

Image Segmentation Techniques in Computer Vision

Goldi Soni

Assistant Professor

Amity University

Raipur, Chhattisgarh

gsoni@rpr.amity.edu

Rahul Sharma

BTech CSE

Amity University

Raipur, Chhattisgarh

Rahul.sharma21@s.amity.edu

Gaurav Sahu

BTech CSE

Amity University

Raipur, Chhattisgarh

gaurav.sahu@amity.edu

Abstract

This review paper provides an in-depth study of the image segmentation algorithms in computer vision, focusing on academic literature from 2020 to 2024. Specifically, this paper aims to analyze the development of image segmentation algorithms, starting from transformer networks and self-supervised learning models to vision-language models and foundation models. A total of 30 articles are selected for systematic literature review, which are highly cited papers in top-tier conferences like CVPR, NeurIPS, ICCV, and ECCV. Based on the results of this study, the years from 2020 to 2024 can be considered a groundbreaking period in the history of image segmentation, when the new era of foundation models has started with the invention of Segment Anything Model (SAM) and its variants, which have enabled promptable and zero-shot image segmentation for the first time. Contemporary algorithms exhibit significantly better performance than their predecessors based on CNNs in terms of accuracy and generalizability, although problems with high computational cost, real-time application, and domain adaptation have yet to be solved.

Keywords

Image Segmentation, Semantic Segmentation, Instance Segmentation, Panoptic Segmentation, Transformer-Based Segmentation, Foundation Models, Segment Anything Model, Self-Supervised Learning, Vision-Language Models, Open-Vocabulary Segmentation, Deep Learning.

1. Introduction

Segmentation of images is one of the basic concepts in the field of computer vision, where a picture is split into regions in order to facilitate further analysis of it. In other words, segmentation allows splitting up a complex image into several simple objects for easier processing. Segmentation serves an essential part in many practical uses of computers, such as self-driving cars (detecting the road, pedestrians, cars, etc.), medical applications (finding out tumors, identifying organs and their locations, and many more), augmented reality applications (analyzing scenes), robotics (navigation and manipulation with objects, etc.), as well as computer graphics (image rendering).

However, the progress in this area is considerable. Previously, traditional approaches were used to perform segmentation, such as thresholding, clustering, and graphical algorithms that were applicable only to clear and easily definable images. The revolution came when CNN appeared; FCN and U-Net networks enabled automatic

feature extraction and higher quality. Currently, transformer models such as Vision Transformer (ViT) and Swin Transformer make segmentation even more accurate due to capturing of global context information. Moreover, recently introduced foundation models such as Segment Anything (SAM) can be applied to any task related to flexible prompt-based segmentation.

Unfortunately, although much research has been done in this area, there are some issues that need to be addressed. Specifically, previous survey papers mainly discuss research until 2021, while the field has moved forward and now focuses on new directions and approaches such as transformer-based algorithms, self-supervised learning, vision-language integration, and foundation models. Hence, the aim of the present survey paper is to discuss 30 papers written from 2020 to 2024. They will be reviewed in detail regarding their approaches, pros and cons, and practical applications.

2. Objective

The main goal of this research is to conduct a thorough review of image segmentation algorithms used in computer vision since 2020 until 2024, paying particular attention to recent achievements that have not been covered in previous surveys. The work focuses on discussing how the trend of replacing the classical CNN-based solutions by the newer transformer-based algorithms and foundation models emerged. Moreover, it highlights emerging trends in image segmentation, such as open-vocabulary segmentation, self-supervised image segmentation, vision-language models, and diffusion models for image segmentation. In this regard, selected methods are compared with each other based on methodology, advantages, disadvantages, and usability. Finally, this research outlines the major challenges facing researchers in the field of image segmentation and provides possible directions for future studies.

3. Literature Review

This essay analyses 30 articles related to image segmentation using computer graphics. The methods applied in the articles are deep learning, transformers, graph theory, and clustering. The majority of the reviewed articles highlight that contemporary methods offer improved results in terms of accuracy and features extracted compared to classical approaches, but they are resource intensive and need huge amounts of data. Thus, the literature reveals some improvement in segmentation methods but also faces problems of efficiency and real-time processing.

Earlier transformer segmentation networks were based on the use of positional encoding, thereby limiting their capacity to work well across varying image resolution; SegFormer solves this issue by designing an effective encoder. There are several types of architectures ranging from B0 to B5, thus offering a trade-off between accuracy and efficiency in computation. From experiments, it is seen that the biggest model performs about 51.8 mIoU on the ADE20K dataset, and at the same time, smaller architectures such as B0 can be used when computing resources are limited. On ADE20K and Cityscapes benchmarks, the network shows good performance and surpasses earlier methods in terms of accuracy despite consuming fewer resources. It has been shown that its hierarchical encoder can detect both small details in the image along with the entire context. However, its light-weighted MLP decoder cannot deliver satisfactory results in tasks needing exact boundaries, like medical imaging.[1]

Swin Transformer won Best Paper Award at ICCV 2021; this work showed that with shifting operation, window-based attention can perform better than global self-attention in segmentation tasks, saving significant computational costs. The main difference from other transformers is the introduction of shifted window-based self-attention where only the computation of the local windows is considered, while the information transfer between the windows is

done using shifts. It results in lower computational complexity than global self-attention used in the initial transformers' architectures. Hierarchical structure provides the creation of multi-scale features and ensures high compatibility with the dense prediction approaches used in image segmentation and object detection tasks. That led to the popularity of the backbone model in various state-of-the-art vision models. However, the problem with such partitioning method is that the window size is fixed, thus constraining the model to recognize objects of different scales .[2]

The first-ever model used for segmentation purposes that was trained using over a billion masks was presented, and as a result, zero-shot segmentation is now available in almost all kinds of scenarios without the need for task-specific fine-tuning. The model utilizes an innovative approach in which promptable segmentation is done through prompting in the form of points, bounding boxes, or even texts. Through this method of interaction, the SAM can generalize well in various domains without the need for fine-tuning specific to each scenario. From an architectural perspective, SAM makes use of an image encoder, prompt encoder, and a mask decoder that is relatively light and allows for the production of segmented outputs efficiently. SAM is a huge milestone in the history of segmentation, as it shows promising results in zero-shot settings while being a reference point for several other works. However, SAM fails in producing precise results in cases where there are overlapping objects and dense scenes..[3]

Before the advent of Mask2Former, each of these segmentation types needed its own special architecture, leading to a lot of redundancy in both the training and deployment process.. The network proposes a masked attention strategy in which cross-attention is only performed in the regions covered by masks that were predicted. This approach makes the segmentation process more effective since feature extraction focuses on important parts, eliminating computation in other regions. As opposed to previous systems that used dedicated networks for different kinds of segmentations, Mask2Former uses an architecture that allows for generalization of segmentation tasks. Experiments show that the proposed network performs very well for all three segmentation tasks, even surpassing dedicated networks for individual segmentations when tested on different benchmark datasets. However, the masked attention method needs good initial mask predictions to work effectively; otherwise, the system may fail to converge properly..[4]

The excessive dependency on labeled data has always been the primary issue facing training of vision transformers for segmentation. BEiT solves this problem by taking the masked token prediction approach of BERT and applying it to image patches. By tokenizing image patches through a pretrained tokenizer, BEiT can predict masked tokens through self-supervised learning. In this way, the model learns semantic representations without much labeled data. These pre-trained representations are further fine-tuned for various applications such as image segmentation and object detection. Experiments prove that BEiT is effective in improving segmentation results, especially under conditions of insufficient labeled data sets. Nevertheless, the use of pretrained image tokenizer by BEiT adds another layer of complexity since the tokenizer may introduce some errors or inconsistencies which could have adverse effects on the downstream segmentation task .[5]

With no labeling used in training, DINO learns attention maps that spontaneously conform to object boundaries, showing the hidden capacity of self-supervised learning for segmentation tasks. It trains a student network by matching the results of a teacher network based on two distinct views of the same image to learn efficient visual representation. One of the findings from this study is that attention maps obtained spontaneously conform to object boundaries, making it an effective method for semantic segmentation without supervision. Such spontaneous behavior reveals the possibility of self-supervised learning in minimizing dependence on labeled data sets. The

study proves that vision transformers can learn meaningful structural information from images. Therefore, DINO has motivated other studies in unsupervised learning and representation learning in segmentation tasks. However, despite attention maps spontaneously conforming to object boundaries in DINO, the challenge is how to generate precise segmentation masks without supervision. It becomes difficult when dealing with small or occluded objects .[6]

Whereas methods like BEiT employ complicated discrete tokenizers in their training procedures, SimMIM suggests the simplicity of learning visual representation via predicting the raw pixel value of the masked patches during training. In contrast to previous works such as BEiT, which used discrete token prediction based on a tokenizer, SimMIM simply learns to predict raw pixel values of the masked patches of the images in question. The design greatly simplifies the process of pre-training while preserving a very robust ability to learn visual representation. From the experimental perspective, the performance of the method shows promising results on a variety of benchmarks. The proposed idea of reconstructing raw pixels as a training objective simplifies the whole pre-training procedure by eliminating the need for tokenizers. Nevertheless, learning from the raw pixels might be biased towards learning low-level texture-related features as opposed to high-level abstract features .[7]

As the bridge between convolutional neural networks and transformers architecture, PVT v2 utilizes the benefits of linear spatial reduction attention and overlapping patches to create an effective multi-scale representation of features for dense prediction. Linear spatial reduction attention is included in the model due to the substantial decrease in computation required without losing relevant spatial data. Overlapping patches are used to increase consistency in features along with capturing more local details. Thus, PVT v2 successfully integrates advantages from both architectures. It provides multi-scale feature representations due to the hierarchical structure of the network, which makes it highly suitable for dense prediction tasks like object detection and image segmentation. The experiments prove high performance on multiple benchmark tests, which led to the extensive use of this network as a backbone of dense prediction models. Nevertheless, linear spatial reduction attention used by PVT v2 aggressively compresses spatial tokens, causing positional information loss .[8]

In an extension of the Segment Anything paradigm to handle video rather than images, SAM 2 adopts a streaming memory mechanism allowing the model to have consistent segmentation of objects over sequential video frames. The model adopts a streaming memory structure that enables it to segment and track objects over sequential video frames. The SAM 2 model has demonstrated its capabilities to have a high degree of zero-shot generalization in video segmentation tasks without any requirement for task-specific training. The ability of SAM 2 to offer real-time segmentation of objects and tracking makes it a powerful tool towards achieving universal image segmentation of both still images and videos. This extends the foundation model concept to include improved scalability in real-world applications. Nevertheless, SAM 2 has a problem of error accumulation due to its streaming memory capability in accumulating temporal information from sequences of videos .[9]

The difficulty in designing an effective pipeline for training and fine-tuning separate models for semantic, instance, and panoptic segmentation tasks has been known for a long time; OneFormer addresses this problem by using a unified framework with task-conditioned joint training inside a single transformer. Contrary to the common practice when separate models are needed to tackle each task independently, OneFormer introduces the idea of task-conditioned joint training, which means that a model changes its performance according to the selected segmentation task. OneFormer makes use of the transformer backbone together with task-dependent conditioning and achieves good performance in capturing local/global context of images. The experiments show that OneFormer

significantly outperforms previous methods designed specifically for certain segmentation tasks and universal segmentation models on several benchmarks. However, OneFormer requires task tokens in order to switch from one segmentation task to another and therefore cannot work in a fully autonomous manner .[10]

Contrary to the conventional wisdom that transformer models perform better than CNNs when it comes to vision applications, InternImage proves that an efficiently crafted CNN with deformable convolutions can provide performance comparable to that of transformers in dense prediction applications. While transformer models make heavy use of self-attentions, this study shows that properly engineered CNNs can be equally competitive and even more powerful than transformers when performing dense predictions, like segmentations of input images. Deformable convolutions enable the model to adaptively change the shapes of its receptive fields to accommodate varying geometries and complex objects. The experiments reveal that InternImage outperforms existing approaches on several benchmark datasets for dense prediction applications while retaining good computational efficiency. Nevertheless, deformable convolution operations have irregular memory access patterns, which makes them challenging to optimize on the standard GPU platforms and thus slow down the inference compared to CNNs of the same sizes .[11]

By methodically adopting transformer architecture techniques to an existing ResNet framework without adding any attention modules, ConvNeXt proves that the architectural optimization is far more important than the basic selection between convolution and transformers. Rather than adding attention modules, the architects in question systematically optimize architectural elements such as normalization modules, kernel dimensions, and training procedures. In turn, their model confirms the ability of an optimized pure CNN to equal, if not surpass, the efficiency of some popular transformer-based architectures, like the Swin Transformer, in certain applications, including image segmentation and classification tasks. ConvNeXt yields powerful multi-level feature maps, which makes it an ideal candidate for dense prediction. Despite the fact that architectural design is proven to matter more than the type of neural network architecture chosen, it should be noted that the model does not have any means to model the dependencies between remote image patches. It can lead to lower effectiveness in segmentation applications where global context plays an important role, for example, in scene parsing..[12]

Plain vision transformers do not include any spatial hierarchy and local features processing since they were not intended for dense prediction tasks; thus, there is a need to integrate additional modules to make ViT capable of handling such tasks, i.e., segmentation tasks. The main idea behind ViT-Adapter lies in integration of spatial prior modules to the existing model without changing the original architecture of ViT. In particular, these adapter modules are capable of introducing locality and multi-scale feature processing. By combining the benefits of transformer models and CNNs, the introduced adapters allow a ViT model to generate hierarchical features that can be used similarly to the CNN segmentation pipeline features. As shown via experiments, the approach provides substantial improvements in the results of segmentation tasks compared to plain ViTs models. Another advantage of using the method under analysis is that no new ViT parameters have to be trained, hence the proposed technique can be viewed as efficient in terms of computational costs. On the other hand, the design of adapter modules depends on the base ViT model used. .[13]

Mask detection and segmentation have always been considered two different goals needing two different methods of training — MaskDINO combines them into one by adding mask prediction capability to the existing DINO detection paradigm. The architecture is based on the existing DINO detection paradigm but adds new capabilities of predicting masks to perform instance and panoptic segmentation. An important achievement of the paper is the effective application of the supervision from detection to improve the quality of predicted masks, revealing the high

potential of cooperation between the detection and segmentation goals. Using shared representation, MaskDINO achieves outstanding results in the task of both localization and segmentation. Experimental results reveal that MaskDINO reaches state-of-the-art results in both instance and panoptic segmentation on several benchmark datasets. The unified approach saves space compared to the two different approaches dedicated to each task separately. However, the joint training of detection and segmentation in MaskDINO may cause gradient conflicts when objects' boundaries are not aligned for both tasks. .[14]

The inability to segment categories of objects that were not present during training has been an inherent drawback in closed vocabulary models. CAT-Seg solves this problem by aggregating costs using CLIP-based image-text similarities. CAT-Seg utilizes existing vision-language pretraining techniques by relying on CLIP-based image-text embedding similarities to link input images and their corresponding query descriptions. One of the major contributions of this paper involves applying cost aggregation to calculate and refine similarities between image segments and text embeddings to accurately produce object segmentation masks. This process enables the model to generalize well to new categories that were not trained during the learning phase without requiring additional training. Experimental evaluations show promising results on open vocabulary datasets, emphasizing the practicality of the algorithm when dealing with applications that require more general object segmentation. Unfortunately, the cost aggregation scheme used by CAT-Seg scales quadratically with the number of candidate categories. .[15]

Pixel-level annotations for thousands of object categories are impractical; OpenSeg solves this problem through a combination of training on image-level labels and vision-language alignment to provide open-vocabulary segmentation at a large scale. Rather than having dense pixel-level annotations for all categories, OpenSeg uses image-level labels for training, greatly decreasing the number of annotations needed. Using vision-language alignment, OpenSeg generalizes to other categories by pairing visual cues with language models, thereby extending its segmentation skills to many categories not encountered during training. This scalable framework proves that large datasets along with vision-language pretraining can be used to provide segmentation skills for thousands of object categories. Experiments conducted reveal the model's capability in providing impressive open-vocabulary segmentation performance on benchmarks with weak supervision. Although this method has proven its capability in segmenting many categories, label ambiguity becomes a major issue because of the use of image-level labels as opposed to pixel-level annotations. .[16]

Even though SAM is very good at generating masks and Grounding DINO is excellent at language-guided detection, none of them can independently perform text-based segmentation of generic objects; therefore, Grounded SAM integrates them to create an adaptable segmentation pipeline in the open world. In this framework, Grounding DINO is used to detect objects according to their description in natural language while SAM generates precise segmentation masks for each detected object region. With such a strategy, Grounded SAM achieves zero-shot segmentation of arbitrary objects with no need for specialized training for different tasks. Grounded SAM shows outstanding results in many real-life situations when there are multiple objects in a scene. It is a considerable achievement since Grounded SAM can unite two types of perception, namely vision and language, into one pipeline, which will be a breakthrough in creating perception systems in the open world. But, at the same time, this sequential procedure implies that any error in the detection process with Grounding DINO will inevitably affect SAM segmentation, which has no recovery mechanisms. .[17]

By generating visual properties that perform efficiently in segmentation, depth prediction, and classification tasks without performing task-specific finetuning, DINOv2 proves to be one of the most effective general-purpose visual

encoders designed using self-supervised training. DINOv2 is trained on image datasets that have been curated very precisely but without any form of manual annotation and learns efficient visual feature representation using self-distillation technique. These generated features perform efficiently in downstream tasks like segmentation, depth prediction, and image classification without being finetuned for task specificity. DINOv2 proves to be a very powerful foundation model due to its effectiveness in vision tasks in a generalized manner. It is among the best universal visual encoders designed by now. Nonetheless, the features produced by DINOv2 are optimized for general-purpose feature representation and not task-specific segmentation, which leads to inefficient performance in highly specialized domains like satellite imagery and histopathology tasks without domain-specific finetuning. [18]

Instead of developing output heads individually for each vision task, Painter reframes the tasks of segmentation, detection, and depth estimation as an image inpainting problem inspired by the in-context learning approach of language models. Unlike conventional approaches where task-specific architectures or output heads are used, Painter employs the idea of in-context learning, wherein it is offered a few examples of inputs and outputs and based on this context, it learns new tasks. Similar to large language models, it transfers knowledge through context, not by re-training the model. Through this paper, Painter shows how a generalist model is able to do various vision tasks without having any architecture changes or task-specific fine-tuning. However, despite its high potential in unifying architectures for computer vision, Painter's in-context learning methodology is extremely dependent on the chosen context examples – poor choices may lead to a wrong learning direction for the model, and there is no existing way to choose the optimal context examples automatically. [19]

Given the capability of large language models to learn new tasks based on a handful of samples, SegGPT proposes an in-context learning strategy for image segmentation by offering example input-output pairs, after which the model learns how to perform similar segmentation for new images based on those pairs. Motivated by the ability of large language models, the strategy lets users specify a few input-output example pairs in order to teach the model what kind of segmentation is needed. Then, using the learned in-context knowledge, the model can do similar segmentation for other images. Such a strategy allows for a general and few-shot learning method for image segmentation across various applications, avoiding further training or fine-tuning. SegGPT achieves excellent results in several image segmentation problems, indicating the possibility of context learning in image processing. Moreover, its uniform strategy simplifies the segmentation process. Nevertheless, SegGPT fails in cases where the target image does not look like or share styles with the input examples used for learning. [20]

Using SAM for video application needs consistent object tracking throughout the frames. It is made possible through the integration of SAM's mask generation and DeAOT tracking technique, where both processes start by receiving prompts from simple user inputs such as clicking the target objects in the first frame. The algorithm integrates SAM and DeAOT tracking method to allow users to easily generate segmentation through simple user clicks to the target object within the first frame of the video. The model will continue segmenting the tracked objects in the following video frames. However, since the algorithm depends on user-provided initial input for segmentation prompts, the accuracy of object identification and tracking may be affected if the user is unable to give accurate prompts for segmentation. [21]

Most segmentation algorithms support one kind of prompt inputs; SEEM goes against the norm by processing text, visual, and spatial prompts altogether through a single framework for multiple types of interactions. The goal is to make a bridge between various segmentation frameworks like referring image segmentation, semantic

segmentation, and interactive segmentation using a single model trained on diverse types of prompts. The use of multiple prompt types in training allows the algorithm to become more versatile in performing different tasks based on different inputs. It becomes evident from the results obtained in SEEM that incorporating diversity in prompt types makes the system more robust in real-life applications. Nevertheless, SEEM has a critical requirement of balanced data sets while training for all kinds of supported prompts; an imbalance in training data may make the model perform better with one type of prompt and poorly with another. [22]

The usage of SAM in latency-sensitive environments was not feasible because of its computational burden; EfficientSAM resolves this problem by applying masked image pretraining and knowledge distillation, resulting in a lightweight yet similar in quality architecture. In essence, the main idea is to keep all of the benefits of SAM's segmentation and optimize it in terms of computational power and memory. Through knowledge distillation from the original SAM model, it is possible to make the segmentation performance equal to the one of SAM but with noticeably fewer computational resources required. It allows using this model in edge devices where it is necessary to reduce computational needs. Despite that, due to knowledge distillation, EfficientSAM cannot go beyond the original SAM performance; therefore, it cannot provide better results in terms of segmentation than those provided by the original SAM. Thus, if there are any systematic errors in SAM, which, for example, do not allow segmenting transparent or reflective objects, these limitations will be carried over to EfficientSAM. [23]

It was not possible to run foundation segmentation models on edge devices like smartphones and embedded systems due to the size and speed of the encoders – MobileSAM addresses the challenge by replacing the huge encoder used in SAM with a lightweight distilled encoder offering more than $60\times$ times speedup. In doing so, the computational complexity is significantly reduced without losing segmentation quality. Results show that the MobileSAM offers more than $60\times$ times speedup compared to the initial SAM encoder, with very little reduction in segmentation accuracy. This makes the model highly feasible for implementation on edge devices due to speed and efficiency requirements. With the use of distilled encoder, MobileSAM provides practical interaction capability in real-world segmentation tasks. However, replacing the heavy image encoder in SAM with the lightweight encoder through the distillation process may result in the representation bottleneck where the encoder may be unable to represent structural details required for segmenting small objects. [24]

In the original DETR architecture, slow convergence and handling high-resolution feature maps were the key challenges; Deformable DETR addresses both problems by implementing sparse deformable attention over a small number of important sampling points rather than global self-attention. In this new approach, self-attention is computed not globally but only with respect to those spatial locations that are deemed interesting by the model, thus drastically lowering the computational cost of the process. This makes it possible to train faster and scales better for dense prediction models, like image segmentation, which require more detailed attention to finer spatial details. The technique is very important because without the use of deformable attention, the transformers cannot be successfully adapted for use in segmentation tasks. On the downside, the effectiveness of deformable attention is critically dependent on sampling offset initialization, since incorrect initialization at the initial stage of training may cause the attention to collapse on non-informative locations. [25]

The computational complexity of global self-attention is quadratic based on image resolution, rendering it unusable for high resolution segmentation – this problem is solved with MaxViT by decomposing the global self-attention operation into local window-based attention operations and dilated attention operations, resulting in linear complexity without compromising context range. This makes MaxViT capable of capturing both long-range

dependencies as well as local ones, using only linear complexity with regard to input image resolution. The alternating use of these two kinds of attention layers allows MaxViT to capture not just the fine-grained features but also the context information, which makes MaxViT very applicable to dense prediction problems like image segmentation. In fact, the model has been proven through experimental testing to perform very well at both segmentation and classification benchmarks, making MaxViT a strong competitor among vision backbones. The scalability of MaxViT enables it to better handle high resolution images compared to traditional global attention-based models, though the design of alternately applying local and dilated attention relies heavily on the assumption that the objects are well-modeled through a fixed grid pattern. [26]

Generative text-to-image diffusion models have the ability to learn extremely rich semantic representations during their training process; ODISE takes advantage of these learned features in the diffusion models for discriminative tasks like open-vocabulary panoptic segmentation. Instead of training the segmentation networks using pixel-level supervision, ODISE uses features derived from the diffusion models, like the Stable Diffusion model, which is trained on large-scale datasets of image and text pairs. By learning to generate images from rich descriptions, the diffusion models have rich semantic and context representation capabilities that enable the successful segmentation of objects belonging to known classes as well as unknown classes. This work shows that features learned in a generative setting do not need segmentation labels to be good at segmentation tasks. However, since ODISE relies on the frozen feature extraction of the pre-trained diffusion model, it cannot specialize its visual representations during the segmentation task, which might hinder its effectiveness in scenarios where high precision is required in the segmentation maps beyond the capabilities of the diffusion model. [27]

Diffusion models trained on billions of image-text pairs learn multi-scale feature representations much more complex than any classifier-trained network, and VPD uses these representations for semantic segmentation and depth estimation without altering the network architecture. It utilizes the complex multi-scale representations learned by diffusion model decoders, trained on massive image-text datasets, and shows that diffusion model features are extremely transferable and can be used in other downstream tasks. In VPD, by adapting existing diffusion models, it is demonstrated how generative models can be successfully employed in dense prediction problems without task-specific architecture alteration. The proposed method is validated using several perception benchmark tests, demonstrating promising performance. The paper proposes an interesting and innovative solution for bridging the gap between generative models and conventional computer vision problems. On the downside, features learned by diffusion models are stochastic because they are learned by training diffusion models to generate samples by adding noise and removing it step-by-step, which makes them non-consistent across multiple runs. [28]

While it is true that CLIP shows amazing performance in vision-language alignment, the self-attention mechanism employed by it creates problems when it comes to localization accuracy for tasks involving dense prediction, and this is where CLIP Surgery finds its niche. This paper analyzes and resolves inconsistencies that arise in the self-attention mechanism used in CLIP models, restricting their efficacy in fine-grained understanding of visuals. Modifications to this mechanism coupled with additional residual connections help to better align CLIP's visual feature maps with semantic information, allowing more precise localization and segmentation of objects utilizing features from CLIP, without the need for extra segment-specific training. Experimental results show substantial improvement in open-vocabulary segmentations, emphasizing the possibility of optimizing vision-language models for specific tasks after pre-training them. While this methodology focuses on increasing the efficacy of already pre-trained models rather than training them anew, the improvements provided by CLIP Surgery have limitations due

to limitations of CLIP architecture itself. Specifically, CLIP models were not trained using any spatial supervision but image-text pairs, imposing limits on localization accuracy irrespective of attention-based optimizations. [29]

However, despite the capabilities of multimodal large language models to describe images at a coarse level, the ability for fine-grained interaction with pixels using the language instruction is yet to be explored – Osprey addresses the problem by including region-level mask information along with natural language instructions. The proposed architecture advances the existing vision-language models by introducing the ability to incorporate mask information as input, leading to a fine-grained interaction between visual and textual inputs. It enables applications such as referring segmentation, region-based description, and visual question answering using pixel-level alignment with natural language instructions. By showing promising results when using multimodal large language models, this work opens new possibilities for building unified AI models that incorporate fine-grained understanding of the scene in addition to higher-level reasoning tasks. However, Osprey relies on the inclusion of region-level mask information as part of the language grounding pipeline. Therefore, having an accurate mask beforehand is required to leverage the potential of the model; otherwise, the end-to-end performance would depend heavily on the accuracy of the masks provided from upstream models. [30]

Based on the review of the literature presented above, there is an obvious trend away from classical image processing algorithms toward deep learning and transformer algorithms. The main problem of classical methods was their heavy reliance on feature extraction, which made them less effective in complicated cases. On the other hand, deep learning algorithms have facilitated the efficiency of image segmentation due to the automatic extraction of features.

Nevertheless, a number of issues are yet to be solved. Contemporary algorithms are based on massive data and high computing power, which makes them inefficient when dealing with urgent cases. Moreover, some algorithms lack sufficient generalization capabilities across different fields. Although transformer algorithms have overcome the problem of global context comprehension, they require higher computing power. Therefore, efficient image segmentation algorithms should be developed.

4. Comparison of Past Published Works

The following table highlights the trend followed by the image segmentation algorithm during the period of 2020 to 2024. The image segmentation algorithms were developed from CNN and graph approaches to sophisticated transformer and self-supervised learning models from 2020 to 2024.

Table1:- Year-wise Comparison (2020 to 2024) of Segmentation Techniques

S. No	Title of Paper	Year of Publication	Proposed System / Objective	Methodology	Contribution / Results
1	SegFormer: Simple and Efficient Design for Semantic Segmentation	2021	Efficient transformer segmentation without positional encoding	MiT encoder + MLP decoder	High accuracy with low compute
2	Swin Transformer: Hierarchical Vision Transformer	2021	Efficient high-resolution transformer	Shifted window attention	Widely used backbone
3	Segment Anything (SAM)	2023	General-purpose segmentation model	Prompt-based foundation model	Strong zero-shot performance
4	Mask2Former	2022	Unified segmentation framework	Masked attention transformer	State-of-the-art multi-task results
5	Osprey: Pixel Understanding with Visual Instruction Tuning	2024	Multimodal segmentation + reasoning	MLLM with visual instruction tuning	Pixel-level reasoning and QA

5. Conclusion

In this paper, an extensive review was conducted regarding 30 important research works in image segmentation that were published between 2020 and 2024. The work includes studies related to transformer-based architectures, self-supervised approaches, vision-language models, and foundation models, thereby highlighting some of the most important trends in the field within the selected timeframe. The discussion makes it evident that there is a shift from task-specific CNN-based models towards more general transformer architectures and foundation models.

It should be noted that the development of the Segment Anything Model is considered as a milestone due to the ability of this model to perform promptable and zero-shot segmentations in various applications. However, despite the advances made, one of the key challenges in the area is achieving a balance between segmentation precision and computation efficiency in tasks like autonomous driving, healthcare applications, and robotics.

6. Future Scope

Future research in image segmentation is expected to address the persistent challenge of achieving high segmentation accuracy while maintaining computational efficiency suitable for real-world deployment. A primary direction is the development of lightweight foundation models for edge and mobile environments, extending the work of MobileSAM and EfficientSAM to enable real-time performance on resource-constrained devices. Large-scale video segmentation also presents significant opportunities, with models such as SAM 2 and SAM-Track establishing strong foundations for consistent object tracking across dynamic and long-duration scenes.

Open-vocabulary and language-guided segmentation is another rapidly growing area, with frameworks such as Grounded SAM, CAT-Seg, and CLIP Surgery demonstrating the potential of vision-language alignment for flexible real-world segmentation. Self-supervised approaches including DINOv2, DINO, and BEiT are expected to further reduce dependency on costly pixel-level annotations. The use of generative models for segmentation, as explored in ODISE and VPD, also represents a promising direction by leveraging diffusion-based representations for dense prediction. Finally, the integration of large language models with pixel-level visual understanding, as demonstrated in Osprey, Painter, and SegGPT, points toward unified multimodal AI systems capable of reasoning and segmenting simultaneously.

References

- [1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.
- [3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.
- [4] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Garg, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1290–1299.

- [5] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.
- [6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9650–9660.
- [7] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9653–9663.
- [8] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVT v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.
- [9] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Komatsu, R. Röhrbach, J. Donahue, K. Rohrbach, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.
- [10] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "OneFormer: One transformer to rule universal image segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2989–2998.
- [11] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14408–14419.
- [12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11976–11986.
- [13] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.
- [14] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and panoptic segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5789–5798.
- [15] S. Cho, W.-D. Jang, S. Kim, and H. Kweon, "CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [16] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *Proc. European Conference on Computer Vision (ECCV)*, 2022, pp. 540–557.
- [17] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Yang, H. Zhang, L. Zhang, and L. Liu, "Grounded SAM: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.
- [18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Joulin, and J. Mairal, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2024.

- [19] X. Wang, Y. Kong, C. Shen, Y. Jiang, and L. Li, "Painter: Images speak in images — a generalist painter for in-context visual learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6660–6671.
- [20] X. Wang, T. Zhang, Y. Jiang, and L. Li, "SegGPT: Segmenting everything in context," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140.
- [21] J. Cheng, H. Tsai, S. Tillman, S. Wang, and J. Tsai, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023.
- [22] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [23] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, R. Krishnamoorthi, and V. Chandra, "EfficientSAM: Leveraged masked image pretraining for efficient segment anything," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [24] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight SAM for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.
- [25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [26] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Proc. European Conference on Computer Vision (ECCV)*, 2022, pp. 459–479.
- [27] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2955–2965.
- [28] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21453–21464.
- [29] Y. Li, F. Wang, J. Xiao, and Y. Qiao, "CLIP surgery for better explainability with enhancement in open-vocabulary tasks," *arXiv preprint arXiv:2304.05653*, 2023.
- [30] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Liu, "Osprey: Pixel understanding with visual instruction tuning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.