BA-SAM: Scalable Bias-Mode Attention Mask for Segment Anything Model

Yiran Song[1]*, Qianyu Zhou[1]*, Xiangtai Li[2], Deng-Ping Fan[3,4], Xuequan Lu[5]†, Lizhuang Ma[1]†
[1]Shanghai Jiao Tong University; [2] Nanyang Technological University;
[3] Nankai International Advanced Research Institute (SHENZHEN FUTIAN);
[4] VCIP, CS, Nankai University; [5] La Trobe University
[1]{songyiran,zhouqianyu,lzma}@sjtu.edu.cn,
[2]xiangtai94@gmail.com, [3,4]dengpfan@gmail.com, [5]b.lu@latrobe.edu.au

Code: https://github.com/zongzi13545329/BA-SAM

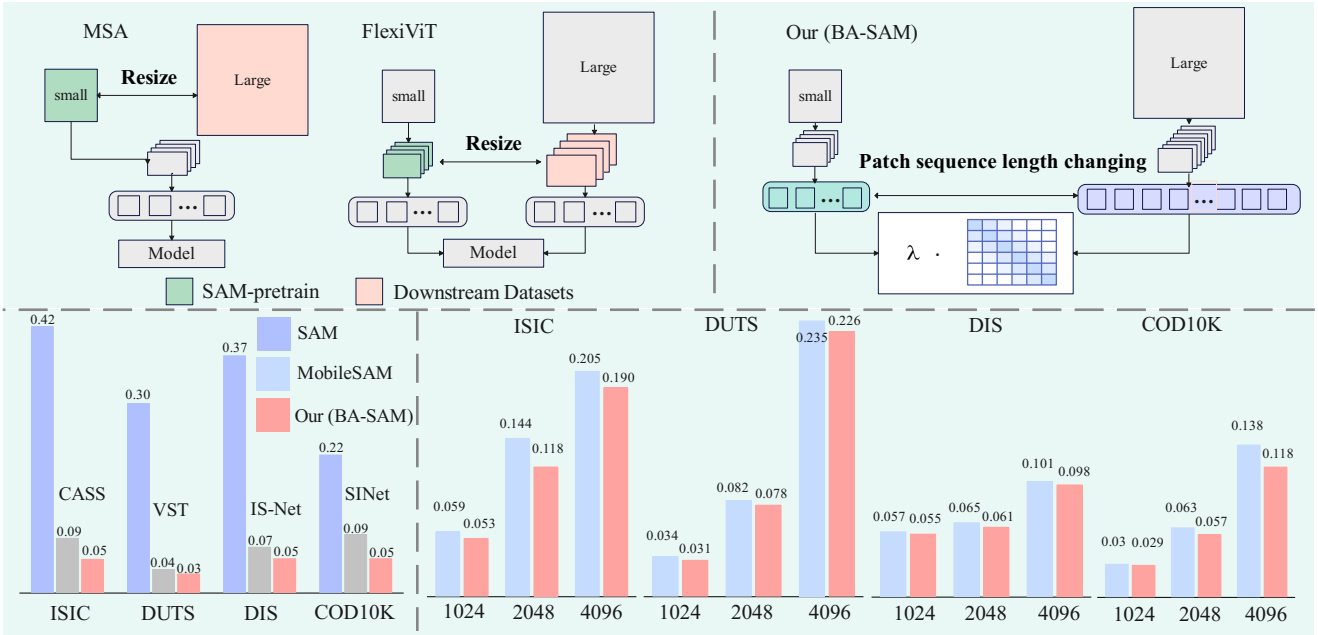

Figure 1. **Top**: contrast between prior methods [4, 82] and BA-SAM. For large-scale datasets, previous approaches often resize images or change patch sizes to handle the issue of varying resolutions. In contrast, we propose a Scalable Bias-Mode Attention Mask (BA-SAM), which enhances SAM's adaptability to varying image resolutions while eliminating structure modifications. **Bottom (left)**: We introduce a generalized model that outperforms state-of-the-art methods across all four datasets. **Bottom (right)**: With resolution variations, prior models' performance degrades drastically. In comparison, our BA-SAM consistently alleviates this issue. The evaluation metric is MAE.

Abstract

In this paper, we address the challenge of image resolution variation for the Segment Anything Model (SAM). SAM, known for its zero-shot generalizability, exhibits a performance degradation when faced with datasets with varying image sizes. Previous approaches tend to resize the image to a fixed size or adopt structure modifications, hindering the preservation of SAM's rich prior knowledge. Besides, such task-specific tuning necessitates a complete retraining of the model, which is cost-expensive and unacceptable for deployment in the downstream tasks. In this paper, we reformulate this challenge as a length extrapolation problem, where token sequence length varies while maintaining a consistent patch size for images with different sizes. To this end, we propose a Scalable Bias-Mode Attention Mask (BA-SAM) to enhance SAM's adaptability to varying image resolutions while eliminating the need for structure modifications. Firstly, we introduce a new scaling factor to ensure consistent magnitude in the attention layer's dot product values when the token sequence length changes. Secondly, we present a bias-mode attention mask that allows each token to prioritize neighboring information, mitigating

*Equal contribution.
†Corresponding authors.

the impact of untrained distant information. Our BA-SAM demonstrates efficacy in two scenarios: zero-shot and fine-tuning. Extensive evaluation of diverse datasets, including DIS5K, DUTS, ISIC, COD10K, and COCO, reveals its ability to significantly mitigate performance degradation in the zero-shot setting and achieve state-of-the-art performance with minimal fine-tuning. Furthermore, we propose a generalized model and benchmark, showcasing BA-SAM's generalizability across all four datasets simultaneously.

1. Introduction

Recently, the computer vision community [9, 22, 26, 28–30, 47, 51, 52, 60, 68, 76, 86–94] has experienced a surge in the development of various foundation models [21, 34, 56]. Notably, Meta has introduced the Segment Anything Model (SAM [37]). SAM can segment any object in an image or video by incorporating a single visual prompt, such as a box or a point, without requiring additional training. SAM is trained on an extensive SA-1B dataset [37], consisting of over 11 million images and one billion masks. Its emergence has undeniably showcased robust generalization capabilities across diverse images and objects, paving the way for new possibilities and avenues in intelligent image analysis and understanding [8, 33, 79, 82]. Based on SAM, some variants have been proposed, such as MobileSAM [79] and SAM-Adapter [8]. These efforts typically focus on improving SAM's performance on specific datasets.

During the pre-training of SAM [37], the input image size is *fixed at 1024*. As a foundational model, SAM is expected to exhibit generalization capabilities across various downstream tasks, each associated with datasets featuring different image sizes. This is particularly crucial for high-resolution (HQ) datasets with larger dimensions and more details. SAM performs well when the resolutions align with its training resolution of 1024. However, significant performance degradation is observed when inference resolutions are larger than 1024. Hence, we aim to study a practical and realistic problem to enhance SAM's adaptability to varying image resolutions of different datasets.

Since SAM adopts the standard Vision Transformer [14] architecture, two previous common approaches address the inconsistency between training and inference sizes for the ViT architecture. As depicted in Fig. 1, the first approach, *e.g.,* MSA [82] and SAM-Adapter [8], involves directly resizing all datasets to match the predefined size. Conversely, the second approach, exemplified by FlexiViT [4], entails adjusting the patch size to accommodate larger image resolutions. Nevertheless, tuning the image or patch size necessitates a complete retraining of the model, which is cost-expensive and unacceptable for deployment in the downstream tasks. Besides, it prevents leveraging the rich prior knowledge reserved in the pre-trained model of SAM. Thus, we aim to explore a solution that enhances SAM's adaptability to datasets of varying resolutions while avoiding structural modifications.

In this paper, we introduce a novel perspective that reframes the challenge of image resolution variation as a length extrapolation problem. Specifically, as shown in Fig. 1, we employ different token sequence lengths for images of varying sizes while keeping a consistent patch size. It has been observed that the inconsistency in token length between training and prediction is a key factor in performance degradation. This inconsistency manifests in two aspects: Firstly, changes in token length lead to variations in the magnitude of attention module values. When the dot product result becomes significantly large in magnitude, it can drive the subsequent Softmax layer into regions with minimal gradients. Consequently, the attention distribution after Softmax becomes highly concentrated, giving rise to the issue of vanishing gradients. Secondly, longer predictions rely on untrained information, such as additional position encodings. The introduction of untrained parameters brings a considerable amount of noise to the model, which, in turn, affects its performance.

To address these issues, we propose a Scalable Bias-Mode Attention Mask (BA-SAM) to enhance the length extrapolation capability of SAM. Our approach introduces two novel designs. Firstly, we present an improved scaling factor to ensure consistency in the attention layer's dot product value. This factor effectively regulates the magnitude of values within the attention layer, mitigating disruptive effects resulting from substantial changes in dot product operations and context length. Secondly, we introduce a novel bias-mode attention mask to maintain consistency in attention focus areas. This attention mask penalizes attention scores between distant query-key pairs, with the penalty increasing as the distance between the key and query grows. Consequently, when the context length varies, the influence of untrained distant information on each token diminishes. We achieve this mask by adding a bias after the query-key dot product. This design is highly lightweight and could be seamlessly integrated into SAM-based models with minimal computational overhead.

Our approach demonstrates efficacy in two scenarios: zero-shot and fine-tuning. Extensive evaluations on datasets from five diverse tasks are conducted, including DIS5K [55], DUTS [66], ISIC [12], COD10K [16], and COCO [46]. These datasets vary in resolution, mostly exceeding SAM's default resolution (1024×1024). In the zero-shot setting, our BA-SAM alleviates the model's performance degradation caused by expanding the inference resolution without requiring additional training. With a few fine-tuning epochs on downstream tasks, our BA-SAM consistently achieves state-of-the-art accuracy across all datasets, as shown at the bottom of Fig. 1. Additionally, to further demonstrate BA-SAM's generalizability, we pro-

pose a generalized model and a new benchmark, which utilize one model to attain state-of-the-art performance across all four datasets simultaneously.

2. Related Work

Visual Foundation Models. Large models that are trained on broad datasets and can be adapted to numerous downstream tasks are called "Foundation Models" [5, 41, 42, 44, 69, 74, 78, 85]. Vision-Language Models (VLM) (CLIP [56] and DALL-E [57]) combine computer vision and natural language processing to understand and generate descriptions or analyze visual content using textual and visual information. Masked Image Modeling [50, 75] (MIM) approaches mask parts of an image during the training to encourage a model to learn contextual information and complete missing regions. SAM [37] is a large vision foundation model designed for segmenting objects or areas in images, offering precise segmentation capabilities. We use a variant of SAM called MobileSAM [79] as the baseline.

Resolution Variation Processing. To enable models to be more adaptable to variations in resolutions, previous works have relied on adjustments to positional embeddings [39] and patch sizes [4, 6, 24, 29, 38, 45, 77]. For example, Patch n' Pack [13] employed sequence packing during the training to handle inputs with arbitrary resolutions and aspect ratios. They all necessitate training from scratch, incurring substantial computational and time costs. In contrast to these methods, we extend the concept of *length extrapolation* from NLP into the context of addressing scale variations in CV. Length extrapolation refers to a model's ability to generalize well to longer inputs than those it was trained on. In NLP, it has been successfully used, such as in ALIBI [54] and KERPLE [11], to enable models to adapt to longer sequences without significant performance degradation. Our approach seamlessly extends to two scenarios: zero-shot and fine-tuning. Our proposed method allows us to leverage prior knowledge embedded in the SAM and significantly reduce training efforts.

Parameter Efficient Tuning. There have been some pioneering works for the Parameter Efficient Tuning (PEFT) of visual models, such as AdaptFormer [7] and visual prompt tuning (VPT) [35]. He et al. [27] analyzed the unified view among PETL techniques such as prefix tuning [43], Prompt-tuning [35], and adapter [7]. Our method belongs to the category of Parameter Efficient Tuning.

Visual Attention Modeling. Various studies have incorporated attention mechanisms into neural network architectures designed for visual tasks [3, 32, 40, 65, 70, 96]. These mechanisms are employed in a channel-wise manner to capture cross-feature information [10, 68, 83]. They are also used for selecting paths in different branches of a network [63], or a combination of both strategies [80]. The advent of transformers has led to hybrid architectures that introduce other modules. Bello's work [2] introduces approximate content attention with a positional attention component. Child *et al.* [71] observe that many early layers in the network learn locally connected patterns akin to convolutions, indicating that hybrid architectures inspired by both transformers and convolutional networks are a compelling design choice. Several recent studies explore this line for various tasks [25, 61, 67, 72]. In contrast to prior work, we do not introduce a new attention structure. Instead, we offer theoretical proof for optimizing existing attention mechanisms. This resulting optimization approach is applicable across various attention designs and demonstrates strong performance across multiple datasets.

3. Preliminaries

SAM. Segment Anything Model (SAM) [37] consists of three core modules: image encoder, prompt encoder, and mask decoder. It has been trained on SA-1B dataset [37], which comprises more than 1 billion automatically generated masks. Consequently, SAM exhibits valuable and robust zero-shot generalization to new data without necessitating further training, and details can be referred to [37]. Our Scalable Bias-Mode Attention Mask (BA-SAM) optimizes the image encoder while keeping the structures of the mask decoder and prompt encoder unchanged.

Attention in Transformer. In this work, we define the input sequence of image patches, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with length N, where $\mathbf{x}_i \in \mathbb{R}^{d_x}$. q_i, k_j, v_j are calculated by $\mathbf{x}_i\mathbf{W}^Q$, $\mathbf{x}_j\mathbf{W}^K$, $\mathbf{x}_j\mathbf{W}^V$. Here, the projections $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_x \times d_k}$ are parameter matrices.

(i) Scaling Factor. The two most commonly used attention functions are additive attention [1] and dot-product attention [64]. The vanilla Transformer chooses dot-product attention for its space efficiency in practice. However, for larger values of d_k, the dot products grow large in magnitude, pushing the Softmax function into regions with minimal gradients. They use *scaling factor* $\lambda_d = \frac{1}{\sqrt{d_k}}$ to scale the dot products, where d_k denotes the dimension. To better analyze the role of the scaling factor, we express the output element $\mathbf{O}_i$ and the weight coefficient $a_{i,j}$ as follows:

$$\mathbf{O}_i = \sum_{j=1}^{N} a_{i,j} v_j, \quad a_{i,j} = \frac{e^{\lambda_d q_i \cdot k_j}}{\sum_{j=1}^{N} e^{\lambda_d q_i \cdot k_j}}, \tag{1}$$

where λ_d represents the scaling factor.

(ii) Absolute & Relative Position Encoding. The original Transformer [64] incorporates absolute non-parametric positional encoding $p = (p_1, \ldots, p_n)$ with x as $x_i = x_i + p_i$. Other works replace them with parametric encoding [23], or adopted Fourier-based kernelized versions [53]. Absolute position encoding enforces a fixed size for inputs. Recent work [58] considers the pairwise relationships between elements, which encodes the relative position between input

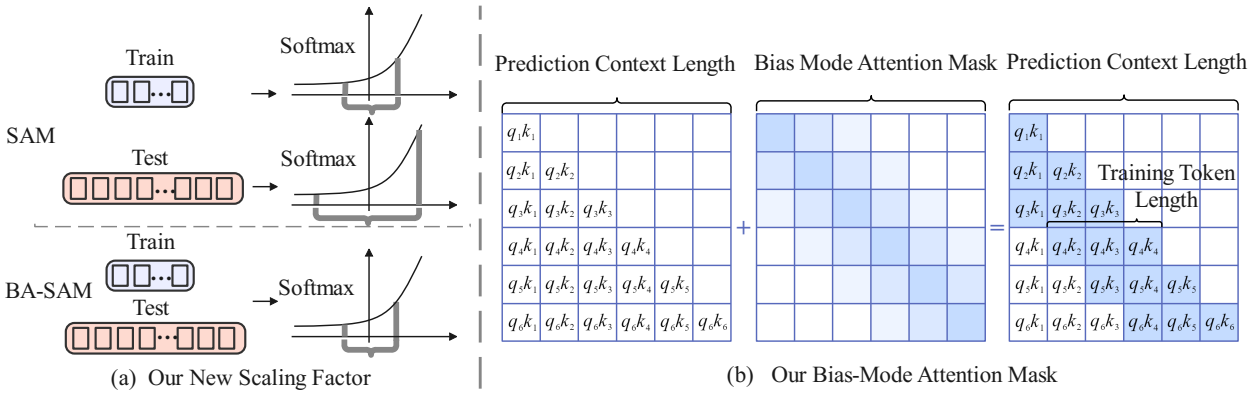

Figure 2. Illustration of the proposed BA-SAM method. (a) In the original SAM, when the length of the input token sequences varies during testing, the magnitude of the Softmax outputs changes drastically. We propose a new scaling factor to address this issue. (b) We introduce a bias-mode attention mask, which increases attention scores' penalties as the distance between the query and key grows.

x_i and x_j into vectors $p^v_{i,j}, p^q_{i,j}, p^k_{i,j} \in \mathrm{R}^{d_k}$. Then, we can reformulate Eq. (1) as follows:

$$\mathbf{O}_i = \sum_{j=1}^{N} a_{i,j} \left(v_j + p^v_{i,j}\right), \tag{2}$$

$$a_{i,j} = \frac{e^{\lambda\left(q_i + p^q_{i,j}\right)\cdot\left(k_j + p^k_{i,j}\right)}}{\sum_{j=1}^{N} e^{\lambda\left(q_i + p^q_{i,j}\right)\cdot\left(k_j + p^k_{i,j}\right)}}, \tag{3}$$

where $p^v_{i,j}, p^q_{i,j}, p^k_{i,j}$ is learned during training.

4. Methodology

Based on the preliminaries, we further analyze that the original SAM sets the input to a fixed resolution of 1024, where it uses absolute position encoding and the dot product. As a result, there are significant limitations in the processing of length extrapolation problems. To address this, as shown in Fig. 2, we present a Scalable Bias-mode Attention Mask (BA-SAM). In Sec. 4.1, we provide a theoretical explanation for the scaling factor used in the original Transformer and introduce a new scaling factor to regulate the magnitude inconsistency caused by length extrapolation. In Sec. 4.2, we design a bias-mode attention mask to place more focus on neighboring tokens, mitigating the impact of untrained distant information. Finally, we explain how we will embed our BA-SAM into the SAM-based structure in Sec. 4.3.

4.1. New Scaling Factor

When the dot product becomes significantly large in magnitude in the original attention module of SAM [64], it can drive the Softmax layer into regions with minimal gradients. This is because the attention distribution after Softmax becomes highly concentrated, giving rise to the issue of vanishing gradients. Upon examination of Eq. (1), it is obvious that the computation of the $q \cdot k$ is intrinsically tied to both the token sequences length N and the dimension d_k. When the token length N and the dimension d_k significantly increase, the overall efficacy of the attention is affected, thus leading to a noticeable performance degradation.

To address this issue, we attempt to design a new scaling factor that allows the model to cope with variations in N and d_k. When N or d_k grows significantly, we expect to regulate the magnitude of the values within the attention layer, maintaining a similar magnitude. [64] introduced a scaling factor $\lambda = \frac{1}{\sqrt{d_k}}$ to counteract the effect of the large growth in magnitude due to the dot products. Below, we will provide a theoretical derivation of this scaling factor and then elaborate on our proposed new one.

The dimension d_k. Following the work [64], we assume the components of q and k are independent random variables with mean 0 and variance 1. The mean of $q \cdot k$ is:

$$\mathrm{E}[q \cdot k] = \mathrm{E}\left[\sum_{i=1}^{d_k} q_i k_i\right] = \sum_{i=1}^{d_k} \mathrm{E}\left[q_i\right] \mathrm{E}\left[k_i\right] = 0 \tag{4}$$

Similarly, we formulate the variance of $q \cdot k$ as follows:

$$\mathrm{var}[q \cdot k] = \mathrm{var}\left[\sum_{i=1}^{d_k} q_i k_i\right] = \sum_{i=1}^{d_k} \mathrm{var}\left[q_i\right] \mathrm{var}\left[k_i\right] = d_k \tag{5}$$

Given this, we can approximately consider the $q \cdot k$ values to be within the range of $-3\sqrt{d_k}$ to $3\sqrt{d_k}$, according to properties of Gaussian distribution. For larger models, d_k is generally a larger positive value, resulting in a significant increase in the magnitude of numerical values of $q \cdot k$, compared to the additive attention option, which has the range of $[-3, 3]$. Consequently, the attention distribution after Softmax becomes highly concentrated. This leads

to severe gradient vanishing, which hampers the effectiveness of the training and induces less desired performance. As the $q \cdot k$ values lie in the range of $[-3\sqrt{d_k}, 3\sqrt{d_k}]$, the scaling factor can be simply defined as $\lambda_d = \frac{1}{\sqrt{d_k}}$, in order to maintain a similar magnitude.

Our new scaling factor. We have provided the interpretation on how the original scaling factor was designed. Now, we explain the design of our new scaling factor.

According to Eq. (4) and Eq. (5), the scale of dot-product attention $q \cdot k$ has the similar magnitude with the additive attention by λ_d, which can be seen as $a_{i,j}$ is independent of d_k. We simplify $\lambda_d q_i \cdot k_j$ with using $x_{i,j}$ and further discuss the effect of length N on $a_{i,j}$.

In Eq. (1) , $a_{i,j}$ can be seen as the conditional distribution with i being the condition and j being the random variable. Inspired by [62], we introduce information entropy to constrain $a_{i,j}$. Specifically, entropy is a measure of uncertainty, and we expect the uncertainty of $a_{i,j}$ to be insensitive to the length N, *i.e.,* the value of each $a_{i,j}$ will change when the token increases, but the entropy value of the overall $a_{i,j}$ can remain relatively stable. The entropy of $a_{i,j}$ is :$\mathcal{H}_i = -\sum_{j=1}^{N} a_{i,j} \log a_{i,j}$ and we substitute Eq. (1):

$$\mathcal{H}_i = \log \sum_{j=1}^{N} e^{\lambda_n x_{i,j}} - \frac{\sum_{j=1}^{N} e^{\lambda_n x_{i,j}} (\lambda_n x_{i,j})}{\sum_{j=1}^{N} e^{\lambda_n x_{i,j}}} \tag{6}$$

Then, we substitute the approximate estimates into Eq. (6):

$$\begin{aligned} &\sum_{j=1}^{N} e^{x_{i,j}} = N \times \frac{1}{N} \sum_{j=1}^{N} e^{x_{i,j}} \approx N \mathbb{E}_j \left[e^{x_{i,j}}\right] \\ &\mathbb{E}_j \left[e^{x_{i,j}} (x_{i,j})\right] \approx 0, \mathbb{E}_j \left[e^{x_{i,j}}\right] = O(1) \end{aligned} \tag{7}$$

Here, we use λ_n to offset the effect of N on $\mathcal{H}_i$. Then, we have the following result:

$$\mathcal{H}_i \approx \log N - k\lambda_n = 0 \Rightarrow \lambda_n = \frac{\log N}{k}, \tag{8}$$

where k is a parameter value. We denote the token sequence length during the training as N_{train} and the token sequence length during the testing as N_{test}, where $N_{test} >> N_{train}$. When $N = N_{train}$, $\lambda_n = 1$ (consistent with the training length). As such, $k = \log N_{train}$ and finally we have $\lambda_n = log_{N_{train}} N_{test}$. Considering both λ_d and λ_n, we can ultimately derive our new scaling factor as:

$$\lambda = \lambda_d \lambda_n = \frac{log_{N_{train}} N_{test}}{\sqrt{d_k}} \tag{9}$$

Our new scaling factor in Eq. (9) ensures attention computation remains consistent, regardless of variations in d_k and N. It will enhance the extrapolative capacity of the model.

4.2. Bias-Mode Attention Mask

Another challenge is that token sequence length changes will lead to positional encoding variations. It is important to ensure the insensitivity of the model when such positional encoding variations occur during the testing.

One possible way is absolute encoding without trainable parameters, such as Sinusoidal [64]. It requires the position encoding to have strong local-to-global inference capabilities. Nevertheless, this assumes that the given function has high-order smoothness (higher-order derivatives exist and are bounded). Commonly used positional encodings are often combined with trigonometric functions. These methods fail to satisfy the requirement of bounded high-order derivatives, making it less accurate to estimate the extrapolated results. Another potential approach is using local attention [49], which constrains the model's field of view and remains insensitive to variations in token sequence length. However, local attention is typically implemented using a local window, necessitating modifications to the SAM structure and further re-training from scratch.

To this end, we propose enabling the attention layer to focus more on the current token's neighboring tokens. In this manner, even with an increase in the length of a token sequence, each token is scarcely affected by the untrained tokens from distant positions. In particular, we design a simple yet effective bias-mode mask, which is achieved by introducing a bias after the query-key dot product.

As shown in Fig. 3, this mask exhibits a bias specified on the distance between the query-key pairs (i.e., $q \cdot k$). We expect that this proposed mask imposes penalties on attention scores between distant query-key pairs, and the penalty increases as the distance between a key q and a query k grows. To achieve this, we simply define the bias as $b_{i,j} = \beta|i-j|$.

$$a_{i,j} = \frac{e^{\lambda(q_i \cdot k_j + b_{i,j})}}{\sum_{j=1}^{N} e^{\lambda(q_i \cdot k_j + b_{i,j})}}, \tag{10}$$

where β is a hyperparameter.

We now discuss the setting of β based on different cases. We set β to a static, non-learned fixed value when conducting zero-shot generalization without fine-tuning. The experimental section will present the specific value setting (Sec. 5). When fine-tuning is required, we make β trainable. Since our Bias-Mode Attention Mask is lightweight, it incurs negligible training overhead.

4.3. BA-SAM Model

As shown in Fig. 3, our BA-SAM is simple to implement, and can be seamlessly integrated into SAM [37] and its variants (such as MobileSAM). Specifically, our design involves a new scaling factor (NSF) for the attention layer and a bias-mode attention mask (BM-AM) in the encoder part. Our method does not involve any alterations to the

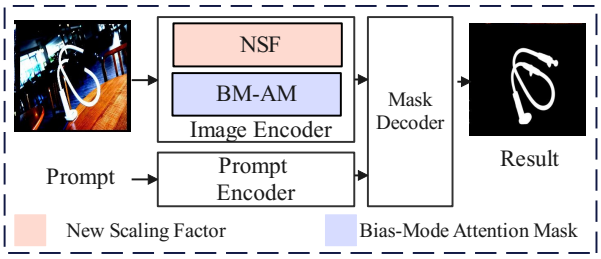

Figure 3. Embedding of our BA-SAM into a SAM backbone. NSF indicates our new scaling factor, and BM-AM denotes our designed bias-mode attention mask.

model structure and is suitable for fine-tuning and zero-shot modes. During fine-tuning, BA-SAM only introduces negligible computation costs, as shown in Tab. 5.

5. Experiments

5.1. Datasets and Implementation

Datasets. For a comprehensive evaluation of our proposed BA-SAM, we conduct extensive experiments on a wide range of segmentation tasks, *i.e.,* salient object segmentation [15, 17, 19], complex object segmentation [18], skin lesion segmentation [96], camouflaged object detection [20, 31], which correspond to four datasets: DUTS [66], DIS-TE4 [55], ISIC [12] and COD10K [16]. Besides, we also verify BA-SAM on the challenging COCO [46] instance segmentation benchmark.

Implementation Details. In the zero-shot setting, we use the original SAM [37] backbone. For fine-tuning scenarios, we employ MobileSAM [79] as a baseline. MobileSAM is a lightweight version of SAM, where its encoder is distilled from the original SAM. For various object segmentation tasks, a random point is extracted from the ground truth as the prompt input during the fine-tuning. We use the ViT-B [14] backbone for the instance segmentation on COCO. In particular, we use the state-of-the-art detector Deformable-DETR [95] trained on the COCO [46] dataset with Swin-L [49] backbone as box prompt generator. *More details are provided in the supplementary material.*

Evaluation Metrics. In the experiments, we use the widely used Mean Absolute Error (MAE) and Average Precision (AP) for evaluation. A lower MAE score and a higher AP score indicate better model performance.

5.2. Main Results

Results of Various Object Segmentation Tasks. Tab. 1 shows the effectiveness of our approach across four diverse segmentation datasets. $\Delta diff$ denotes the value of the performance degradation due to resolution changes during the inference. The upper and lower parts of the table indicate the results without and with fine-tuning. We have three observations: Firstly, our proposed BA-SAM consistently outperforms both SAM [37] and MobileSAM [79] baselines on all four datasets. This is mainly because these baselines do not consider the issue of varying image resolutions. In contrast, our presented scaling factor and bias-mode attention mask explicitly handle this issue and further alleviate the performance degradation. Secondly, when using higher-resolution images than the training images, SAM [37] and MobileSAM [79] baselines show less desirable results than the original image size. In contrast, our BA-SAM incurs significantly less performance drop in different datasets. Thirdly, we observe negligible computational overhead, whether fine-tuning is applied or not, which supports the claim in the method section. See Sec. 5.3 for details.

Results of Instance Segmentation. In Tab. 3, we evaluate the performance of our method on the COCO benchmark. For a fair comparison, all experiments are conducted in a zero-shot manner, with the same initialized parameters for the comparative methods and without the use of any additional training data. Our proposed BA-SAM consistently outperforms SAM [37] and MobileSAM [79] baselines, demonstrating better zero-shot generalization capability on instance segmentation.

Comparisons with State-of-the-Art Methods: To further demonstrate the superiority and generalizability of our method, we compare our method with the state-of-the-art approaches in Tab. 2. From the table, we have two following observations: Firstly, all the state-of-the-art approaches [16, 36, 48, 55, 59, 81, 84, 96] show less-desirable performance in each dataset. In comparison, our BA-SAM (specialized models) consistently outperforms these methods when fine-tuned on each downstream dataset. Secondly, almost all of these state-of-the-art techniques are specifically designed for one task and cannot be generalized well to other tasks. Due to the strong zero-shot generalization capability of SAM [37], our proposed BA-SAM can also be employed as a generalized model, which fine-tunes with all these downstream datasets in a unified and shared model. Importantly, unlike [12, 48, 59, 81], we eliminate the need for employing additional techniques to enhance the performance further. As shown in Tab. 2, our generalized model also consistently promotes the performance of SAM on all datasets, demonstrating its remarkable generalizability.

5.3. Ablation Study and Analysis

In this section, we first conduct ablation study to study the contribution of each component. Then, we investigate the impact of the new scaling factor (NSF) and the bias-mode attention mask (BM-AM) with a more detailed analysis.

Ablations Studies of Each Component. Tab. 4 summarizes the effect of each component on the settings with and without fine-tuning, respectively. The baseline means using the MobileSAM [79] as the base network that uses the

Method	Train Size	Test Size	ISIC [12]	Δ*diff*	DUTS [66]	Δ*diff*	DIS-TE4 [55]	Δ*diff*	COD10K [16]	Δ*diff*
Without fine-tuning										
SAM [37]	-	1024	0.421	-	0.298	-	0.362	-	0.217	-
	-	2048	0.601	18.0%	0.360	6.2%	0.411	4.9%	0.391	17.4%
Ours (w [37])	-	1024	**0.417**	-	**0.294**	-	**0.356**	-	**0.208**	-
	-	2048	0.589	17.2%	0.348	5.4%	0.406	5.0%	0.387	17.9%
MobileSAM [79]	-	1024	0.463	-	0.502	-	0.544	-	0.465	-
	-	2048	0.641	17.8%	0.437	6.5%	0.427	11.7%	0.346	11.9%
		4096	0.693	23.0%	0.328	17.4%	0.355	18.9%	0.300	16.5%
Ours (w [79])	-	1024	0.452	-	0.486	-	0.515	-	0.440	-
	-	2048	0.611	15.9%	0.413	7.3%	0.406	10.9%	0.321	11.9%
		4096	0.657	20.5%	0.283	20.3%	0.361	15.4%	0.246	19.4%
With fine-tuning										
MobileSAM [79]	1024	1024	0.059	-	0.034	-	0.057	-	0.030	-
		2048	0.144	8.5 %	0.082	4.8%	0.065	0.8%	0.063	3.3%
		4096	0.205	14.6 %	0.235	20.1%	0.101	4.4%	0.138	10.8%
	2048	2048	0.083	-	0.045	-	0.056	-	0.036	-
		4096	0.227	14.4%	0.091	4.6%	0.066	1.0%	0.059	2.3%
Ours (w [79])	1024	1024	**0.053**	-	**0.031**	-	0.055	-	**0.029**	-
		2048	0.118	6.5 %	0.078	4.4%	0.061	0.6%	0.057	2.5%
		4096	0.190	13.7%	0.226	19.2%	0.098	4.3%	0.118	8.6%
	2048	2048	0.080	-	0.043	-	**0.053**	-	0.033	-
		4096	0.214	13.4%	0.088	4.4%	0.061	0.8%	0.056	2.3%

Table 1. Performance comparisons in varying image resolutions. We employed the widely used MAE (Mean Absolute Error) score. Lower MAE scores indicate better model performance. Δ*diff* denotes the performance degradation due to resolution changes. Compared to the SAM [37] and MobileSAM [79] baselines, our proposed BA-SAM achieves smaller degradation when encountering token sequence length changes. The best MAE performance is highlighted in bold, and the smallest performance degradation is underlined.

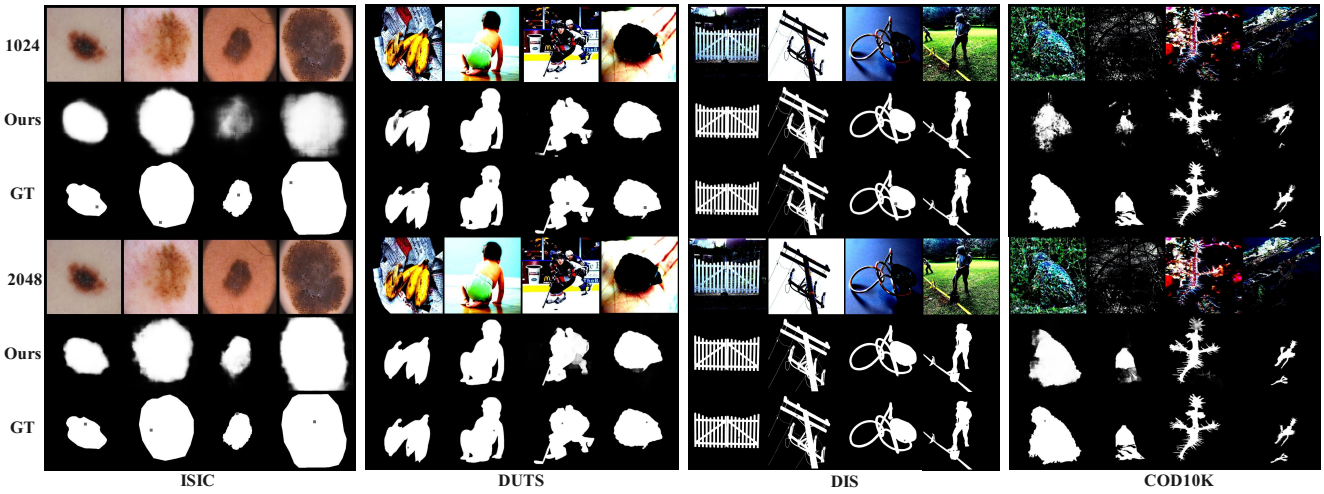

Figure 4. Visualization results of our BA-SAM on four object segmentation tasks, *i.e.,* skin lesion segmentation, salient object segmentation, complex object segmentation, camouflaged object detection, which correspond to four datasets: ISIC [12], DUTS [66], DIS-TE4 [55], and COD10K [16]. Our BA-SAM can accurately handle the issue of varying image resolutions and segments in different tasks.

vanilla scaling factor (VSF) in the attention layer [64]. New Scaling Factor and Bias-Mode Attention Mask are abbreviated as NSF and BM-AM, respectively. The table shows that NSF performs better compared to the VSF baseline. This is because the vanilla attention in SAM [37] and MobileSAM [79] does not consider maintaining the magnitude consistency when Softmax outputs change drastically due to varying input resolutions during the testing. In contrast, our NSF explicitly maintains the magnitude consistency and alleviates the performance degradation. Furthermore, by

Methods	ISIC [12]	DUTS [66]	DIS-TE4 [55]	COD10K [16]
Specialized models				
CASS [59]	0.086	-	-	-
DINO [81]	0.081	-	-	-
MSA [73]	0.049	-	-	-
VST [48]	-	0.037	-	-
ICONet [96]	-	0.037	-	-
Gate [84]	-	-	0.109	-
IS-Net [55]	-	-	0.072	-
SINet [16]	-	-	-	0.092
SegMaR [36]	-	-	-	0.034
The same framework, 4 Specialized models				
Ours (BA-SAM)	**0.053**	**0.031**	**0.055**	**0.029**
Generalized model				
SAM [37]	0.419	0.298	0.373	0.217
Ours (BA-SAM)	**0.054**	**0.030**	**0.054**	**0.054**

Table 2. Comparison results (MAE) with state-of-the-art specialized models on various segmentation tasks.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SAM [37]	42.5	69.6	44.7	29.7	47.0	56.7
Ours (w [37])	43.0	70.0	45.4	30.0	47.4	57.1
MobileSAM [79]	40.8	68.4	41.6	26.0	44.4	57.6
Ours (w [79])	41.2	69.0	42.1	26.2	44.8	58.2

Table 3. Results (AP) on COCO [46] instance segmentation.

adding BA-SAM, the performance could be further boosted when extrapolating to a larger test length. The improvements confirm that these individual components are complementary, which results in mutual benefits.

Impact of Slope in Bias-Mode Attention Mask. In the Bias-Mode Attention Mask, the magnitude of the slope β determines penalty rates in different heads. We found that the best performance is achieved when $\beta = 0.1$. Besides, our method is robust to different slope choices. We use a fixed slope $\beta = 1$ by default in the zero-shot setting.

Computational Efficiency. In Tab. 5, we analyze the computational efficiency between the baselines and our BA-SAM. All the experiments are conducted on the same NVIDIA RTX 4090GPU to ensure fair comparisons. The table shows that our BA-SAM is highly lightweight, incurring negligible computational overhead to the models. The reasons are two-fold: firstly, the NSF exhibits nearly identical computational complexity to the vanilla one. In addition, the BM-AM is seamlessly incorporated by adding a mask matrix to the query-key dot product before applying the Softmax operation. Although there is a slight increase in memory usage, it remains negligible compared to the memory occupied by large models.

Visualization. In Fig. 4, we present several visual examples on various datasets. Our BA-SAM shows fine-grained segmentation results on various high-resolution inputs.

Methods	ISIC [12]	DUTS [66]	DIS-TE4 [55]	COD10K [16]
Without fine-tuning				
Baseline [79]	17.8	6.5	11.7	11.9
+ NSF	16.4	7.5	11.3	12.0
+ BM-AM	16.8	**7.2**	11.7	11.9
+ Both	**15.9**	7.9	**10.9**	**11.9**
With fine-tuning (1024)				
Baseline [79]	42.2	4.8	0.8	3.3
+ NSF	40.9	4.6	0.7	3.0
+ BM-AM	41.2	4.5	0.8	2.7
+ Both	**40.4**	**4.4**	**0.6**	**2.5**
With fine-tuning (2048)				
Baseline [79]	14.4	4.6	1.0	2.3
+ NSF	**13.1**	4.6	0.8	2.4
+ BM-AM	13.7	4.5	0.9	2.3
+ Both	13.4	**4.4**	**0.8**	**2.3**

Table 4. Ablation study of each component on the settings with and without fine-tuning. Numbers indicate the performance degradation, $\Delta diff$. A lower $\Delta diff$ means a better performance.

Model	Params (M)	Speed (ms)	Train Hours (h)
SAM [37]	81	113.9	-
Ours (w [37])	81	114.0	-
MobileSAM [79]	9.66	16.2	0.64
Ours (w [79])	9.67	16.5	0.65

Table 5. Comparisons of computational efficiency between the baselines and our BA-SAM. Params: number of parameters. Speed: inference speed. The top part uses the zero-shot setting, and the bottom part uses fine-tuning.

6. Conclusion

In this paper, we address the important problem of varying image resolutions in SAM models by reformulating it as a problem of length extrapolation. To enhance the length extrapolation capability of SAM, we propose the Scalable Bias-mode Attention Mask for SAM (BA-SAM). A new scaling factor is introduced to maintain the consistent magnitude of attention. In addition, a bias-mode attention mask is designed to prioritize neighboring information, mitigating the impact of untrained distant information. Extensive evaluation on diverse datasets reveals its ability to significantly alleviate performance degradation in the zero-shot setting and achieve state-of-the-art performance with minimal fine-tuning. Furthermore, we propose a generalized model and benchmark, showcasing BA-SAM's generalizability across all four datasets.

Acknowledgement

The work is supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), National Natural Science Foundation of China (No. 72192821), YuCaiKe [2023] (No.14105167-2023).

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015. 3

[2] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *CVPR*, 2021. 3

[3] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. In *ICCV*, 2019. 3

[4] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *CVPR*, 2023. 1, 2, 3

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 3

[6] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once for all: Train one network and specialize it for efficient deployment. *ICLR*, 2020. 3

[7] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *NeurIPS*, 2022. 3

[8] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. In *arXiv preprint arXiv:2304.09148*, 2023. 2

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 2

[10] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, 2020. 3

[11] Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *NeurIPS*, 35, 2022. 3

[12] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI*, 2018. 2, 6, 7, 8

[13] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *NeurIPS*, 2023. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 2, 6

[15] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018. 6

[16] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020. 2, 6, 7, 8

[17] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *TNNLS*, 2021. 6

[18] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *TPAMI*, 2022. 6

[19] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *TPAMI*, 2022. 6

[20] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *VI*, 2023. 6

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 2

[22] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *PR*, 2022. 2

[23] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017. 3

[24] Chengyue Gong, Dilin Wang, Meng Li, Xinlei Chen, Zhicheng Yan, Yuandong Tian, Vikas Chandra, et al. Nasvit: Neural architecture search for efficient vision transformers with gradient conflict aware supernet training. In *ICLR*, 2021. 3

[25] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *ICCV*, 2021. 3

[26] Qiqi Gu, Qianyu Zhou, Minghao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Pit: Position-invariant transform for cross-fov domain adaptation. In *ICCV*, 2021. 2

[27] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *ICLR*, 2021. 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3

[30] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *ACM MM*, 2021. 2

[31] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *MIR*, 2023. 6

[32] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Ming-Ming Cheng, Bowen Zhou, and Luc Van Gool. Sam struggles in concealed scenes–empirical study on" segment anything". *SCIS*, 2023. 3

[33] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *CVPR Workshop*, 2023. 2

[34] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[35] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 3

[36] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR*, 2022. 6, 8

[37] 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. *ICCV*, 2023. 2, 3, 5, 6, 7, 8

[38] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representations for adaptive deployment. *arXiv preprint arXiv:2205.13147*, 2022. 3

[39] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, 2023. 3

[40] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, 2019. 3

[41] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv pre-print*, 2023. 3

[42] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? *CVPR*, 2024. 3

[43] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021. 3

[44] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022. 3

[45] Mingbao Lin, Mengzhao Chen, Yuxin Zhang, Ke Li, Yunhang Shen, Chunhua Shen, and Rongrong Ji. Super vision transformer. *IJCV*, 2022. 3

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 6, 8

[47] Fengqi Liu, Jingyu Gong, Qianyu Zhou, Xuequan Lu, Ran Yi, Yuan Xie, and Lizhuang Ma. Cloudmix: Dual mixup consistency for unpaired point cloud completion. *TVCG*, 2024. 2

[48] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, 2021. 6, 8

[49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5, 6

[50] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 3

[51] Shaocong Long, Qianyu Zhou, Chenhao Ying, Lizhuang Ma, and Yuan Luo. Diverse target and contribution scheduling for domain generalization. *arXiv preprint arXiv:2309.16460*, 2023. 2

[52] Shaocong Long, Qianyu Zhou, Chenhao Ying, Lizhuang Ma, and Yuan Luo. Rethinking domain generalization: Discriminability and generalizability. *arXiv preprint*, 2023. 2

[53] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 3

[54] Ofir Press, Noah A Smith, Mike Lewis, et al. Train short, test long: Attention with linear biases enables input length extrapolation. *ICLR*, 2021. 3

[55] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 2, 6, 7, 8

[56] 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. 2, 3

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[58] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of NAACL-HLT*, 2018. 3

[59] et al Singh, Pranav. Cass: cross architectural self-supervision for medical image analysis. *NeurIPS*, 2022. 6, 8

[60] Yiran Song, Qianyu Zhou, and Lizhuang Ma. Rethinking implicit neural representations for vision learners. In *ICASSP*, 2023. 2

[61] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *CVPR*, 2021. 3

[62] JianLin Su. Transformer upgrade road: 7, length extrapolation and local attention, 2023. 5

[63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 3

[64] 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. 3, 4, 5, 7

[65] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 3

[66] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2, 6, 7, 8

[67] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 3

[68] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3

[69] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *MIR*, 2023. 3

[70] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 3

[71] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Generating long sequences with sparse transformers. *arXiv preprint arXiv:2006.03677*, 2020. 3

[72] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[73] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 8

[74] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *TPAMI*, 2024. 3

[75] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 3

[76] Hongyi Xu, Fengqi Liu, Qianyu Zhou, Jinkun Hao, Zhijie Cao, Zhengyang Feng, and Lizhuang Ma. Semi-supervised 3d object detection via adaptive pseudo-labeling. In *ICIP*, 2021. 2

[77] Jiahui Yu, Pengchong Jin, Hanxiao Liu, Gabriel Bender, Pieter-Jan Kindermans, Mingxing Tan, Thomas Huang, Xiaodan Song, Ruoming Pang, and Quoc Le. Bignas: Scaling up neural architecture search with big single-stage models. In *ECCV*, 2020. 3

[78] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint*, 2024. 3

[79] 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. *arXiv preprint arXiv:2306.14289*, 2023. 2, 3, 6, 7, 8

[80] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *CVPR*, 2022. 3

[81] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ICLR*, 2023. 6, 8

[82] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. In *arXiv preprint arXiv:2304.13785*, 2023. 1, 2

[83] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020. 3

[84] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, 2020. 6, 8

[85] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*, 2023. 3

[86] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *CVIU*, 2022. 2

[87] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *TCSVT*, 2022.

[88] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *ACM MM*, 2022.

[89] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *ECCV*, 2022.

[90] Qianyu Zhou, Chuyun Zhuang, Ran Yi, Xuequan Lu, and Lizhuang Ma. Domain adaptive semantic segmentation via regional contrastive consistency regularization. In *ICME*, 2022.

[91] Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Xuequan Lu, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *TPAMI*, 2023.

[92] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *TPAMI*, 2023.

[93] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *CVPR*, 2023.

[94] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *CVPR*, 2024. 2

[95] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 6

[96] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *TPAMI*, 2022. 3, 6, 8