J. Vis. Commun. Image R. 97 (2023) 103946

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Short Communication

MFS enhanced SAM: Achieving superior performance in bimodal few-shot segmentation

Ying Zhao, Kechen Song *, Wenqi Cui, Hang Ren, Yunhui Yan

School of Mechanical Engineering and Automation, Northeastern University, Shenyang, Liaoning 110819, China
The National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang 110819, China
Key Laboratory of Data Analytics and Optimization for Smart Industry (Northeastern University), Ministry of Education, China

ARTICLE INFO

Keywords:
Segment anything
Few-shot segmentation
RGB-T SAM
Gated prediction selection

ABSTRACT

Recently, Segment Anything Model (SAM) has become popular in computer vision field because of its powerful image segmentation ability and high interactivity of various prompts, which opens a new era of large vision foundation models. But is SAM really omnipotent? In this letter, we establish a comprehensive bimodal few-shot segmentation indoor dataset VT-840-5^{i}, and compare SAM with eight state-of-the-art few-shot segmentation (FSS) methods on two benchmark datasets. Qualitative and quantitative experiment results show that although SAM is very effective in general object segmentation, it still has room for improvement in some challenging scenarios. Therefore, we introduce thermal infrared auxiliary information into the segmentation task and provide multiple fusion strategies (MFS) for readers to choose the most suitable approach for the specific task. Finally, we discuss several potential research trends about SAM in the future. Our test results are available at: https://github.com/VDT-2048/Bi-SAM.

1. Introduction

In recent years, the popularity of large pre-trained models, namely foundation models (FMs) [1], marks the transformation from artificial intelligence (AI) system to artificial general intelligence (AGI) system. AGI means that facing different tasks, it does not need to train a specific model to adapt, but relies on transfer learning or fine-tuning a model to perform various realistic general tasks. This new paradigm has made pioneering work in the fields of language and image vision. In the natural language processing (NLP) community, large language foundation models (LLMs) refer to pre-trained language models (PLMs) with billions of parameters, such as BERT [2], LLAMA [3], InstrucGPT [4] and GPT-3 [5].

In the last six months, ChatGPT [6], a language model developed by OpenAI, has gained more than 100 million active users worldwide. Its powerful language understanding ability and reasoning ability have been used in various industries. On the other hand, Meta AI Research recently released the first large vision foundation model Segment Anything Model (SAM) [7] in history. SAM uses the data engine to train the largest segmented dataset named SA-1B so far, so this general perception model shows strong zero-shot transferability ability in a series of tasks such as edge detection and instance segmentation. As of May 2, SAM's GitHub warehouse had a Star count of 30.9 k. Different from the above-mentioned general foundation models, which require huge data, few-shot segmentation (FSS) [27] can quickly model and distinguish different categories with only a couple of samples. It can segment new categories from the background without fine-tuning, thus bridging the gap between human intelligence and AGI. This is very helpful for scenarios with insufficient data and expensive marking.

Because of the powerful performance of SAM in many tasks, many researchers have tested the effect of SAM in other fields, including medical image [8], surface defect detection [9], camouflaged object detection [10], image restoration [11] and so on. The above research shows that SAM does not perform well on datasets in many specific fields. This is reasonable, because SAM's training set mainly includes 11 million natural images. And it does not include a vast amount of medical images or industrial defect images, which makes SAM challenging in these fields.

Different from the above research, in this letter, we compare the performance of SAM and bimodal FSS algorithms in complex and changeable natural scenes (e.g., weak illumination, transparent objects, exposure, low contrast, continuous branch, clutter and out of focus, etc.). The results indicate that it is necessary to continue to study bimodal FSS. And SAM still has limitations in natural images.

* Corresponding author.
E-mail address: songkc@me.neu.edu.cn (K. Song).

https://doi.org/10.1016/j.jvcir.2023.103946
Received 9 June 2023; Received in revised form 25 August 2023; Accepted 23 September 2023
Available online 26 September 2023

Meanwhile, thermal infrared modality is introduced to make up for the shortcomings of visible (RGB) images, which is helpful to further improve the robustness of SAM in complex environments and better distinguish various objects. Our contributions can be summarized as follows:

(1) To evaluate the applicability of SAM in real-world natural scenario, we created a new bimodal few-shot semantic segmentation dataset called VT-840-5^i for indoor environment to include challenging cases from diverse indoor environments.
(2) There is still a gap between SAM (RGB) and the SOTA RGB-T FSS. We select the optimal fusion strategy to enhance the overall semantic segmentation performance by evaluating the model's performance with multiple fusion suggestions (MFS) in different scenarios and considering the characteristics of the task comprehensively.
(3) Our discussion delved into the current limitations, available prospects and future research trends of SAM, with the aim of providing inspiration to future researchers.

2. Experiment

2.1. Datasets

In order to evaluate the performance of SAM in natural scenarios more comprehensively, we conducted experiments on two FSS benchmark datasets, indoor and outdoor.

(1) VT-840-5^i: To create a high-quality dataset and make up for the vacancy of indoor dataset, we established a comprehensive FSS evaluation benchmark, VT-840-5^i. It includes 840 pairs of aligned RGBT image pairs, with 390 pairs from VT821 [12], VT1000 [13], and VT5000 [14], and the remaining 450 pairs from VI-RGBT1500 [15]. The dataset comprises 20 object categories, and the image size is standardized at 640 × 480. The dataset includes four different illumination conditions, including bright illumination, uneven illumination, weak illumination and dark illumination, to fully demonstrate the superiority of bimodal image fusion. Moreover, the dataset includes image clutter and out of focus.

(2) Tokyo Multi-Spectral-4^i: For outdoor road scenes, we select the Tokyo Multi-Spectral-4^i dataset [16], which includes some Tokyo Multi-Spectral images and annotations [17]. There are 1126 pairs of RGB-T images in the dataset, with 16 different semantic categories, and the image size is 200 × 200. It contains two different environments, day and night, and the visibility of RGB images in the night environment is very low, sometimes with severe glare.

2.2. Mask selection strategy

The SAM project supports four different prompt types to produce accurate results, including foreground/background points, bounding boxes, mask, and text. Moreover, SAM supports three main segmentation settings: automatic segmentation setting, bounding box setting, and click setting. To ensure scalable evaluation and minimize subjective human participation, we choose the automatic segmentation setting to evaluate SAM: SAM will automatically identify all objects in the image and generate multiple binary masks. For images that contain multiple objects of the same category, we employ a selection strategy that superimposes or negates the multiple masks generated by SAM for each image to produce masks that are closest to the ground truth.

2.3. Evaluation metrics

Following the previous work on FSS [20,23], this letter uses mean intersection-over-union (mIoU) and foreground and background intersection-over-union (FB-IoU) as performance metrics to compare SAM with other advanced FSS algorithms.

2.4. Learning paradigm

For the FSS algorithms, we refer to OSLSM [18] to cross-validate using 4 folds. During testing, we use the same random seed to sample 500 pairs of images [36]. The final result is the average of four-fold metrics. Since SAM only requires testing without retraining the model, we directly input the corresponding validation fold data into the frozen SAM, as shown in Fig. 1. Note that all models are evaluated on PyTorch 1.8 using the NVIDIA 3060 GPU.

2.5. Quantitative evaluation

To ensure fairness, we compared the best performance version of SAM (ViT-H) with eight state-of-the-art FSS methods, including CANet [19], PGNet [20], PFENet [21], ASGNet [22], SAGNN [23], HSNet [24], ASNet [25], and V-TFSS [16]. The first seven algorithms belong to single-modality RGB methods, and the last belongs to a dual-modality RGB-T FSS. We used the late-fusion strategy to convert single-modality models into dual-modality models.

(1) Table 1 summarizes the mIoU and FBIoU results of all methods on the indoor dataset under 1-shot and 5-shot settings. The results show that SAM (RGB) performs outstandingly in indoor semantic segmentation tasks, achieving 83.2 % mIoU and 89.9 % FBIoU, respectively. Compared with the previous SOTA ASNet, SAM's mIoU and FBIoU increase by 15.4 % and 9.1 % under 1-shot setting, and by 10.6 % and 5.7 % under 5-shot setting, respectively. This result can be understood, as SAM is trained on a large-scale natural image dataset.
(2) Table 2 summarizes the mIoU and FBIoU results of all methods on the road dataset with 1-shot and 5-shot settings. The results show that the segmentation models directly trained on the Tokyo Multi-Spectral-4^i dataset (HSNet, ASNet) provide higher mIoU and FBIoU results than SAM (RGB). The main reason may be that SAM's average training image resolution is as high as 3300 × 4950, which limits its generalization ability to low-resolution images (only 200 × 200) in the road dataset. Moreover, the targets and backgrounds have high color similarity and low visibility in night-time environments, which also presents challenges to SAM. Meanwhile, it also shows that it is valuable to continue to study bimodal few-shot segmentation.

2.6. Qualitative evaluation

(1) Fig. 2(a) shows the qualitative results of indoor scenes. In the first four columns, SAM performs well in facing complex indoor illumination changes, including bright illumination, uneven

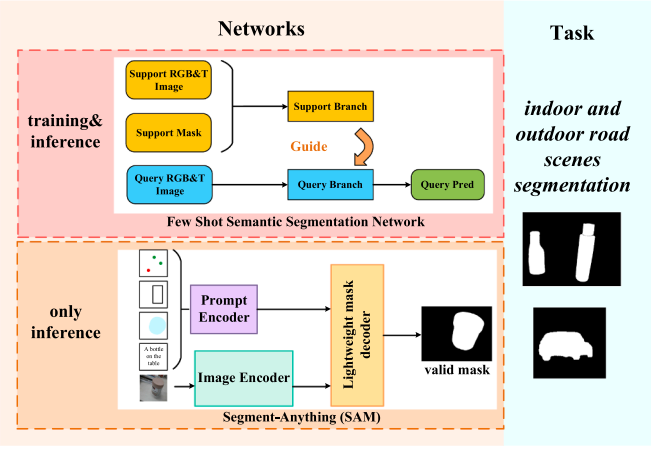

Fig. 1. We compare bimodal few-shot segmentation net with SAM.

Table 1
Quantitative comparison results of different methods on VT-840-5^i. The best results in each column are marked in red.

Methods	1-way 1-shot						1-way 5-shot					
	fold = 0	fold = 1	fold = 2	fold = 3	mIoU	FB-IoU	fold = 0	fold = 1	fold = 2	fold = 3	mIoU	FB-IoU
CANet	51.7	61.5	55.4	67.3	59.0	75.3	55.3	65.2	57.8	72.0	62.6	78.0
PGNet	56.4	67.2	61.9	73.8	64.8	79.5	57.1	67.1	62.5	75.7	65.6	80.1
PFENet	34.4	39.2	46.9	50.9	42.8	66.0	35.0	39.6	46.8	52.5	43.5	66.5
ASGNet	33.8	38.8	47.0	50.6	42.6	65.4	35.4	38.3	44.9	51.1	42.4	65.3
SAGNN	32.2	35.2	41.0	48.4	39.2	62.3	31.6	35.6	41.6	49.1	39.5	62.6
HSNet	56.6	72.4	62.8	77.4	67.3	80.8	62.1	78.1	68.6	81.6	**72.6**	**84.2**
ASNet	57.1	72.1	63.8	78.0	**67.8**	**80.8**	62.2	76.0	68.0	82.3	72.1	83.6
V-TFSS	56.1	57.4	51.1	70.7	58.8	75.8	58.2	63.7	54.6	71.6	62.0	77.9
SAM (RGB)	85.3	86.8	83.7	77.1	**83.2**	**89.9**	85.3	86.8	83.7	77.1	**83.2**	**89.9**
SAM (T)	50.0	49.3	48.3	53.6	50.3	67.8	50.0	49.3	48.3	53.6	50.3	67.8
SAM (RGB-T)	65.6	78.3	70.9	75.6	**74.6**	**85.6**	65.6	78.3	70.9	75.6	**74.6**	**85.6**

Table 2
Quantitative Comparison Results of Different Methods on Tokyo Multi-Spectral-4^i. The best results in each column are marked in red.

Methods	1-way 1-shot						1-way 5-shot					
	fold = 0	fold = 1	fold = 2	fold = 3	mIoU	FB-IoU	fold = 0	fold = 1	fold = 2	fold = 3	mIoU	FB-IoU
CANet	27.6	9.4	28.5	33.8	24.8	57.0	32.0	10.6	30.1	35.3	27.0	57.2
PGNet	24.9	6.1	29.9	39.4	25.2	57.5	31.2	6.0	29.3	43.3	27.4	58.9
PFENet	32.2	6.3	26.8	33.6	24.7	57.8	34.1	13.9	27.1	41.6	29.2	60.2
ASGNet	33.5	5.4	30.3	34.7	26.0	59.2	36.4	10.9	31.8	44.8	31.0	62.0
SAGNN	23.7	5.6	24.4	33.4	21.8	55.2	26.8	4.8	23.5	38.1	23.3	56.4
HSNet	41.1	17.9	36.6	40.2	**34.0**	**63.1**	48.2	23.1	42.2	45.9	**39.8**	**66.2**
ASNet	40.9	17.6	36.6	45.4	**35.1**	**63.8**	44.5	24.0	44.7	49.7	**40.7**	**66.4**
V-TFSS	29.3	7.6	27.7	42.5	26.8	58.6	27.0	8.1	29.9	47.9	28.2	59.2
SAM (RGB)	33.6	35.5	30.7	31.0	32.7	55.0	33.6	35.5	30.7	31.0	32.7	55.0
SAM (T)	22.1	23.5	22.6	20.7	22.2	41.2	22.1	23.5	22.6	20.7	22.2	41.2
SAM (RGB-T)	40.9	43.5	38.8	40.6	**40.9**	**62.9**	40.9	43.5	38.8	40.6	**40.9**	**62.9**

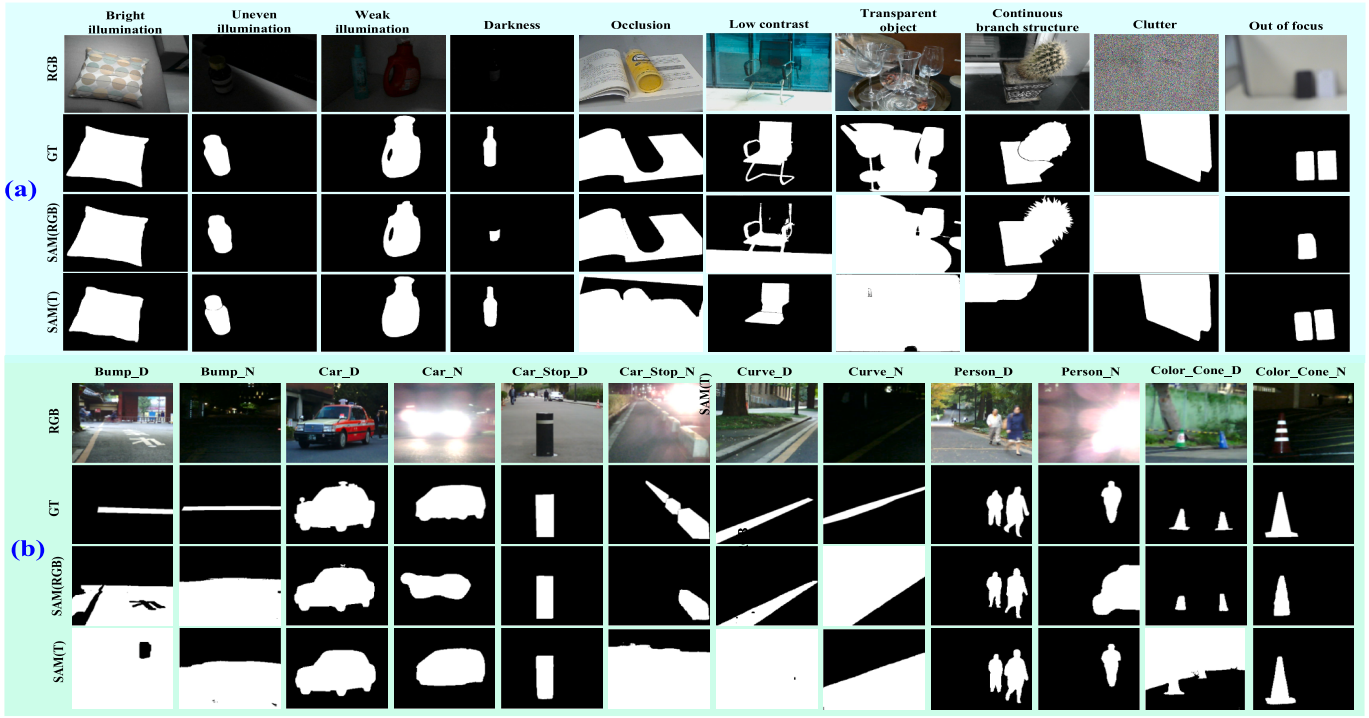

Fig. 2. Visual comparison of several representative and challenging realistic indoor and road scenes.

illumination and weak illumination conditions. However, when the target is in a completely dark environment, SAM can only recognize a small part of it. In the fifth column, SAM can handle object occlusion well, but in low-contrast situations, such as when the color pattern of chair legs is consistent with that of the white floor under strong light (the sixth column), SAM has difficulty detecting the target. In addition, transparent objects may refract light in unpredictable ways, making it difficult for SAM to distinguish between target and background (the seventh column). At the same time, we speculate that SAM may not handle

continuous branch structures well, such as the thorns of a cactus ball (the eighth column). Unfortunately, when the RGB image is affected by clutter or the image is out of focus and the target contour is unclear (the reciprocal two columns), SAM can't deal with it.

(2) Fig. 2(b) shows some qualitative results of road scenes. As shown in the third, fifth, seventh, and ninth columns, SAM can accurately segment targets under good illumination and clear boundaries, showing its strong localization ability. Similar to indoor scenes, when the target is completely invisible in the dark, SAM cannot distinguish between foreground and background well (such as in the second, eighth, and last columns). It should be noted that due to the extremely bright car lights at night and the extreme brightness contrast with the surroundings, resulting in exposure, SAM may fail to detect the target or can only detect part of it (such as in the fourth, sixth, and last columns). We also found that SAM still has great room for improvement in handling small objects (Such as the first column and the penultimate column).

2.7. Bimodal SAM structure

(1) Motivation. Visible imaging is easily affected by environmental or human factors. If we only rely on low-quality RGB images for segmentation, it will lead to segmentation failure. This also explains the limitations of SAM mentioned in Section F. In contrast, thermal infrared imaging (T) focuses on the thermal radiation emitted by objects, which has the characteristics of strong anti-interference and is more suitable for low illumination. Therefore, the combination of the two spectra can improve the accuracy and robustness of segmentation.

(2) Specific Operation. From the results in Tables 1 and 2 and Fig. 2, we see that SAM predictions relying solely on RGB or T information do not consistently perform well on all images. How to adaptively combine the advantages of each predictor and surpass each predictor through fusion technology is a key challenge. We add an additional SAM (T) branch in parallel to the existing SAM (RGB) branch and complete the late-fusion of the probability maps of the two predictors. Meanwhile, we provide multiple fusion suggestions (MFS) so that readers can choose the most suitable method for the task. The overall pipeline is shown in Fig. 3. These integration proposals include:

(1) Element-level fusion: The probability graphs generated by SAM (RGB) and SAM (T) are added element by element. This method is simple and direct.

(2) Commonality enhancement fusion: Multiplying the probability maps of SAM (RGB) and SAM (T) at pixel level. This method can enhance the commonness and is suitable for the two predictors with high accuracy in different aspects.

(3) Average fusion: The probability maps of SAM (RGB) and SAM (T) are averaged pixel by pixel. This simple fusion method can balance the outputs of the two predictors.

(4) Residual connection fusion: The probability maps of two are connected by residual connection to introduce additional information.

(3) Comparison Experiments. The experimental results of various fusion strategies are shown in Table 3. The results also prove our hypothesis: the same fusion method has different performance in different

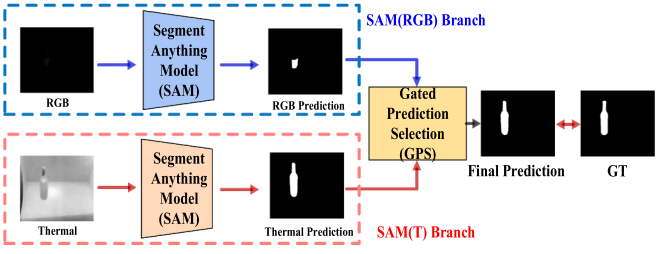

Fig. 3. The overall pipeline of multiple fusion suggestions for RGB-T. SAM.

Table 3
Quantitative comparison results of proposed module.

Fusion Method	VT-840-5[i]		Tokyo Multi-Spectral-4[i]	
….	mIoU	FB-IoU	mIoU	FB-IoU
SAM(RGB)	**83.2**	**89.9**	32.7	55.0
SAM(T)	50.3	67.8	22.2	41.2
SAM(RGB) + SAM(T)	12.4	48.8	5.8	28.7
[SAM(RGB) + SAM(T)]/2	39.1	63.0	15.1	33.4
SAM(RGB)*SAM(T)	74.6	85.6	**40.9**	**62.9**
SAM(RGB)*SAM(T) + SAM(RGB)	13.1	54.6	5.4	41.4

dataset scenarios. This reminds us that when choosing fusion suggestions, we need to consider the characteristics of tasks, the attributes of datasets and the performance of models.

On indoor datasets, we find that compared with SAM (RGB), merging the results of SAM (T) branches directly will lead to a certain degree of performance degradation. This may be because SAM has achieved very accurate results on RGB images, while the overall performance of SAM (T) is poor, and direct fusion will produce huge interference. It should be noted that the commonality enhancement fusion has achieved suboptimal results.

On outdoor datasets, the reverse commonality enhancement fusion has achieved the best results. Compared with SAM (RGB), the mIoU and FB-IoU are increased by 8.2 % and 7.9 % respectively, and the gains are obtained by 18.7 % and 21.7 % respectively compared with SAM (T). It is worth noting that the commonality enhancement fusion has achieved good results on both datasets. We speculate that this is because their respective predictions contain large noise, and the common high confidence area can be emphasized by multiplying them, thus reducing the error of the final prediction.

3. Discussion and outlook

Based on the experimental analysis above, we found that there is still a performance gap between SAM(RGB) and the SOTA RGB-T FSS. We believe that bimodal FSS still has significant research value, and at the same time, enchancing SAM is also necessary. Therefore, we will discuss the potential future research directions of SAM.

3.1. Multi-modal SAM

The current SAM has powerful zero-shot transferability for visible light images with clear boundaries and obvious targets. However, its performance is limited in challenging scenarios such as low-light conditions, exposure, clutter, and out of focus. Therefore, it is helpful and necessary to consider adding other auxiliary modalities such as thermal infrared [28–30], depth [31–33], and radar [34,35], etc [37].

3.2. Multi-domain SAM

SAM can usually achieve comparable or better performance than fully supervised/transfer learning methods on natural images. However, there is still a significant performance gap between SAM and SOTA when applied to fields such as medical image segmentation [8], surface defect detection [9], and camouflaged object detection [10]. Therefore, fine-tuning with a couple of data or adding Adapters [26] to combine specific domain knowledge.

3.3. Multi-functional SAM

SAM can achieve very fine-grained image segmentation, and it can be connected with powerful models of other machine vision tasks to provide new functions that SAM does not have at present, so as to be easily applied to more downstream tasks.

4. Conclusion

In this letter, we compared the performance of SAM with bimodal few-shot segmentation algorithms on two different benchmark datasets. We conducted a more in-depth analysis of the strengths and weaknesses of SAM in various challenging scenarios of natural images, and put forward some potential development directions, which provides a powerful guidance for the next optimization. We are also the first to study the practical effect of bimodal SAM, and we plan to explore a universal fusion stage and method to fully realize the value of SAM in the future. We hope that our evaluation and findings can present a fresh insight.

CRediT authorship contribution statement

Ying Zhao: Methodology, Writing – original draft, Writing – review & editing. **Kechen Song:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition. **Wenqi Cui:** Investigation, Visualization, Writing – review & editing. **Hang Ren:** Formal analysis, Writing – review & editing. **Yunhui Yan:** Conceptualization, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] R. Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv: 1810.04805, 2018.

[3] H. Touvron et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] L. Ouyang, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Proces. Syst. 35 (2022) 27730–27744.

[5] T. Brown, et al., Language models are few-shot learners, Adv. Neural Inf. Proces. Syst. 33 (2020) 1877–1901.

[6] OpenAI, Introducing chatgpt, https://openai.com/blog/chatgpt, 2023b. Accessed: 2023-04-19.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," *arXiv preprint arXiv: 2304.02643*, 2023.

[8] C. Hu, and X. Li, "When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation," *arXiv preprint arXiv:2304.08506*, 2023.

[9] W. Ji, J. Li, Q. Bi, W. Li, L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," *arXiv preprint arXiv: 2304.05750*, 2023.

[10] G. P. Ji et al, "SAM Struggles in Concealed Scenes–Empirical Study on" Segment Anything"," *arXiv preprint arXiv:2304.06022*, 2023.

[11] Q. Shen, X. Yang and X. Wang, "Anything-3d: Towards single-view anything reconstruction in the wild," *arXiv preprint arXiv:2304.10261*, 2023.

[12] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, B. Luo, RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach, in: Proc. Chin. Conf, Image Graph. Technol, 2018, pp. 359–369.

[13] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, J. Tang, RGB-T image saliency detection via collaborative graph learning, IEEE Trans. Multimedia 22 (1) (Jan. 2020) 160–173.

[14] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," 2020, *arXiv:2007.03262*. [Online]. Available: http://arxiv.org/abs/2007.03262.

[15] K. Song, L. Huang, A. Gong and Y. Yan, "Multiple Graph Affinity Interactive Network and A Variable Illumination Dataset for RGBT Image Salient Object Detection," in *IEEE Transactions on Circuits and Systems for Video Technology*, doi: 10.1109/TCSVT.2022.3233131.

[16] Y. Bao, K. Song, J. Wang, L. Huang, H. Dong, Y. Yan, Visible and thermal images fusion architecture for few-shot semantic segmentation, J. Vis. Commun. Image Represent. 80 (2021), 103306.

[17] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108-5115, 2017.

[18] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[19] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5217-5226, 2019.

[20] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9587-9595, 2019.

[21] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, J. Jia, Prior Guided Feature Enrichment Network for Few-Shot Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2) (2022) 1050–1065.

[22] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8334-8343, 2021.

[23] G.-S. Xie, J. Liu, H. Xiong, and L. Shao, "Scale-aware graph neural network for few-shot semantic segmentation," in *Proc. IEEE Conferenceon Computer Vision and Pattern Recognition*, 2021, pp. 5475–5484.

[24] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6941-6952, 2021.

[25] D. Kang, and M. Cho, "Integrative Few-Shot Learning for Classification and Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9979-9990, 2022.

[26] J. Wu et al., "Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[27] Z. Fang, G. Gao, Z. Zhang, A. Zhang, Hierarchical context-agnostic network with contrastive feature diversity for one-shot semantic segmentation, J. Vis. Commun. Image Represent. 90 (2023), 103754.

[28] K. Song, Y. Zhao, L. Huang, Y. Yan, Q. Meng, RGB-T image analysis technology and application: A survey, E*Ngineering Applications of Artificial Intelligence* 120 (2023), 105919.

[29] M. Feng, K. Song, Y. Wang, J. Liu, Y. Yan, Learning discriminative update adaptive spatial-temporal regularized correlation filter for RGB-T tracking, J. Vis. Commun. Image Represent. 72 (2020), 102881.

[30] C. Jiang, Y. Liu, J. Sun, J. Guo, W. Lu, Illumination-based adaptive saliency detection network through fusion of multi-source features, J. Vis. Commun. Image Represent. 79 (2021), 103192.

[31] G. Xu, W. Zhou, X. Qian, L. Ye, J. Lei, L. Yu, CCFNet: Cross-complementary fusion network for RGB-D scene parsing of clothing images, J. Vis. Commun. Image Represent. 90 (2023), 103727.

[32] H. Liu, M. Philipose, M.T. Sun, Automatic objects segmentation with RGB-D cameras, J. Vis. Commun. Image Represent. 25 (4) (2014) 709–718.

[33] B. He, G. Wang, C. Zhang, Iterative transductive learning for automatic image segmentation and matting with RGB-D data, J. Vis. Commun. Image Represent. 25 (5) (2014) 1031–1043.

[34] G. Chen, Z. Jiang, M.M. Kamruzzaman, Radar remote sensing image retrieval algorithm based on improved Sobel operator, J. Vis. Commun. Image Represent. 71 (2020), 102720.

[35] B. Ding, G. Wen, Sparsity constraint nearest subspace classifier for target recognition of SAR images, J. Vis. Commun. Image Represent. 52 (2018) 170–176.

[36] Y. Zhao, K. Song, Y. Zhang, Y. Yan, "BMDENet: Bi-directional Modality Difference Elimination Network for Few-shot RGB-T Semantic Segmentation," in: IEEE Transactions on Circuits and Systems II: Express Briefs, doi: 10.1109/TCSII.2023.3278941.

[37] K. Song, Y. Zhang, Y. Bao, Y. Zhao, Y. Yan, Self-Enhanced Mixed Attention Network for Three-Modal Images Few-Shot Semantic Segmentation, Sensors 23 (14) (2023) 6612.