Adapting Lightweight SAM with Gradient Map for Mirror Object Segmentation

Dongshen Han[a], Chaoning Zhang[a,*], Fachrina Dewi Puspitasari[a], Shuxu Chen[a], Feng Qiao[c], Sheng Zheng[a], Sungyoung Lee[b], Choong Seon Hong[b] and Yang Yang[a]

[a]University of Electronic Science and Technology of China, Chengdu, 611731, China
[b]Kyunghee University, Yongin-si, 17104, Republic of Korea
[c]Washington University in Saint Louis, Saint Louis, 63130, United States of America

ARTICLE INFO

Keywords:
Mirror Object Segmentation
Foundation Models
Segment Anything Model(SAM)
Gradient Adapter

ABSTRACT

Mirror object segmentation has been challenging due to their inapparent features, making them difficult to extract compared to non-mirror objects. To address this challenge, inspired by the success of foundation models (FMs) in numerous vision tasks, this work pioneers applying a segmentation FM to segment mirror objects. Given our preliminary investigation that the recent SAM families (powerful segmentation FMs) fail to accurately segment mirror objects, we propose fine-tuning them on images with mirror objects. We identify that mirror-specific features tend to lie in the mirror boundary, and we exploit these features via a gradient map to adapt the SAM to enhance its mirror segmentation capability. Specifically, we utilize an effective gradient adapter module with the SAM backbone frozen, allowing the model to fuse critical boundary cues from gradient maps with the comprehensive knowledge of foundation models. Experimental results demonstrate that our method, GMSM, achieves competitive performance against existing methods on standard mirror segmentation benchmarks. Moreover, GMSM achieves notable efficiency, requiring 2.24× fewer FLOPs, 4.03× fewer parameters, and running 4.27× faster than EBLNet, making it well-suited for edge deployment.

1. Introduction

Mirror objects are widely used in daily facilities and pose safety risks for unmanned technologies (e.g., robots and drones) [1]. Specifically, the existence of reflective surfaces in mirror objects may jeopardize the vision-based navigation system as it tends to detect and segment objects reflected in the mirror instead of the mirror itself [2]. The mirrors' inapparent features make them difficult to extract compared to non-mirror objects, making the mirror segmentation challenging. Prior works have experimented with various techniques to learn these subtle features. One of the most sought methods is to find the contextually contrasted information between the mirror and surrounding objects, which is extracted hierarchically [3]. While this technique produces rich semantic representations, it still relies on the conventional feature extractors pre-trained on relatively limited data, which may hinder the mining of more subtle features in the mirror.

Foundation models (FMs), such as CLIP [4] and DINOv2 [5] have shown robust feature extraction capabilities across a wide range of vision tasks by learning transferable features from large-scale and diverse data [6]. They have been successfully applied to anomaly detection [7], video segmentation [8], and 3D scene understanding [9]. Inspired by their success, we explore the use of segmentation foundation models (SAM, MobileSAM) [10, 11] and leverage their powerful feature extraction capabilities to address the challenging task of mirror object segmentation.

Despite SAM's proven capability across a wide range of segmentation tasks, we observe that its performance on mirror objects remains unsatisfactory, often lagging behind traditional non-foundation model approaches. We attribute this limitation to the unique properties of mirror objects, which exhibit strong reflective characteristics that can confuse the model and make feature extraction more challenging. While it is common practice to address such issues through fine-tuning on specific datasets [12], naive fine-tuning may fail to accurately capture the distinctive features of mirror objects and can risk compromising the generalization ability of the foundation model [13]. To mitigate this issue, we argue that we can fine-tune SAM on the mirror segmentation dataset using an adapter that supplies mirror-specific

*This document is the result of the research project funded by the National Science Foundation.
chaoningzhang1990@gmail.com (C. Zhang)
ORCID(s):

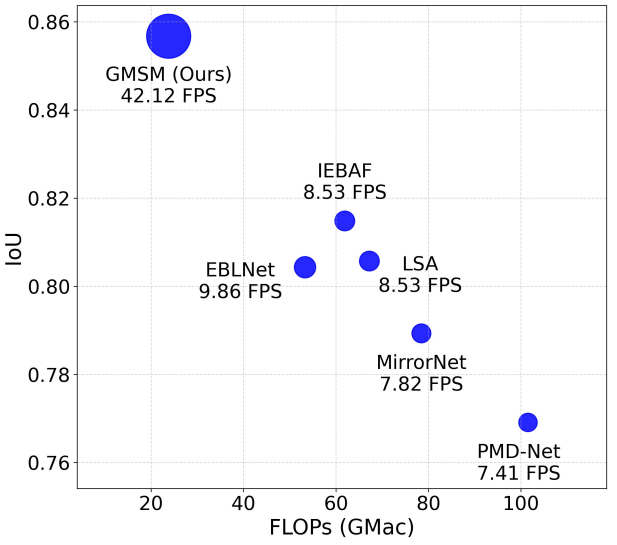


Figure 1: Comparison of IoU, FLOPs, and FPS between GMSM and other mirror object segmentation models on the MSD benchmark. GMSM achieves state-of-the-art performance while offering significant advantages in computational efficiency and inference speed.

representation to the frozen SAM's backbone, integrating domain-specific information without jeopardizing SAM's already-rich representation. These mirror-specific features are particularly the boundary discontinuity that can be extracted from the gradient map.

Unlike traditional edge-aware mechanisms that primarily detect sharp edges, mirror boundaries are more challenging because they often exhibit subtle and smooth gradient transitions caused by reflections. These nuanced boundaries are easily overlooked by conventional edge detectors. In contrast, gradient maps can capture gradual changes in pixel intensity, providing richer and complementary boundary information. By integrating these informative gradient cues with extensive knowledge embedded in the foundation model, our approach enables more accurate feature extraction and segmentation of mirror objects.

Based on the above motivation, we propose the Gradient-based Mirror Segmentation Model (GMSM) that integrates an image gradient learning module into the foundation model SAM. The framework of GMSM consists of a gradient encoder and G-transformer adapters. The gradient encoder efficiently extracts the boundary discontinuity features from the gradient map. The G-transformer adapter then supplies this information to the SAM's frozen image encoder to facilitate learning on both image and mirror features during training. We train our model with a joint weighted intersection over union (IoU) loss and internal boundary loss to encourage learning on image features and the pixel information influenced by local reflections around the internal mirror boundary. In practical implementation, considering the model size and speed, we experiment with SAM two lightweight backbones: ViT-B (proposed in the original SAM paper [10]) and TinyViT (proposed in MobileSAM [14]). As a result of our effective adapter design and the use of a lightweight backbone, the proposed GMSM achieves a 4.1% improvement in IoU over previous methods, along with a 4× speed-up in inference and a 2.24× reduction in FLOPs (see Figure 1). This combination of high accuracy and efficiency is particularly valuable for mirror object segmentation tasks commonly employed on resource-constrained platforms such as UAVs and robotics.

This work focuses on the mirror object segmentation task by utilizing a segmentation FM. Overall, we summarize the contributions of our work as follows.

- We demonstrate that SAM families (segmentation FMs) fail in the mirror segmentation task due to the inability to capture mirror-specific features, and we propose to solve this by fine-tuning with an effective adapter.
- Motivated by the observation that mirror-specific features are often concentrated along boundaries, we propose to exploit the gradient map for adapting the SAM to mirror object segmentation.

- Extensive experiments show that GMSM achieves state-of-the-art accuracy on mirror object segmentation benchmarks, outperforming prior methods with a 4.1% improvement in IoU. Moreover, it is 4× faster and 2.24× more efficient in FLOPs, making it highly suitable for real-time deployment on edge devices.

2. Related works

2.1. Mirror Object Segmentation

Mirror object segmentation is a task that involves recognizing mirror regions within a single RGB image. Yang *et al.* [2] first perform the first method for automatic mirror object segmentation in RGB images by leveraging neural network backbones to extract multi-level and contextual contrast information between regions inside and outside the mirror. Lin *et al.* [3] gathered mirror images to establish a more rigorous PMD benchmark and introduced PMD-Net, a deep network utilizing a refinement network to extract mirror regions based on relational contextual contrast features and edge information. Guan *et al.* [15] propose a semantic side path to utilize semantic information to recognize mirror object regions and employ neural network contextual relationships to segment mirror regions. Mei *et al.* [16] and Zhou *et al.* [17] utilize additional depth information to assist in segmenting mirror objects from RGB images using neural networks. Several advanced transformer-based [18] and diffusion-based segmentation methods [19] have been proposed, but they often overlook the crucial boundary regions necessary for mirror object identification, leading to confusion with reflected content.

In addition, boundary discontinuity is a crucial feature in recognizing mirror objects. He *et al.* [20] argue that the inner object appearance of mirror objects and their surrounding backgrounds, as utilized in previous works, do not significantly contribute to segmentation due to their inherent confounding nature; instead, they emphasize the crucial importance of boundary information. Han *et al.* [21] find that mirror objects often accompany external boundary features like a frame, differing from internal boundaries and thus design a network to utilize these features for mirror object segmentation. While they both succeeded in automating mirror object detection [3], limitations in training data and methods led to inaccurate extraction of object features and the inclusion of irrelevant contextual features, confusing the training process. Furthermore, these methods [2] rely on large parameter neural networks to extract discontinuous context and boundary features, which hinders their application in real scenarios where automatic mirror detection is required, such as autonomous vehicles. In this work, we address these challenges by leveraging gradient map information, which provides both abrupt and smooth transitions in intensity, providing a more comprehensive representation of mirror boundaries than conventional edge-aware approaches. We further leverage extensive representation knowledge from the foundation model, thereby inheriting its robust and precise feature extraction capability to accurately extract features of challenging mirror objects.

2.2. Foundation Model and Fine-tuning Adaptation

Foundation models, which are trained on vast amounts of data and possess strong generalization capabilities, are receiving widespread attention from researchers. Brown *et al.* [22] proposed GPT-3, which has been widely recognized as one of the most prominent foundation models for natural language processing (NLP) and serves as a key component behind the success of ChatGPT [23]. Radford *et al.* [4] introduced the foundation model Contrastive Language–Image Pre-training (CLIP), which enables the generation of outputs based on textual instructions [24]. In the realm of vision, Meta AI's public vision foundation model SAM, which exhibits exceptional generalization capabilities for object segmentation, has gained substantial attention [25]. Its remarkable capabilities have enabled its application in various domains, including image editing [26], inpainting [27]. Recently, Ravi *et al.* [28] proposed the Segment Anything Model 2 (SAM 2), which is developed as a foundational video model for advanced visual segmentation. SAM 2 leverages a data engine that improves performance through user interactions with the largest video segmentation dataset.

To enable efficient fine-tuning of large pre-trained models with only a small number of additional parameters, Houlsby et al. [29] first introduced adapters for NLP tasks. Pan *et al.* [30] propose ST-Adapter for efficient fine-tuning in video tasks, enabling pre-trained image models to handle dynamic video content with minimal parameter overhead. Stickland *et al.* [31] investigated multi-task learning using a shared BERT model combined with adapters specific to each task. Gao *et al.* [32] introduce CLIP-Adapter, a method for enhancing vision-language models through feature adapters and residual-style blending, offering an alternative to prompt tuning. Wu *et al.* [33] propose the Medical SAM Adapter (Med-SA), which integrates domain-specific medical knowledge into the foundation model for application in the field of medical image segmentation. Recently, it has been found that upsampling and downsampling modules can serve as components of plain ViT for object detection [34] and Video Depth Estimation [35]. The adapter, a lightweight

Table 1
Qualitative results of various SAM families in the MSD benchmark.

Methods	Backbone	IoU ↑	F_β ↑	Accuracy ↑	MAE ↓
MobileSAM	TinyViT	42.93	0.7705	85.59	0.14409
SAM	ViT-B	35.54	0.7128	84.20	0.15801
SAM	ViT-H	51.57	0.8176	81.74	0.12418

module, modifies the features extracted by models trained on large datasets to make them suitable for downstream tasks. We utilize lightweight adapter modules to effectively fuse mirror-specific features into the well-established segmentation representations of the foundation model.

3. Method

Figure 2: Visualization of SAM predicted results with mirror object images. Highlighted points and regions indicate the SAM-predicted mask and the prompt area, respectively.

3.1. SAM for Mirror Object Segmentation Without Adaptation

SAM, as a segmentation foundation model, is not capable of handling mirror object segmentation. We conduct an initial experiment to prove this limitation. Specifically, we select a number of images with mirrors present and supply them to the SAM together with five high-quality prompts at different positions. In the predicted output masks (Figure 2), we notice that SAM prefers to segment the objects being reflected in the mirror, but not the mirror itself. We conjecture that this is largely caused by the reflective nature of the mirror, which makes the extraction of the subtle true mirror feature challenging. This qualitative evidence is further supported by the IoU evaluation metric, which employs huge ViT-H as its backbone and only yields IoU of 51.57 (Table 1), indicating its poor proficiency in mirror segmentation.

3.2. Gradient-based Adaptation

Based on previous investigations, the SAM has unsatisfactory performance for segmenting mirror objects. Based on the above finding, we propose a Gradient-based Mirror Segmentation Model (GMSM) that adapts a well-trained SAM with a gradient adapter. The GMSM modifies the image encoder of the foundation model by coupling both transformer adapters [36] and gradient adapters into it. Figure 3 illustrates the construction of GMSM.

3.2.1. Gradient Encoder

The gradient encoder extracts the boundary discontinuity (Figure 4) information from the gradient map. Specifically, given an input image I, we generate the gradient map and supply it as an input to the gradient encoder. The gradient map refers to the differences between adjacent pixels that we calculate with the following formula.

$$\begin{aligned} \nabla I_x(\mathbf{x}) &= I(x+1, y) - I(x-1, y), \\ \nabla I_y(\mathbf{x}) &= I(x, y+1) - I(x, y-1), \\ G(I) &= \|(\nabla I_x(\mathbf{x}), \nabla I_y(\mathbf{x}))\|_2, \end{aligned} \tag{1}$$

where $G(I)$ denotes the extraction function with its elements representing the magnitudes of gradients for pixels located at coordinates (x, y).

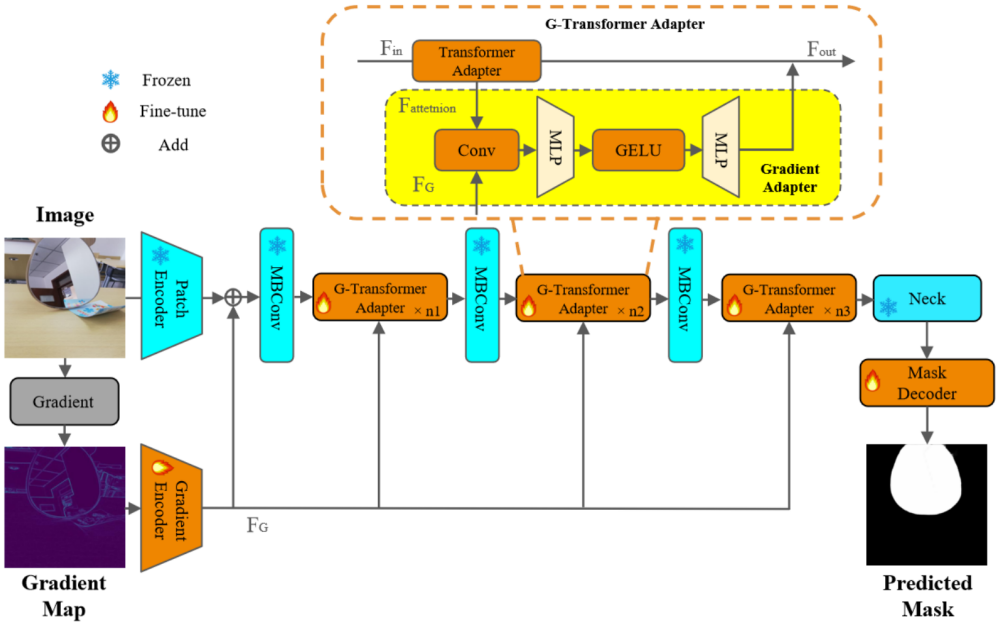
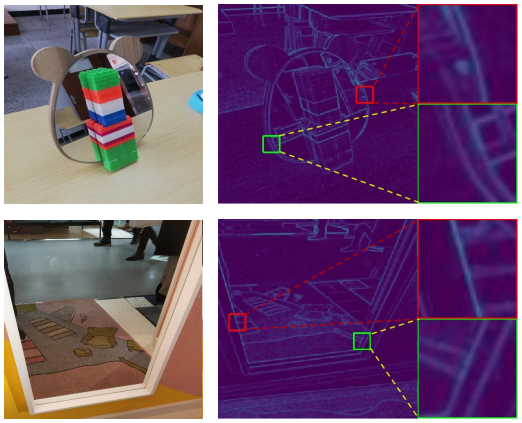


Figure 3: Overview of our proposed method. We employ a lightweight foundation model as the backbone, fine-tuning the G-Transformer, gradient encoder, and MobileSAM mask decoder, while keeping the remaining modules frozen.



(a) Image (b) Gradient Map

Figure 4: (a) The original image containing a mirror object. (b) The corresponding gradient map, with zoomed-in details of the boundary region. The zoomed-in areas of the gradient map reveal high values and display chaotic, discontinuous features, especially along the mirror's boundary.

Our gradient encoder consists of a gradient extractor block and an upsampling depth-wise convolution block. On the extractor block, we use 3×3 convolution layers, the same as the patch embedding on the image feature extractor route, while we halve the channel dimension of its output. To match the dimension of the patch embedding output, we design our upsampling depthwise convolution to comprise a GELU activation layer and a 1×1 convolution layer. Then, we represent our gradient encoder as follows:

$$\begin{aligned} F_{embedding} &= \mathrm{Conv}_{3\times3}\left(\mathrm{GELU}\left(\mathrm{Conv}_{3\times3}\left(G(I)\right)\right)\right), \\ F_g &= \mathrm{Conv}^{i}_{1\times1}\left(\mathrm{GELU}\left(F_{embedding}\right)\right). \end{aligned} \tag{2}$$

Finally, we add the boundary discontinuity feature, F_g, to the output of the patch embedding and supply it as input to the image encoder. The specific network structure of our gradient encoder is illustrated in Figure **??**. The input image is

first processed to compute gradients along its three channels, capturing pixel value changes. These gradient features, F_g, are then passed through two convolutional layers to increase the feature dimension to 96. Subsequently, we employ an MBConv layer and use interpolation to align both the channel dimension and spatial resolution with those required by the corresponding backbone module.

3.2.2. G-Transformer Adapter

The main goal of the G-Transformer adapter is to improve the capacity of SAM's image encoder to learn mirror-specific features alongside the overall image features by integrating the boundary discontinuity features F_g with the image features. G-Transformer adapters replace the transformer blocks in the image encoder at every stage. The construction of the G-Transformer adapter follows the basic transformer adapter, coupled with the gradient adapter. The basic transformer adapter comprises basic attention and multilayer perceptrons (MLPs), which are inspired by the transformer's multi-head attention and MLPs that adjust the feature maps' sparsity and channel dimensions. On the other hand, the gradient adapter consists of a sequence of channel fusion depth-wise convolution layers, downsampling MLP, GELU activation layer, upsampling MLP, and residual connection. We concatenate the output of the transformer adapter $F_{attention}$ with that of the gradient encoder F_g in the channel fusion and generate fusion features F_{fusion}. We further reduce the dimension of F_{fusion} with a 1×1 convolution layer and batch normalization. Subsequently, we send the F_{fusion} to the sequence of downsampling MLP, GELU layer, and upsampling MLP to produce the final output features F_{out}. We employ simple MLP layers to do this final processing to ensure lightweight parameterization while also learning the dimensional features. Overall, we express the operations mentioned above as follows.

$$\begin{aligned} F_{fusion} &= \mathrm{Conv}^{i}_{1\times1}\left(\mathrm{Concat}\left(F_{attention}, F_g\right)\right), \\ F_{out} &= \alpha \cdot \mathrm{MLP}^{i}_{up}\left(\mathrm{GELU}\left(\mathrm{MLP}^{i}_{down}\left(F_{fusion}\right)\right)\right), \end{aligned} \tag{3}$$

with i representing the i-th block of transformers in each stage and the MLP_{down} and MLP_{up} respectively performing downsampling and upsampling by a factor of two. We set the hyperparameter α to 0.2. Note that we freeze all parameters of the image encoder during the optimization of the gradient encoder and adapter. The network consists of three G-Transformer Adapter blocks, with embedding dimensions of $N \in 64, 128, 160$ at each block, respectively. At every block, the spatial resolution of the feature map is reduced by half, for example, from 256 at the input to 64 at the final stage. In each block, we concatenate the feature gradient F_g with the input embedding, and use a convolutional layer to reduce the combined dimension from $N + 64$ to N. This is followed by two MLP layers: the first reduces the dimension by half, and the second restores it to N before producing the final output.

3.2.3. Total Loss Function

The complete objective function for segmenting mirror objects is the combination of learning the global extremities of all objects and the local internal boundary of the mirror in an image.

The internal boundary loss L_{inBCE} is designed to make the network focus more on learning the boundary pixels, with particular emphasis on the internal boundaries of mirrors where reflections cause prominent discontinuities. These boundary pixels, although important, constitute a small portion of the image and are underemphasized during the learning process, especially in pixel-wise binary classification tasks. To address this, we introduce a spatial weight map W^C within the cross-entropy loss, which assigns higher weights to pixels in the mirror's internal regions. The spatial weight map is defined as follows:

$$W^C = 2 - G^{th}_{In} \odot Gaussian(G^{th}_{In}), \tag{4}$$

where G^{th} represents the ground truth mirror region. The map is generated by applying a Gaussian smoothing filter with a kernel size of 11 and a sigma of 7, resulting in W^C values ranging from just above 1 to less than 2, with higher values at the boundaries. Incorporating W^C into the binary cross-entropy (BCE) loss gives us the internal boundary loss:

$$\begin{aligned} L_{inBCE} = -\sum_{x,y} W^C(x,y) \cdot (P(x,y) \cdot \log G^{th}(x,y) \\ +(1 - P(x,y)) \cdot \log(1 - G^{th}(x,y))), \end{aligned} \tag{5}$$

where (x, y) represents the pixel coordinates and P is the predicted map. This approach ensures that the network gives appropriate attention to the boundary pixels during training.

Moreover, IoU loss functions, widely used in semantic segmentation, provide macro-level supervision by focusing on foreground regions to ensure complete and accurate segmentation results. We formulate the IoU loss as follows,

$$L_{IoU} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} P(x,y)G(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W}[P(x,y) + G(x,y) - P(x,y)G(x,y)]}. \tag{6}$$

Here, $P(x, y)$ and $G(x, y)$ represent the predicted and ground truth values at pixel (x, y), respectively.

Comprising of internal boundary loss L_{inBCE} and IoU loss L_{IoU}, our hybrid loss function is shown as follows:

$$L_{hybird} = \beta L_{IoU} + \lambda L_{inBCE}, \tag{7}$$

where the balancing weights β and λ are set to 3 and 1, respectively.

3.3. Backbone Choice in Practical Implementation

Our proposed GMSM adopts the image encoder in SAM as the backbone. In practice, mirror segmentation methods are mainly used in unmanned technologies such as robots, drones, and autonomous vehicles [1], thus requiring these methods to be lightweight and real-time. The SAM family provides a wide variety of image encoders with different sizes. Specifically, the original SAM paper introduces three image encoders, out of which we experiment with ViT-b (the most lightweight one). In addition, MobileSAM [14] is one of the pioneering frameworks that improve the efficiency of SAM by distilling its image encoder into the lightweight TinyViT [37]. To this end, we further experiment with its image encoder TinyViT, which includes a downsampling module with patch embedding and three-stage blocks comprising MBConv [38] blocks followed by a transformer [39] block, as depicted in Figure 3. The neck block, positioned at the end of the image encoder, adjusts the input dimensions to align with those required by the mask decoder. As shown in Table 2, TinyViT from MobileSAM is empirically found to be more effective than ViT-b from the original SAM in our investigation. Considering its effectiveness and efficiency, we choose the TinyViT image encoder as the final backbone. We leave the search for the optimal image encoder to future work. It is worth mentioning that we also use the mask decoder in the SAM family to inherit the robust decoding capability provided by SAM. We discard the prompt encoder since GMSM employs an automatic segmentation mode,

4. Experiments

4.1. Experimental setup

4.1.1. Dataset

We experiment on two mirror segmentation datasets, MSD [2] and PMD [3]. MSD contains 3,063 training and 955 test images with mirrors occupying large pixel regions, as the pictures were taken close to the mirror. The scenes captured in MSD are of common daily life, which pose fewer challenges for the network to learn. PMD consists of 5,095 training and 571 test images of mirrors in diverse indoor and outdoor scenes.

4.1.2. Evaluation Metrics

We evaluate our results on mirror segmentation using four common evaluation metrics: intersection over union (IoU), pixel accuracy (Acc), weighted F_β [48], and mean absolute error (MAE). F_β is a harmonic mean of average precision and average recall defined as follows.

$$F_\beta = \frac{(1+\beta^2)(Precision \times Recall)}{\beta^2 Precision + Recall}, \tag{8}$$

where β^2 is set to 0.3 [49]. MAE is commonly used in foreground-background segmentation tasks to calculate the average pixel-wise error between the predicted mask P and the ground truth mask G.

$$MAE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} \mid P(i,j) - G(i,j) \mid, \tag{9}$$

where $P(i, j)$ indicates predicted probability at location (i, j).

Table 2
A comparison of various methods on the MSD dataset, with $*$ indicating adaptations for derivative models of SAM. The best results are indicated in **bold**, and dashed lines separate methods from different task domains.

Method	Public. Year	IoU ↑	F_β ↑	Accuracy ↑	MAE ↓
Semantic segmentation					
CPNet [40]	CVPR2020	69.86	0.8314	92.44	0.07603
GloRe [41]	CVPR2019	71.95	0.8406	92.33	0.06957
PSPNet [42]	CVPR2017	67.99	0.8459	91.29	0.07875
Camouflage object and shadow segmentation					
BDRAR [43]	CVPR2018	75.37	0.8619	93.50	0.06510
DSC [44]	ECCV2018	75.36	0.8479	92.82	0.07206
Salient object segmentation					
MINet [45]	CVPR2020	66.42	0.8172	92.78	0.08842
VST [46]	ICCV2021	79.57	0.8772	93.89	0.05421
EGNet [47]	ICCV2019	69.61	0.8238	91.54	0.08479
Mirror segmentation					
MirrorNet [2]	CVPR2020	78.93	0.8597	93.55	0.07257
PMD-Net [3]	CVPR2020	76.94	0.8691	93.94	0.06130
EBLNet [20]	ICCV2021	80.33	0.8839	93.64	0.04953
LSA [15]	CVPR2022	79.85	0.8887	94.63	0.05421
IEBAF [21]	Pub. 2023	81.48	0.8990	95.27	0.04733
Segmentation foundation models					
SAM [10]	ICCV2023	35.54	0.7128	84.20	0.15801
MobileSAM [11]	Pub. 2023	42.93	0.7705	85.59	0.14409
GMSM (SAM*)	-	**84.62**	**0.9198**	**95.72**	**0.04520**
GMSM (MobileSAM*)	-	**85.67**	**0.9244**	**96.28**	**0.04401**

4.2. Main Results

4.2.1. Performance Comparison

We validate our method by comparing it with 18 methods from related fields, including semantic segmentation, salient object segmentation, camouflage object and shadow segmentation, mirror segmentation, and segmentation foundation models. To ensure a fair comparison, we use either their publicly available codes or implementations with recommended parameter settings. Except for the zero-shot foundation models, all models are retrained on their respective training sets as described in their papers, and all prediction maps are evaluated using the same code.

Table 2 presents these results on the MSD dataset and Table 3 on the PMD dataset. Our method consistently outperforms all compared approaches. In particular, it achieves improvements of 4.19% in IoU, 0.0323 in F_β, 1.01% in accuracy, and a reduction of 0.0033 in MAE compared to the previous best method, IEBAF, on the MSD dataset. In addition, compared to MobileSAM, our GMSM (MobileSAM*) significantly enhances performance on mirror object segmentation. It achieves a 42.74% improvement in IoU, a 0.1539 increase in F_β, a 10.69% boost in accuracy, and a 0.1001 reduction in MAE. These results confirm the effectiveness of our gradient-guided adaptation in enhancing segmentation quality while retaining the lightweight efficiency of SAM. Moreover, compared with models designed for related tasks such as camouflage, shadow, and salient object segmentation, GMSM demonstrates superior performance. This advantage stems not only from leveraging the broad visual knowledge of foundation models (FMs), but also from the use of gradient maps, which effectively emphasize boundary information crucial for accurate mirror object recognition.

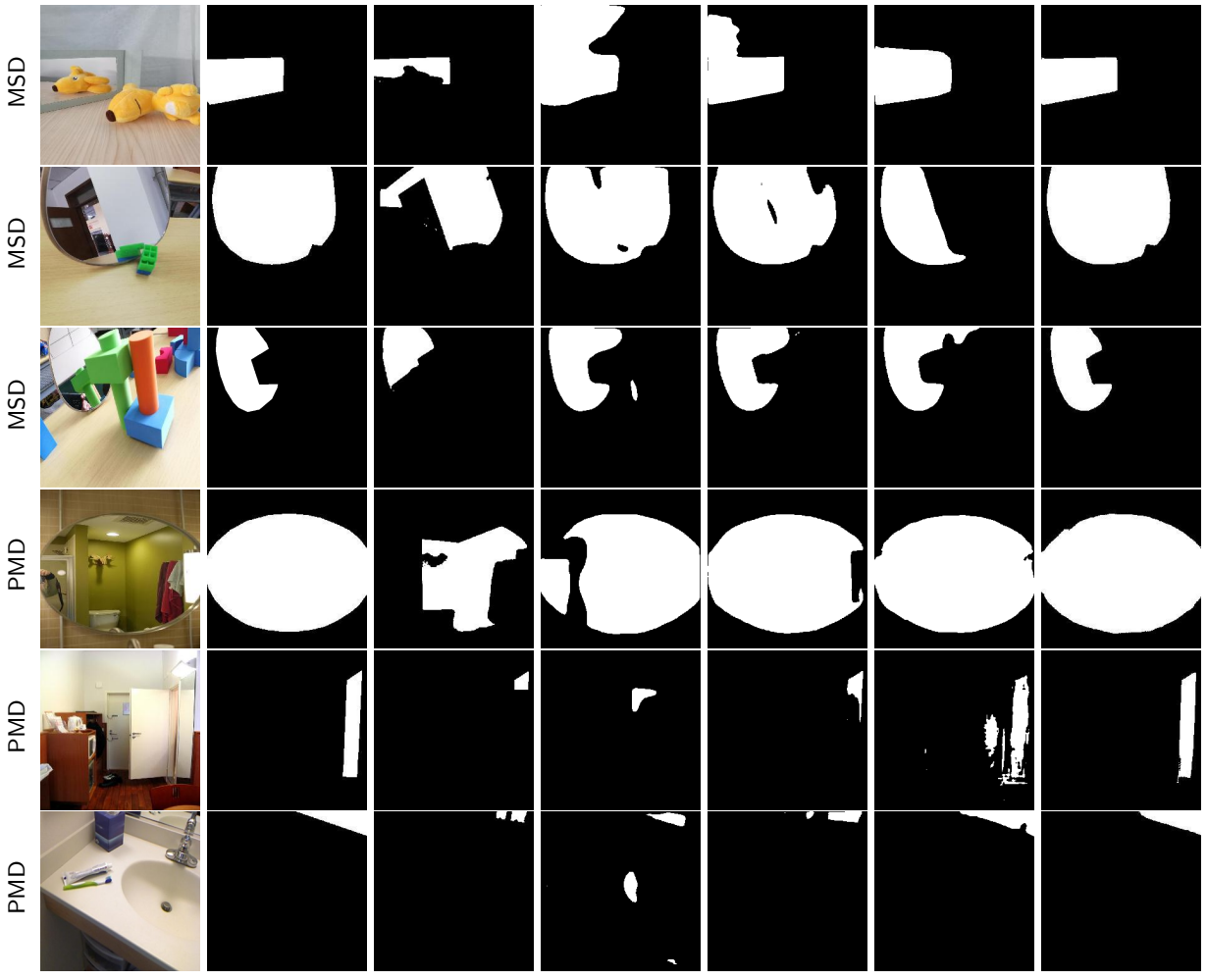


Figure 5: Qualitative comparison results on MSD and PMD benchmark.

4.2.2. Efficiency Comparison

To evaluate the efficiency of GMSM, we conduct a comprehensive comparison of FLOPs (computational complexity), FPS (frames per second), and model parameters with nine state-of-the-art methods, including both mirror detection and classical salient object detection approaches, as shown in Table 4. To ensure fairness, all experiments are conducted on the same device, as the hardware environment influences FPS, and each method is run 100 times to calculate the average.

Despite our method having a significantly larger image input compared to other methods, it surpasses others in both FPS and the number of parameters, making it suitable for deployment on edge devices. This also indicates that our method can potentially be applied in real-time detection for unmanned technologies.

4.2.3. Qualitative Results

Figure 5 shows the qualitative evaluation of GMSM on MSD and PMD datasets. Overall, compared to other methods, our approach accurately identifies mirror regions. The first four rows present close-up images with large mirror regions, resulting in rich and complex reflections. In these cases, our method correctly detects the mirrors, whereas other methods are affected by the reflections. The last two rows show distant mirrors occupying smaller regions in the images. In such scenarios, both the human eye and other methods struggle to recognize mirror regions, but our method excels in accurately identifying them, outperforming other approaches.

Table 3
Comparison of different methods on PMD Dataset, with $*$ indicating adaptations for derivative models of SAM. The best results are indicated in **bold**, and dashed lines separate methods from different task domains.

Method	IoU ↑	MAE ↓	Acc ↑
CPNet [40]	56.36	0.051	94.85
GloRe [50]	61.25	0.044	95.61
PSPNet [42]	60.44	0.039	96.13
BDRAR[43]	58.43	0.043	95.66
MirrorNet [2]	62.50	0.041	96.27
PMD-Net [3]	62.40	0.055	96.80
LSA [15]	66.84	0.049	96.82
EBLNet [20]	67.15	0.042	96.21
MobileSAM [11]	56.25	0.072	93.15
SAM [10]	64.75	0.053	94.75
GMSM (ours)	**67.79**	**0.031**	**96.87**

Table 4
Quantitative comparison of efficiency. We compare our model with relevant state-of-the-art models in terms of the number of parameters (Param.), FLOPs, and FPS.

Method	FLOPs↓	Param.↓	FPS↑
EGNet [47]	156.27	111.64	10.76
MINet [45]	89.12	162.38	3.55
VST [46]	25.88	44.48	6.05
MirrorNet [2]	78.48	121.77	7.82
PMD-Net [3]	101.54	147.66	7.41
LSA [15]	67.23	104.80	8.53
EBLNet [20]	53.31	46.21	9.86
IEBAF [21]	61.91	65.43	8.53
GMSM (ours)	**23.81**	**11.47**	**42.12**

4.3. Ablation Study

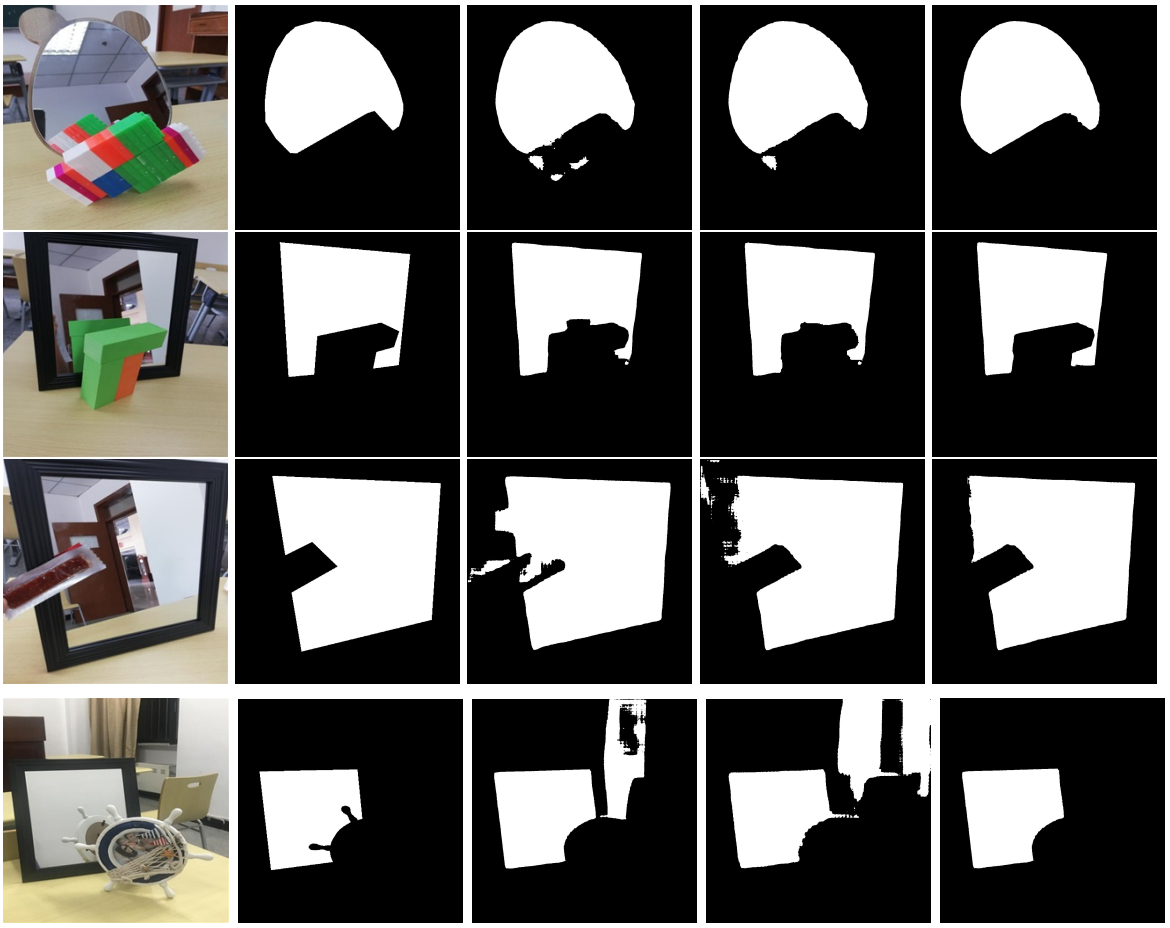
To analyze the importance of each component and loss function in our method, we conduct an ablation study on the MSD benchmark.

4.3.1. Attaching Adapter and Freezing Image Encoder

In this ablation study, we would like to learn the effect of the combination of attaching an adapter and choosing whether or not to freeze the image encoder. Consequently, we devise three scenarios as follows:

- "**MobileSAM**" refers to fine-tuning all MobileSAM image encoders' parameters based on the pre-trained checkpoint.
- "**MobileSAM + A**" refers to performing the same action as the above scenario and attaching adapters.
- "**MobileSAM* + A**" refers to freezing the MobileSAM's image encoder and attaching adapters.

Table 5 shows that attaching the adapter module improves the performance of the standalone foundation model. Specifically, IoU, F_β, Acc, and MAE improve by 0.97, 0.016, 0.8, and 0.0087, respectively. This improvement proves that our proposed adapter-based network is helpful for the segmentation foundation model to learn the subtle mirror features from an input image. Further, we found that in addition to attaching the adapter, freezing MobileSAM's image encoder during training further improves the performance indicated by IoU, F_β, Acc, and MAE, which improve by 2.19, 0.021, 2.31, 0.0143, respectively. Note that the scale of improvements is noticeably larger than solely attaching an adapter to a non-frozen image encoder. We conclude that this result is primarily caused by preserving the existing formidable segmentation knowledge from the foundation model.



(a) Image (b) GT (c) TinyViT-5M (d) TinyViT-11M (e) TinyViT-21M

Figure 6: Qualitative results of more challenging images with different image encoders as backbone.

Table 5
Effect of attaching adapters and freezing the image encoder. Here, * indicates that the MobileSAM image encoder is frozen during training, A denotes adapter-based fine-tuning, and G refers to gradient features.

	IoU↑	F_β↑	Acc↑	MAE↓
MobileSAM	79.96	0.8721	92.87	0.07231
MobileSAM + A	80.93	0.888	93.67	0.06360
MobileSAM* + A	83.12	0.909	95.98	0.04932
MobileSAM* + A + G	85.67	0.9244	96.28	0.04401

4.3.2. Effectiveness of Gradient Encoder

We verify the effectiveness of the gradient encoder in GMSM that supplies the boundary discontinuity features needed for learning. We experiment by removing the gradient encoder from GMSM so that we force our model to extract the subtle mirror-specific features from only the RGB input image during training. As shown in Table 5, all learning scenarios without gradient encoder are capped by the scenario with gradient encoder ("**MobileSAM* + A + G**"). The presence of a gradient encoder improves the best scenario from the previous ablation study in IoU, F_β, Acc, and MAE by 2.55, 0.0154, 0.3, and 0.005, respectively. The largest improvement in IoU indicates that boundary discontinuity features extracted from the gradient map are indeed a substantial piece of information for the mirror segmentation task, as the reflections presented by mirrors only add to the confusion.

Table 6
Effect of different loss terms on MSD.

	IoU↑	F_β↑	Acc↑	MAE↓	BF ↑
L_{BCE}	81.28	0.9073	95.17	0.04976	0.8071
$L_{BCE} + L_{IoU}$	84.78	0.9113	95.98	0.04803	0.8322
$L_{inBCE} + L_{IoU}$	85.67	0.9244	96.28	0.04401	0.8418

Table 7
Comparison on different backbone image encoders on the MSD benchmark.

Backbone	IoU↑	F_β↑	Acc↑	MAE↓
TinyViT-5M	85.67	0.9244	96.28	0.04401
TinyViT-11M	86.97	0.9289	96.57	0.03821
TinyViT-21M	88.21	0.9353	96.75	0.03329

4.3.3. Effectiveness of Joint Loss Function

We explore the impact of the joint loss function L_{joint} by incrementally fixing individual elements starting from the classic BCE L_{BCE}. In the second scenario, we couple it with the IoU loss L_{IoU} where we achieve a gain of 3.5 in IoU. Next, we fix the spatial weight map into the BCE to produce the internal boundary loss L_{inBCE} as presented in Equation 5. Thus, in the third scenario, we apply the complete joint loss function as formulated in Equation 7. This setup yields a gain of 0.89 in IoU compared to the second scenario, which supports our conjecture that giving more attention to the subtle mirror features helps to balance how the model treats all pixels in the image during learning for the specific mirror segmentation task. Furthermore, we report the Boundary F1-score (BF score) in our experiments. The results demonstrate that our proposed L_{inBCE} loss significantly improves boundary accuracy compared to baseline methods, further highlighting the effectiveness of L_{inBCE} in enhancing edge-aware segmentation. We summarise these results in Table 6.

4.3.4. Choice on Backbone Image Encoder

We conduct experiments using various types of image encoders provided by MobileSAM, including TinyViT-11M and TinyViT-21M. These image encoders inherit the knowledge of the foundation model through knowledge distillation, with the main difference being the embedding dimensions for each stage of convolution. Specifically, TinyViT-5M, TinyViT-11M, TinyViT-21M have embedding dimension of {64, 128, 160, 320}, {64, 128, 256, 448}, and {96, 192, 384, 576}, respectively. Larger embedding dimensions can enhance the model's ability to extract object features effectively as shown in Table 7. Notice that as the embedding dimensions increase, the resulting IoU also improves. We further validate this performance by handpicking challenging images where mirrors are occluded. Qualitative results in Figure 6 show that image encoders with more parameters provide better mask predictions and are less affected by reflective patterns and object occlusions. Although this improvement comes with an increase in model parameters and a reduction in FPS, we argue that our lightweight architecture can still meet the requirement of low-latency mirror segmentation when deployed on edge devices.

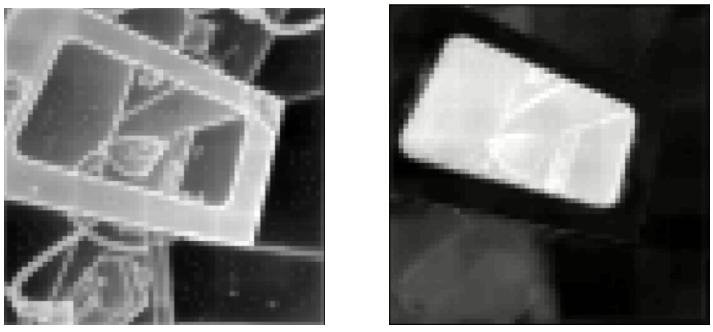
4.4. Ablation Study of Model Hyperparameters

We conduct an ablation study to investigate the effects of key model parameters, including the loss function kernel size σ, as well as the hyperparameters β and λ, which control the relative contributions of individual loss components. As shown in Table 8, the choice of these parameters significantly affects the model's performance. Based on the results, we set the kernel size, β, and λ to 7, 3, and 1, respectively, for all subsequent experiments. As illustrated by the results, an excessively large λ causes the model to focus too much on fine details while neglecting the global structure, whereas an overly large β leads the model to ignore important details. Additionally, a large kernel size results in a significant amount of non-boundary information being included in the loss computation, which can negatively impact segmentation accuracy.

5. Discussion

Table 8
Sensitivity analysis of kernel size, β, and λ. The numbers in parentheses indicate the corresponding hyperparameter values.

Parameter	Value 1	Value 2	Value 3
Kernel size	84.12 (5)	**85.67 (7)**	84.55 (9)
β	85.17 (2)	**85.67 (3)**	85.46 (4)
λ	**85.67 (1)**	85.03 (2)	84.92 (3)



(a) w gradient adapter (b) w/o gradient adapter

Figure 7: Visualization of principal components in feature maps with and without gradient adapter fine-tuning.

The superior performance of GMSM can be attributed to its ability to guide the foundation model (e.g., SAM) toward mirror-specific visual cues. Mirror objects pose unique challenges such as internal reflections and ambiguous textures, which often confuse general-purpose segmentation models. By introducing a gradient map as an auxiliary supervisory signal, GMSM explicitly encourages the model to focus on boundary regions rather than misleading reflective appearances. To further support our analysis, we conduct principal component analysis (PCA) and visualize the feature space before and after fine-tuning with our method. The results shown in Figure 7, GMSM effectively guides the model to extract more discriminative, mirror-specific features, highlighting the benefits of our adaptation strategy. Moreover, our results demonstrate that while foundation models are powerful, they may overlook subtle yet critical low-level features in domain-specific tasks. GMSM bridges this gap through a lightweight, gradient-based adaptation, resulting in consistent performance improvements across multiple benchmarks.

6. Conclusion

Our work is the first to adapt segmentation foundation models (e.g., SAM) for mirror object segmentation and reveals their limitations on this challenging task. To address this, we propose GMSM, a lightweight adapter that incorporates gradient-based boundary information to enhance mirror-specific representations. Extensive experiments on two benchmark datasets demonstrate that GMSM achieves state-of-the-art performance with significantly reduced computational complexity (2×), faster inference speed (4×), and fewer parameters (4×). Comprehensive ablation studies further validate the effectiveness of key components, including the frozen image encoder, the gradient-guided adapter, and the hybrid loss. This study opens a new direction for applying foundation models to mirror segmentation, and future work may explore additional mirror-specific priors to further improve adaptation.

CRediT authorship contribution statement

Dongshen Han: Writing – original draft, Validation, Methodology, Conceptualization,Software. **Chaoning Zhang**: Writing – review ς editing, Supervision, Investigation, Visualization, **Fachrina Dewi Puspitasari**: Conceptualization, Formal analysis, Data curation, Writing, Visualization. **Shuxu Chen**: Writing – review ς editing, Supervision, Software. **Sheng Zheng**: Writing – review ς editing, Supervision, Investigation. **Sungyoung Lee**: Writing – review ς editing, Supervision, Funding acquisition, Validation. **Choong Seon Hong**: Conceptualization, Project administration, Supervision. **Yang Yang**: Conceptualization, Project administration, Supervision, Validation.

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

The data used for this work is publicly available and can be found at GitHub.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62220106008. It was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation). In addition, this work was supported by the Ministry of Science and ICT (MSIT), Korea, through the ITRC (Information Technology Research Center) support program (RS-2023-00259004), supervised by the IITP.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[2] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8809–8818, 2019.

[3] Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3697–3705, 2020.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[6] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[7] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

[8] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. Clip-vis: Adapting clip for open-vocabulary video instance segmentation. *arXiv preprint arXiv:2403.12455*, 2024.

[9] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[11] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023.

[12] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023.

[13] Fengqian Ding, Chen Xu, Han Liu, Bin Zhou, and Hongchao Zhou. Bridging pre-trained models to continual learning: A hypernetwork based framework with parameter-efficient fine-tuning techniques. *Information Sciences*, 674:120710, 2024.

[14] Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579*, 2023.

[15] Huankang Guan, Jiaying Lin, and Rynson WH Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2022.

[16] Fengze Li, Jieming Ma, Zhongbei Tian, Ji Ge, Hai-Ning Liang, Yungang Zhang, and Tianxi Wen. Mirror-yolo: An attention-based instance segmentation and detection model for mirrors. *arXiv preprint arXiv:2202.08498*, 2022.

[17] Wujie Zhou, Yuqi Cai, Liting Zhang, Weiqing Yan, and Lu Yu. Utlnet: Uncertainty-aware transformer localization network for rgb-depth mirror segmentation. *IEEE Transactions on Multimedia*, 2023.

[18] Maxim Khomiakov, Michael Riis Andersen, and Jes Frellsen. Geoformer: A multi-polygon segmentation transformer. *arXiv preprint arXiv:2411.16616*, 2024.

[19] Zhihao Shuai, Yinan Chen, Shunqiang Mao, Yihan Zho, and Xiaohong Zhang. Diffseg: a segmentation model for skin lesions based on diffusion difference. *arXiv preprint arXiv:2404.16474*, 2024.

[20] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15859–15868, 2021.

[21] Dongshen Han and Seungkyu Lee. Internal-external boundary attention fusion for glass surface segmentation. *arXiv preprint arXiv:2307.00212*, 2023.

[22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

[23] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, et al. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*, 2023.

[24] Defeng Xie, Ruichen Wang, Jian Ma, Chen Chen, Haonan Lu, Dong Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing, 2023.

[25] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[27] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[29] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[30] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

[31] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019.

[32] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[33] J Wu, Y Zhang, R Fu, H Fang, Y Liu, Z Wang, Y Xu, and Y Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. arxiv 2023. *arXiv preprint arXiv:2304.12620*.

[34] Y Li, S Xie, X Chen, P Dollar, K He, and R Girshick. Benchmarking detection transfer learning with vision transformers. arxiv 2021. *arXiv preprint arXiv:2111.11429*.

[35] Bowen Zhao, Hongdou He, Hang Xu, Peng Shi, Xiaobing Hao, and Guoyan Huang. Lda-mono: A lightweight dual aggregation network for self-supervised monocular depth estimation. *Knowledge-Based Systems*, 304:112552, 2024.

[36] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[37] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 68–85. Springer, 2022.

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[40] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12416–12425, 2020.

[41] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 433–442, 2019.

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[43] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.

[44] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7454–7462, 2018.

[45] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9413–9422, 2020.

[46] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4722–4732, 2021.

[47] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788, 2019.

[48] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014.

[49] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009.

[50] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, pages 433–442, 2019.