) metric, that can represent a better quality of an image parsing on stuff classes and for large object instances.

Pose2Seg: Detection Free Human Instance Segmentation

Instance segmentation has a typical approach that detects the object first, and with the help of bounding box detection, the objects are segmented. So, Mask R-CNN [28] using the deep learning method follows the same approach. However, the uniqueness of human categories can be explored when the pose skeleton has been unexplored. The human pose can be exploited to differentiate instance masks along with occlusion rather than using bounding boxes. In [79], the author proposes a new technique that works for pose-based instance segmentation for humans based on their human poses to separate instances without using the region proposal detection. An affine-align module has been proposed, which is a pose-based align branch that allows for scaling, translation, and flipping operations to correct their odd position through human poses. Additionally, the skeleton features are concatenated for the human pose to guide the segmentation branch, which in return improves the segmentation accuracy. Furthermore, a new benchmark exists, such as OCHuman or occluded human-introduced to solve the problem of high occlusion having complete annotations that consist of bounding boxes, human poses, and instance masks.

PolyTransform: Deep Polygon Transformer for Instance Segmentation

In [70], the author proposes a new technique for instance segmentation that helps to tackle the problem of instance segmentation and interactive annotation that occur due to large overlaps or if the object disconnects into multiple parts. The technique is used to generate instance masks with the help of a segmentation network. Later, the instance mask is converted into a set of polygons that are fed to a deforming network which transforms the polygon points to fit the object edges accurately. This method outperforms the model performance and improves the backbone of instance segmentation. It is also ranked 1st on the Cityscapes leaderboard.

CenterMask: Real-Time Anchor-Free Instance Segmentation

Compared to the two popular frameworks such as Mask R-CNN [28] represents a two-stage framework where mask accuracy is high but the speed is relatively low, and YOLACT [51] is a one-stage network that outperform the execution speed but with low mask accuracy. To achieve high speed and high mask accuracy, the author proposes [71], a real-time anchor-free instance segmentation that uses an anchor-free One-Stage Object Detector (FCOS) [69] to generate a predicted object in a bounding box. However, each detected box is fed to a novel Spatial Attention-Guided Mask (SAG-MASK) branch to predict the instance mask using the spatial attention map, which helps to concentrate

on meaningful pixels and discard the uninformative part (noise). Furthermore, it uses an improved version of the backbone in VoVNetV2 with two effective approaches; the first is to maximize the optimization using the residual connection, while the second is to prevent information and performance losses by merging the FC layer in a single unit with effective squeeze excitation. Moreover, the proposed framework SAG-Mask and VoVNetV2 are used to design CenterMask and CenterMask-lite, which are marked for each large and small model respectively. The models outperform the state-of-the-art in both speed and accuracy.

SOLO: Segmenting Objects by Locations

Existing methods such as dense prediction tasks, i.e, semantic segmentation, must deal with the simple fact that the random number of instances is challenging for the task of instance segmentation. Several techniques are used to predict instance masks by following detect and segment strategy or via predicting embedded vectors and using a clustering technique to group pixels into single instances. The author proposes [72] a new method by presenting the concept of instance categories where these are assigned to each pixel in an instance that quantifies the center locations and their object sizes which enables it to segment objects by locations. This approach outperforms performance along with mask accuracy correlated with the existing two-stage framework and achieves good accuracy compared to single-shot techniques.

SOLOv2: Dynamic and Fast Instance Segmentation

Following the existing method SOLO [72], the author proposes [73] a framework to make the process performance stronger and more efficient. SOLOv2 is split into two approaches; Firstly, the dynamic approach segments the object locations dynamically. However, the mask is learned in two crucial ways: Convolutional Kernel Learning and Feature Learning. Secondly, it implements the matrix NMS technique to suppress duplicate prediction and improve the mask AP. This matrix NMS technique work in parallel to perform matrix operations in a single shot and generates good results.

Instance Segmentation Using Reinforcement Learning

Reinforcement Learning (RL) is an active research field that shows promising results in solving complex tasks [81–83]. Reinforcement Learning is still unexplored outside its conventional domain, but solving instance segmentation problems using RL is a challenging answer to computer vision tasks, and little research has been conducted to date. We have used the actor-critic instance segmentation [84] framework to illustrate the RL approach in Fig. 6. There are a few methods proposed using RL as mentioned below:

Actor-Critic Instance Segmentation

Existing approaches give more importance to parallel processing by visualizing and analyzing the image from its elements. However, there is a single area that can work sequentially to segment multiple occluded objects in a scene. To solve this difficult task, the recurrent formulation was used in the context of reinforcement learning. The author proposes [84], in order to overcome the limitation of the max-matching assignment, to implement an actor-critic model where the actor repeatedly predicts individual instances at a time, and the critic network is trained simultaneously to employ the gradient. Hence, the basic elements of reinforcement learning are formulated, such as state, action, and rewards which can be used for the long-term effect from the present state of prediction, producing information that can be used for the gradient signal. The author introduces an encoder–decoder baseline that can be exploited for pixel-wise prediction by scaling the input image. Furthermore, to allow reinforcement learning for instance segmentation connected with high-dimensional output space, the conditional variational auto-encoder is employed to learn the compact representations and merge them with the recurrent prediction pipeline.

Reinforced Coloring for End-to-End Instance Segmentation

Delineating an individual object is a demanding task in computer vision and, naturally, is a research area for instance segmentation. The existing techniques for instance segmentation, work on diverse images that produce good results, but the output has some drawbacks, such as splitting or merging while segmenting multiple objects, which necessitates post-processing methods. On the other side, some other methods are used to extract individual objects at a time with the help of certain approaches, such as using shapes, boundaries, etc., avoiding previous techniques. In [85], the author introduces a novel instance segmentation technique using a reinforcement learning agent which learns to separate multiple objects concurrently. Using asynchronous Actor-Critic algorithms (A3C), the RL agent is trained to group pixels that belong to a similar object via a graph coloring algorithm. The proposed technique is used for multiple object segmentation that avoids postprocessing. In Table 1, we have

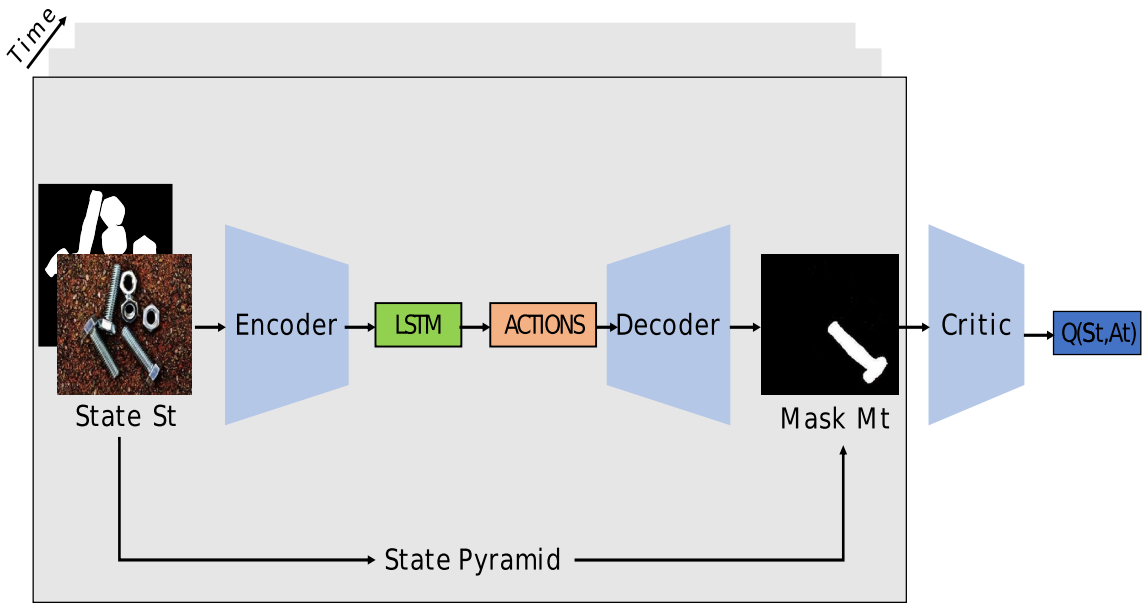

Fig. 6 Architecture of actor-critic instance segmentation: the input of the network contains the image and the binary mask chosen randomly from the ground truth. The encoder learns the instance features and decoder will receive the instance mask. LSTM helps to perform the prediction ordering and then generates actions. State pyramid generates high resolution information from different scales to the decoder to minimize the resolution loss. Finally, the critic reduces the reward errors to maximizes Q value

Table 1 Reinforcement learning results for instance segmentation

Method	Year	CVPPP	KITTI	CREMI
Actor-critic instance segmentation [84]	2019	**80.5**	71.9	–
Reinforced coloring for end-to-end IS [85]	2020	80.0	**77.0**	Type I-0.41, type II-0.379

shown the results for the above RL models. The best result is shown by bold font in the table.

Instance Segmentation Based on Transformers

Transformers have shown strong performance on natural language processing (NLP), which helps researchers to focus on these techniques to address computer vision problems. However, several algorithms have been introduced that address different tasks in computer vision, such as object recognition, object detection, semantic segmentation, and some others. The technique used in transformers uses a multi-head self-attention module which does not require any image-specific biases. The images are split into several patches to process with the help of a transformer encoder. We have used Segmenting Objects with Instance-Aware Transformers (SOIT) diagram in the Fig. 7. There are several approaches introduced to address the instance segmentation technology that are explained below:

ISTR: End-to-End Instance Segmentation with Transformers

To improve the accuracy of end-to-end models like object detection task has replaced components such as non-maximum suppression (NMS) with bipartite matching to reduce the redundant results. However, this upgradation is not applicable to instance segmentation due to high dimensional output. The author proposed End-to-End Instance Segmentation with Transformers (ISTR) [124], which helps to predict low-level mask embeddings rather than high-level embeddings, which makes it easy to train the model with a small number of datasets and also motivates measuring the bipartite matching cost for instance masks. Additionally, the recurrent refinement strategy work in parallel to detect and segment objects which creates a new approach to boost the performance for instance segmentation as compared to the bottom-up and top-down models.

SOTR: Segmenting Objects with Transformers

Previous research has significantly seen some drawbacks using CNN techniques, such as the absence of top-level

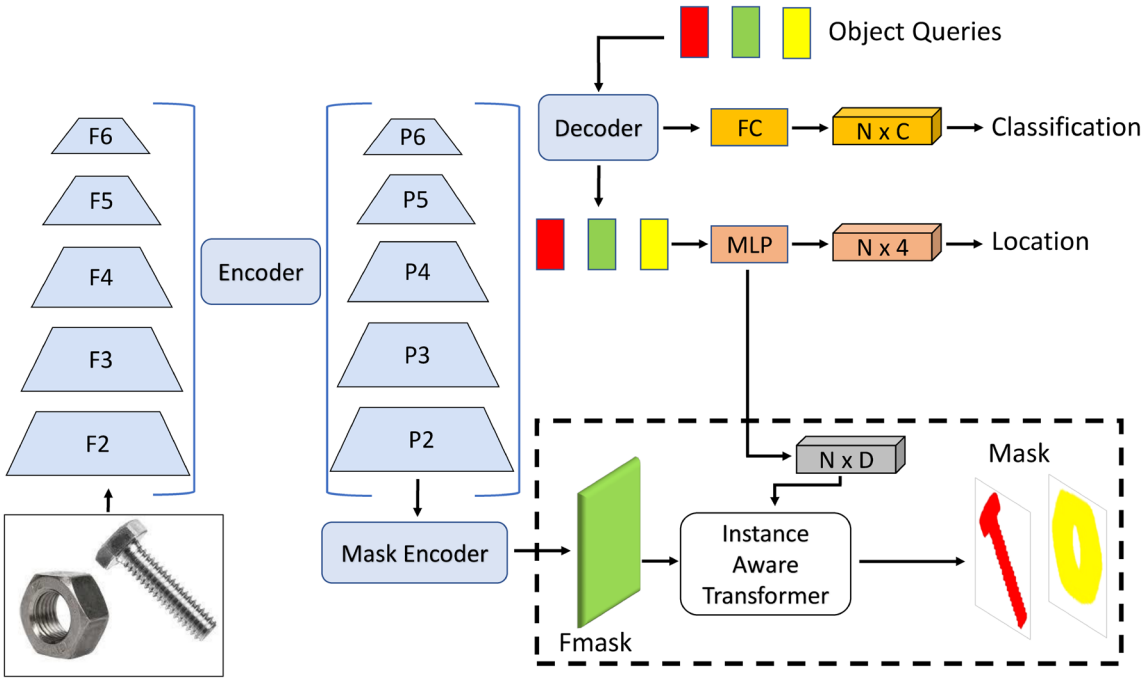

Fig. 7 Architecture of SOIT (Segmenting Objects with Instance-Aware Transformers): the multi-scale feature maps from F2 to F6 are extracted from the backbone such as Resnet-50. Transformer encoder refined the multi-scale feature memory from F2 to F6. Mask encoder generates the mask features for Fmask. Instance aware transformer are constructed by D-dimensional dynamic parameters from the mask branch. In the black dashed box, the pixel-wise mask are created by instance-aware transformer

visual information related to instances due to a shortage of receptive fields that leads to insignificant results on large-scale objects and also reduces the segmentation quality and performance in the complex framework. This drawback has been addressed using bottom-up techniques [125, 126], but the limitation of this technique shows unstable clustering (i.e., disintegrated and joint masks) and overfitting problems on a dataset with different scenes. So, to overcome such issues, the author proposed Segmenting Objects with Transformers (SOTR) [127] that, successfully learns the location-sensitive features and generates object masks by following the basic idea of [73], by isolating the clustering and bounding box techniques. On the other hand, a transformer has played crucial role in natural language processing (NLP) [128–130] that has replaced convolution operation or combining the CNN architectures with transformer for computer vision tasks [117, 120, 131] that can capture global-range features and distant semantic dependencies. But still are several challenges in transformer-based techniques, such as low-level feature extraction, memory, and time greedy due to the large feature map. To deal with these weaknesses, the author proposed a creative model called SOTR that integrates the advantages of both CNN and transformer that effectively predicts the object mask of each instance without the help of object detectors. However, the author has come up with a twin attention mechanism, which helps to reduce the computation and memory usage compared to the initial transformer. The model was evaluated on the MS COCO dataset that outperforms the instance segmentation techniques.

SOIT: Segmenting Objects with Instance-Aware Transformers

To overcome the set prediction problem for instance segmentation that can eliminate several hand-crafted components such as RoI cropping, one-to-many label assignment and non-maximum suppression (NMS) post-processing eliminates paired instances during the testing phase. As a result, these instance segmentation techniques do not satisfy an end-to-end optimization that leads to poor sub-optimal results. The author proposed a new transformer-based technique called Segment Objects with Instance-aware Transformer (SOIT) [132] that revise segmentation task as a set prediction problem and creates an end-to-end framework. The framework helps to generate pixel-wise mask directly for each instance in the absence of RoI cropping or NMS post-processing. However, this method helps to learn queries that can encode different object representations concurrently, such as categories, positions and

pixel-wise masks. This multi-task learning model helps to create a connection between object detection and instance segmentation that encourage both tasks to aid one another. Finally, this is a single-phase segmentation framework that has been conducted on the MS COCO dataset, which surpasses state-of-the-art instance segmentation techniques significantly.

A Simple Single-Scale Vision Transformer for Object Localization and Instance Segmentation

To avoid using traditional CNN's hierarchical pyramid and highly customized architecture of transformer for the dense prediction tasks for object detection and instance segmentation. Moreover, to minimize the computational cost for dense vision tasks that contain high-resolution input images. Although the combination of CNNs and Vision transformer (ViT) [136] performance in recent works achieves outstanding performance on vision tasks, it is still unclear about its true benefits. However, the author comprehensively studied the three architecture options for ViTs such as spatial reduction, doubled channels, and multi-scale features that help to propose a new single-scale transformer called Universal Vision Transformer (UViT) [135] without using the handcrafting feature pyramid of CNNs but only to support the fixed feature resolution and hidden size all over the encoder blocks and generates a single-scale feature map. Later, UViTs introduced a new scaling rule (depth, width, input size) for the dense vision tasks to improve performance efficiency. In order to minimize the computation cost, a split window strategy has been adopted in attention layers where the window size gets increased as the attention layers get deeper, leading to drop the computation cost. Finally, UViTs achieve strong performance for the task of object detection and instance segmentation on the COCO dataset.

Results

In this section, we will compare the instance segmentation results based on different techniques respective to their datasets. However, the results of the proposal-based and proposal-free approaches have been grouped and shown according to their datasets. In Table 2, the major datasets such as Coco [31], Cityscapes [32] and PascalVOC [86] are used for the experiments for instance segmentation. Furthermore, the methods and publication date are also mentioned accordingly. In Table 3, the other datasets that are not so much used for the experiments for instance segmentation technology.

Datasets

The datasets contain grayscale or RGB images, which are used to perform instance segmentation research. This two-dimensional dataset contains limited space for the location of pixels and their intensity values. The different dataset images are shown in Fig. 8. There are several datasets used for the task of instance segmentation as mentioned below:

MS COCO Dataset

The MS COCO, also known as Microsoft common objects in context datasets [31] is a large-scale image dataset used for several tasks that include object detection, segmentation, and captioning. This dataset consists of 80 classes where 82,783 are training images, 40,504 are validation images, and 80,000 are testing images. The COCO datasets are important for the computer vision community due to their large size and that they are also popularly used for instance segmentation.

Cityscapes Dataset

The cityscapes [32]is a large-scale dataset that contains urban road images that concentrate on semantic segmentation of the street scenes. This dataset contains annotations on several tasks such as semantic, instance-specific, and pixelwise with 30 classes which are grouped into 8 categories such as flat, human, vehicle, sky, etc. The images are collected from 50 different cities along with high-quality annotations where 5000 images and 20,000 images belong to coarse annotations. In forming the cityscapes dataset, several months were needed to collect the images in good weather conditions before being manually selected in various aspects like numbers of active objects, diverse scenes, and changing backgrounds.

The Mapillary Vistas Dataset (MVD)

The MVD [39] contains a large street-level images dataset with 66 object categories which total 25,000 annotated images. The annotation has been completed with a dense, fine-grained physically using the polygons to delineate the objects individually. The mapillary dataset is 5 times larger than the cityscapes with fine annotations, and the images are collected around the world and recorded under different weather conditions, seasons, and times of day, using discrete devices like mobile phones, cameras, tablets, etc. to capture the images and with the help of multiple photographers. The objective of creating this database is to add further progress toward state-of-the-art research to understand street scenes.

Table 2 Results on instance segmentation based on different techniques along with datasets

Methods	Year	Coco	Cityscapes	Pascalvoc
Instance-sensitive FCN [26]	2016	39.2	–	38.8
Deep watershed transformer IS [58]	2016	–	19.4	–
InstanceCut [59]	2016	–	13	–
Recurrent instance segmentation [60]	2016	–	–	43.7
Iterative instance segmentation [61]	2016	–	–	43.3
Mask R-CNN [28]	2017	31.5	36.5	–
Pixelwise IS with a DIN [62]	2017	–	38.8	48.6
End-to-end IS with recurrent attention [63]	2017	–	27.5	–
SGN [64]	2017	–	25	47.2
Distance to center of mass encoding [77]	2017	–	5.7	–
Boundary aware instance segmentation [34]	2017	–	36.7	**65.69**
Proposal-free Network (PFN) [76]	2017	–	–	58.7
FCN instance-aware semantic segmentation [33]	2017	29.2	–	–
Path Aggregation Network (PANet) [38]	2018	42	36.5	–
Masklab [40]	2018	35.4	–	–
Weakly supervised IS using class peak response [75]	2018	-	–	44.3
Tensormask [47]	2019	37.1	–	–
Shapemask [49]	2019	40	–	–
RetinaMask [74]	2019	39.4	–	–
YOLACT [51]	2019	29.8	–	–
Hybrid task cascade (HTC) for IS [53]	2019	38.4	–	–
Mask scoring RCNN [54]	2019	35.4	–	–
SSAP [66]	2019	–	32.7	–
Polytransform [70]	2019	–	**44.6**	–
Polarmask [69]	2020	32.1	–	–
Centermask [71]	2020	39.8	–	–
BlendMask [55]	2020	37	–	–
Solo [72]	2020	36.8	–	–
SoloV2 [73]	2020	38.8	–	–
Mask encoding for SS IS [56]	2020	33.9	–	–
Boundary preserving Mask RCNN [57]	2020	34.7	–	–
PolarMask++ [107]	2021	38.7	–	–
ISTR [124]	2021	**48.1/38.6**	–	–
SOTR [127]	2021	40.2	–	–
SOIT [132]	2021	42.5	–	–
A simple SS VT for object localization and IS [135]	2021	46.1	–	–

Pascal VOC 2012 Dataset

The pascal VOC 2012 [86] is a popular dataset used to perform several computer tasks like classification, object detection, and segmentation. It contains 20 classes like household, animal, airplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, sofa, etc. Each image contains 3 segments of annotations, such as pixel-level segmentation, bounding box, and object class annotations. The dataset consists of 2913 total images which are split into 2 subsets like 1464 train images and 1449 test images.

CVPPP Dataset

THE CVPPP [87], or to give it its full name, computer vision problems in plant phenotyping datasets is popular for the task of instance segmentation. The dataset has been collected from 5 different plants. The common sequence of AI contains an important number of baselines. It consists of 128 top-down image visualizations of size 530×500 each pixel with 128 training images and a hidden test set comprised of 33 images in the same sequence. The most challenging task for instance segmentation is that they contain a high diversity of leaf contours and occlusion problems among leaves.

Table 3 Results on instance segmentation for other datasets

Methods	Year	Datasets	AP
Path Aggregation Network (PANet) [38]	2018	Mapillary Vista [39]	26.5
Deeplab single-shot image parser [80]	2019	Mapillary Vista [39]	PC-31.95 PQ-55.26
Recurrent instance segmentation [60]	2016	CVPPP [87]	66.6
End-to-end IS with recurrent attention [63]	2017	CVPPP [87]	84.9
End-to-end IS with recurrent attention [63]	2017	KITTI [88]	80
Pixelwise IS with a DIN [62]	2017	SBD [90]	44.8
Ternaus-NetV2 [78]	2018	SpaceNet [91]	74
Pose2seg [79]	2019	OCHuman dataset	54.4
Polytransform [70]	2020	Self-driving datasets	35.3
PolarMask++ [107]	2021	ICDAR2015 [108]	85.4
PolarMask++ [107]	2021	DBS2018 [109]	74.2
S4Net [37]	2017	MSRA-B,HKUIS	86.7

KITTI Dataset

The KITTI dataset [88] is a car segmentation dataset that contains 3714 images for training, 144 images for validation, and 120 images for testing. The trained labels are created from [89] which consist of the unrefined resolution, and the images have high resolution in testing and validation. Other datasets are available for instance segmentation purpose such as Semantic boundaries dataset [90] (SBD), ICDAR 2015 [108] dataset, DBS 2018 [109] and SpaceNet [91].

Discussion, Challenges and Future Scope

In this section, we would like to discuss the several backbones used for the task of instance segmentation using deep learning and reinforcement learning, their challenges, and the future scope for instance segmentation.

Backbone Networks

In recent years, the popularity of deep learning has reached many milestones by using/implementing several deep backbones. These backbones include ResNet [92] which contains 50 and 101 layers, ResNext [93] with 101 layers, and VGG-16 [94] containing 16 layers, etc., each of which has been used successfully, although the computing can be too expensive and time-consuming.

Challenges

Instance segmentation has benefited greatly from using deep learning techniques, but still, many challenges lie ahead. On the other hand, we provide several challenges using transformers for instance segmentation technology. We will discuss here some of the research areas that will help to further advance instance segmentation technology.

Challenges Using Deep Learning Techniques

More complex datasets For the task of instance segmentation, many large-scale datasets were created but still there remains a need for more complex datasets along with different kinds of images. Existing still images datasets have many objects and often contain overlapping objects that are important. During the training of the models, these large-scale datasets with overlapping objects have played a vital role in being better at understanding the dense object along with their masks. Another complex challenge occurs with aerial images, a field that still needs to be explored. However, less research and fewer datasets have been available for instance segmentation. Aerial images contain several problems such as occlusion, scaling, segments in tiny objects, etc.

Enhancement of object representation Object recognition or segmentation has various taxing factors such as object size, background, blur, resolution, noise, etc. The following techniques can be used to handle these challenges:

- Resizing or scaling the objects using image pyramids, extracting features from different convolutional layers, up-sampling/down-sampling, etc.
- A spatial transformer network is used to handle the problem of occlusion, deformation, etc.

Challenges Using Transformers Techniques

High Computational Cost In the computer vision community, transformers have outperformed over CNNs, due to their high parametric complexity. This remarkable property helps to train large-sized models that, results in high

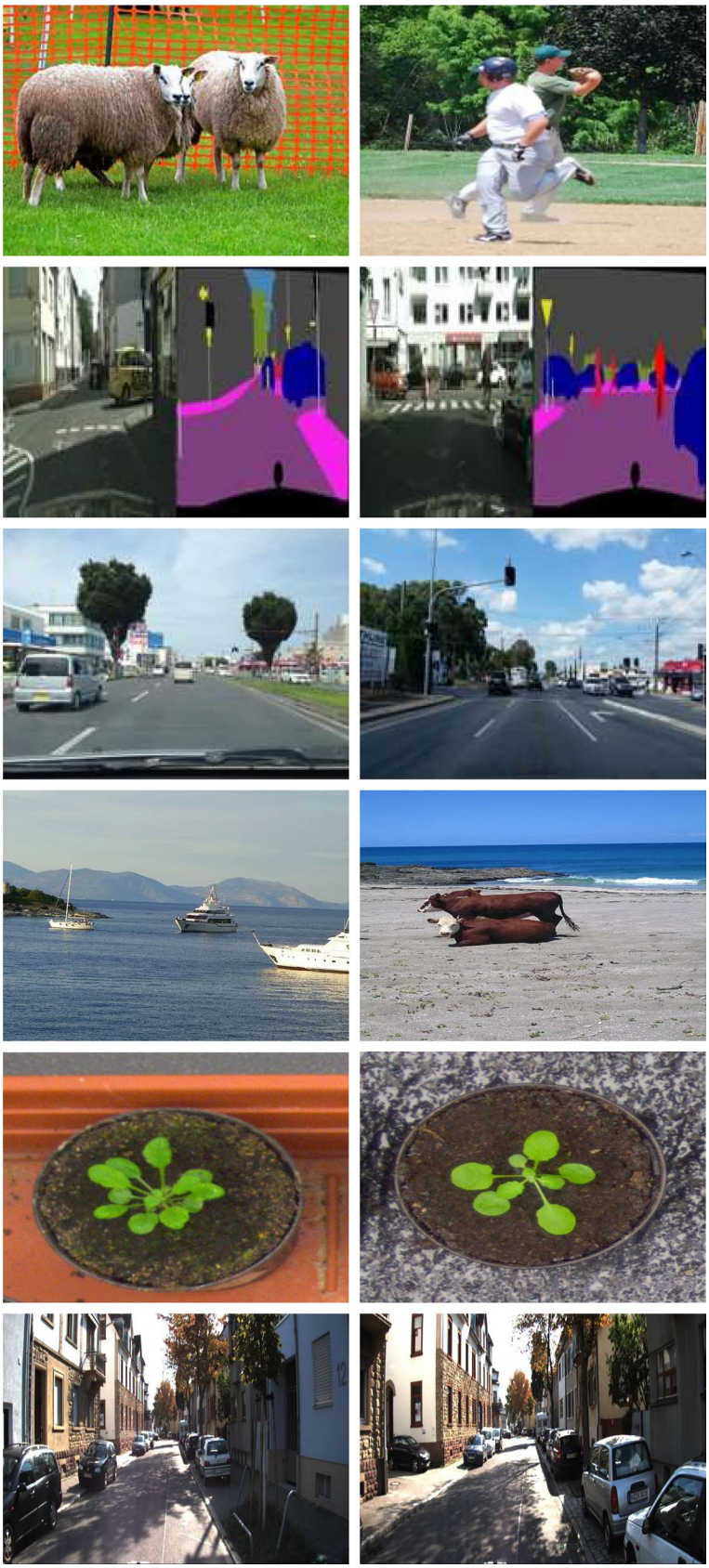

Fig. 8 (**a**) First row: Sample images from MS COCO, (**b**) Second row: Sample images from Cityscapes, (**c**) Third row: Sample images from Mapillary, (**d**) Fourth row: Sample images from Pascal VOC 2012, (**e**) Fifth row: Sample images from CVPPP, (**f**) Sixth row: Sample images from KITTI datasets respectively

training and inference cost. As a result, (a) Scaling up the computation, models and training datasets will boost the performance. (b) Having more training datasets will benefit the larger models. However, to this design, the transformers models are too expensive and time-consuming too.

Requirement of Large-scale data Transformers do not have any prior knowledge of dealing with visual data, so they require a large amount of training data to figure out the algorithm rules. For example, a CNNs architecture contains pre-processing tools such as weight sharing, fixed parameters, etc. performed by pooling operations and multi-scale blocks. However, Transformer pipeline needs to identify these image-specific rules on its own from the training dataset. As a result, this process needs longer training times, computational requirements increase significantly and large training datasets are beneficial for processing.

Scope of Future Work

Instance segmentation is still a challenging task using deep learning, reinforcement learning, and transformers. Deep learning techniques have been explored a lot, but there is still plenty of room for improvement in terms of speed v/s accuracy trade-offs and hardware vs. model complexity. The task of performing instance segmentation using deep learning is computationally expensive, memory demanding, and data greedy. Despite this, real-time instance segmentation also remains an open challenge in terms of speed v/s accuracy, which can be beneficial to many applications such as video surveillance, autonomous driving, traffic management, etc. Furthermore, Instance segmentation still remains challenging in the field, such as segmenting small objects, aerial object segmentation, real-time segmentation using drone technology, etc. Meanwhile, the annotation task for instance segmentation is also very monotonous, which requires not only pixel-level labeling such as semantic segmentation but also to differentiate individual instances of the same class, which is quite expensive.

The implementation of reinforcement learning for instance segmentation can solve more complex problems outside its domain. The RL techniques are not explored sufficiently for this task. Existing drawbacks can be overcome by using RL. A system can be trained with the minimum dataset and minimize the prediction error with the help of continuous updating feedback, i.e., reward function, but the implementation of reinforcement learning is a bit tricky. There remain ample scopes to explore this area with RL techniques. On the other hand, transformers have overcome many deep learning techniques in terms of accuracy v/s speed but lack training time. This can be improved using high-quality systems, i.e., GPUs and TPUs. In the near future, more different techniques can be applied to boost performance and accuracy. Recently, significant research has been carried out on vision transformers for classification and object detection. However, there is still not enough research on Vision Transformers for end-to-end instance segmentation. Future work should include various advanced architecture of transformers and evaluate their performance on instance segmentation.

Conclusion

In this survey paper, an overview of instance segmentation technology based on different techniques is given. We have surveyed more than 40 research papers based on deep learning techniques for instance segmentation which has been explained from 2016 to date. Also, we have discussed 2 research papers based on reinforcement learning techniques for instance segmentation from 2019 to date. The recent work has shown significant success using the transformers techniques that overcome the existing techniques in terms of accuracy and speed. We have mentioned the mean average precision results in a tabular form along with their accuracy. We have also discussed several commonly used datasets for instance segmentation technology. In this paper, their challenges and future scope have been considered. This survey paper will impart information about the state-of-the-art in the field of instance segmentation using deep learning, reinforcement learning, and transformers.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research did not receive any specific funding.

Availability of data and materials Publicly available datasets are used in this study.

Code availability Survey Paper.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Since no experiments are performed on humans or animals (dead or alive) in this research, therefore, Ethical approval is not required.

References

1. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
2. Zeiler M, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision, pp. 818–833. 2014
3. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
4. Albawi S, Mohammed T, Al-Zawi S. Understanding of a convolutional neural network. In: 2017 International conference on engineering and technology (ICET), pp. 1–6. 2017.
5. O'Shea K, Nash R. An introduction to convolutional neural networks (2015). arXiv preprint arXiv:1511.08458.
6. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing atari with deep reinforcement learning. 2013. arXiv preprint arXiv:1312.5602.
7. Mnih V, Badia A, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: International conference on machine learning, pp. 1928–1937. 2016.
8. 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, pp. 770–778. 2016.
9. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
10. Tang P, Wang X, Huang Z, Bai X, Liu W. Deep patch learning for weakly supervised object classification and discovery. Pattern Recogn. 2017;71:446–59.
11. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587. 2014.
12. Huang L, Yang Y, Deng Y, Yu Y. Densebox: unifying landmark localization with end to end object detection. 2015. arXiv preprint arXiv:1509.04874.
13. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg A . Ssd: Single shot multibox detector. In: European conference on computer vision, pp. 21–37. 2016.
14. 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, pp. 779–788. 2016.
15. Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. 2015. arXiv preprint arXiv:1506.01497.
16. Tang P, Wang C, Wang X, Liu W, Zeng W, Wang J. Object detection in videos by high quality object linking. IEEE Trans Pattern Anal Mach Intell. 2019;42(5):1272–8.
17. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M. Deep learning for generic object detection: a survey. Int J Comput Vis. 2020;128(2):261–318.
18. Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE international conference on computer vision, pp. 1134–1142. 2015.
19. Stewart R, Andriluka M, Ng A. End-to-end people detection in crowded scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2325–2333. 2016.
20. Dai J, Li Y, He K, Sun J. R-fcn: object detection via region-based fully convolutional networks. 2016. arXiv preprint arXiv:1605.06409.
21. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille A. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell. 2017;40(4):834–48.
22. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 603–612. 2019.
23. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440. 2015.
24. 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, pp. 2881–2890. 2017.
25. Dai J, He K, Li Y, Ren S, Sun J. Instance-sensitive fully convolutional networks. In: European conference on computer vision, pp. 534–549. 2016.
26. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440. 2015.
27. Pinheiro P, Lin TY, Collobert R, Dollár P. Learning to refine object segments. In: European conference on computer vision, pp. 75–91. 2016.
28. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969. 2017.
29. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448. 2015.
30. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16.
31. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C. Microsoft coco: common objects in context. In: European conference on computer vision, pp. 740–755. 2014.
32. Cordts M, Omran M, Ramos S, Scharwächter T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset. In: CVPR workshop on the future of datasets in vision. 2015.
33. Li Y, Qi H, Dai J, Ji X, Wei Y. Fully convolutional instance-aware semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2359–2367. 2017.
34. Hayder Z, He X, Salzmann M. Boundary-aware instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5696–5704. 2017.
35. Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3150–3158. 2016.
36. Everingham M, Winn J, The pascal visual object classes challenge,. (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning. Tech Rep. 2012;8:2011.
37. Fan R, Cheng MM, Hou Q, Mu TJ, Wang J, Hu SM. S4net: Single stage salient-instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6103–6112. 2019.
38. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference

on computer vision and pattern recognition, pp. 8759–8768. 2018.
39. Neuhold G, Ollmann T, Rota Bulo S, Kontschieder P. The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision, pp. 4990–4999. 2017.
40. Chen LC, Hermans A, Papandreou G, Schroff F, Wang P, Adam H. Masklab: Instance segmentation by refining object detection with semantic and direction features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4013–4022. 2018.
41. Uhrig J, Cordts M, Franke U, Brox T. Pixel-level encoding and depth layering for instance-level semantic labeling. In: German conference on pattern recognition, pp. 14–25. 2016.
42. Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 447–456. 2015.
43. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. 2017. arXiv preprint arXiv:1706.05587.
44. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y. Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp. 764–773. 2017.
45. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G. Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp. 1451–1460. 2018.
46. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G, Davis A, Dean J, Devin M, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. 2016. arXiv preprint arXiv:1603.04467.
47. Chen X, Girshick R, He K, Dollár P. Tensormask: a foundation for dense object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 2061–2069. 2019.
48. Lin TY., Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125. 2017.
49. Kuo W, Angelova A, Malik J, Lin TY. Shapemask: learning to segment novel objects by refining shape priors. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9207–9216. 2019.
50. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988. 2017.
51. Bolya D, Zhou C, Xiao F, Lee Y. Yolact: real-time instance segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9157–9166. 2019.
52. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell. 2015;38(1):142–58.
53. Chen K, Pang J, Wang J, Xiong Y, Li X, Sun S, Feng W, Liu Z, Shi J, Ouyang W, et al. Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4974–4983. 2019.
54. Huang Z, Huang L, Gong Y, Huang C, Wang X. Mask scoring r-cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6409–6418. 2019.
55. Chen H, Sun K, Tian Z, Shen C, Huang Y, Yan Y. BlendMask: top-down meets bottom-up for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8573–8581. 2020.
56. Zhang R, Tian Z, Shen C, You M, Yan Y. Mask encoding for single shot instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10226–10235. 2020.
57. Cheng T, Wang X, Huang L, Liu W. Boundary-preserving mask R-CNN. In: European conference on computer vision, pp. 660–676. 2020.
58. Bai M, Urtasun R. Deep watershed transform for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5221–5229. 2017.
59. Kirillov A, Levinkov E, Andres B, Savchynskyy B, Rother C. Instancecut: from edges to instances with multicut. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5008–5017. 2017.
60. Romera-Paredes B, Torr P. Recurrent instance segmentation. In: European conference on computer vision, pp. 312–329. 2016.
61. Li K, Hariharan B, Malik J. Iterative instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3659–3667. 2016.
62. Arnab A, Torr P. Pixelwise instance segmentation with a dynamically instantiated network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 441–450. 2017.
63. Ren M, Zemel R. End-to-end instance segmentation with recurrent attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6656–6664. 2017.
64. Liu S, Jia J, Fidler S, Urtasun R. Sgn: Sequential grouping networks for instance segmentation. In: Proceedings of the IEEE international conference on computer vision, pp. 3496–3504. 2017.
65. Liang X, Lin L, Wei Y, Shen X, Yang J, Yan S. Proposal-free network for instance-level object segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;40(12):2978–91.
66. Gao N, Shan Y, Wang Y, Zhao X, Yu Y, Yang M, Huang K. Ssap: single-shot instance segmentation with affinity pyramid. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 642–651. 2019.
67. 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, pp. 2881–2890. 2017.
68. Xie E, Sun P, Song X, Wang W, Liu X, Liang D, Shen C, Luo P. Polarmask: single shot instance segmentation with polar representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12193–12202. 2020.
69. Tian Z, Shen C, Chen H, He T. Fcos: fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9627–9636. 2019.
70. Liang J, Homayounfar N, Ma WC., Xiong Y, Hu R, Urtasun R. Polytransform: deep polygon transformer for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9131–9140. 2020.
71. Lee Y, Park J. Centermask: real-time anchor-free instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13906–13915. 2020.
72. Wang X, Kong T, Shen C, Jiang Y, Li L. Solo: segmenting objects by locations. In: European conference on computer vision, pp. 649–665. 2020.
73. Wang X, Zhang R, Kong T, Li L, Shen C. Solov2: dynamic, faster and stronger. 2020. arXiv preprint arXiv:2003.10152.
74. Fu CY., Shvets M, Berg A. RetinaMask: learning to predict masks improves state-of-the-art single-shot detection for free. 2019. arXiv preprint arXiv:1901.03353.
75. Zhou Y, Zhu Y, Ye Q, Qiu Q, Jiao J. Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3791–3800. 2018.

76. Liang X, Lin L, Wei Y, Shen X, Yang J, Yan S. Proposal-free network for instance-level object segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;40(12):2978–91.
77. Watanabe T, Wolf D. Distance to center of mass encoding for instance segmentation. In: 2018 21st International conference on intelligent transportation systems (ITSC), pp. 3825–3831. 2018.
78. Iglovikov V, Seferbekov S, Buslaev A, Shvets A. Ternausnetv2: fully convolutional network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 233–237. 2018.
79. Zhang SH., Li R, Dong X, Rosin P, Cai Z, Han X, Yang D, Huang H, Hu SM. Pose2seg: detection free human instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 889–898. 2019.
80. Yang TJ., Collins M, Zhu Y, Hwang JJ., Liu T, Zhang X, Sze V, Papandreou G, Chen LC. Deeperlab: single-shot image parser. 2019. arXiv preprint arXiv:1902.05093.
81. Lillicrap T, Hunt J, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. 2015. arXiv preprint arXiv:1509.02971.
82. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing atari with deep reinforcement learning. 2013. arXiv preprint arXiv:1312.5602.
83. Mnih V, Badia A, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: International conference on machine learning, pp. 1928–1937. 2016.
84. Araslanov N, Rothkopf C, Roth S. Actor-critic instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8237–8246. 2019.
85. Anh T, Nguyen-Tuan K, Jeong WK. Reinforced coloring for end-to-end instance segmentation. 2020. arXiv preprint arXiv:2005.07058.
86. Vicente S, Carreira J, Agapito L, Batista J. Reconstructing pascal voc. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 41–48. 2014.
87. Dobrescu A, Valerio Giuffrida M, Tsaftaris S. Leveraging multiple datasets for deep leaf counting. In: Proceedings of the IEEE international conference on computer vision workshops, pp. 2072–2079. 2017.
88. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the kitti dataset. Int J Robot Res. 2013;32(11):1231–7.
89. Papandreou G, Chen LC., Murphy K, Yuille A. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp. 1742–1750. 2015.
90. Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J. Semantic contours from inverse detectors. In: 2011 International conference on computer vision, pp. 991–998. 2011.
91. Van Etten A, Lindenbaum D, Bacastow T. Spacenet: a remote sensing dataset and challenge series. 2018. arXiv preprint arXiv:1807.01232.
92. 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, pp. 770–778. 2016.
93. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500. 2017.
94. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:1409.1556.
95. Cheng MM, Mitra N, Huang X, Torr P, Hu SM. Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell. 2014;37(3):569–82.
96. Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S. Salient object detection: a discriminative regional feature integration approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2083–2090. 2013.
97. Zhu W, Liang S, Wei Y, Sun J. Saliency optimization from robust background detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2814–2821. 2014.
98. Rother C, Kolmogorov V, Blake A. GrabCut interactive foreground extraction using iterated graph cuts. ACM Trans Graphics (TOG). 2004;23(3):309–14.
99. Hou Q, Cheng MM., Hu X, Borji A, Tu Z, Torr P. Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3203–3212. 2017.
100. Li G, Yu Y. Deep contrast learning for salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 478–487. 2016.
101. Wang L, Lu H, Ruan X, Yang MH.. Deep networks for saliency detection via local estimation and global search. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3183–3192. 2015.
102. Dai J, He K, Sun J. Convolutional feature masking for joint object and stuff segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3992–4000. 2015.
103. Hariharan B, Arbeláez P, Girshick R, Malik J. Simultaneous detection and segmentation. In: European conference on computer vision, pp. 297–312. 2014.
104. Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 447–456. 2015.
105. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587. 2014.
106. Dai J, Li Y, He K, Sun J. R-fcn: object detection via region-based fully convolutional networks. 2016. arXiv preprint arXiv:1605.06409.
107. Xie E, Wang W, Ding M, Zhang R, Luo P. PolarMask++: enhanced polar representation for single-shot instance segmentation and beyond. IEEE Trans Pattern Anal Mach Intell. 2021.
108. Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V, Lu S, et al. ICDAR 2015 competition on robust reading. In: 2015 13th International conference on document analysis and recognition (ICDAR), pp. 1156–1160. 2015.
109. B. A. Hamilton Kaggle. 2018 data science bowl: find the nuclei in divergent images to advance medical discovery. Kaggle. 2018.
110. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25.
111. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:1409.1556.
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, pp. 770–778. 2016.
113. 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, pp. 779–788. 2016.
114. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY., Berg A. Ssd: single shot multibox detector. In: European conference on computer vision, pp. 21–37. 2016.

115. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, et al. A survey on visual transformer. arXiv e-prints, arXiv-2012. 2020.
116. Khan S, Naseer M, Hayat M, Zamir S, Khan F, Shah M. Transformers in vision: a survey. In: ACM computing surveys (CSUR). 2021.
117. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020. arXiv preprint arXiv:2010.11929.
118. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers distillation through attention. In: International conference on machine learning, pp. 10347–10357. 2021.
119. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European conference on computer vision, pp. 213–229. 2020.
120. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: deformable transformers for end-to-end object detection. 2020. arXiv preprint arXiv:2010.04159.
121. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr P, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6881–6890. 2021.
122. Doersch C, Gupta A, Zisserman A. Crosstransformers: spatially-aware few-shot transfer. Adv Neural Inf Process Syst. 2020;33:21981–93.
123. Kumar M, Weissenborn D, Kalchbrenner N. Colorization transformer. 2021. arXiv preprint arXiv:2102.04432.
124. Hu J, Cao L, Lu Y, Zhang S, Wang Y, Li K, Huang F, Shao L, Ji R. Istr: end-to-end instance segmentation with transformers. 2021. arXiv preprint arXiv:2105.00637.
125. Liu Y, Yang S, Li B, Zhou W, Xu J, Li H, Lu Y. Affinity derivation and graph merge for instance segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 686–703. 2018.
126. Neven D, Brabandere B, Proesmans M, Gool L. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp. 8837–8845. 2019.
127. Guo R, Niu D, Qu L, Li Z. Sotr: segmenting objects with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7157–7166. 2021.
128. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
129. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint arXiv:1810.04805.
130. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
131. Chen M, Radford A, Child R, Wu J, Jun H, Luan D, Sutskever I. Generative pretraining from pixels. In: International conference on machine learning, pp. 1691–1703. 2020.
132. Yu X, Shi D, Wei X, Ren Y, Ye T, Tan W. SOIT: segmenting objects with instance-aware transformers. 2021. arXiv preprint arXiv:2112.11037.
133. Hafiz A, Bhat G. A survey on instance segmentation: state of the art. Int J Multimedia Inf Retriev. 2020;9(3):171–89.
134. Tian D, Han Y, Wang B, Guan T, Gu H, Wei W. Review of object instance segmentation based on deep learning. J Electron Imaging. 2021;31(4): 041205.
135. Chen W, Du X, Yang F, Beyer L, Zhai X, Lin TY, Chen H, Li J, Song X, Wang Z, et al. A simple single-scale vision transformer for object localization and instance segmentation. 2021. arXiv preprint arXiv:2112.09747.
136. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai, X., Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020. arXiv preprint arXiv:2010.11929.