2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)

Open-set image segmentation based on an improved Segment Anything model

1st Haiyang Wu
School of electronic information
Xijing University
Xi'an, Shaanxi
2010530510@qq.com

2nd Lingyun Kong*
School of electronic information
Xijing University
Xi'an, Shaanxi
1400100383@qq.com*

3rd Zhipeng Xue
School of electronic information
Xijing University
Xi'an, Shaanxi
3213924681@qq.com

4th Denghui Liu
School of electronic information
Xijing University
Xi'an, Shaanxi
1470824183@qq.com

5th Songyang Li
School of electronic information
Xijing University
Xi'an, Shaanxi
1165434445@qq.com

6th Qi Zhang
School of electronic information
Xijing University
Xi'an, Shaanxi
1409163868@qq.com

2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI) | 979-8-3315-4169-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/RICAI64321.2024.10911698

***Abstract*—Image segmentation is crucial in computer vision, but traditional approaches are limited by dataset and label diversity. Segment Anything introduces open-set segmentation capabilities but struggles with clause-level prompts and precise mask edges. This paper presents SAM-TP, a new algorithm combining Grounding DINO for clause-level prompt handling and Segment Anything for segmentation. By integrating non-maximum suppression and connected component denoising, we refine segmentation accuracy. Additionally, incorporating ViT-Matte enhances edge precision through a trimap-based structure. Experiments on the COCO2017 dataset show that SAM-TP improves Mean Pixel Accuracy by 6.04%, achieving 24.74%, demonstrating its enhanced foreground-background segmentation ability.**

***Index Terms*—Segment Anything, Grounding DINO, ViT-Matte, Image Segmentation, Open-set**

I. Introduction

Image segmentation is a classic challenge in computer vision research and has become a focal point in image understanding. Serving as the first step in image analysis, image segmentation forms the foundation of computer vision and is a crucial component of image understanding. It is also one of the most difficult tasks in image processing. Image segmentation involves dividing an image into multiple non-overlapping regions based on features like grayscale, color, spatial texture, and geometric shape, ensuring feature consistency or similarity within the same region and distinct differences across regions.

In fields such as intelligent security, autonomous driving, satellite remote sensing, and medical image processing, image segmentation provides straightforward and reliable image feature information, significantly enhancing the efficiency of subsequent vision tasks. In practical applications, depending on the specific context, it is necessary to flexibly choose different segmentation methods to meet the requirements of various tasks.

Early image segmentation tasks mainly focused on distinguishing between background and target regions, where the substantial contrast between the two often allowed for simpler segmentation methods. Traditional segmentation methods include widely used and classic approaches based on thresholding [1], edges [2], regions [3], clustering [4], and specific theoretical frameworks [5] [6].

Threshold-based image segmentation [1] works by setting different grayscale thresholds to classify the grayscale histogram of an image, treating pixels within the same grayscale range as belonging to the same category, indicating a certain similarity. This method is suitable for images with a uniform grayscale distribution in the target and a significant difference in grayscale between the foreground and background. However, it only considers pixel grayscale values without incorporating semantic or spatial features, making it susceptible to noise. Thus, threshold segmentation is less effective for complex images.

Edge-based image segmentation [2] can be categorized into two methods depending on the strategy: serial edge detection and parallel edge detection. In serial edge detection, the process starts by identifying edge initiation points, which are then linked to adjacent edge points through similarity searches. Parallel edge detection, on the other hand, uses spatial differentiation operators, applying them as templates convolved with the image to perform segmentation. Common differentiation operators include Sobel, Canny, and LoG filters. Since segmentation results vary depending on the operator, even with the same image, the segmentation quality is heavily influenced by the choice of operator, leading to inconsistent results.

Region-based image segmentation [3] leverages spatial information to segment an image by grouping pixels based on similarity, forming distinct regions. Common region-based methods include region growing and split-and-merge techniques. In region growing, similar pixels are aggregated to form independent regions, achieving segmentation. The split-and-merge method divides and merges regions iteratively to

produce image subregions. While region growing is computationally simple, it is sensitive to noise and may result in missing regions. The split-and-merge method is more computationally complex and can potentially disrupt boundary integrity. Both methods have advantages and limitations.

Clustering-based segmentation [4] groups similar pixels into the same region by iteratively clustering pixels into distinct categories, thus achieving segmentation. However, these traditional methods have significant limitations: they cannot segment specific content, recognize particular foregrounds, or achieve high segmentation precision. They are unsuitable for tasks that require substantial semantic information, thus failing to meet practical needs.

With the advent of deep learning, computer vision has achieved groundbreaking progress, and convolutional neural networks (CNNs) have become a crucial tool in image processing. To address the increasing complexity of segmentation tasks, a series of deep learning-based semantic segmentation methods have been developed.

Traditional deep learning methods for semantic segmentation include four primary models: FCN, PSPNet, DeepLab, and Mask R-CNN.

The Fully Convolutional Network (FCN) [7] was the pioneering deep learning framework for semantic segmentation, establishing a foundational approach for segmenting images through network models. The Pyramid Scene Parsing Network (PSPNet) [8] enhances this by considering contextual information, leveraging global feature priors to analyze various scenes and segment scene objects semantically. The core innovation of the DeepLab model [9] is its use of atrous (or dilated) convolution, where spacing in the convolution kernel increases the receptive field, allowing it to capture more feature information without adding parameters and explicitly control the resolution of feature responses.

Mask R-CNN, developed by He et al. as an extension of Faster R-CNN [10], is a deep convolutional network tailored for image segmentation. It performs both object detection and high-quality segmentation. In the first stage, it uses a region proposal network to identify candidate object bounding boxes. In the second stage, it performs class prediction and bounding box regression to predict each pixel's category, achieving segmentation. Unlike the other models, Mask R-CNN builds upon semantic segmentation to achieve instance segmentation, allowing it to distinguish between individual objects of the same class.

Each of these models has strengths and weaknesses in segmentation tasks, but none can effectively handle multi-task, complex scene segmentation. In 2023, Meta introduced the Segment Anything Model (SAM) [11], a state-of-the-art open-set segmentation model. Trained on an unprecedented dataset of 11 million images and over 1 billion masks, SAM has learned general object concepts, enabling it to generate masks for any object in any image, including unseen objects and image types that were not part of its training set.

II. RELATED WORK

With the rapid development of computer technology, image segmentation has become increasingly crucial across various real-world applications. Traditional image segmentation methods, however, are generally limited to a closed-set environment [12], where both training and testing categories remain fixed. In this setting, models can only recognize known classes encountered during training. Yet, real-world scenarios are often filled with unknown classes, creating a demand for models that can handle open-set segmentation tasks [13], also known as open-vocabulary segmentation. Open-vocabulary segmentation requires that models not only segment known classes from the training set but also handle new, unseen categories, posing significant challenges for existing segmentation methods.

Open-set image segmentation is an extension of traditional segmentation, aiming to enable models to segment objects belonging to previously unseen categories. To meet the challenges of open-set segmentation, researchers have explored various approaches, including Vision-Language Models (VLMs) [14], general feature learning [15] [16], and label hierarchy with semantic space alignment [17]. In recent years, VLMs have seen rapid development and remarkable achievements. By incorporating pretrained text encoders such as CLIP and ALIGN [18] [19], VLMs can align visual information with language concepts, allowing the model to recognize and process unseen categories. These models can perform segmentation tasks guided by text prompts or labels, achieving effective open-set segmentation.

Segment Anything, proposed by Meta AI in 2023, represents a pioneering open-set image segmentation model. Leveraging a data collection loop, the research team developed the largest segmentation dataset to date, containing 11 million images and over a billion masks. This model's design and prompt-based training enable zero-cost adaptation across new image distributions and segmentation tasks.

The authors aim to create a foundational model for image segmentation, one that is promptable and pretrained on a large, diverse dataset to ensure robust generalization. They address three key factors: task, model, and data. To this end, they define a versatile promptable segmentation task as a strong pretraining objective supporting various downstream tasks. This task requires a model that can output real-time segmentation masks based on flexible prompts, facilitating interactive use.

To support zero-shot generalization, the authors develop a promptable task, aiming to output valid segmentation masks for any prompt, including spatial or textual hints. A simple model architecture was designed, comprising an image encoder, a prompt encoder, and a lightweight mask predictor. By decoupling image encoding from prompt encoding, SAM allows efficient reuse of image embeddings for multiple prompts. SAM supports point, box, mask, and free-form text prompts, predicting multiple masks for a single ambiguous prompt, thus handling ambiguity effectively.

To enhance generalization across diverse data distributions,

the authors assembled a large-scale dataset of billions of text-mask pairs. This extensive data enables SAM to generate object subset masks by prompting for object locations, improving SAM's robustness and general applicability while providing valuable resources for training other segmentation models.

The SAM model underwent extensive evaluation by the research team, and the results demonstrated its outstanding performance across various datasets and segmentation tasks. First, the authors utilized a new set of 23 segmentation datasets, finding that SAM could generate high-quality masks from a single foreground point, with performance typically just slightly below the manually labeled ground truth. Second, under the zero-shot transfer protocol, SAM consistently delivered excellent results across various downstream tasks, including edge detection, object proposal generation, instance segmentation, and preliminary explorations into text-to-mask prediction. These findings suggest that, in addition to its training data, SAM can be used in real-time tasks to solve a wide range of problems involving objects and image distributions.

Building on the same concept, open-set object detection tasks have also been proposed, with the representative model being Grounding DINO [20]. This model combines a Transformer-based detector (DINO) with ground truth pre-training, enabling the detector to detect any object based on human input (such as category names or indicative expressions). The key solution for open-set object detection is integrating language into a closed-set detector for open-set concept generalization. To effectively fuse the language and visual modalities, the paper conceptually divides the closed-set detector into three stages and proposes a tightly integrated solution, including a feature enhancer, language-guided query selection, and cross-modal fusion.

Grounding DINO performed exceptionally well across all three stages, including benchmark tests on the COCO dataset, LVIS dataset, ODinW dataset, and RefCOCO/+/g datasets. On the COCO zero-shot detection transfer benchmark, Grounding DINO achieved an AP of 52.5%, even without using any training data from the COCO dataset. After fine-tuning on COCO data, Grounding DINO's AP on the COCO dataset reached 63.0%.

However, the SAM model also has some limitations. The pre-trained language encoder used in SAM is unable to recognize clause-level prompts, which limits its ability to identify more precise locations within images. Moreover, the model still operates at the level of segmenting image categories, meaning that the results primarily rely on the inference and label files from the training set. As a result, the segmentation boundaries are not always precise. Additionally, the inference results from SAM do not generate a single mask but instead output multiple masks simultaneously, which makes it difficult to meet the requirements for batch processing of segmentation tasks. These limitations hinder SAM's effectiveness in certain applications where accurate edge detection and batch processing are crucial.

III. MODEL DESIGN

To address the limitations of the Segment Anything model, this paper proposes the following three improvements:Integrating the Output of Grounding DINO as Input to Segment Anything;Combining the ViTmatte Model to Implement a Trimap-based Structure;Introducing Non-Maximum Suppression (NMS) and Connected Component Filtering.

These three improvements enhance the accuracy, robustness, and batch processing capability of the Segment Anything model, particularly in handling complex images and multi-object scenarios.

A. Grounding DINO + Segment Anything

Segment Anything accepts three types of input prompts: point selection, box selection, and text prompts. When using point or box selection for segmentation of arbitrary images, batch segmentation cannot be performed. To achieve batch segmentation for the same target, text prompts are required. However, the text prompt input encoder used in Segment Anything is the CLIP encoder, which cannot achieve precise segmentation using sentence-based input. Therefore, we propose a variant of Segment Anything that uses box selection as input, where Grounding DINO, as a zero-shot object detection model, can accept sentence-based input as text prompts for localized open-vocabulary object detection.

By using the output from Grounding DINO along with the image as input to Segment Anything, the final segmentation mask can be obtained. For example, in the case of dog detection, suppose we input an image containing two dogs, and we want to generate the mask for the dog on the left. When using only text prompts, the model cannot accurately localize the specific target to be segmented, and thus, it will segment all the dogs in the image. However, by incorporating the Grounding DINO model, we use box selection as input. The text prompt is fed into the Grounding DINO model, which automatically detects the specific location of the object to be segmented and outputs the bounding box information. This bounding box information is then used as input to Segment Anything, allowing for precise segmentation. The specific workflow is illustrated in Figure 1.

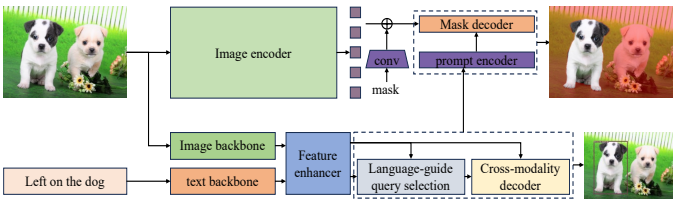

Fig. 1. segment anything + grounding DINO model combined with flowcharts.

B. Integration of ViTmatte Model for Trimap-based Structure

In the field of image segmentation, there are two main refinement methods: Trimap-based [22] and Trimap-free [21] segmentation techniques. Upon reviewing the Segment Anything model, it was observed that the model does not employ

Trimap-based refinement after the image segmentation process. To address this, we developed a local solution using the OpenCV package to generate Trimap images and subsequently applied the ViTmatte [23] model for segmentation refinement. By combining the ViTmatte model with Segment Anything, we constructed a Trimap-based image segmentation model. The specific structure of this model is shown in Figure 2.

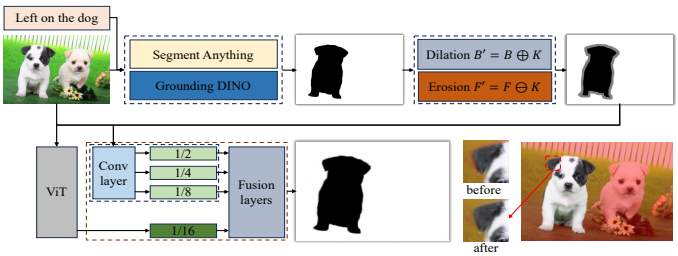

Fig. 2. ViTmatte model combined with flowcharts.

C. Batch Automatic Segmentation of the Model

During prediction, due to the uncertain number of detected objects and the input image resolution, the SAM model may sometimes output multiple prediction boxes. When the coordinates of multiple prediction boxes are input into the Segment Anything model, multiple masks are generated, which prevents the automation of segmentation. To address this issue, after obtaining the output boxes from Grounding DINO, Non-Maximum Suppression (NMS) is first applied to filter out boxes with low classification probabilities and small areas. The output prediction box coordinates are given by the lower-left corner (x_0, y_0) and upper-right corner (x_1, y_1). The NMS formula can be expressed as:

$$\sigma = \max\left(\left|(x_1^k - x_0^k)(y_1^k - y_0^k)\right|\right), \quad k \in (1, n)$$

where n is the total number of prediction boxes, and σ represents the final output prediction box coordinates. By applying NMS, boxes with lower probabilities and smaller areas are filtered out, resulting in a single, more accurate prediction box, thereby automating the object detection process.

Due to varying input image sizes, when the input image is smaller, the output masks may exhibit unclear edges or poor segmentation quality, often resulting in significant noise in the segmentation results. To mitigate this, after obtaining the masks, all connected components within the mask are detected, and any components smaller than 50 pixels are filtered out (assuming these small connected regions are noise). This process provides an initial denoising of the mask.

Based on the three improvements mentioned above, this paper proposes a new open-set image segmentation model, SAM-TP. The process consists of the following steps:

- Input to Grounding DINO:
  The image and prompt are input into the Grounding DINO model. The image encoder and text encoder are used to process the inputs, which are then fed into the object detection model to obtain the predicted bounding box coordinates.
- Non-Maximum Suppression (NMS) on Predicted Boxes:
  The predicted bounding box coordinates are processed using NMS to filter out the boxes with smaller areas and lower prediction probabilities, leaving the box with the largest area and highest probability as the final prediction.
- Segmentation with Segment Anything:
  The original image and the predicted bounding box coordinates are input into the Segment Anything model, performing image segmentation on the selected target.
- Denoising with NMS:
  The resulting segmentation mask is processed using NMS to remove small connected components (less than 50 pixels in size), effectively eliminating noise from the mask.
- Refinement with ViTMatte:
  Finally, the ViTMatte model is applied to further refine the segmentation mask, improving the accuracy and sharpness of the mask boundaries.

The specific process is shown in Figure 3.

IV. Experimental Results and Analysis

The experimental environment is configured as follows: the operating system is Ubuntu 22.04, and the deep learning framework used is PyTorch. The hardware setup includes an NVIDIA GeForce RTX 3060 graphics card with 8 GB of VRAM, an Intel i7-12700H processor, and 16 GB of RAM.

A. Evaluation Metrics

In image segmentation tasks, common evaluation metrics include Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Mean Intersection over Union (MIOU), and Frequency Weighted Intersection over Union (FWIOU). These four metrics are used to evaluate the performance of segmentation models at the pixel level. In this paper, we also incorporate the Kullback-Leibler (KL) divergence to assess the model's performance. The definitions and formulas of these metrics are as follows:

1) Pixel Accuracy (PA): Pixel Accuracy (PA) is one of the simplest image segmentation evaluation metrics. It represents the ratio of correctly classified pixels to the total number of pixels. The formula is:

$$PA = \frac{\sum_i TP_i}{\sum_i (TP_i + FN_i)}$$

Where: - TP_i is the number of pixels correctly classified as class i. - FN_i is the number of pixels incorrectly classified as another class.

2) Mean Pixel Accuracy (MPA): Mean Pixel Accuracy (MPA) is the average of pixel accuracy for each class. It first calculates the pixel accuracy for each class and then computes the mean of those values. The formula is:

$$MPA = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}$$

Where: - N is the total number of classes.

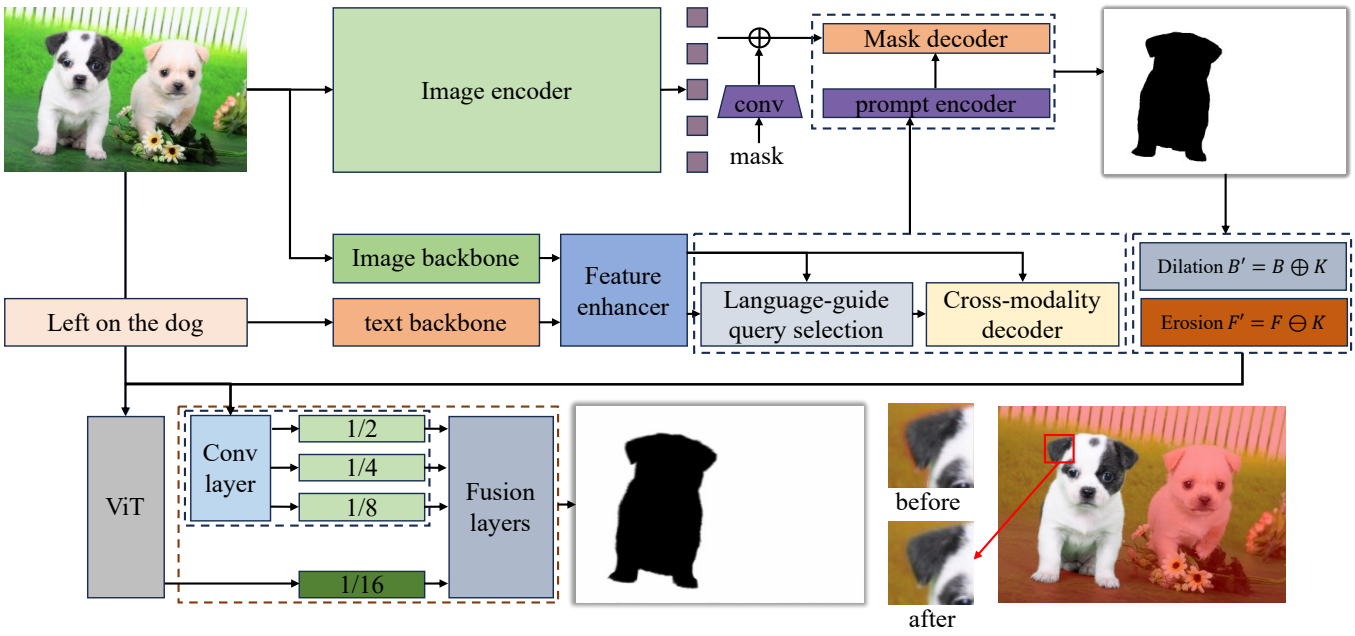

Fig. 3. Network of SAM-TP.

3) Mean Intersection over Union (MIOU): Mean Intersection over Union (MIOU) is one of the most commonly used metrics for image segmentation evaluation. It represents the ratio of the intersection of predicted and ground truth segmentation to the union of the two. The formula is:

$$MIOU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i}$$

Where: - FP_i is the number of pixels incorrectly classified as class i.

4) Frequency Weighted Intersection over Union (FWIOU): Frequency Weighted Intersection over Union (FWIOU) is an extension of MIOU that accounts for the proportion of pixels in each class, giving more weight to frequently occurring classes. The formula is:

$$FWIOU = \sum_{i=1}^{N} \left(\frac{TP_i + FN_i}{\sum_j TP_j + FN_j} \right) \frac{TP_i}{TP_i + FP_i + FN_i}$$

5) Kullback-Leibler Divergence (KL Divergence): Kullback-Leibler divergence (KL Divergence), also known as relative entropy, is a measure of the difference between two probability distributions, P and Q. Although referred to as a "divergence" or "distance", it does not satisfy all properties of a true distance metric (especially symmetry and the triangle inequality), so it is more accurately described as a measure of difference. The mathematical expression for KL divergence is:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

For continuous probability distributions, the expression is given by the integral form:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Where $P(i)$ and $Q(i)$ (or $p(x)$ and $q(x)$ for continuous distributions) represent the probability densities or masses of distributions P and Q at a given point (or interval).

These five metrics help assess the performance of segmentation models and provide insights into the model's performance across different categories and overall accuracy.

B. Ablation Study

In this paper, the improvements proposed focus on three main aspects:

- Integration of the Grounding DINO model with Segment Anything: This allows the model to accept clause-level prompts, enabling more precise object localization.
- Combination with the ViTMatte model to implement a Trimap-based structure: This improves the accuracy of segmentation mask edges.
- Incorporation of Non-Maximum Suppression (NMS) and connected component denoising techniques: This facilitates batch segmentation of masks.

Among these improvements, the integration of clause-level prompts will be demonstrated using visual methods, and batch segmentation via NMS and connected component denoising is mainly an enhancement in the segmentation pipeline, which will not be shown here. The focus of this section is to evaluate the impact of the ViTMatte model in enhancing the Trimap-based structure for segmentation performance.

In this study, we innovatively introduce the ViTMatte model as an enhancement strategy for the Trimap component of the original model, specifically designed for generic object region segmentation tasks. To comprehensively assess the practical effectiveness of this improvement, we have designed a series of comparative experiments using the authoritative COCO 2017 dataset.

COCO (Common Objects in Context) is a flagship dataset developed by Microsoft, widely regarded as one of the most important benchmarks in the field of image recognition. It is particularly known for its expertise in object recognition and understanding in complex scenes. The dataset is vast, covering 91 object categories, over 328,000 images, and 2.5 million finely annotated objects. COCO is renowned for its realistic depiction of everyday scenes and precise segmentation annotations, providing a solid foundation for semantic segmentation research.

In particular, COCO 2017, as a standout in the COCO series, focuses on the recognition and segmentation of 80 common object categories, encompassing over 330,000 images, with 200,000 of them featuring detailed annotations. The total number of annotated instances exceeds 1.5 million, making it one of the largest and most comprehensive datasets in the field of semantic segmentation.

To verify the superiority of the proposed improvement (the integrated model of Segment Anything and ViTMatte), we designed the following experimental process: First, the Segment Anything model was applied to a portion of the training set in the COCO 2017 dataset for inference, successfully generating high-quality masks for 76 categories, totaling 24,000 images. These masks were labeled as "Original Mask." Then, we applied the improved model—an integrated inference system of Segment Anything and ViTMatte—to the same dataset, generating another set of masks for the same 76 categories and 24,000 images. These masks, enhanced with ViTMatte optimizations, were labeled as "Improved Mask."

Finally, we utilized a series of scientifically rigorous evaluation metrics to compare both sets of masks with the original ground truth annotations in the COCO 2017 dataset. The goal was to quantify and clearly demonstrate the significant improvements in object region segmentation accuracy and edge detail optimization brought by our proposed approach. This comparison aims to fully validate the effectiveness and advancement of our improvement strategy.

In our in-depth study, we performed a comprehensive performance evaluation on two independent datasets, each containing 24,000 meticulously annotated mask images, spanning 76 diverse categories. We calculated the specific values for key metrics such as Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Mean Intersection over Union (MIOU), and Frequency Weighted Intersection over Union (FWIOU). Furthermore, we innovatively used visualization techniques to convert these complex data into intuitive heatmaps. By comparing the heatmaps generated from the original and improved models across both datasets, we were able to visually observe performance differences and distribution characteristics. The heatmaps are shown at Figure4 and Figure5:

Fig. 4. Average Metrics per Category (Improved).

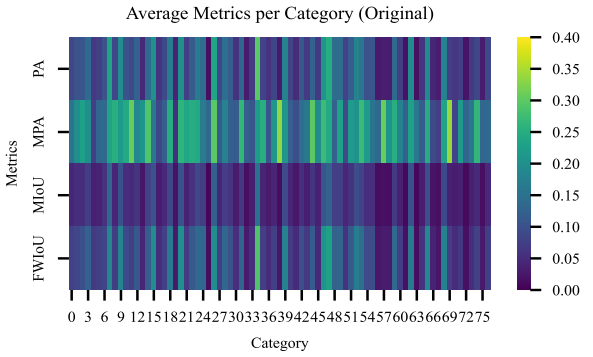

Fig. 5. Average Metrics per Category (Original).

To further quantify and highlight the significant improvements brought by the modified model, we overlaid the two heatmaps and conducted an in-depth comparative analysis. The difference between the two heatmaps is within ±0.1, so the range of the new heatmap is set to 0.1 to -0.1. In the new heatmap, areas that show improvement with the modified model are displayed in brighter, lighter colors, while areas that show a decrease in performance are represented in darker, deeper colors, as shown in Figure 6.

From Figure 6, we can observe that most of the regions are presented with brighter colors, indicating that the proposed improved model outperforms the original model not only overall but also in terms of each category across different metrics. Notably, there are significant improvements in Pixel Accuracy (PA) and Mean Pixel Accuracy (MPA), which further validate the effectiveness of the proposed improvements.

It is important to note that this experiment only demonstrates the improvement of the modified model compared to the original model under the same prompt. A major enhancement in the improved model is the integration of Grounding DINO, which allows the model to accept clause-level prompts and specify the target segmentation areas, rather than performing a general segmentation across the entire image. This leads to more precise segmentation regions.

The heatmap in Figure 6 displays the improvements of both models across 76 categories in the segmentation task, evaluated using four different metrics.

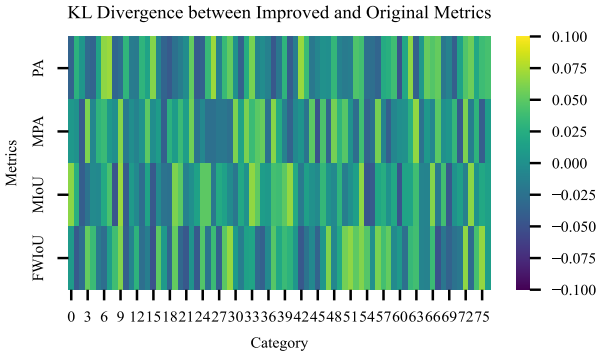

Fig. 6. KL Divergence between Improved and Original Metrics.

To provide a more comprehensive and intuitive presentation of the improvement, we further calculated the average values of all categories across the four evaluation metrics and displayed the comparison in the form of bar charts, as shown in Figure 7. This chart clearly illustrates the leap in overall performance achieved by the improved model, particularly with a significant increase in Mean Pixel Accuracy (MPA), a key metric.

This result not only strongly validates the effectiveness and rationality of the proposed improvements but also provides solid data support and theoretical foundation for future research and applications.

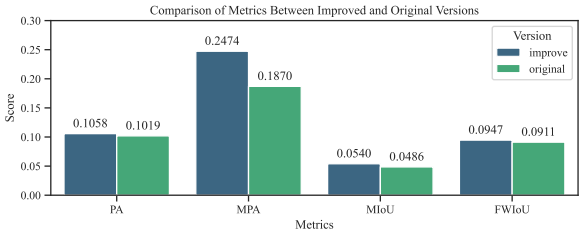

Fig. 7. Comparison of Metrics Between Improved and Original Versions.

Upon further analysis of the data presented in the bar chart, we observed that while the overall improvements in Pixel Accuracy (PA), Mean Intersection over Union (MIOU), and Frequency Weighted Intersection over Union (FWIOU) appeared to be subtle at first glance, a more precise and detailed analysis was necessary to capture and display the incremental yet comprehensive improvements brought about by the proposed strategy.

Specifically, we calculated the Kullback-Leibler (KL) Divergence values for both the Original Mask and the Improved Mask datasets relative to the ground truth masks. This step not only considered the difference between predicted and true masks but also quantified the degree of this difference using KL divergence, an information-theoretic measure. Subsequently, we innovatively visualized these KL divergence values in 3D space, creating two separate 3D plots,as shown in Figure 8 and Figure 9.

To make the comparison more intuitive, we cleverly overlaid the two 3D plots and differentiated the regions using color encoding: regions where the improved model exhibited significant optimization (i.e., a reduction in KL divergence) were highlighted in red, symbolizing progress and enhancement. On the rare occasions where there was a slight decline in performance (although this was almost non-existent in our study), we marked them in blue for distinction.

When the two 3D plots were overlaid and merged, it became evident that the red regions dominated the majority of the space, virtually eclipsing the occasional blue dots. This stark contrast vividly demonstrated that the proposed improvements led to a comprehensive and substantial optimization across multiple key metrics, including PA, MIOU, and FWIOU. The results are shown in Figure 10.

This finding not only deepened our understanding of the performance improvements brought about by the enhanced model but also provided solid data support and confidence for future research directions.

C. Experimental Results Visualization

To verify the effectiveness of the proposed model improvement by combining Grounding DINO and Segment Anything, this study selects three different prompts for multi-object image segmentation tasks. The chosen prompts are ”cat on the left,” ”dog on the left,” and ”girl in the middle.” Experiments were conducted using both the Segment Anything model and the combined Segment Anything + Grounding DINO model to evaluate the results. The experimental outcomes are shown in the Figure 11, Figure 12 and Figure 13.

As shown in the above figures, the Segment Anything model is unable to recognize the location-based terms in the prompts, meaning it can only identify individual words but not comprehend the entire sentence. However, after integrating the Grounding DINO model, the system is able to segment specific targets within the designated area, enabling the model to perform clause-level prompt-based segmentation.

Furthermore, after incorporating the ViTMatte model to refine the segmentation masks, the improvement in segmentation quality is quite evident. The specific results are shown in the Figure 14, Figure 15 and Figure 16.

As seen in the above figure, after incorporating ViTMatte, the edges of the segmentation become much clearer. It is now able to more precisely delineate the foreground, such as hair and skin edges, thus improving the distinction between the foreground and background. Compared to the original model, the segmentation performance of the improved model is significantly better.

V. CONCLUSION

Based on an in-depth analysis of existing image segmentation methods, we propose a new image segmentation approach, SAM-TP, which accepts more accurate prompts and produces clearer edge segmentation. SAM-TP addresses two key

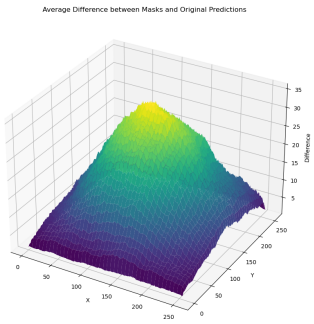

Fig. 8. Average Difference between Masks and Original Predictions

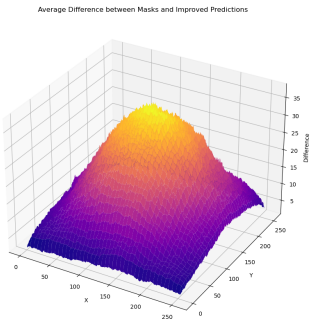

Fig. 9. Average Difference between Masks and Improved Predictions

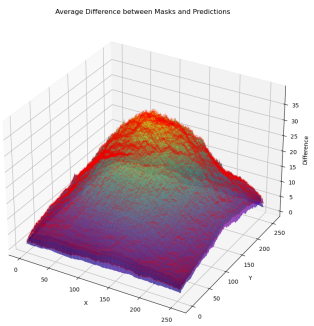

Fig. 10. Average Difference between Masks and Predictions

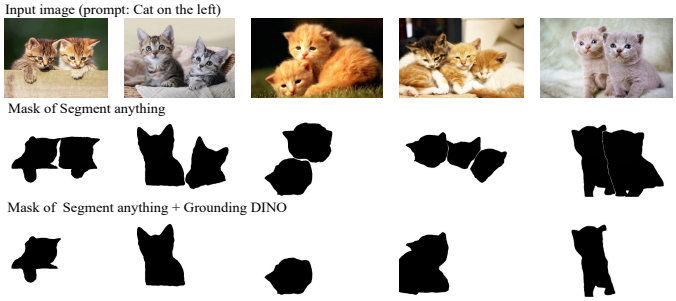

Fig. 11. segment anything + Grounding DINO visualization of experimental results(cat).

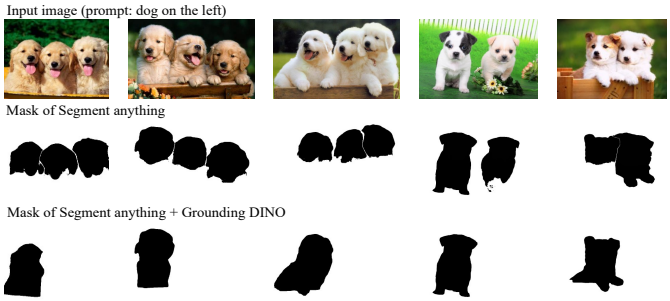

Fig. 12. segment anything + Grounding DINO visualization of experimental results(dog).

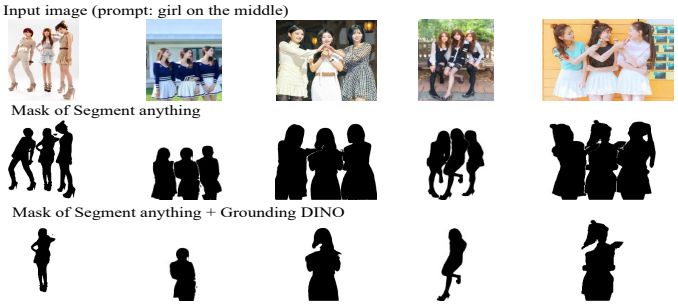

Fig. 13. segment anything + Grounding DINO visualization of experimental results(girl).

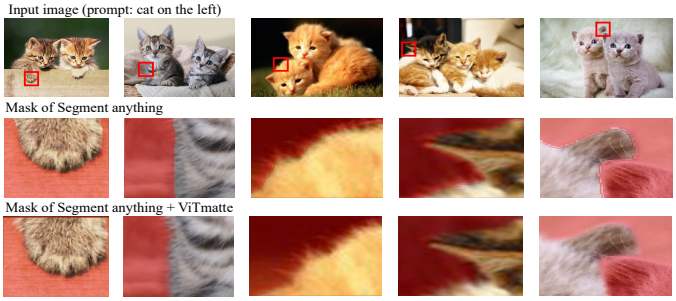

Fig. 14. segment anything + ViTmatte visualization of experimental results(cat)

Fig. 15. segment anything + ViTmatte visualization of experimental results(dog)

Fig. 16. segment anything + ViTmatte visualization of experimental results(girl)

challenges in current open-set image segmentation methods: the inability to accept clause-level prompts and the issue of unclear edge segmentation. This paper combines two models, Grounding DINO and ViTMatte, to resolve the limitations of the original model—specifically, the lack of precise clause-level prompts and poor edge segmentation.

Firstly, we use the Grounding DINO model to perform foreground object detection, followed by non-maximum suppression to obtain the most relevant foreground and the largest bounding box. This output is then used as input for the Segment Anything model. By feeding the original image and bounding box coordinates into Segment Anything, we obtain the segmentation result, which is then refined through connected component denoising to produce the largest and most complete mask. Next, morphological operations such as dilation and erosion are applied to the mask to generate a trimap. Both the original image and the trimap are input into the ViTMatte model to obtain the final output.

The experimental phase is divided into ablation experiments and visualization. The ablation study shows that the improved model outperforms the original model in terms of PA, MPA, MIOU, and FWIOU, with significant improvements. Notably, there is a 6.04% increase in MPA compared to the original model. The results of the visualization experiments indicate that, without the integration of the Grounding DINO model, the original model cannot recognize positional words such as "left" or "right" in the prompt. However, after adding Grounding DINO, the model accurately segments the foreground in the specified region. Furthermore, before incorporating ViTMatte, the model struggled to accurately segment the edges of hair in the image. After the addition of ViTMatte, the model can more accurately segment details such as skin and hair edges in the foreground, improving the precision of the segmentation and resulting in clearer boundaries.

In summary, the improved model shows significant advantages in terms of segmentation accuracy and edge clarity, demonstrating superior performance in open-set image segmentation tasks compared to the original model. This research provides an important theoretical reference for open-set image segmentation and opens up new possibilities for enhancing the accuracy of segmentation in future research.

REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst., Man, Cybern., vol. 9, no. 1, pp. 62-66, 1979, doi: 10.1109/TSMC.1979.4310076.

[2] J. F. Khan, S. M. A. Bhuiyan, and R. R. Adhami, "Image segmentation and shape analysis for road-sign detection," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 1, pp. 83-96, 2011, doi: 10.1109/TITS.2010.2073466.

[3] W. Cui, Z. Guan, and Z. Zhang, "An Improved Region Growing Algorithm for Image Segmentation," 2008 Int. Conf. Comput. Sci. Softw. Eng., Wuhan, China, 2008, pp. 93-96, doi: 10.1109/CSSE.2008.891.

[4] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 11, pp. 2274-2282, 2012, doi: 10.1109/TPAMI.2012.120.

[5] J. Crespo, R. W. Schafer, J. Serra, and C. Gratin, "The flat zone approach: A general low-level region merging segmentation method," Signal Process., vol. 62, no. 1, pp. 37-60, 1997.

[6] H. H. Liu, Z. H. Chen, X. H. Chen, et al., "Multiresolution medical image segmentation based on wavelet transform," in Proc. IEEE Eng. Med. Biol. 27th Annu. Conf., New York, 2005, pp. 3418-3421, doi: 10.1109/IEMBS.2005.1617212.

[7] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3431-3440.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 6230-6239.

[9] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Rethinking Atrous Convolution for Semantic Image Segmentation," arXiv preprint arXiv:1706.05587, 2017.

[10] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017.

[11] J. Xu, M. Zhang, and T. Yao, "A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-Language Model," in Eur. Conf. Comput. Vis. (ECCV), 2021.

[12] N. M. Zaitoun and M. J. Aqel, "Survey on Image Segmentation Techniques," Procedia Comput. Sci., vol. 65, pp. 797-806, 2015, doi: 10.1016/j.procs.2015.09.027.

[13] F. Liang, K. Li, and L. Zhang, "Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP," in 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 7061-7070.

[14] M. Xu, J. Guo, Y. Wang, and W. Zheng, "A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-Language Model," in Eur. Conf. Comput. Vis. (ECCV), 2021.

[15] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in Eur. Conf. Comput. Vis. (ECCV), 2018.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 15979-15988.

[17] J. Xu, M. Li, and D. Wei, "Learning Open-Vocabulary Semantic Segmentation Models From Natural Language Supervision," in 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 2935-2944.

[18] M. Li, L. Zhang, C. Wang, and S. Huang, "CLIP-Event: Connecting Text and Images with Event Structures," in 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16399-16408.

[19] C. Jia, Y. Gao, and T. Yao, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in Int. Conf. Mach. Learn. (ICML), 2021.

[20] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C.-y. Li, J. Yang, H. Su, J.-J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," ArXiv, vol. abs/2303.05499, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257427307

[21] N. Xu, B. L. Price, S. D. Cohen, and T. S. Huang, "Deep Image Matting," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 311-320. [Online]. Available: https://api.semanticscholar.org/CorpusID:14061786

[22] C. Rhemann, C. Rother, A. Rav-Acha, and T. Sharp, "High resolution matting via interactive trimap segmentation," in 2008 IEEE Conf. Comput. Vis. Pattern Recognit., 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587441.

[23] J. Yao, X. Wang, S. Yang, and B. Wang, "ViTMatte: Boosting Image Matting with Pretrained Plain Vision Transformers," ArXiv, vol. abs/2305.15272, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258866204