HAISTA-NET: Human Assisted Instance Segmentation Through Attention

by

Muhammed Korkmaz

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Engineering

KOÇ ÜNİVERSİTESİ

September 18, 2023

HAISTA-NET: Human Assisted Instance Segmentation Through Attention

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Muhammed Korkmaz

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. T. Metin Sezgin (Advisor), Koç University

_____

Prof. Yücel Yemez, Koç University

_____

Assist. Prof. Alexandra Bonnici, University of Malta

Date: _____

to my family

ABSTRACT

HAISTA-NET: Human Assisted Instance Segmentation Through Attention

Muhammed Korkmaz

Master of Science in Computer Engineering

September 18, 2023

Instance segmentation is a fundamental computer vision task with a wide range of applications. Some instance segmentation tasks such as medical image analysis, and image/video editing require high levels of precision. However, this precision is often beyond the reach of what even state-of-the-art, fully automated instance segmentation algorithms can deliver. The performance gap becomes particularly prohibitive for small and complex objects. Practitioners typically resort to fully manual annotation, which can be a laborious process. In order to overcome this problem, we propose a novel approach to enable more precise predictions and generate higher-quality segmentation masks for high-curvature, complex and small-scale objects. Our human-assisted segmentation model, HAISTA-NET, augments the existing Strong Mask R-CNN network to incorporate human-specified partial boundaries. We also present a dataset of hand-drawn partial object boundaries, which we refer to as "human attention maps." In addition, the Partial Sketch Object Boundaries (PSOB) dataset contains hand-drawn partial object boundaries which represent curvatures of an object's ground truth mask with several pixels. Through extensive evaluations, we show that HAISTA-NET outperforms state-of-the art methods such as Mask R-CNN, Strong Mask R-CNN, and Mask2Former, achieving respective increases of +36.7, +29.6, and +26.5 points in AP_{Mask} metrics for these three models. Our novel approach sets a baseline for future human-aided deep learning models by combining fully automated and interactive instance segmentation architectures.

ÖZETÇE

Yüksek Lisans Tez Başlığı
Muhammed Korkmaz
Bilgisayar Mühendisliği, Yüksek Lisans
18 Eylül, 2023

Obje segmentasyon, tıbbi görüntü analizi ve görüntü/video düzenleme gibi çeşitli alanlardaki uygulamalarında yüksek doğruluk gerektirir; bu mevcut popüler tam otomatik obje segmentasyon modellerinin performansını da dikkate almamızı gerektirir. Özellikle küçük ve karmaşık şekillere sahip nesneler için yeni modeller geliştirilmesi ya da mevcut modellerin iyileştirmesi gereksinimi son yıllarda hızla artmaktadır. Bu nedenle, bu alanlardaki uygulayıcılar zahmetli bir iş olan tamamen manuel etiketleme metoduna başvurmaktadır. Bu sorunların üstesinden gelmek amacıyla, daha kesin tahminler yapmak ve yüksek eğrilikli, karmaşık ve küçük ölçekli nesneler için daha kaliteli segmentasyon maskeleri oluşturmak amacıyla yeni bir yaklaşım öneriyoruz. İnsan destekli segmentasyon modelimiz HAISTA-NET, Strong Mask R-CNN ağını insan tarafından belirlenen kısmi sınırları dikkate alacak şekilde genişletir. Ayrıca İnsan Dikkat Haritaları (HAM) olarak adlandırdığımız elle çizilmiş kısmi nesne sınırları veri setini de sunmaktayız. PSOB (Kısmi Çizim Nesne Sınırları) veri seti, bir nesnenin temel gerçek maskesinin sınırlarını birkaç pikselle temsil eden, elle çizilmiş kısmi vuruşları içerir. Kapsamlı değerlendirmeler sonucunda, PSOB veri seti ile eğittiğimiz HAISTA-NET'in, sırasıyla Mask R-CNN, Strong Mask R-CNN ve Mask2Former modellere kıyasla AP_{Mask} metriğinde +36,7, +29,6 ve +26,5 puanlık artış elde ederek bu gelişmiş yöntemlerden daha iyi performans gösterdiğini deneylerimizle sunuyoruz. Yaklaşımımızın, Tam Otomatik ve Etkileşimli Obje Segmentasyon mimarilerini birleştirerek gelecekteki insan destekli derin öğrenme modelleri için bir temel oluşturacağını umuyoruz. Kod ve PSOB veri seti herkese açık olarak paylaşılacaktır.

ACKNOWLEDGMENTS

First and foremost, I would like to extend my sincere gratitude to my Dear Lord, who is the most forgiving and generous. Secondly, I want to express my gratitude to my advisor, Professor T. Metin Sezgin, for his essential advice and support. His knowledge of the subject and encouragement have always inspired and guided me throughout my research. Special thanks to Prof. Yücel Yemez and Assist. Prof. Alexandra Bonnici, who allocated valuable time to participate in my thesis defense as a committee member. I would also like to thank Reyhan Eren, who has been very valuable to me and has supported me under all circumstances. On the other hand, I would like to express my gratitude to my dear brother Tolga Büyükyazı, a 2000s conventional computer vision expert, and M. Emin Karaismail, with whom we fought together. I hope that we can achieve many more successes together. Finally, I wish to make the world a better place to live in and to do good things for the benefit of humanity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANOVA	Analysis of Variance
AP	Average Precision
BBOX	Bounding Box
CONV1	First Convolution of the Network
C-AP	Curvature-based Average Precision
CNN	Convolutional Neural Network
FPN	Feature Pyramid Network
GUI	Graphical User Interface
HAM	Human Attention Map
HAISTA	Human Assisten Instance Segmentation Through Attention
INT	Interaction
IOU	Intersection Over Union
LS	Length of Sketch
MIOU	Mean Intersection Over Union
PSOB	Partial Sketch Object Boundaries
PP	Perimeter of Polygon
RGB	Red Green Blue
R-CNN	Region Based Convolutional Neural Network
RPN	Region Proposal Network
SCP	Simple Copy Paste
RES	ResNet Architecture
RESX	ResNeXt Architecture
RITM	Reviving Iterative Training with Mask
SOTA	State-of-the-art
UI	User Interface

Chapter 1

INTRODUCTION

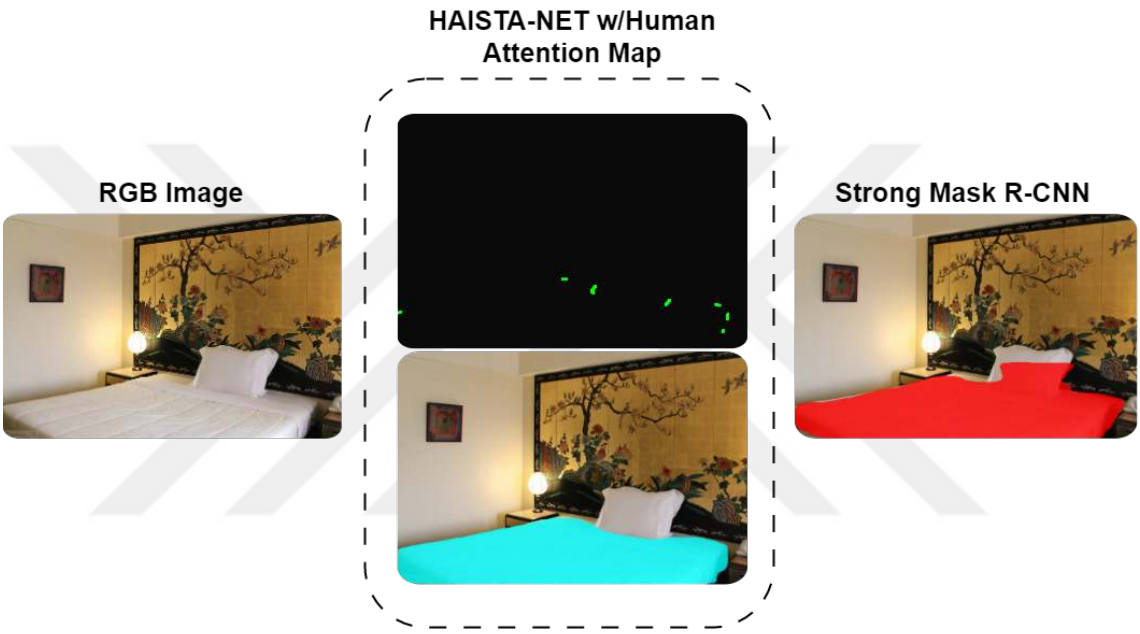

Figure 1.1: First glance of more precise mask prediction of HAISTA-NET through Human Attention Map (Middle). HAISTA-NET outperforms the mask prediction of Strong Mask R-CNN (Right) on high-curvature objects.

In recent years, demand has grown for deep-learning-based, fully automated instance segmentation models [He et al., 2017, Bolya et al., 2019, Fang et al., 2021, Kirillov et al., 2020, Cheng et al., 2022, Qiao et al., 2021, Fang et al., 2019, Tian et al., 2021, Wang et al., 2020]. Due to rapid progress in their development, these image detection tools have become the preferred method for applications such as object refinement [Chen et al., 2018], medical image screening [Wang et al., 2018a, Wang et al., 2018b], and image/video editing [Benard and Gygli, 2017, Li et al., 2004]. However, they often perform poorly when objects are small-scale or have

a pronounced curvature, causing inaccuracies in the instance segmentation mask, such as boundary outflows or under-covering (Figure 1.3) [Tang et al., 2021]. To improve mask precision, users may resort to manual annotation [Andriluka et al., 2018], though this has several drawbacks, including the added time cost.

As a result, researchers have been studying alternative approaches to interactive segmentation [Benenson et al., 2019, Li et al., 2018, Liew et al., 2017, Acuna et al., 2018], to remedy existing deficiencies in segmentation masks while reducing the time needed for manual annotation. One method involves deep-learning-based interactive segmentation models that use corrective action techniques for object mask retrieval (Figure 1.4) [Benenson et al., 2019, Liew et al., 2019]. Another proposed technique involves mask editing based on mouse clicks [Papadopoulos et al., 2017], whereby the user helps to guide the automated model manually by detecting boundary outflows or under-covering (Figure 1.5).

With HAISTA-NET, we bring a new approach to instance segmentation by combining two main areas: fully automated instance segmentation, and interactive instance segmentation. With this proposed method, users convey the intended boundary by marking it manually. A deep learning network, previously trained with such markings, uses this input to drive the instance segmentation process. Our novel approach eliminates mask errors by training the model with user inputs to fix the mask where other models stumble and fail.

We also present our new Partial Sketch Object Boundaries (PSOB) dataset. This dataset has images, objects, and categories taken from LVIS [Gupta et al., 2019] and extended with additional hand-drawn sketches. It contains raster images, which we refer to as human attention maps, drawn by users with a few pixels (minimal input) of the high-curvature regions of the object boundary, where segmentation masks are most erroneous. For the PSOB dataset, 30 users annotated 18,677 objects of different scales and various high-curvature sections.

HAISTA-NET architecture uses human attention maps both for the training phase and for running inference. Our model can be easily integrated with various deep learning-based segmentation and detection models (see Section 3.4). HAISTA-NET uses a Strong Mask R-CNN baseline [He et al., 2017, Ghiasi et al., 2021]. In

our model, we removed the first convolution layer of the Mask R-CNN [He et al., 2017] backbone and added a new one fed by a combination of the Human Attention Map and a three-channel RGB image. As a result of our experiments, the HAISTA-NET architecture outperforms the Strong Mask R-CNN baseline with a more precise mask for high-curvature and/or small-scale objects (Figure 1.1).

We also developed a user-friendly interface that allows users to interact with objects to create or edit human attention maps using partial strokes [Olsen et al., 2009]. Using this interface, users can either directly annotate images without prior knowledge of the segmentation results or perform the annotation after viewing the Strong Mask R-CNN outputs.

Our model achieves more accurate results when using the PSOB dataset with human attention maps than existing state-of-the-art models. Evaluations conducted with the PSOB dataset show that our model achieves a performance that is 36.7 points better than Mask R-CNN, according to the AP_{Mask} metric, and +29.6 points compared with Strong Mask R-CNN. Using the AP_{Bbox} metric, our model also demonstrates an increase of +33.5 and +31.1 points, respectively, versus Mask R-CNN and Strong Mask R-CNN (see Section 4.1). Moreover, HAISTA-NET achieves a +26.5-point increase in AP_{Mask} versus Mask2Former [Cheng et al., 2022], which is the current state-of-the-art model for instance segmentation on the COCO [Lin et al., 2014] dataset.

Our main contributions can be summarized as follows:

- We have developed HAISTA-NET based on the Mask R-CNN architecture to combine three-channel RGB images with a human attention map.

- We propose a sketch-based representation of user-defined, high-curvature sections of objects of interest called the human attention map.

- We present the Partial Sketch Object Boundaries (PSOB) dataset, which will be valuable in driving further research in user-assisted instance segmentation.

- We propose the Adaptive Curvature Number Detector (ACND) for detecting and classifying the curvatures of segmentation masks.

Figure 1.2: **Method Outline.** Users draw a couple of pixels according to their attention to the object. Then, a Human Attention Map is generated to concatenate with the RGB image to feed the model. We denote the concatenate operator as $\otimes$.

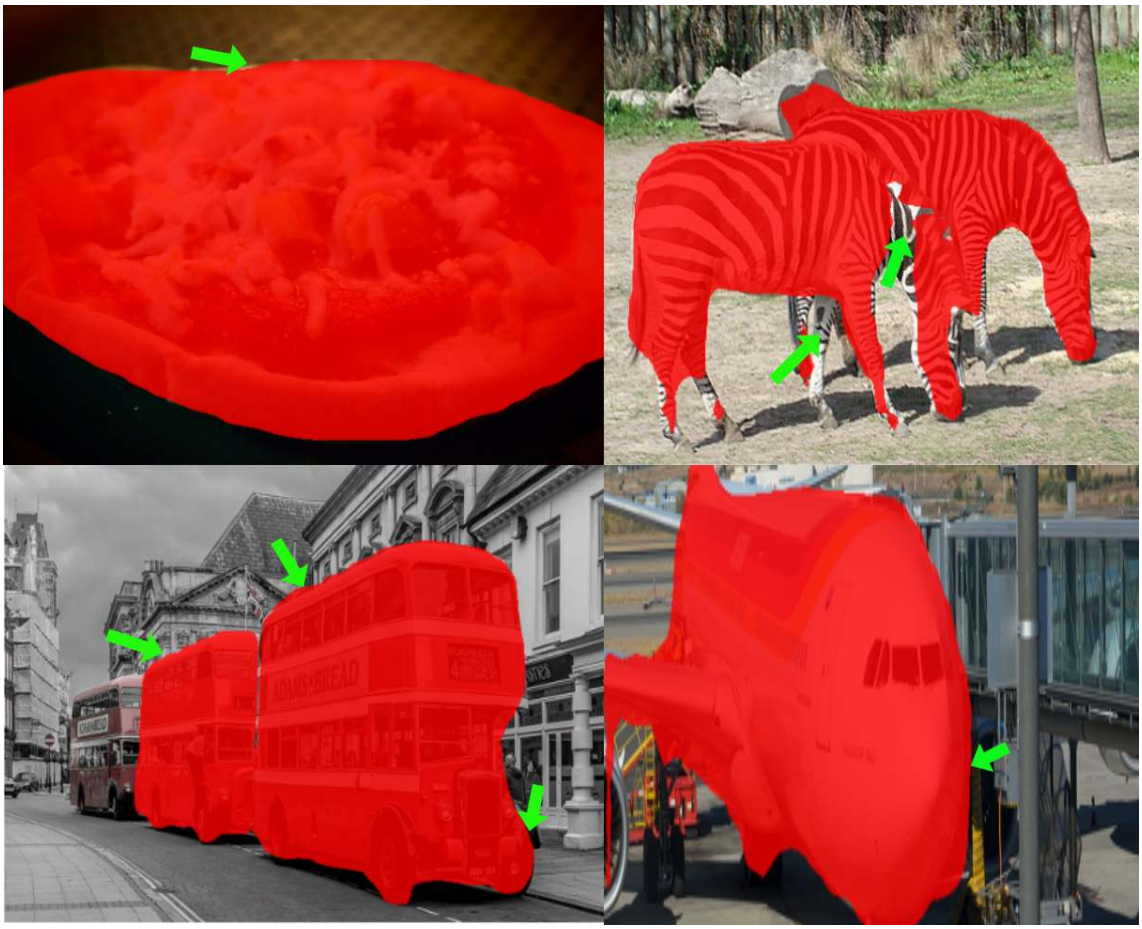

Figure 1.3: **Examples of boundary outflows and under-covering.** The figure demonstrates that Instance Segmentation architectures may detect inaccurate segmentation masks such as boundary outflows or under-covering. In this image, Strong Mask R-CNN is utilized.

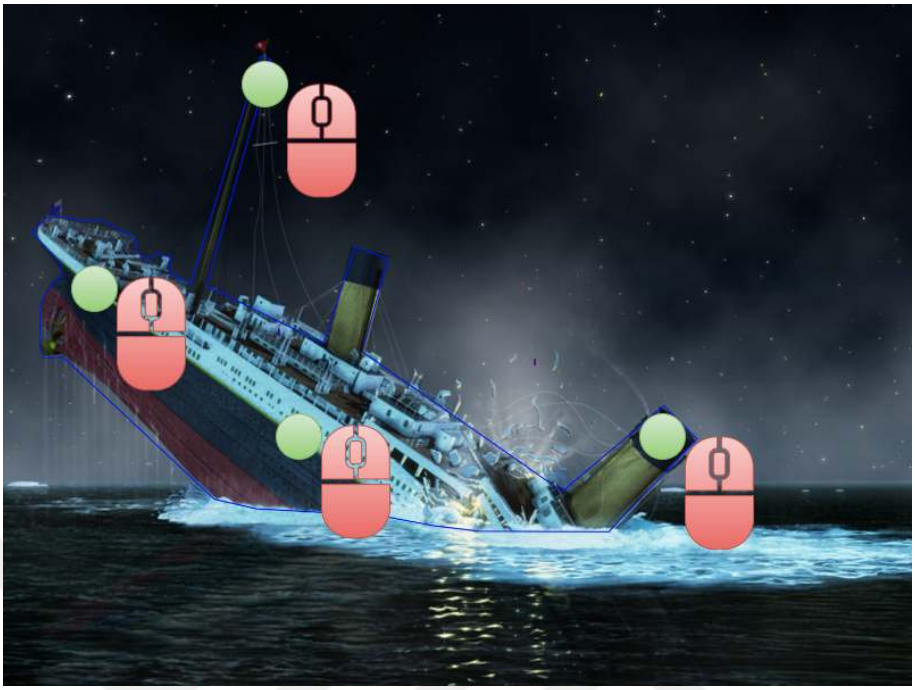

Figure 1.4: **Example of click-based mask obtaining.** In this image, Adobe Interactive Tool is utilized in order to retrieve partial masks by mouse clicks.

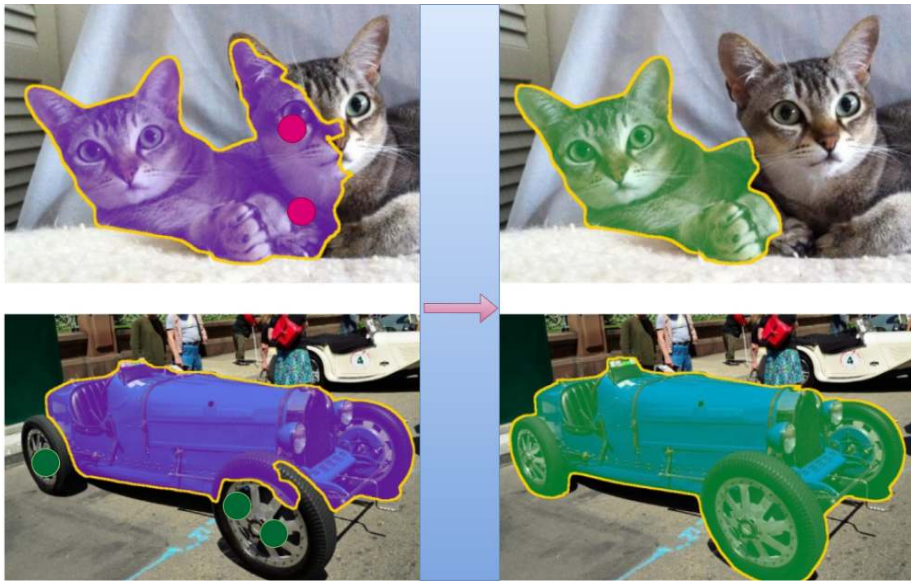

Figure 1.5: **Example of boundary correction using mouse clicks.** The figure demonstrates that RITM [Sofiiuk et al., 2022] can correct an inaccurate segmentation mask. Green-colored clicks represent the missing parts, and pink-colored pieces are redundant parts.

- Applying multiple factor analysis, we have reported our model results for objects according to different scales, curvature numbers, and user drawing characteristics.

- Our user study also analyzes the cost of sketch-based manual annotations in LVIS dataset quality.

Chapter 2

RELATED WORK

Instance segmentation is a demanding computer vision task that has experienced noticeable progress in recent years. This task goes beyond detecting objects to determining the precise boundaries of objects within an image. In this section, we will review the advancements of instance segmentation methods, dividing them into two main approaches: Fully Automated Instance Segmentation [Fang et al., 2021] and Interactive Instance Segmentation [Benenson et al., 2019, Li et al., 2018, Liew et al., 2017, Acuna et al., 2018, Sofiiuk et al., 2022].

2.1 Fully Automated Instance Segmentation

Fully Automated Instance Segmentation methods attract significant attention in the computer vision community. These methods are designed to process large datasets effectively and have been applied in many fields, such as autonomous driving, medical image analysis, and robotics. An essential feature of Fully Automated Instance Segmentation approaches is the ability to create instance masks for objects within images automatically. These methods distinguish objects from the background by analyzing information at the pixel level and provide precise object segmentation. Furthermore, they are particularly advantageous if there is a need for real-time processing. The most popular model that comes to mind is Mask R-CNN [He et al., 2017] when fully automated instance segmentation is investigated. Because the paper of [He et al., 2017] was published six years ago and propagated to a wide range. The Strong Mask R-CNN [He et al., 2017, Ghiasi et al., 2021] model is one of the state-of-the-art models that achieves high precision. This architecture extends from the Faster R-CNN [Ren et al., 2015] model and sets a baseline for many deep learning models with its simplicity and ease of implementation. With the contributions

of [Ghiasi et al., 2021], the simple copy-paste method has improved the performance of the basic Mask R-CNN architecture and transformed it into the Strong Mask R-CNN model. Strong Mask R-CNN, our baseline, became prominent with its practical use by obtaining high-precision masks with fast-forward training and low memory usage.

2.2 Interactive Instance Segmentation

Interactive Instance Segmentation methods are generally interested in two primary goals. They aim to correct the inaccurate masks and obtain partial masks, which practitioners interact with a pen or mouse. Notably, Some researchers [Benenson et al., 2019, Li et al., 2018, Liew et al., 2017, Acuna et al., 2018, Benard and Gygli, 2017, Li et al., 2004, Liew et al., 2019] use less-time-costly manual annotation techniques to avoid segmentation failures in object boundaries. Interactive segmentation is another alternative that has emerged recently and promises superior results to manual annotation. Some studies suggest using mouse clicks or free-style painting [Lin et al., 2016] to correct the boundaries of segmentation masks, assuming this reduces annotation time. According to [Benenson et al., 2019], the point and click-based [Xu et al., 2016] method is an effective technique to improve mask precision rates. Using mouse clicks requires two types of marking for redefining object boundaries. If the estimated segmentation mask overflows the actual boundary of the object, the user sets negative markers to ignore those segments. If the boundary under-covers the ground truth, the user sets the targeted distance by placing positive markers. However, the mouse click technique may fail for small-scale objects as it requires a high level of detail to adjust in bounding box limits.

2.3 Strong Mask R-CNN

Among the Fully Automated Instance Segmentation methods, the Strong Mask R-CNN architecture deriving from Mask R-CNN [He et al., 2017] with the addition of Simple Copy Paste Method [Ghiasi et al., 2021] stands out as a state-of-the-art model. This architecture extends the Faster R-CNN [Ren et al., 2015] and is known

for its simplicity and applicability. With its reputation for achieving high precision and accuracy in instance segmentation tasks, Strong Mask R-CNN has become a benchmark for various deep-learning models in this field.

The Strong Mask R-CNN architecture combines object detection and Instance Segmentation in a single unified framework. It uses convolutional neural networks (CNNs) to extract features from input images and uses region proposal networks (RPNs) to generate candidate object regions. These regions are then enhanced to produce precise instance masks, distinguishing individual objects within the image. Strong Mask R-CNN training involves supervised learning using labeled datasets to optimize the model's parameters. During training, the model learns to simultaneously predict object classes, bounding boxes, and instance masks. This training process usually requires significant labeled data and large computational resources. In terms of inference, Strong Mask R-CNN exhibits outstanding efficiency. Once trained, the model can detect objects from frames in real time. Strong Mask R-CNN is a powerful model that has made significant progress in the areas of instance segmentation and object detection. If we make a comparative analysis:

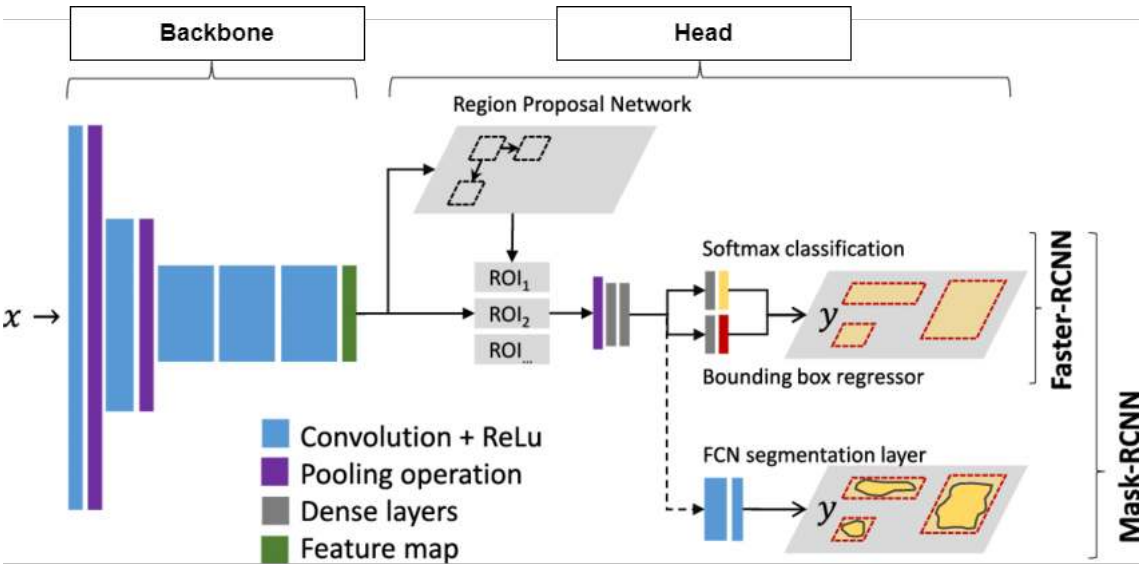

Figure 2.1: **The difference between Mask R-CNN and Faster R-CNN on network architecture.** Mask R-CNN extends the Faster R-CNN by adding a head branch.

- **Faster R-CNN** stands out as a basic object detection model. It offers a wide range of applications, runs fast, and can be used in real-time applications. However, it lacks the ability for direct instance segmentation and requires special layers to create a segmentation mask.

- **Mask R-CNN** supports both object detection and instance segmentation tasks by adding a mask head to Faster R-CNN. It sets a baseline for many models as well as Strong Mask R-CNN.

- **Strong Mask R-CNN** enables object detection and Instance Segmentation tasks as well as Mask R-CNN. It works fast and can be effective for real-time applications, but since it uses a data augmentation technique, this model must be trained longer than standard Mask R-CNN (Figure 2.1).

In conclusion, Strong Mask R-CNN is a model that has taken an essential step towards providing high precision and accuracy in instance segmentation tasks. It is suitable for real-time applications, but training and resource requirements must be considered. It is crucial to choose the most appropriate architecture depending on the application.

2.4 The Simple Copy Paste Data Augmentation Technique

The Simple Copy Paste Technique (SCP) [Ghiasi et al., 2021], introduced by Ghiasi and his team in 2021, emerged as a method that significantly improves the performance of Mask R-CNN-based models. This methodology not only increases the capabilities of these models but also leads to the development of robust augmentation techniques.

The Strong Mask R-CNN architecture was designed based on the Simple Copy Paste Method and has received significant attention in the computer vision community. This architecture stands out for its capacity to deliver higher-quality masks. This property is fundamental in various practical applications, especially in cases where objects must be identified precisely and fine-grained. SCP creates a new image, which is in actual size, on the training phase and randomly selects ground truth

masks of an object. After randomly selecting the ground truth, it copies all of them and pastes them on the same image. For instance, most of the dataset contains ordinary images and ground-truths. We do not see an image that involves a couple of people playing football and a giraffe watching them. However, it is possible with SCP, so the vision community represents them as strong.

As a result, the Simple Copy Paste Data Augmentation Method [Ghiasi et al., 2021] has increased the performance of Mask R-CNN-based models and significantly contributed to the data augmentation method.

2.5 Mask2Former

Transformer-based architectures [Carion et al., 2020, Dosovitskiy et al., 2020, Liu et al., 2021] are increasing in popularity due to their high image feature learning and attention mechanisms. For example, Mask2Former [Cheng et al., 2022] architecture uses transformers in place of the conventional RPN backbone of Faster R-CNN, and surpasses both standard Mask R-CNN [He et al., 2017] and Strong Mask R-CNN [He et al., 2017, Ghiasi et al., 2021] in results with its transformer-based structural design. These new-generation transformers are replacements of CNN-based traditional architectures. Mask2Former is a universal image segmentation architecture. It is based on the DETR model [Carion et al., 2020], which uses a transformer decoder to estimate masks for each object in a frame. Mask2Former is a universal architecture for various tasks, such as semantic, panoptic, and instance segmentation. It is more accurate than previous image segmentation architectures because it uses a multi-scale decoder and a masked attention mechanism. The multi-scale decoder in Mask2Former allows it to attend to local and global features. The masked attention mechanism in Mask2Former limits the decoder's attention to the foreground region of each object. The masked attention mechanism in Mask2Former works by only allowing the decoder to attend to the foreground region of each object. So, it supports Mask2Former to focus on each object's relevant features and prevents the decoder from joining the background noise. Yet such models still have a high memory and time cost [Han et al., 2021], not only for training but also for inference.

Also, they tend to under-perform in fine-grained applications such as medical image analysis [Razzak et al., 2018].

2.6 Mouse Click-Based Interaction Technique

Mouse-click-based correction and interaction[Papadopoulos et al., 2017] is becoming increasingly popular as it offers an effective solution to segmentation challenges, especially at object boundaries. Traditional methods often rely on time-consuming manual annotation techniques to obtain accurate segmentations, but these methods can be impractical and require more resources. However, interactive instance segmentation techniques, particularly mouse clicks, have emerged as promising alternatives to complete manual annotation.

A distinct advantage of the Mouse-Click Technique is effectively refining the boundaries of segmentation masks. Benenson and other researchers [Benenson et al., 2019] have demonstrated the effectiveness of point-and-click-based methods [Xu et al., 2016] in increasing mask sensitivity rates. Users can precisely redefine object boundaries using mouse clicks. When the estimated segmentation mask exceeds the actual object boundary, users can set negative markers to instruct these segments to be ignored. Conversely, users can place positive markers to indicate the desired boundary distance if the boundary does not satisfy the actual bounding constraint. This interactive process enables fine-tuning, increasing the precision of object segmentation.

Despite this, the difficulties that the Mouse-Click Technique may encounter when working with small-scale objects should be addressed. Achieving a high level of detail for adjusting bounding box boundaries can be more difficult for smaller objects, potentially leading to sub-optimal results. However, the Mouse-Click Technique is still considered a popular option in interactive instance segmentation, offering a practical and user-friendly method to refine segmentation masks and reduce dependence on time-consuming manual annotations. As technology advances, further improvements and developments are expected to address object size and detail limitations.

2.7 Reviving Iterative Training with Mask (RITM)

Interactive segmentation involves the process of refining and correcting segmentation masks with user guidance, and this is often done through techniques such as mouse clicks or strokes. Reviving Iterative Training with Mask Guidance for Interactive Segmentation study aims to bring iterative training back to the center in the interactive segmentation process.

This approach has many advantages. First, this approach offers the potential for a significant increase in segmentation accuracy [Sofiiuk et al., 2022]. Iterative training refines the segmentation mask with each iteration, reducing errors and providing a more accurate result. Additionally, the inclusion of mask guidance gives users more control over the segmentation process. Users can provide corrections and guidance to fine-tune the mask, meaning user input is effectively integrated. Additionally, this method has the potential to make the process more efficient and reduce users' workload by reducing the need for manual input by users. At the same time, mask guidance allows users to shape this enhancement by providing precise corrections and orientations for each object. This method has the potential to obtain more accurate and user-friendly results in the field of instance segmentation and can make a significant contribution to applications in this field.

However, this innovative approach needs some help. Iterative training has increased computational resource requirements, especially involving multiple iterations. This situation may limit its usability on lower-power hardware. Additionally, there may be a potential learning curve as users become accustomed to the interactive training process, which may create initial adoption challenges. Further research and development in this area may help overcome these challenges and improve the methodology for practical applications.

2.8 Overview

This research suggests a more effective way to partially mark object boundaries before the Instance Segmentation step instead of mouse clicks, which can lead to time-wasting to correct faulty masks. Our approach involves marking boundaries,

especially in regions with high curvature, where segmentation has the most significant difficulties. In this way, it aims to reduce errors by making the instance segmentation process more efficient and precise.

Chapter 3

PROPOSED APPROACH

We propose a methodology to generate more precise segmentation masks by integrating human attention maps with fully automated segmentation models [He et al., 2017, Ghiasi et al., 2021]. Our study includes the following steps respectively; a review of dataset collection, data annotations, selection of input representation techniques, converting input representations to human attention maps, demonstration of network architecture, fine tuning training parameters, and running inference mode.

3.1 Partial Sketch Object Boundaries Dataset

In the stage of creating our dataset and deciding on the annotation technique we explored failure conditions of the instance segmentation models. First of all, we trained a Strong Mask R-CNN [He et al., 2017, Ghiasi et al., 2021]model with the LVIS dataset to investigate segmentation mask failures. Following the training, we selected 3070 test images from the LVIS [Gupta et al., 2019] dataset to run inference in order to determine object mask failures. The object mask prediction was not precise in following conditions:

- If the object scale is small $area < 32^2$.
- If medium-scale $32^2 < area < 96^2$ and large-scale $area > 96^2$ objects have a high number of curvatures.
- If the object is located behind another object, causing an occlusion problem.
- If the object's shape and curvature differ in test images from training images.

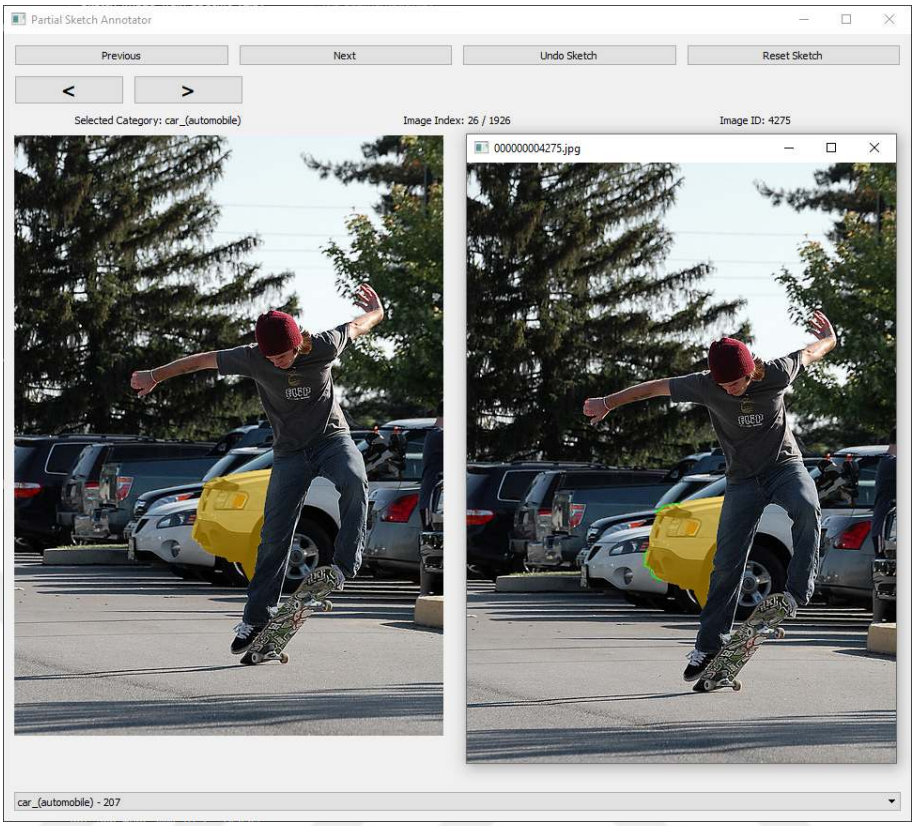

Figure 3.1: **Interactive interface.** Our interface supports both sketching and processing image at the same time. Practitioners can see prediction results while sketching.

Regarding the analysis results, we propose a minimal sketch-based technique to extend the LVIS dataset with user input. Our newly constituted dataset, PSOB, contains hand-drawn partial object boundaries in high-curvature points (see section 3.2) for 115 categories of objects taken from the LVIS dataset. 30 users annotated data, using hand-drawn partial sketches to represent object high-curvature points. Our aim was to minimize human effort while interacting with the object. For data distribution to be reasonable, we annotated 10450 training objects, 4109 validation objects, and 4118 test objects. For experimental reasons, we also recorded the object's total sketch time [sec], time for each stroke [sec], and pixel-wise stroke's corresponding values in the xy coordinate frame.

The important factor of this dataset is Human Attention Points representing user inputs to form Human Attention Maps. With these attention points, some parts

of the boundaries of the objects of interest are sketched by users. These sketched points represent human assistance in our work. We identify the human assistance with following rules:

- If sketch input covers less than 25% of the object's boundary, human assistance is designated as **minor assistance**.
- If sketch input covers the object's boundary between 25% and 50% human assistance is designated as **medium assistance**.
- If sketch input covers greater than 50% of the object's boundary, human assistance is designated as **major assistance**.

3.2 Adaptive Object Curvature Detector

Object masks consist of polygons having different angles between line segments. Angles of intersecting line segments constitute a high-curvature or low-curvature point depending on their degree. We use the Ramer-Douglas-Peucker (RDP) [Douglas and Peucker, 1973] algorithm to refine curvature points for each object boundaries.

The standard RDP algorithm uses a static epsilon value to simplify the number of the object's corners. Designated static RDP epsilon values require parameter tuning for every scale of the object. In order to fine tune this value, we calculate the perimeter of the polygon from the coordinates and set the epsilon value to three percent of the perimeter. This method returns more accurate results while calculating number of curvature points for objects from different scales, compared to classical RDP. We use the following formula where ϵ is adaptive epsilon for RDP, n is line segment count, and $\mathcal{I}$ is line segments:

$$\epsilon = 0.03 \cdot \sum_{k=1}^{n} \sqrt{(\mathcal{I}_{k_{x_2}} - \mathcal{I}_{k_{x_1}})^2 + (\mathcal{I}_{k_{y_2}} - \mathcal{I}_{k_{y_1}})^2} \tag{3.1}$$

After obtaining the results of the adaptive object curvature detector (eq. 3.1), we classify the objects according to their number of curvature points by using following rules:

- If an object's number of curvature points is less than 6, an object's curvature type is designated as **low curvature**.

- If an object's number of curvature points is between 6 and 10, an object's curvature type is designated as **medium curvature**.

- If an object's number of curvature points is greater than 10, an object's curvature type is designated as **high curvature**.

3.3 Representation of Human Attention Map

As briefly described in previous sections, high-curvature points on the objects of interest are partially annotated by human users in order to create human attention maps. These user strokes' x and y pixel coordinates are used to form binary images with the exact dimensions as input images. These binary images are representations of the human attention map. They are later supplied to the fourth channel of the first convolution layer, where the other three channels are reserved for the original RGB image. Attention points annotated by practitioners as high curvature points are set to the pixel value of 255, while other regions are left initially as zero. This algorithm results in a sparse representation where most of the image is zero while a few pixels at the location of interest are 255. Random initialization of the fourth channel weights was used during the training phase.

The effects of representing unrelated regions with various forms were investigated during the preliminary analysis stage. Since, during back-propagation, randomly multiplying the initialized weights of the fourth channel with zeros prevents optimization of these weights, alternative strategies were tested, such as using a small fixed value or a randomly generated small number instead of zeros.

The results show an insignificant difference in performance among the three methods proposed for representing unrelated regions, obviously using zeros, a small fixed pixel value ($\mathcal{P}$) such as 10, and a randomly generated number. So, a constant

value of 10 was selected to represent outside the interested boundaries. (eq. 3.2).

$$\mathcal{P} = \begin{cases} 255, & \text{if location is attention point} \\ 10, & \text{otherwise} \end{cases} \tag{3.2}$$

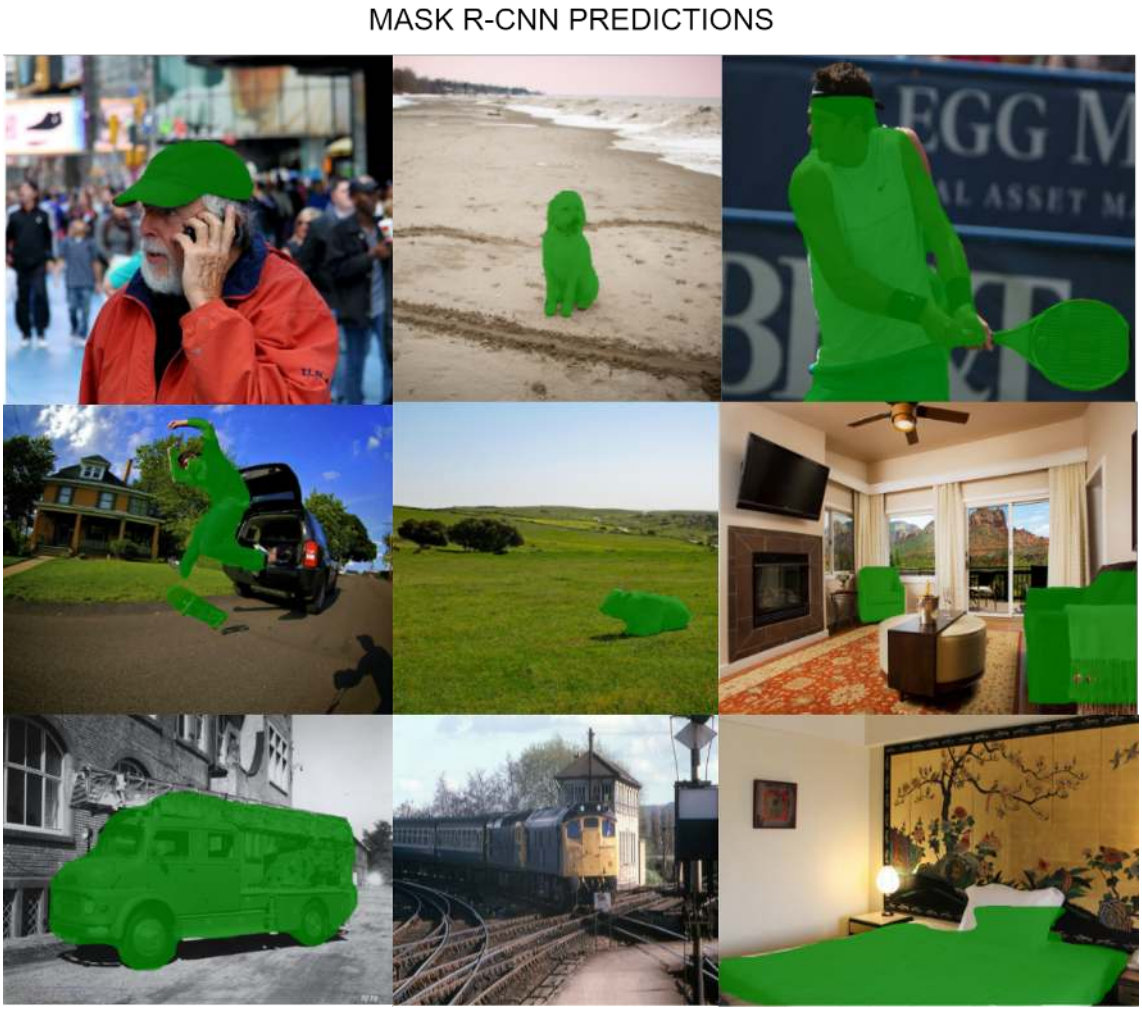

Figure 3.2: **Visualization of the Predictions of Mask R-CNN.** We present images with different scales and curvature numbers.

3.4 Network Architecture

HAISTA-NET adds a head branch to Mask R-CNN [He et al., 2017] architecture. We remove the first convolution from the backbone of Mask R-CNN to set our 2D convolution. We merge the RGB image with our Human Attention Map in order to

Figure 3.3: **Visualization of the Predictions of Strong Mask R-CNN.** We present images with different scales and curvature numbers.

form 4-channel image inputs (figure 1.2). We set the new convolution that overrides the first convolution of the current backbone with an input channel parameter equal to 4 and an output channel parameter equal to 64, as represented in standard ResNet [He et al., 2016] architecture. The new convolution kernel size is 7x7, the stride is 2x2, and the padding is 3x3. While adopting the new weights of the 4-channel convolution, for the first 3-channel, instead of randomly initializing weights, we use pre-trained weights of the existing 3-channel convolution. Because the Human Attention Map channel is new, we randomly initialize the weights between a range of (0, 0.001). As we described earlier, Mask R-CNN [He et al., 2017] architecture extends from Faster R-CNN [Ren et al., 2015]. The region proposal network back-

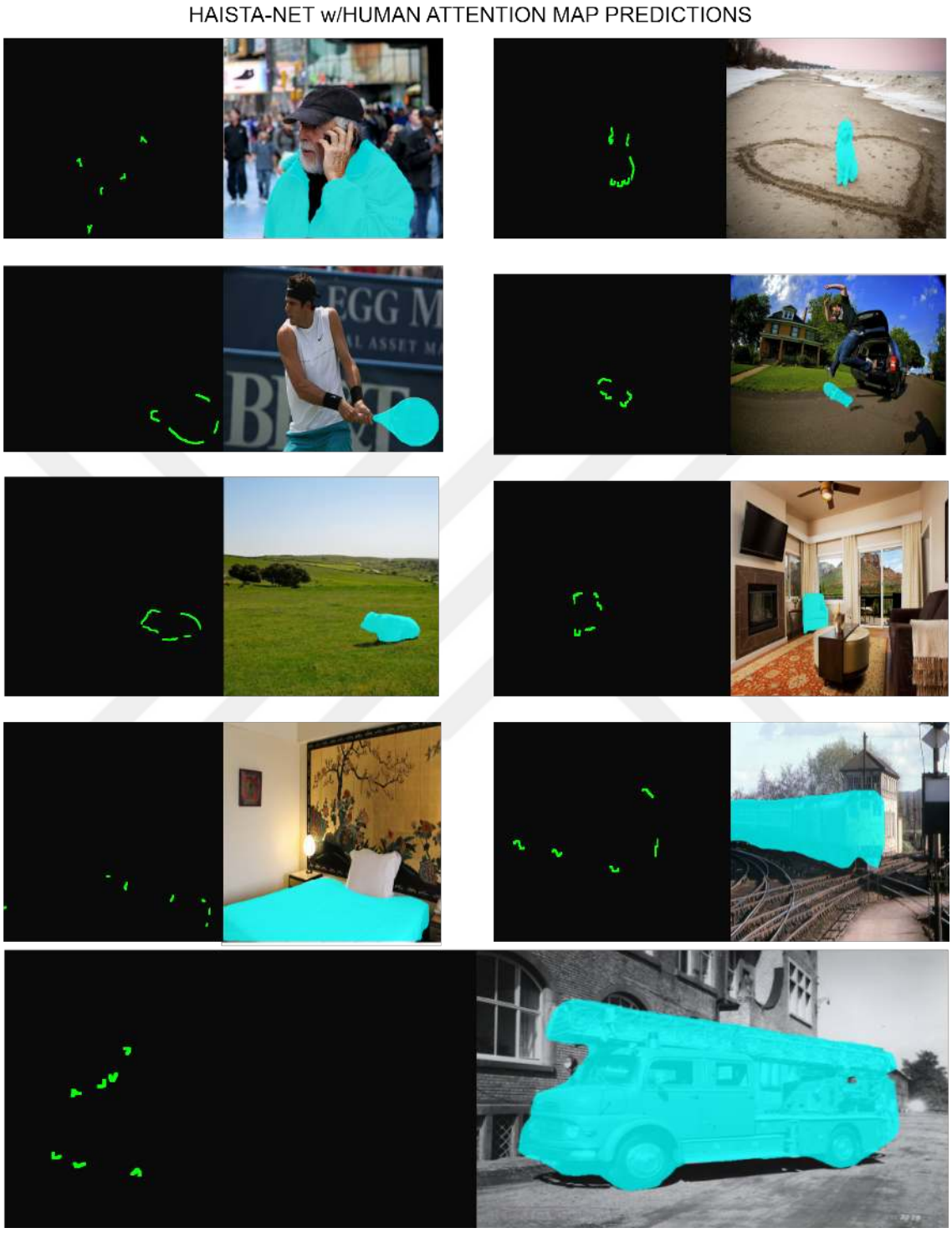

Figure 3.4: **Visualization of the Predictions of HAISTA-NET.** We present images with different scales, curvature numbers, and hand-drawing-based assistance types.

bone of Faster R-CNN uses data transformation parameters of mean and standard deviation vectors to adopt new input values when the channel number increases or decreases. So, we need to adjust the final channel's mean and standard deviation vectors. We set the mean of the final channel RGB image as 0.5. The first 3-channel RGB image mean is (0.485, 0.456, 0.406).

Additionally, we set the standard deviation of the final channel to 0.2, where the standard deviation for the first 3-channel is (0.229, 0.224, 0.225). We evaluate Res-Net [He et al., 2016] and Res-NeXt [Xie et al., 2017] architectures of depth 50, 101, and 152 layers with feature pyramid networks (FPN) [Lin et al., 2017]. We achieved the best results with Res-NeXt-101-FPN.

3.5 Data Augmentation

In order to develop enhanced model, we perform different training techniques [Du et al., 2021]. Firstly, we flip the image randomly and assign the probability value of being flipped to 0.5. As a result of this, sample diversity increases. Secondly, we use the Large Scale Jitter augmentation technique to randomly resize the image and image bounding box. We resize the image and boundary within the scale range (0.1, 2). The transform of the target size is 1024x1024. We also use bilinear interpolation as a parameter. Finally, we use the fixed-size crop to scale the image to the target size. Before feeding the model with our dataset, we use the Simple Copy Paste [Ghiasi et al., 2021] data augmentation method to create images from rare categories. We take object ground truth annotations from two images; create a new image using those annotations to reproduce samples. We compare the original 3-Channel Mask R-CNN and our model regarding the effectiveness of the Simple Copy Paste Data Augmentation method. As a result, Mask R-CNN achieved ≈7 AP improvements while we achieved less than ≈3 AP.

3.6 Training Parameters

We fine tune HAISTA-NET with 26 epochs for both ResNet [He et al., 2016] and ResNeXt [Xie et al., 2017] backbones using SGD optimizer. The learning rate of

Table 3.1: **Mask-based AP Metrics.** The results of the models with different backbones and augmentation techniques. SCP represents the Simple Copy Paste data augmentation [Ghiasi et al., 2021] technique. We train HAISTA-NET 26 Epochs without Simple Copy Paste and 50 epochs with Simple Copy Paste. The training combinations of Mask R-CNN are same as HAISTA-NET. However, both versions of Mask2Former are trained with 25 epochs.

		Mask				
Model	Backbone	AP	AP_{50}	AP_S	AP_M	AP_L
Mask R-CNN	RES-50 + FPN	18.2	30.3	4.6	15.0	25.7
Mask R-CNN w/SCP	RES-50 + FPN	25.3	36.3	12.4	23.0	34.4
Mask2Former	RES-50 + FPN	28.4	36.4	13.0	25.7	38.6
Mask2Former w/SCP	RES-50 + FPN	30.3	39.4	12.6	27.0	39.5
HAISTA-NET	RES-50 + FPN	51.2	68.5	41.8	53.5	57.2
HAISTA-NET w/SCP	RES-50 + FPN	51.2	68.9	41.4	53.5	57.7
HAISTA-NET	RES-101 + FPN	51.0	69.1	40.0	52.7	60.5
HAISTA-NET w/SCP	RES-101 + FPN	52.2	69.8	41.6	54.4	59.2
HAISTA-NET	RES-152 + FPN	52.8	71.0	39.8	55.4	60.1
HAISTA-NET w/SCP	RES-152 + FPN	53.1	71.3	45.2	56.6	59.4
HAISTA-NET	X-101 + FPN	52.0	70.9	40.1	54.9	58.1
HAISTA-NET w/SCP	X-101 + FPN	**54.9**	**73.8**	**47.8**	**58.4**	**61.5**

the optimizer is 0.0025. We set multi-step learning rate scheduling to decrease the learning rate by 10 at 16th and 22th epoch. Also, we set weight decay of 0.0001 and momentum of 0.9. Since we present a new dataset, PSOB, we must analyze the performance of other popular models such as Strong Mask R-CNN and Mask2Former on this dataset. For Mask R-CNN training, we use the same hyperparameters as HAISTA-NET because our model is based on MASK R-CNN. However, Since Mask2Former is a transformer-based architecture, we tune this model with 25 epochs, and the AdamW optimizer is used with a learning rate of 0.0025. Also, we set weight decay of 0.05, epsilon of 1e-08, and betas of (0.9, 0.999).

Table 3.2: **Bounding-Box-based AP Metrics.**The results of the models with different backbones and augmentation techniques. SCP represents the Simple Copy Paste data augmentation [Ghiasi et al., 2021] technique. We train HAISTA-NET 26 Epochs without Simple Copy Paste and 50 epochs with Simple Copy Paste. The training combinations of Mask R-CNN are same as HAISTA-NET. However, both versions of Mask2Former are trained with 25 epochs.

		Bbox				
Model	Backbone	AP	AP_{50}	AP_S	AP_M	AP_L
Mask R-CNN	RES-50 + FPN	23.3	34.4	14.0	23.5	27.4
Mask R-CNN w/SCP	RES-50 + FPN	25.7	36.6	19.1	27.6	30.6
Mask2Former	RES-50 + FPN	27.7	35.0	16.7	27.2	35.1
Mask2Former w/SCP	RES-50 + FPN	30.1	38.2	14.0	30.1	35.5
HAISTA-NET	RES-50 + FPN	52.1	69.3	51.4	56.8	53.1
HAISTA-NET w/SCP	RES-50 + FPN	53.3	69.2	53.4	58.4	54.1
HAISTA-NET	RES-101 + FPN	50.5	69.9	45.2	55.5	53.9
HAISTA-NET w/SCP	RES-101 + FPN	54.5	70.0	54.5	59.9	55.6
HAISTA-NET	RES-152 + FPN	52.6	71.7	49.6	57.3	54.6
HAISTA-NET w/SCP	RES-152 + FPN	55.3	71.7	55.4	61.5	54.4
HAISTA-NET	X-101 + FPN	51.5	71.4	45.3	57.5	52.7
HAISTA-NET w/SCP	X-101 + FPN	**56.8**	**74.4**	**56.0**	**62.3**	**57.2**

We trained these three different models on the PSOB dataset in order to compare their performances. Since the PSOB dataset contains Human Attention Maps, RGB images, and annotation features such as category, bounding box, and mask polygon, we can apply the same process to these 3 models. However, only HAISTA-NET uses Human Attention Maps due to model configuration. Mask R-CNN and Mask2Former only use RGB images to train.

PyTorch Framework is used for all implementations and configurations. In all experiments for training, we choose NVIDIA RTX 3090 TI single GPU with a batch size of 2. The image shuffling technique randomly selects images from the dataset

in the training phase. We use two subprocesses for data loading in training.

3.7 Inference

We performed inference for 4118 annotations after the training. Inference runs in NVIDIA RTX 3090 TI with a single GPU with a batch size of 1. We obtained AP [Lin et al., 2014], AP_{50}, AP_S, AP_M, and AP_L metrics for all models. HAISTA-NET, which has Strong Data Augmentation [Ghiasi et al., 2021] and Res-NeXt-101-FPN backbone, results in the best object mask and bounding box predictions. In order to interact with images during inference time, we present a user-friendly graphical user interface (figure 3.1). With the help of the interface, users may upload an image in seconds and create hand-drawn sketches to use as a human attention map. Following the sketch, user can observe object mask results as well as the bounding box of the object. Model guidance is possible by using the interface in two ways. In first case, the image goes into the Mask R-CNN model for the user to see the results and to create a Human Attention Map. In this case, a user may detect the missing object parts. Users may use it as a reference to create more precise partial sketches for HAISTA-NET. In second case, the image is not fed into the Mask R-CNN model, and the user creates heuristic sketches without any prior knowledge of the segmentation results. So, the HAISTA-NET makes the predictions concerning the Human Attention Map.

Chapter 4

EXPERIMENT

For measuring the performance of the models, we report experimental results, including AP metrics, mask/bounding box predictions, and analysis of the PSOB dataset. Additionally, we present the relational analyses for the user's sketch characteristics with curvature type and object scales used in PSOB dataset creation.

4.1 Main Results

We evaluated our HAISTA-NET model with 4118 test annotations of the PSOB dataset. Our results significantly improve over state-of-the-art models such as Strong Mask R-CNN and Mask2Former. The analysis demonstrates apparent progress on mask precision deficiencies (figure 3.2, figure 3.3), like object boundary outflow or under-covering. Results exhibit the Human Attention Map's effectiveness in overcoming mask prediction issues standard models face in small-scale and high-curvature objects. Our experiments support the potential of our architecture for acquiring high-precision segmentation masks (figure 3.4). We demonstrated that HAISTA-NET produced better results than other models in all AP, AP_{50}, AP_S, AP_M, AP_L results. Among Res-Net-50-FPN, Res-Net-101-FPN, Res-Net-152-FPN, and Res-NeXt-101-FPN backbones, we achieved the best performance with Res-NeXt-101-FPN. (see Table 3.1 and Table 3.2). We also report that AP values improve if our model uses the Simple Copy Paste data augmentation method.

We benchmark the HAISTA-NET Strong Mask R-CNN baseline by utilizing output images. HAISTA-NET demonstrates better results with challenging high-curvature objects and small-scale objects (table 4.1).

4.2 Multiple Factor Analysis

In Multiple Factor Analysis [St et al., 1989], we classified the object given in test data concerning scale, curvature, and level of assistance. In order to perform the analysis, we first run the inference with HAISTA-NET. Later on, we retrieved the mask predictions and split results into classified partitions such as small-scale objects with low-curvature and having major user assistance and large-scale objects with medium-curvature and having minor user assistance.

We conducted a factorial ANOVA (two-way) to compare the main effects on size, curvature, and user assistance. We also compared the correlated interaction effects on size $\otimes$ curvature, size $\otimes$ assistance, and curvature $\otimes$ assistance. A statistically significant difference exists between size, curvature, and assistance at p-value $(p) < 0.05$. The main effect of size (F (2,5) = 123.59, $p = 0.001$), curvature (F (2,5) = 9.13, $p = 0.021$), and assistance (F (2,5) = 9.21, $p = 0.021$) in $\mathrm{AP}_{\mathrm{Mask}}$ are significant such that objects having large size, low curvature, and major or medium assistance received higher scores. For $\mathrm{AP}_{\mathrm{Mask}}$ analysis, there is a significant interaction between size and curvature (F (4,5) = 14.05, $p = 0.006$) as well as curvature and assistance (F (4,5) = 5.28, $p = 0.048$). The main effect of size (F (2,5) = 16.04, $p = 0.007$), curvature (F (2,5) = 6.23, $p = 0.044$), and assistance (F (2,5) = 9.87, $p = 0.018$) in $\mathrm{AP}_{\mathrm{Bbox}}$ are significant such that objects have a large or medium size, high or low curvature, and minor or medium assistance received higher scores. For $\mathrm{AP}_{\mathrm{Bbox}}$ analysis, there is a significant interaction between size and curvature (F (4,5) = 15.82, $p = 0.005$), size and assistance (F (4,5) = 10.54, $p = 0.012$), as well as curvature and assistance (F (4,5) = 7.16, $p = 0.027$) (figure 4.1).

4.3 Curvature-Based Average Precision

We report the results of the average precision (AP) values of objects in all scales according to curvature classification. In order to analyze the results of curvature-based AP, we split the objects of the PSOB test set considering low curvature, medium curvature, and high curvature types. Also, we compare HAISTA-NET AP results with Strong Mask R-CNN. For $\mathrm{AP}_{\mathrm{low\text{-}curvature}}$, $\mathrm{AP}_{\mathrm{medium\text{-}curvature}}$ and $\mathrm{AP}_{\mathrm{high\text{-}curvature}}$, which

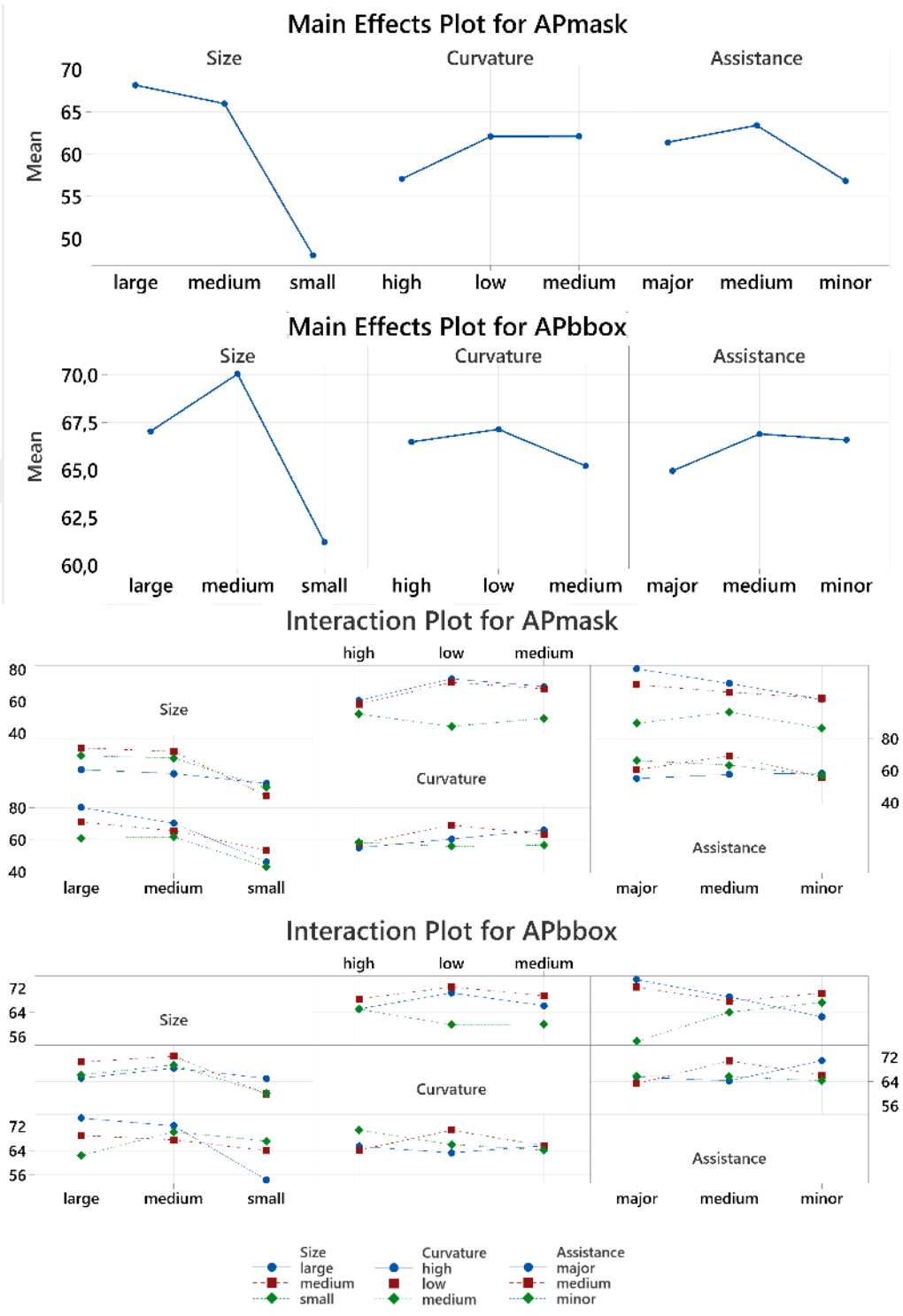

Figure 4.1: The graphs demonstrate the main and interaction effects of AP_{Mask} and AP_{Bbox} values generated using HAISTA-NET.

demonstrates that HAISTA-NET is more successful than Strong Mask R-CNN for all curvature types (table 4.1). Curvature-based average precision analyses do not exist in the literature. We propagate these analyses as C-AP to image recognition tasks in the literature.

Table 4.1: Results of the AP values of the models in different curvature types. M is Mask R-CNN w/SCP + Res-50-FPN, H_1 is HAISTA-NET w/SCP + Res-50-FPN, and H_2 is HAISTA-NET w/SCP + X-101-FPN

	Mask		
Model	$AP_{low\text{-}curvature}$	$AP_{med.\text{-}curvature}$	$AP_{high\text{-}curvature}$
M	25.5	21.4	21.4
H_1	58.1	53.3	50.4
H_2	**61.2**	**58.2**	**51.9**

4.4 PSOB Interaction Time Analysis

We annotated 18677 objects from 115 categories in three scales (small, medium, and large). This data is collected from 30 users. We store critical numerical metrics such as total interaction time, area, perimeter (length) of segmentation polygon (mask), length of sketch input, percentage of covering object boundaries with the sketch, stroke count, and sketching time except latency. One of the most crucial factors of this analysis is the time benchmark. Since our model takes extra user input during training and inference, shorter interaction is better. We developed sketch-based interaction considering this situation. Time analysis shows that total interaction time is ≈3 times more than sketching time because interaction time has latency. The human reasoning process causes an extension of time for detecting an object's exact boundaries. Furthermore, we make multiple regression [Berry et al., 1985] analysis to determine which factors affect total interaction time. We set total interaction time as a response (dependent variable) and the object's curvature number, length of segmentation polygon, stroke count, and length of sketch input

as predictors (independent variables). As a result, all the independent variables significantly predict total interaction time (table 4.2).

4.5 Multiple Regression Model

Since our PSOB dataset has continuous values, Regression analyses show the factors that affect the dependent variable. In this dataset, we set the total time sketch (sec) as the dependent variable and determine which conditions affect this variable. During the labeling process, we collect partial annotation time, stroke count, area of object, perimeter of object, and total pixels of sketched fields. We use the multiple regression technique because it has one dependent and multiple independent variables. The regression analysis demonstrates that stroke count, total pixels of sketch, perimeter of the object, and edge count of the object affect the sketching time because the P-value of the object is less than 0.05. In order to determine whether the regression model is valid, the R^2 value is checked. The R^2 of the current model is 76.49%, which means these independent variables significantly predict the dependent variable. Furthermore, our regression model generates graphs demonstrating outliers and unequal error variances (figure 4.2).

Table 4.2: PSOB dataset stores the pieces of information about sketched objects. LS/PP represents "Length of Sketch" over "Perimeter of Polygon." This ratio represents the number of pixels that are drawn on the object's boundaries as a percentage.

	Train	Validation	Test
Interaction Time [sec]	7.2	8.3	8.4
Sketching Time [sec]	2.0	3.2	3.3
No. Of Curvatures	7.6	7.5	8.0
Perimeter of Polygon	695.4	679.6	688.2
LS/PP (%)	19.9	31.2	31.2
Stroke Count	6.5	6.2	6.0
Annotation Count	10450	4109	4118

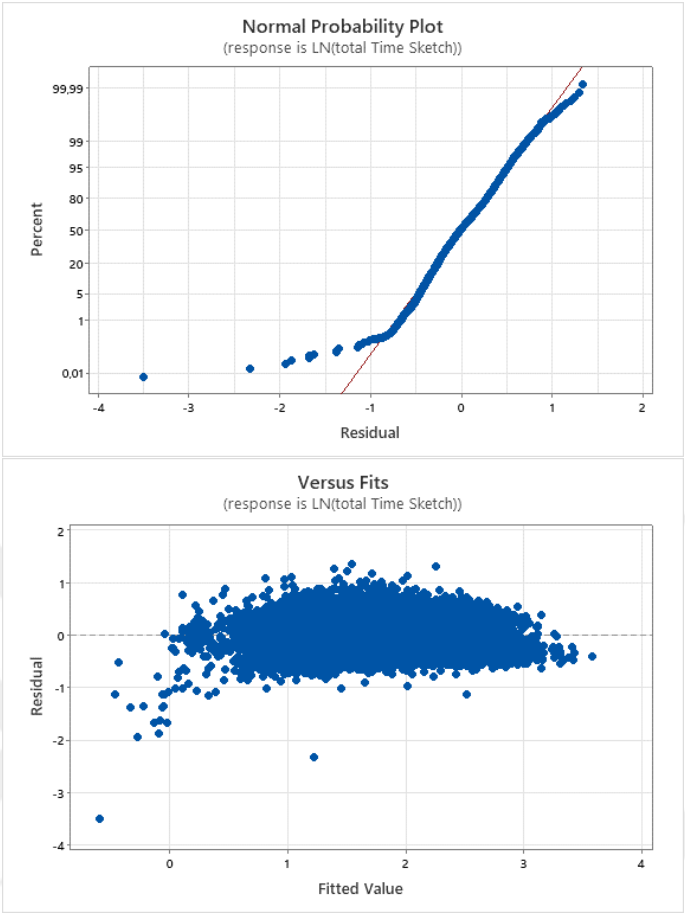

Figure 4.2: The graphs demonstrate the differences between observed and predicted values, unequal error variances, and outliers of the sketching time.

4.6 User Study: PSOB vs Manual Annotation Methods

Users can annotate objects manually if the model output is not satisfying. However, manual annotation is time consuming compared to PSOB's partial annotation. Particularly, retrieving high-detail object polygons like LVIS [Gupta et al., 2019] annotations requires extra attention and focus. Since LVIS contains fine-grained polygons of object boundaries, we were curious about interaction time for annotating an object's mask with hand-drawn sketches. In order to perform this experiment, we collected 643 objects in different scales (small, medium, and large) to analyze average interaction time and mIOU values of LVIS-Like sketches and rough sketches. Then we compared the results with PSOB average interaction time (table 4.3) and

mIOU value (table 4.4). The following steps are specified for performing the study:

- Users draw the boundaries of ground truth masks with fine-grained characteristics like LVIS annotations and interaction time[sec] is recorded.
- Users draw the boundaries of ground truth masks as a rough sketch and interaction time[sec] is recorded.
- We feed the model with Human Attention Map from PSOB dataset to determine how close our model results to LVIS mask.

Results can be summarized as follows;

- LVIS-Like interactions have the best quality but take the longest time with highest effort.
- Rough sketches are fast but do not cover the curvatures of an object due to its characteristics.
- Since PSOB annotations are partial data, they have the shortest time. As a result, we acquire mask qualities closest to LVIS-Like, without complete manual annotation when we feed the model with Human Attention Map.

Object		Average Sketch Time [sec]		
Scale	Number	LVIS-Like	Rough	PSOB
small	27	257	15	9
medium	199	116	17	8
large	238	117	16	9

Table 4.3: Results of the average sketch time differences of LVIS-Like, Rough, and PSOB sketches for different scale objects.

	mIOU		
Interaction Type	Small	Medium	Large
LVIS-Like	99%	97%	96%
Rough	91%	82%	78%
PSOB	98%	93%	91%

Table 4.4: Results of the mean intersection over union (mIOU) values of LVIS-Like, Rough, and PSOB sketches for different scale objects.

4.7 Comparison with SOTA Interaction Method

HAISTA-NET method can become an accuracy improver for any segmentation architecture with the help of Human Attention Maps (HAM). It has more potential than click-based interactive segmentation methods because of its easy implementation and more accurate results with less annotation cost (table 4.5). As a matter of fact, HAM is indispensable for HAISTA-NET, it is the intended usage. Removing the HAM data from the HAISTA-NET method devolves it into a traditional three-channel architecture like Mask R-CNN. HAISTA-NET always takes four input channels (RGB+HAM), but Mask R-CNN and Mask2Former take three.

Interactive Instance Segmentation methods like RITM [Sofiiuk et al., 2022] necessitate or call for additional segmentation model outputs to correct segmentation mask errors. On the other hand, HAISTA-NET does not use a mask obtained from an additional model. Instead, it uses human inputs at first to improve the performance of Mask R-CNN. For this reason, we base our study on the more convenient usage of HAISTA-NET versus other conventional approaches, including the mouse click technique. Evaluation of HAISTA-NET and click-based interactive segmentation methods differ respectively; while HAISTA-NET relies on COCO AP metrics, other interactive segmentation methods - that we studied in the paper- rely on the NoC (number of click-based mIoU evaluation) metric. Moreover, click-based methods are time-consuming compared to HAISTA-NET because they need more user

interaction. Our method requires no additional explanations such as: "Label the missing object parts with the left-click as positive and label the boundary outflows with the right-click as negative."

Our approach does not require users to draw complete precise boundaries of an objects with HAISTA-NET. Our empirical experiment results on annotation time, and boundary labeling accuracy supports the case where HAM pan out with higher mask IoU in less annotation time compared to results of click-based interactive segmentation. Results to validate the interactive segmentation fail in small objects: We made an experiment to compare our results with RITM [Sofiiuk et al., 2022] click-based interactive segmentation method using Mask R-CNN outputs of test images of PSOB dataset. When we try to make corrections of inaccurate masks of Mask R-CNN outputs with RITM, we achieve a much longer annotation time (Table 4.5) than Human Attention Maps. Also, if Mask R-CNN or other external segmentation model does not predict the mask of an object that is interested, RITM [Sofiiuk et al., 2022] will find no way out.

Table 4.5: **Results of the average interaction time differences and mIoU of HAISTA-NET and RITM for different scale objects.** For RITM, we only allowed a maximum of one hundred mouse clicks per object (positive + negative). For HAISTA-NET, the number of maximum allowed strokes is seven. Interaction time is limited to thirty seconds maximum for both methods.

Object		Avg. Int. Time [sec]		mIoU	
Scale	Number	HAISTA-NET	RITM	HAISTA-NET	RITM
small	27	9	18	98%	86%
medium	199	8	21	93%	82%
large	238	9	23	91%	77%

Chapter 5

DISCUSSIONS, INTERPRETATIONS AND LIMITATIONS

This section discusses the research interpretations, general evaluations, comparisons with previous research, and its usability. We discuss the difficulty/ease of segmenting with HAISTA-NET and the limitations of its use. We will describe how practitioners reacted when compared to previous interactive segmentation methods.

5.1 General Discussion

Considering the results of all these experiments, we see that fully automated instance segmentation methods need further development. We have shown in experiments that our idea, combined with fully automatic segmentation methods, makes a significant difference. In particular, even the representation of heuristically drawn parts with a few pixels allowed us to predict unreachable masks. We started with thirty practitioners to conduct the experiments and create the dataset. All of these thirty practitioners were very comfortable drawing small objects and did not have any difficulties. Even if the vertices of the small objects were complex, the difficulty level did not change. The most challenging objects for the practitioners were those with large scales and high curvature. Since the object was large and had much curvature, people wondered from which point to start drawing. This consideration time costs practitioners at least two seconds. HAISTA-NET obtained more accurate masks in about 7 seconds on average. This time is for small, medium, and large-scale objects. The object's scale is the most crucial factor affecting the drawing time. Our statistical analysis proved this. On the other hand, curvature and assistance types were also among the factors directly affecting the time. However, these factors are less effective than the scale factor. The most important factor affecting the duration is

scale.

5.2 Interpretation of Usage

The accuracy of the mask is ultimately related to the input from the practitioner and the object's scale. Also, it is essential that the practitioner draws on the boundaries. Because in HAISTA-NET's training, all user inputs are on the object's boundaries. If irrelevant places are marked instead of marking the boundaries of the object, the mask estimation will be seriously distorted. Our research idea is based on pen-based UI, so we must compare our method with its main competitor, the mouse. The interface developed to run HAISTA-NET can also work with a mouse. However, drawing with a pen is more precise. Pens are usually used with touch tablets or graphics tablets. These devices are designed for pen interaction. A mouse is a common input device for general computing. It is, therefore, not as sensitive as a pen. For this reason, scribbling with a mouse is not considered a feasible method. Interaction with the mouse is usually click-based. Most research at this level has been done to combine segmentation with mouse clicks. Our proposed approach explains that the points given with the mouse do not represent the boundaries well enough, and when we experimented with the mouse click-based system, we observed that the number of clicks increased very much. This issue did not satisfy the practitioner. Because when we want to segment a large-scale object, there are too many positive (+) and negative (-) encodings. This situation needs to be clarified for the user.

5.3 UI Tool Limitations

The annotation interface, the heart of HAISTA-NET, allows practitioners to draw the boundaries of the objects. Tablet computers are generally preferred in order to use pen-based UI. While creating this idea, we collected all the data with a graphics tablet. We used the Wacom brand Cintiq 27QHD (figure 5.1) touch product as a graphics tablet. All data was collected through the Wacom tablet. For practitioners who will use HAISTA-NET, we recommend using a graphics tablet. In this case, it is possible to use a mouse for those who need access to the tablet, but we have

doubts about the efficiency. As mentioned in the previous section, scribbling with a mouse takes work. HAISTA-NET has certain limitations for efficient use. These are the minus side of our idea, but we should also consider that the use of tablets has increased intensively in recent times. The screen of standard tablets may not be as sensitive as graphics tablets. This state can cause difficulties with fine-grained details or color transitions. Tablet screens are limited in size and may not give the required space to segment objects. Zooming in or changing the scale may be needed. Also, some tablets may have less pressure sensitivity than others. This issue can make it challenging to control strokes or thickness.

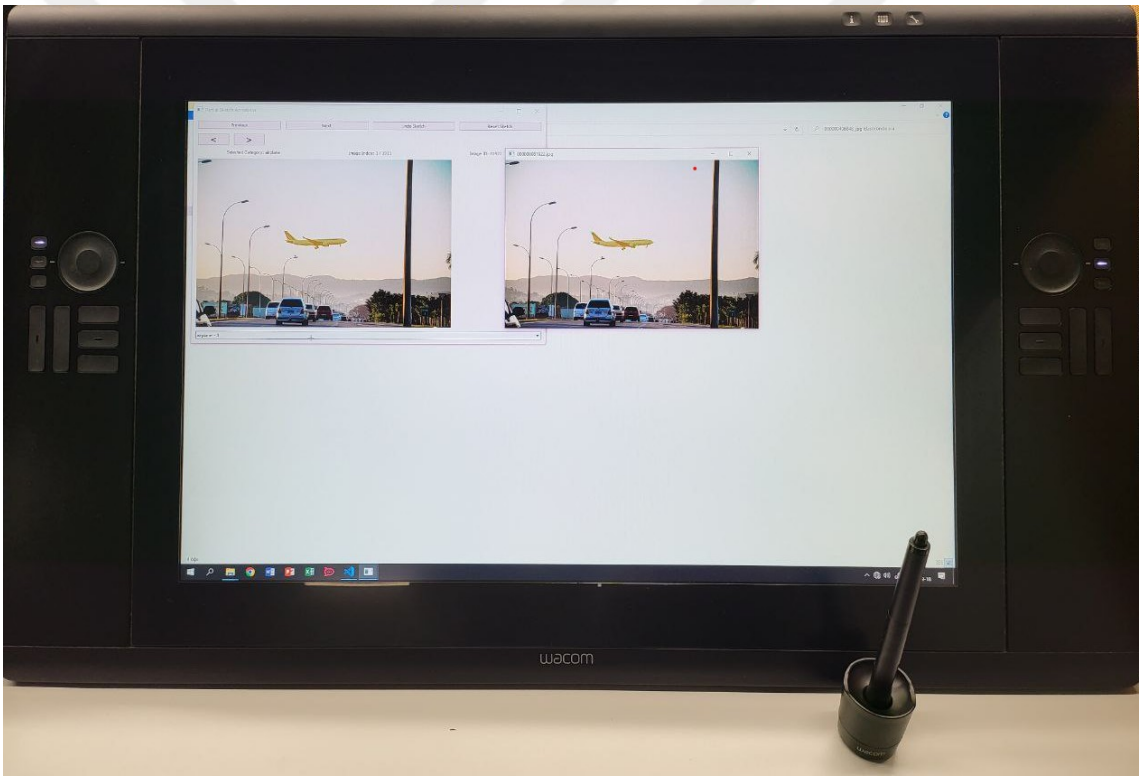

Figure 5.1: The graphics tablet, where the PSOB dataset is collected, helps draw boundaries of the object with its wide screen and high sensitivity.

5.4 Impact on Practitioners

We had the practitioners of the PSOB dataset, which is required to train HAISTA-NET, perform a test. We asked them whether the segmentation mask resulting from

this test was satisfactory and request them to answer. In the experiment with 161 test objects, the general comments were that they would prefer to use HAISTA-NET. The following questions were sent to the practitioners;

1.) How simple is interactive segmentation with HAISTA-NET?

[•] Extremely simple

[•] Simple

[•] Neutral

[•] Difficult

[•] Extremely difficult

2.) What is the quality of the segmentation mask produced by HAISTA-NET?

[•] Extremely good

[•] Good

[•] Neutral

[•] Poor

[•] Extremely poor

3.) Which object scale is the most comfortable for you to draw?

[•] Small

[•] Medium

[•] Large

[•] Nothing

4.) How does the mask behave when multiple objects are overlapped?

[•] Found them separately

[•] Found but insufficient

[•] Nothing found

5.) How is drawing with the current Wacom graphics tablet?

[•] Extremely comfortable

[•] Comfortable

[•] Neutral

[•] Uncomfortable

[•] Extremely uncomfortable

6.) You drew about three different objects per person; how was the output produced after drawing?

[•] Very satisfied

[•] Satisfied

[•] Neutral

[•] Dissatisfied

[•] Very dissatisfied

7.) Would you prefer to use HAISTA-NET instead of fully automatic segmentation methods?

[•] Absolutely yes

[•] Yes

[•] Neutral

[•] No

[•] Absolutely no

8.) Is the HAISTA-NET interface difficult to use with the mouse?

[•] Extremely difficult

[•] Difficult

[•] Neutral

[•] Simple

[•] Extremely simple

According to the answers, HAISTA-NET generally produces high-quality masks and is easy to use. In addition, the mask outputs were higher than the users' expectations. The majority of users stated that they were more comfortable drawing small objects. They specify that annotation with a mouse is not proper.

5.5 Real-World Adaptation

The primary benefit of HAISTA-NET will be in the area of automatic data labeling and image editing. Automatic labeling tools are available in existing labeling interfaces. This state is enabled by applying deep learning based segmentation and object detection methods. However, mask generation problems, which are the main topic of this thesis, also exist here. HAISTA-NET can help in automatic labeling and mask retrieval with minimal user interaction. In this way, manual annotation will come to an end.

Chapter 6

CONCLUSION

Finding correct boundaries for small scale objects and objects having high curvature points is a challenging task in instance segmentation. In this study, we presented a novel approach, HAISTA-NET, and Human Attention Maps that achieves significant improvements in these areas compared to current fully automated state-of-the-art algorithms. Our method requires a minimal amount of input from users and does not increase training and inference time in any noticeable manner. Moreover, we also present a new dataset, PSOB (Partial Sketch Object Boundaries), that combines Human Attention Maps with LVIS dataset for instance segmentation. Our user-friendly interface brought a new perspective to annotation and interaction techniques. We also provide an extensive analysis on factors affecting the performance of our architecture and state-of-the-art methods. Multiple Factor Analysis, with AP_{Bbox} and AP_{Mask} metrics, shows the most affecting factors.

HAISTA-NET is easy to use and extendable to other Computer Vision tasks such as Object Detection, Panoptic Segmentation etc. It can be easily implemented into other Computer Vision applications that use CNNs with minimal amount of user input. We hope that evidence we provided in this study encourages other authors to incorporate HAISTA-NET in their own research to improve performance.

Chapter 7

FUTURE WORK

In the future, we aim to upgrade HAISTA-NET to a transformer-based architecture. We are closely following the recent popularity of transformer-based architectures. We have explained the problems with transformers in the related work section, and it is not possible to work fully effectively with current GPUs. Interactive segmentation methods should produce fast results and be user-oriented. In order to increase user focus and make the architecture work fast, we will serve HAISTA-NET with a distributed GPU infrastructure. Thus, HAISTA-NET will become a tool that is easily accessible to users, including touchscreen mobile devices. Intelligent segmentation will be possible via touchscreen tablets, graphics tablets, smartphones, and personal computers. After implementing our research idea, Meta AI Research introduced the Segment Anything [Kirillov et al., 2023] framework, whose paper will be officially published soon. Segment Anything is an interactive segmentation model, indicating that we are working in the right field. Therefore, we must benchmark the results against the Segment Anything framework. Segment Anything, like its predecessors, is mouse-based, but the dataset it is trained on is extensive and focuses only on mask generation. With the prompt encoder architecture developed, bounding boxes and points can be given as input. We also aim to build a sketch-based prompt encoder and increase the contents of the PSOB dataset (Currently collected about twenty thousand objects). With HAISTA-NET, we believe that data labeling or annotation processes will be easier to manage, so we will make it a platform accessible to everyone. We will furthermore make mask generation more effective, particularly vital for medical image analysis. For this reason, HAISTA-NET will be trained with the medical image data in addition to the daily life data.

BIBLIOGRAPHY

[Acuna et al., 2018] Acuna, D., Ling, H., Kar, A., and Fidler, S. (2018). Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*.

[Andriluka et al., 2018] Andriluka, M., Uijlings, J. R., and Ferrari, V. (2018). Fluid annotation: a human-machine collaboration interface for full image annotation. In *ACM Multimedia*.

[Benard and Gygli, 2017] Benard, A. and Gygli, M. (2017). Interactive video object segmentation in the wild. *arXiv preprint arXiv:1801.00269*.

[Benenson et al., 2019] Benenson, R., Popov, S., and Ferrari, V. (2019). Large-scale interactive object segmentation with human annotators. In *CVPR*.

[Berry et al., 1985] Berry, W. D., Berry, W. D., Feldman, S., and Stanley Feldman, D. (1985). *Multiple regression in practice*. Sage.

[Bolya et al., 2019] Bolya, D., Zhou, C., Xiao, F., and Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In *ICCV*.

[Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV*.

[Chen et al., 2018] Chen, L.-C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., and Adam, H. (2018). Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*.

[Cheng et al., 2022] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *CVPR*.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Douglas and Peucker, 1973] Douglas, D. H. and Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*.

[Du et al., 2021] Du, X., Zoph, B., Hung, W.-C., and Lin, T.-Y. (2021). Simple training strategies and model scaling for object detection. *arXiv preprint arXiv:2107.00057*.

[Fang et al., 2019] Fang, H.-S., Sun, J., Wang, R., Gou, M., Li, Y.-L., and Lu, C. (2019). Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *ICCV*.

[Fang et al., 2021] Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., and Liu, W. (2021). Instances as queries. In *ICCV*.

[Ghiasi et al., 2021] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*.

[Gupta et al., 2019] Gupta, A., Dollar, P., and Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*.

[Han et al., 2021] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. In *NeurIPS*.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *ICCV*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

[Kirillov et al., 2023] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything. *arXiv:2304.02643*.

[Kirillov et al., 2020] Kirillov, A., Wu, Y., He, K., and Girshick, R. (2020). Pointrend: Image segmentation as rendering. In *CVPR*.

[Li et al., 2004] Li, Y., Sun, J., Tang, C.-K., and Shum, H.-Y. (2004). Lazy snapping. *ACM Transactions on Graphics (ToG)*.

[Li et al., 2018] Li, Z., Chen, Q., and Koltun, V. (2018). Interactive image segmentation with latent diversity. In *CVPR*.

[Liew et al., 2017] Liew, J., Wei, Y., Xiong, W., Ong, S.-H., and Feng, J. (2017). Regional interactive image segmentation networks. In *ICCV*.

[Liew et al., 2019] Liew, J. H., Cohen, S., Price, B., Mai, L., Ong, S.-H., and Feng, J. (2019). Multiseg: Semantically meaningful, scale-diverse segmentations from minimal user input. In *ICCV*.

[Lin et al., 2016] Lin, D., Dai, J., Jia, J., He, K., and Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*.

[Lin et al., 2017] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *CVPR*.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

[Olsen et al., 2009] Olsen, L., Samavati, F. F., Sousa, M. C., and Jorge, J. A. (2009). Sketch-based modeling: A survey. *Computers & Graphics*.

[Papadopoulos et al., 2017] Papadopoulos, D. P., Uijlings, J. R., Keller, F., and Ferrari, V. (2017). Extreme clicking for efficient object annotation. In *ICCV*.

[Qiao et al., 2021] Qiao, S., Chen, L.-C., and Yuille, A. (2021). Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*.

[Razzak et al., 2018] Razzak, M. I., Naz, S., and Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making*.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

[Sofiiuk et al., 2022] Sofiiuk, K., Petrov, I. A., and Konushin, A. (2022). Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*.

[St et al., 1989] St, L., Wold, S., et al. (1989). Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems*.

[Tang et al., 2021] Tang, C., Chen, H., Li, X., Li, J., Zhang, Z., and Hu, X. (2021). Look closer to segment better: Boundary patch refinement for instance segmentation. In *CVPR*.

[Tian et al., 2021] Tian, Z., Shen, C., Wang, X., and Chen, H. (2021). Boxinst: High-performance instance segmentation with box annotations. In *CVPR*.

[Wang et al., 2018a] Wang, G., Li, W., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S., et al. (2018a). Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*.

[Wang et al., 2018b] Wang, G., Zuluaga, M. A., Li, W., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S., et al. (2018b). Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Wang et al., 2020] Wang, X., Zhang, R., Kong, T., Li, L., and Shen, C. (2020). Solov2: Dynamic and fast instance segmentation. In *NeurIPS*.

[Xie et al., 2017] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *CVPR*.

[Xu et al., 2016] Xu, N., Price, B., Cohen, S., Yang, J., and Huang, T. S. (2016). Deep interactive object selection. In *CVPR*.