Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 8 No. 3, November 2023, pp. 310 – 319
DOI:

Conditional Matting For Post-Segmentation Refinement Segment Anything Model

Al Birr Karim Susanto[1], **Moch Arief Soeleman**[2]
[1,2]Informatics Engineering, Computer Science Faculty, Dian Nuswantoro University
E-mail : albirkarim1@gmail.com[1], arief22208@gmail.com[2]
**Corresponding author*

Fikri Budiman[3]
[3]Informatics Engineering, Computer Science Faculty, Dian Nuswantoro University
E-mail: fikri.budiman@dsn.dinus.ac.id [3]

Abstract - Segment Anything Model (SAM) is a model capable of performing object segmentation in images without requiring any additional training. Although the segmentation produced by SAM lacks high precision, this model holds interesting potential for more accurate segmentation tasks. In this study, we propose a Post-Processing method called Conditional Matting 4 (CM4) to enhance high-precision object segmentation, including prominent, occluded, and complex boundary objects in the segmentation results from SAM. The proposed CM4 Post-Processing method incorporates the use of morphological operations, DistilBERT, InSPyReNet, Grounding DINO, and ViTMatte. We combine these methods to improve the object segmentation produced by SAM. Evaluation is conducted using metrics such as IoU, SAD, MAD, Grad, and Conn. The results of this study show that the proposed CM4 Post-Processing method successfully improves object segmentation with a SAD evaluation score of 20.42 (a 27% improvement from the previous study) and an MSE evaluation score of 21.64 (a 45% improvement from the previous study) compared to the previous research on the AIM-500 dataset. The significant improvement in evaluation scores demonstrates the enhanced capability of CM4 in achieving high precision and overcoming the limitations of the initial segmentation produced by SAM. The contribution of this research lies in the development of an effective CM4 Post-Processing method for enhancing object segmentation in images with high precision. This method holds potential for various computer vision applications that require accurate and detailed object segmentation.

Keywords - Segment Anything Model, Post-Processing, High Precision, Image Segmentation.

1. INTRODUCTION

Image segmentation is a fundamental technique within the realm of computer vision, serving the purpose of distinguishing objects from their backgrounds in images. In this context, the Segment Anything Model (SAM) emerges as a notable deep learning tool for image segmentation [1]. SAM's prowess lies in its ability to perform image segmentation by utilizing engineering prompts. Trained on an extensive dataset of 11 million images and 1.1 billion masks, SAM has acquired a profound understanding of common object characteristics, enabling it to generalize to new objects without the need for additional training, a phenomenon referred to as "zero-shot segmentation." This capability is particularly

remarkable, as SAM can accurately group objects in images it has never encountered before. Recent studies have showcased SAM's effectiveness, achieving an impressive 85% accuracy in object segmentation for previously unseen image datasets, marking a substantial advancement compared to conventional methods that necessitate extensive training for each new object class.

While SAM demonstrates proficiency in producing segmentation results, there remain cases where further refinement is necessary. It's worth noting that SAM primarily generates binary masks, while high-precision image segmentation often requires an alpha channel for detailed and accurate delineation. In response to this challenge, previous research, exemplified by Matte Anything or "Mat Any" [2], has introduced an interactive image matting model capable of producing high-quality alpha mattes. This approach combines SAM for contour prediction with user interaction and employs Open Vocabulary (OV) detection to determine the transparency of objects. Furthermore, it leverages VitMatte [3], a pre-trained image matting model that generates alpha mattes through the utilization of a pseudo trimap. Notably, Mat Any's methodology also encompasses an Open Vocabulary (OV) detector, which generates bounding boxes with associated text, serving as input for SAM and enabling the detection of common transparent objects, such as glass, lenses, crystals, diamonds, bubbles, bulbs, webs, and grids. Grounding DINO [4] is the pivotal OV detector employed within the Matte Anything framework.

In the broader context of image segmentation, many researchers have explored post-segmentation refinement techniques [5] [6] [2], mirroring the approach we undertake in our research. Some of these techniques incorporate graphical models, such as Conditional Random Fields (CRF) [7], and region growing [8]. Diverging from these post-processing refinement approaches, we introduce HQ-SAM [9], a system designed to maintain the zero-shot segmentation proficiency of SAM. HQ-SAM directly predicts high-quality masks by reusing SAM's image encoder and mask decoder, eliminating the need for intermediate coarse masks and images. This architectural shift represents a departure from prior high-quality segmentation studies and showcases the effectiveness of HQ-SAM in zero-shot experiments.

Inspired by the innovative Matte Anything framework and its utilization of an Open Vocabulary (OV) detector, we present Conditional Matting 4 (CM4) as a post-segmentation refinement method for SAM. CM4 amalgamates a range of computer vision and neural network techniques, incorporating morphological operations, Recognize Anything Model (RAM) [10], DistilBERT [11], InSPyReNet [12], Grounding DINO [4], and ViTMatte [3]. To evaluate CM4's efficacy, we employ established metrics such as Intersection over Union (IoU), Sum of Absolute Differences (SAD), Mean Absolute Difference (MAD), Grad, and Conn.

2. RESEARCH METHOD

2.1. Proposed Method

The CM4 method can be likened to how a good restaurant team works together. It's like having a waiter, a chef manager, and a specialized chef. The waiter, or "The Waiter" in CM4, is like someone who listens to what you want, even if you're not very clear about it. Just like a waiter might suggest "pizza" when you say you want something with bread, CM4 recommends things based on what you want. The chef manager, or "The Chef Manager" in CM4, is like the person who decides which chef should cook your food. They choose the right chef, just like CM4 selects the best way to process information. In CM4, they use models like Recognize Anything Model (RAM) [10] and DistilBert [11] to help with this.

For example, if you show a picture of fire, RAM acts like a waiter and says, "This looks like 'burn,' 'ember,' 'fire,' and 'flame.'" Then, DistilBert acts like a chef manager and says, "Let's use the transparent mode to refine this image." It's like having the right chef for the right dish. CM4 makes sure your information is handled well, just like a restaurant team makes sure you get the food you want. It's all about providing a smooth and efficient experience in handling information.

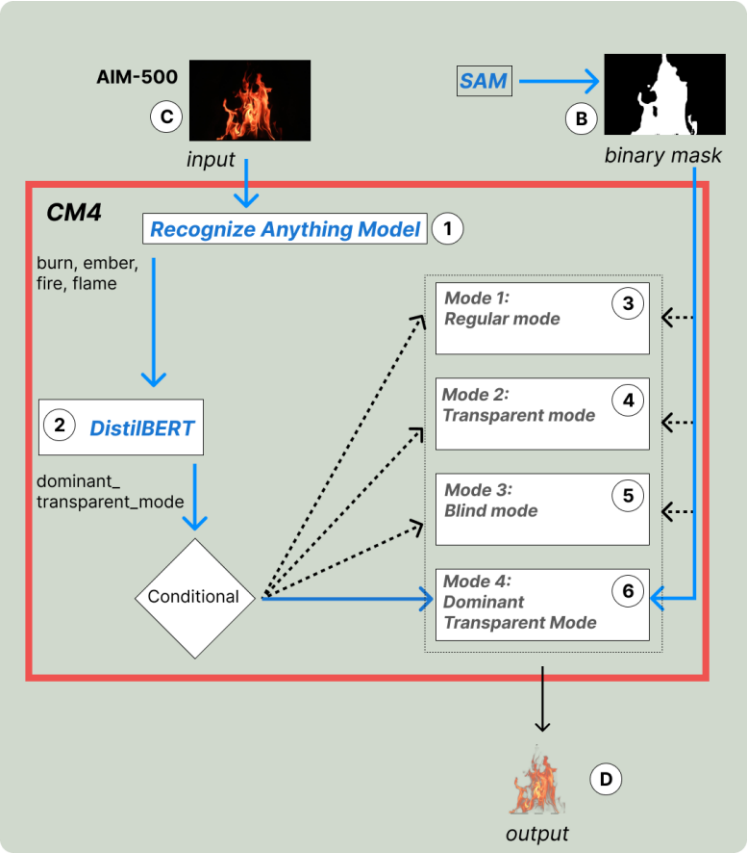

Figure 1. CM4

Conditional Matting 4 (CM4) is our proposed method, Number 4 is here because we use 4 matting modes as follows:

1. Regular Mode

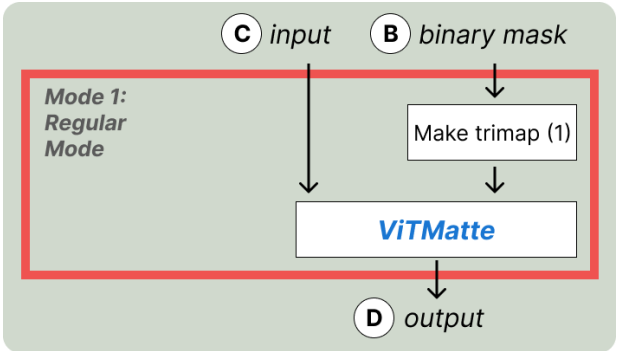

Figure 2. The regular mode of CM4

The normal mode is suitable for non-transparent objects. Such as animals, humans, etc. The regular mode is similar to the transparent mode, so why do we separate it as a new mode? we want to make efficient ways, when it can be classified with DistilBERT and directly given into mode 1. So the process doesn't through GroundingDINO which add more execution time.

2. Transparent mode

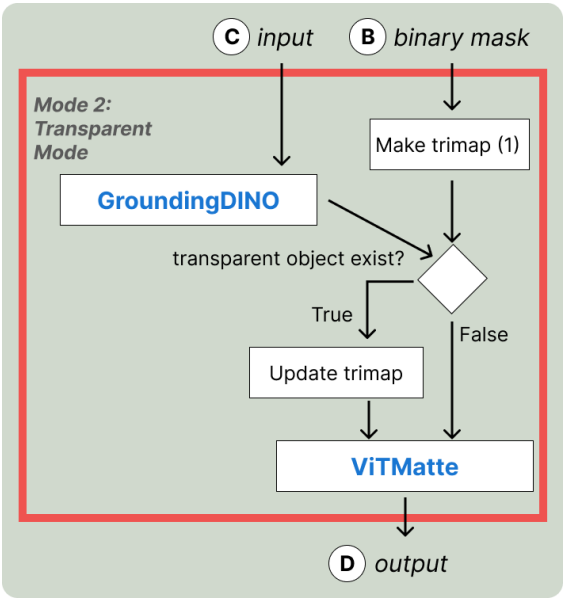

Figure 3. Transparent mode of CM4

This transparent mode is suitable for transparent objects such as glasses, glasses, and light bulbs. This mode 2 is similar to the post-processing that Matte Anything [2] uses.

3. Blind Mode

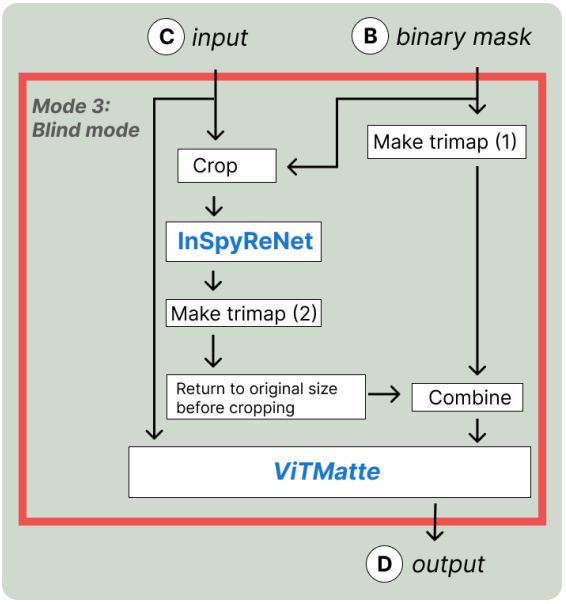

Figure 4. Blind mode of CM4

Blind Mode / blind mode is suitable if the other modes fail to recognize/know the location of objects. This mode doesn't know what the object is. But this mode can distinguish the foreground, background, and transitions with precision and good. In this mode, we use InSPyReNet [12].

4. Dominant Transparent Mode

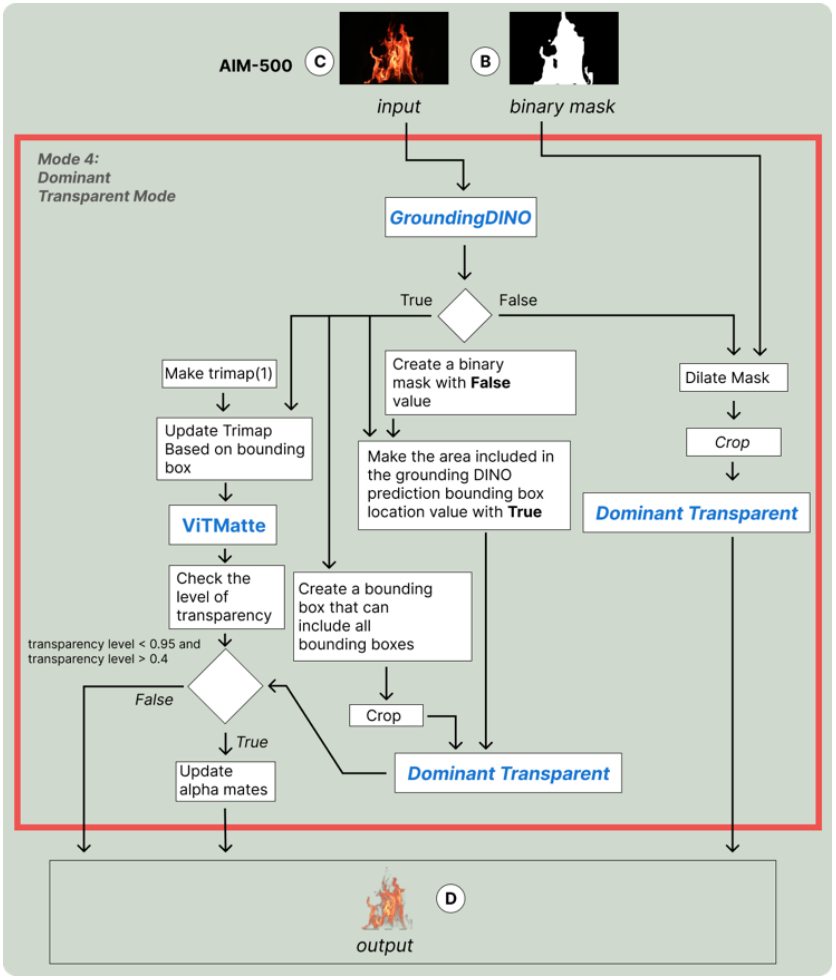

Figure 5. Dominant Transparent Mode of CM4

There are cases where ViTMatte [3] fails to distinguish foreground, background, and transitions with a given trimap. Like stars in the sky, galaxies, smoke, fire, and liquid. The dominant transparent function is a function that changes the alpha channel of all pixels, if the pixel value is closer to the pixel that appears the most (dominant) then it will be changed closer to 0. In non-linear changes, we apply easeOutQuad, because it is a good equation imitating particle transparency like smoke, fire, etc.

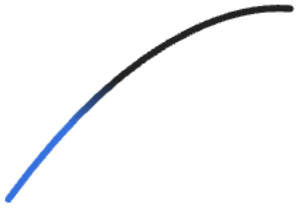

Figure 6. Ease Out Quad

Here the ease-out quad equation

$$1 - (1 - x) * (1 - x) \qquad (1)$$

Each of the modes has its advantages and disadvantages. Then how can we direct the input image according to the mode that gives good matting results? We use the Recognize Anything Model (RAM) [10] and DistilBERT [11]. We use RAM to provide tags, tags are information about an image, such as what object is in the image. We use DistilBERT [11] as a classification. A tag generated by the RAM can be classified into what mode. We use DistilBERT so that the classification can be dynamic according to the training data.

The algorithms we used in CM4 is the state of the art of its task, for example, ViTMatte is the best in image matting, InSPyReNet is the best in Dichotomous Image Segmentation (DIS) [13] task, also RAM is the best in task image tagging, DistilBERT is also good performance for text classification. When there are the best latest algorithms in your time, you can just replace them with that, for each task Image Tagging, Text Classification, Zero-Shot Object Detection, DIS, and Image Matting.

2.2. Dataset

We use AIM-500 introduced in [14]. AIM-500 contains 7 categories.

Table 1. Category In AIM-500

Name	Number of images
Portrait	100
Animal	200
Transparent	34
Plant	75
Furniture	45
Toy	36
Fruit	10

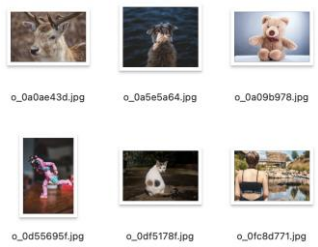

Figure 7. Example original images from AIM-500

Figure 8. Example ground truth mask images from AIM-500

Evaluations that we use are Intersection over Union (IOU), Mean Square Error (MSE), Mean Absolute Difference (MAD), Sum of Absolute Difference (SAD), Connectivity (Conn), and Gradient (Grad). Like the evaluation metrics used in [14].

3. RESULTS AND DISCUSSION

In this section, we present a comprehensive analysis of the outcomes obtained with Conditional Matting 4 (CM4), comparing its performance against the Segment Anything Model (SAM) and Matte Anything [2]. Figure 8 visually summarizes the comparison between CM4 and Matte Anything, highlighting the strengths of our approach, particularly in handling abstract transparent images like galaxies, fire, and smoke. To delve into the details, we conducted a thorough evaluation using various metrics, such as Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Mean Absolute Difference (MAD), Connectivity (Conn), Gradient (Grad), and Intersection over Union (IoU).

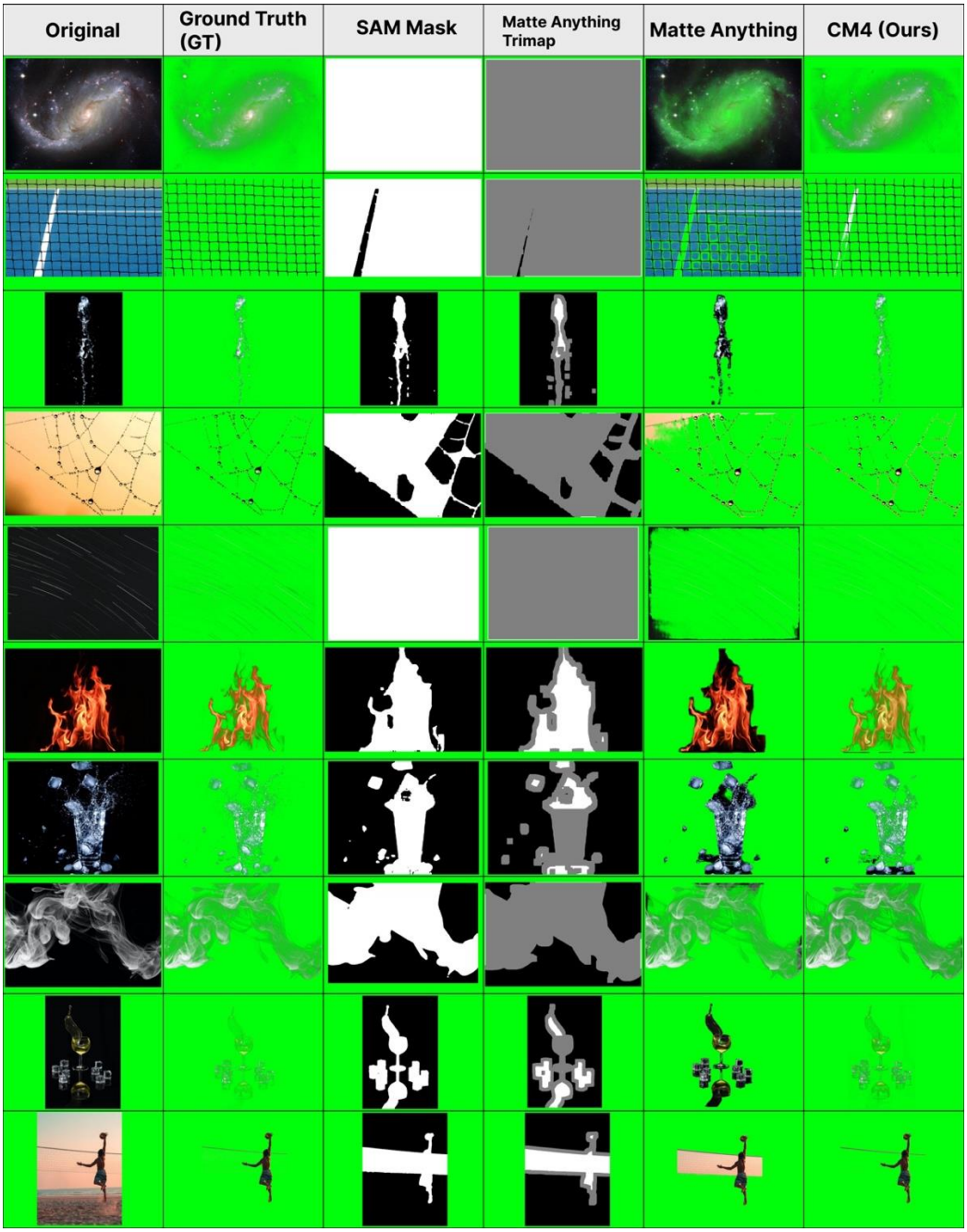

Figure 9. Comparison result between CM4 and Matte Anything

Table 2. Table Comparison SAM and CM4

Method	Evaluation Metrics					
	SAD	**MSE**	**MAD**	**Conn**	**Grad**	***IoU***
SAM	70,43	36,08	43,109	72,164	86,739	0,922
CM 4 (Ours)	20,42	5,176	12,085	19,474	21,122	0,887

Table 2 reveals a compelling outcome: CM4 significantly enhances the alpha channel, resulting in substantial improvements in SAD, MSE, MAD, Conn, and Grad metrics when compared to SAM. However, it's important to note that IoU is adversely affected. This

decrease in IoU is attributed to CM4's removal of the alpha channel on the object's edges, resulting in IoU scores that are slightly lower. SAM, on the other hand, maintains the edge area, preserving ground truth equivalence.

Table 3. Tabel Matte Anything and CM4

Method	SAD			MSE		
	All	transparent	opaque	All	transparent	opaque
MatAny [2]	27,83	110,50	16,98	9,36	35,30	5,95
CM 4 (ours)	20,42	76,39	15,96	5,17	19,14	3,97

Table 3 presents a comparative evaluation between CM4 and Matte Anything, highlighting the improvements brought by our method. In this assessment, CM4 outperforms Matte Anything in terms of SAD and MSE across all image categories. Notably, CM4 exhibits a remarkable increase in SAD, particularly in the transparent image category, where it achieves a 31% improvement, and in the opaque object category, where it boosts performance by 6%. Table 4 further elaborates on the percentage increase in CM4's evaluation results concerning Matte Anything, providing a more detailed breakdown of improvements:

Table 4. Table Of The Percentage Of CM4 (Our) Evaluation Results On Matte Anything

Method	SAD			MSE		
	All	transparent	opaque	All	transparent	opaque
CM 4	27%	31%	6%	45%	46%	33%

Furthermore, in our quest for optimization, it is crucial to address the memory requirements for implementing CM4 and SAM. The total size of the model checkpoints, as outlined in Table 5, is substantial, with SAM's model alone demanding 2.56 GB. To mitigate memory-related issues, we recommend future research endeavors explore methods for optimization. We try to optimize by code, and talk about memory to start an web server (Contain SAM+CM4) will cost about 7,38 GB. Load all the model at once on start.

Table 5. Model Checkpoint Sizes

Checkpoint	Size
sam_vit_h_4b8939.pth	2,56 GB
ViTMatte_B_Com.pth	386 MB
InSPyReNet_SwinB_Large.pth	367 MB
groundingdino_swint_ogc.pth	694 MB
ram_swin_large_14m.pth	5,63 GB
distilBERT_tags_aim_500_CM4.pth	267,9 MB

When using many model it comes with memory issue, so we suggest for the future research to do optimization. In summary, our results demonstrate the effectiveness of CM4 in enhancing object segmentation, particularly in challenging scenarios involving abstract transparent images. While CM4 showcases improvements in several key metrics, it's essential to consider its impact on IoU, which is slightly reduced due to the removal of the alpha channel on object edges.

4. CONCLUSION

In conclusion, the innovative CM4 method has demonstrated remarkable success in advancing the field of object segmentation, as evidenced by the significant reduction in the Sum of Absolute Differences (SAD) evaluation metric to 20.42. This represents an impressive

27% improvement when compared to the Matte Anything framework [2], as evaluated on the AIM-500 dataset. Moreover, the Mean Squared Error (MSE) score of 21.64 showcases a notable 45% increase from its precursor, further reinforcing CM4's efficacy in this context.

As we look ahead, it is paramount to emphasize the necessity of a more extensive and diversified dataset for evaluation purposes. A larger, more representative dataset has the potential to provide a robust validation of the method's performance, thereby enhancing confidence in the derived evaluation outcomes. By conducting assessments across a broader spectrum of data, researchers can acquire a more comprehensive understanding of the method's capabilities, paving the way for further advancements in the realm of object segmentation. This pursuit of broader and more diverse datasets will undoubtedly play a pivotal role in propelling the field to new heights.

REFERENCES

[1] A. Kirillov *et al.*, "Segment Anything." arXiv, Apr. 05, 2023. doi: 10.48550/arXiv.2304.02643.

[2] J. Yao, X. Wang, L. Ye, and W. Liu, "Matte Anything: Interactive Natural Image Matting with Segment Anything Models." arXiv, Jun. 06, 2023. doi: 10.48550/arXiv.2306.04121.

[3] J. Yao, X. Wang, S. Yang, and B. Wang, "ViTMatte: Boosting Image Matting with Pretrained Plain Vision Transformers." arXiv, May 24, 2023. doi: 10.48550/arXiv.2305.15272.

[4] S. Liu *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection." arXiv, Mar. 20, 2023. doi: 10.48550/arXiv.2303.05499.

[5] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang, "CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement." arXiv, May 05, 2020. doi: 10.48550/arXiv.2005.02551.

[6] T. Shen *et al.*, "High Quality Segmentation for Ultra High-resolution Images." arXiv, Dec. 26, 2021. doi: 10.48550/arXiv.2111.14482.

[7] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." arXiv, Oct. 20, 2012. doi: 10.48550/arXiv.1210.5644.

[8] P. A. Dias and H. Medeiros, "Semantic Segmentation Refinement by Monte Carlo Region Growing of High Confidence Detections." arXiv, Feb. 21, 2018. doi: 10.48550/arXiv.1802.07789.

[9] L. Ke *et al.*, "Segment Anything in High Quality." arXiv, Jun. 02, 2023. doi: 10.48550/arXiv.2306.01567.

[10] Y. Zhang *et al.*, "Recognize Anything: A Strong Image Tagging Model." arXiv, Jun. 09, 2023. doi: 10.48550/arXiv.2306.03514.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv, Feb. 29, 2020. doi: 10.48550/arXiv.1910.01108.

[12] T. Kim, K. Kim, J. Lee, D. Cha, J. Lee, and D. Kim, "Revisiting Image Pyramid Structure for High Resolution Salient Object Detection." arXiv, Nov. 16, 2022. doi: 10.48550/arXiv.2209.09475.

[13] X. Qin, H. Dai, X. Hu, and D.-P. Fan, "Highly Accurate Dichotomous Image Segmentation".

[14] J. Li, J. Zhang, and D. Tao, "Deep Automatic Natural Image Matting." arXiv, Jul. 15, 2021. Accessed: May 20, 2023. [Online]. Available: http://arxiv.org/abs/2107.07235