Segment Any-Quality Images with Generative Latent Space Enhancement

Guangqian Guo[1,2] Yong Guo[2] Xuehui Yu[3] Wenbo Li[4] Yaoxing Wang[1] Shan Gao[1]
[1]Northwestern Polytechnical University [2]Huawei [3]Tencent [4] Huawei Noah's Ark Lab
{guogq21, wangyx24}@mail.nwpu.edu.cn {guoyongcs, fenglinglwb}@gmail.com
xuehuiyu@tencent.com gaoshan@nwpu.edu.cn

Abstract

Despite their success, Segment Anything Models (SAMs) experience significant performance drops on severely degraded, low-quality images, limiting their effectiveness in real-world scenarios. To address this, we propose ***GleSAM****, which utilizes Generative Latent space Enhancement to boost robustness on low-quality images, thus enabling generalization across various image qualities. Specifically, we adapt the concept of latent diffusion to SAM-based segmentation frameworks and perform the generative diffusion process in the latent space of SAM to reconstruct high-quality representation, thereby improving segmentation. Additionally, we introduce two techniques to improve compatibility between the pre-trained diffusion model and the segmentation framework. Our method can be applied to pre-trained SAM and SAM2 with only minimal additional learnable parameters, allowing for efficient optimization. We also construct the LQSeg dataset with a greater diversity of degradation types and levels for training and evaluating the model. Extensive experiments demonstrate that GleSAM significantly improves segmentation robustness on complex degradations while maintaining generalization to clear images. Furthermore, GleSAM also performs well on unseen degradations, underscoring the versatility of our approach and dataset.*

1. Introduction

Accurate object detection and segmentation [3, 12, 15, 16, 18, 26, 51, 55] in diverse scenarios is a fundamental task for various high-level visual applications, such as robotics and autonomous driving. The recently developed Segment Anything Models (SAMs), including SAM [27] and SAM2 [44], serving as a foundational model, have gained significant influence within the community [4, 5, 10, 16, 36, 37] due to their outstanding zero-shot segmentation abilities.

Figure 1. The comparison of qualitative results on low-quality images with varying degradation levels from an unseen dataset. To generate images with different degradation levels, we progressively added Gaussian Noise, Re-sampling Noise, and more severe Gaussian noise to an image. Results indicate that the baseline SAM [27] shows limited robustness to degradation. Although RobustSAM [8] retains some resilience against simpler degradations, it struggles with more complex and unfamiliar degradations. In contrast, our method consistently demonstrates strong robustness across images of varying quality.

Despite their success, SAMs perform poorly on common low-quality images, such as those degraded by noise, blur, and compression artifacts [22, 42, 47, 56], which are often encountered in real-world scenarios [19, 34, 35]. Previous methods [8, 11] have employed distillation-based consistent learning to enhance degradation-robust features.

---

[*] This work was done during Guangqian Guo's internship at Huawei, supervised by Yong Guo. The first two authors contribute equally.

[†] Corresponding author: Shan Gao

Nonetheless, they still face challenges in handling severely degraded low-quality images, as illustrated in Fig. 1. As degradations become more complex (*e.g*. combining various types of degradation or increasing the level of degradation), the existing SAMs [8, 27] struggle to accurately segment edges and complete target areas, leading to incorrect segmentation. We analyze that it is caused by the limited feature representation for degraded images. The visualizations in Fig. 2 reveal that SAM's latent features from severely degraded images contain excessive noise, compromising the original representations and subsequently impacting the predictions of the decoder. Furthermore, the large gap between low-quality and high-quality features complicates consistency learning [38] in previous works [8], hindering performance improvement. Thus, achieving high-quality latent feature representations and robust segmentation across varying image quality, especially for degraded images, remains challenging.

The recently developed generative Diffusion Models (DM) [20, 50], especially the large-scale pre-trained Latent Diffusion Models (LDM) [45] have demonstrated powerful content generation capabilities. Having been trained on internet-scale data [46], LDM that proceed diffusion and denoising in latent space, possess powerful representation prior, which can be well explored to enhance the latent representation of segmentation models. This inspires us to take full advantage of the generative ability of pre-trained diffusion models and incorporate it into the latent space of SAMs to enhance low-quality features, thus promoting accurate segmentation in low-quality images.

To this end, we propose ***GleSAM***, which reconstructs high-quality features (Fig. 2 (d)) in SAM's latent space through generative diffusion, enabling accurate segmentation across any-quality images. Starting with low-quality features, high-quality representations are generated through single-step denoising. To integrate LDM generative knowledge, we incorporate a pre-trained U-Net from LDM with learnable LoRA layers [21] to align with segmentation-specific features. Furthermore, to improve compatibility between the pre-trained diffusion model and the segmentation framework, we introduce two effective techniques: ***Feature Distribution Alignment*** (FDA) and ***Channel Replicate and Expansion*** (CRE). These techniques bridge feature distribution and structural alignment gaps between models. Built upon SAM and SAM2, GleSAM leverages the generalization of pre-trained segmentation and diffusion models, with a few learnable parameters added, and can be efficiently trained within 30 hours on four GPUs.

In terms of data, we constructed ***LQSeg*** based on existing datasets [9, 17, 31, 32, 49] to train and assess segmentation models on low-quality images. LQSeg incorporates a greater diversity of degradation types than previous methods [8], combining basic degradation models (e.g., noise and blur) to simulate complex and real-world noise [54, 63]. We also introduce three degradation levels for a more comprehensive evaluation. We hope LQSeg will inspire the development of more robust segmentation models and contribute to future research. Overall, our contributions are summarized as:

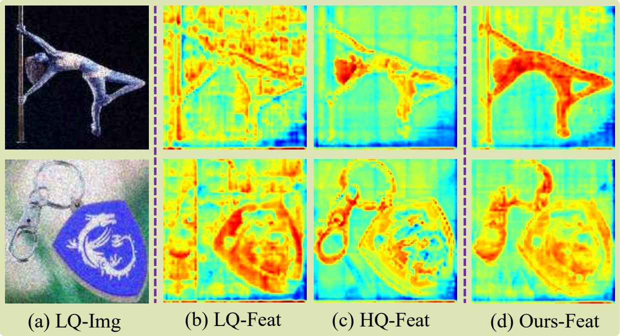


Figure 2. The visualization of latent features: (a) low-quality (LQ) images, (b) the SAM's latent features extracted from LQ images, which contain excessive noise and compromise the original representations, (c) the high-quality (HQ) features of the corresponding clear images, which are more salient than LQ ones, and (d) enhanced representation by our GleSAM.

- We propose GleSAM, a SAM-based framework incorporating generative latent space enhancement, to generalize across images of any quality. GleSAM exhibits significantly improved robustness, particularly for low-quality images with varying degradation levels.
- Two effective techniques: FDA and CRE, are introduced to bridge feature distribution and structural gaps between the pre-trained latent diffusion model and SAM.
- We also construct the LQSeg dataset which includes a wide range of degradation types and levels, to effectively train and evaluate the model.
- Extensive experiments show that our method performs excellently on low-quality images with varying degrees of degradation while maintaining generalization to clear images. Additionally, our method achieves strong performance on unseen degradations, highlighting the adaptability of both our framework and dataset.

2. Related Work

2.1. Segmentation on Low-Quality Images

Executing robust segmentation across various scenarios is a critical issue. Numerous studies [8, 24, 43, 56] have highlighted significant performance degradation in conventional segmentation models and foundational SAMs when confronted with low-quality images with degradation. Many related studies [11, 14, 24, 29, 43] have been proposed to enhance the robustness of segmentation models against low-quality data. These methods primarily consider a single type of degradation. Recently, RobustSAM [8] is introduced to enhance the robustness of the SAM against multiple image degradations through anti-degradation feature learning. However, its performance also struggles when dealing with complex degradations. The real-world image noise is often too complex to be modeled by a single degradation [19, 34, 35]. Therefore, robustly segmenting images of any quality remains challenging.

2.2. Diffusion Models for Perception Tasks

Recently, diffusion models [20, 39, 45, 50, 62] have garnered significant attention in research, due to their powerful generation capabilities. Numerous studies [1, 2, 6, 23, 33, 40, 53, 60, 64] explore how to extend their applications to a broader range of tasks, such as detection, segmentation, and image reconstruction, *etc*. For diffusion-based perception tasks, one category of methods [1, 7, 58, 59] reformulate the perception tasks as progressive denoising from random noise. such as DiffusionDet [6] and DiffusionInst [6]. Another route employs the denoising autoencoder pre-trained on the text-to-image generation as a backbone for downstream perception tasks [13, 23, 28, 40, 64]. For example, VPD [64] passes the image through a pre-trained diffusion model and extracts intermediate features for task prediction. Diverging from these existing works, we preserve the original segmentation structure and fine-tune a generative diffusion to enhance the segmentation model's latent representations for accurate segmentation of any quality images.

2.3. Segment Anything Model and Variants

Segment Anything Models (SAMs) [27, 44] have gained significant influence within the community due to their outstanding zero-shot segmentation capabilities. SAM [27] can interactively segment any object in an image using visual prompts such as points and bounding boxes. Most recently, the updated SAM2 [44] has been released, showing improved segmentation accuracy and inference efficiency. Their generalization abilities have led to breakthroughs and new paradigms in various downstream tasks [25, 30, 36, 48, 52, 57, 61, 65]. Although SAM is powerful, its performance decreases when facing complex scenarios, such as degraded images [22, 42, 56] and adverse weather conditions [47], which significantly hinders the real-world applications of SAM. Enhancing SAM's capability in such challenging scenarios is a worthwhile research topic.

3. Generative Latent Space Enhancement for Any-Quality Image Segmentation

In the following, we explore how to improve SAM's robustness for low-quality images through generative latent space enhancement, thus enabling it to generalize across varying image qualities. The overall framework of the proposed ***GleSAM*** is shown in Fig. 3. To begin, in Sec. 3.1, we propose incorporating diffusion models' generative capabilities into SAM's latent space to effectively and efficiently enhance low-quality feature representations. Next, to improve the compatibility of feature distribution and architecture between the pre-trained diffusion model and SAM, we introduce two techniques: ***Feature Distribution Alignment*** and ***Channel Replicate and Expansion***, which are detailed in Sec. 3.2 and Sec. 3.3, respectively. Finally, the overall training method is outlined in Sec. 3.4.

3.1. Latent Denoising Diffusion in Segmentation

Recall that diffusion models [20, 45, 50] are a class of probabilistic generative models that progressively add noise to the latent space, and then they learn to reverse this process by predicting and removing the noise. Formally, in LDMs, the forward noise process iteratively adds Gaussian noise with variance $\beta_t \in (0, \mathrm{I})$ to the variable z. The sample at each time point is defined as:

$$z_t = \sqrt{\overline{\alpha}_t} z + \sqrt{1 - \overline{\alpha}_t}\epsilon, \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\epsilon \in \mathcal{N}(0, \mathrm{I})$. While the inverse diffusion process is modeled by applying a neural network $\epsilon_\theta(z_t, t)$ to predict the noise $\hat{\epsilon}$ and recover the original input $\hat{z}$. LDMs model the above process in a latent space using a pre-trained Variational AutoEncoder (VAE) and then up-sample the latent output to the original resolution using the VAE decoder, enabling more efficient computations in the training and inference phases.

A similar idea motivates us to introduce the generative latent space denoising process into the SAMs' framework to reconstruct low-quality segmentation features. Let's denote $\mathcal{E}_\theta$ and $\mathcal{D}_\theta$ the segmentation encoder and decoder of SAMs, respectively. As shown in Fig. 3, the input LQ image is first compressed by $\mathcal{E}_\theta$ and generates LQ feature z_L. We consider z_L to be a noisy version of z_H, containing sufficient information to reconstruct a high-quality feature. Instead of the complex multi-step denoising from random noise, we start directly from z_L and forward with a single denoising step. Specifically, based on Eq. 1, the clean latent variable z can be directly predicted from the model's predicted noise $\hat{\epsilon}$, as:

$$\hat{z} = \frac{z_t - \sqrt{1 - \overline{\alpha}_t}\hat{\epsilon}}{\sqrt{\overline{\alpha}_t}}, \tag{2}$$

where $\hat{\epsilon}$ is the prediction of the network ϵ_θ with given z_t and t: $\hat{\epsilon} = \epsilon_\theta(z_t; t)$. We re-parameterize the above generative denoising process to adapt low-quality latent space enhancement in segmentation, as:

$$\hat{z}_H = \mathrm{GLE}(z_L) = \frac{z_L - \sqrt{1 - \overline{\alpha}_T}\epsilon_\theta(z_L; T)}{\sqrt{\overline{\alpha}_T}}, \tag{3}$$

where we consider low-quality feature x_L as the noised feature and perform one-step denoising at the T-th diffusion

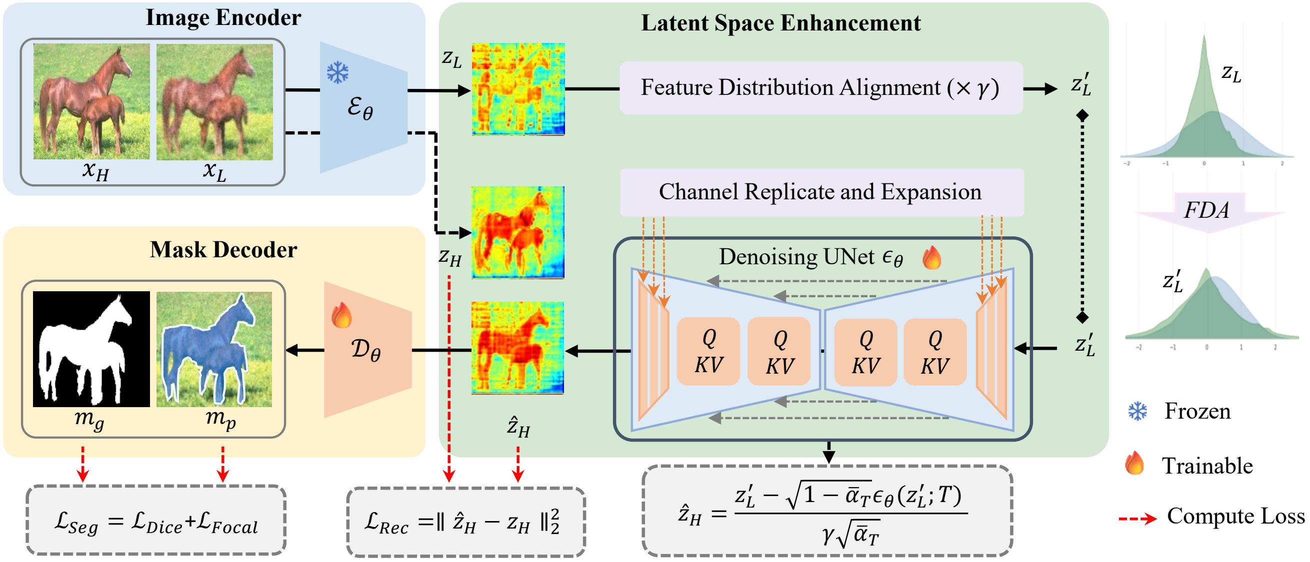


Figure 3. Given an input image, GleSAM performs accurate segmentation through image encoding, generative latent space enhancement, and mask decoding. During training, HQ-LQ image pairs are fed into the frozen image encoder to extract the corresponding HQ and LQ latent features. We then reconstruct high-quality representations in the SAM's latent space by efficiently fine-tuning a generative denoising U-Net with LoRA. Subsequently, the decoder is fine-tuned with segmentation loss to align the enhanced latent representations. Built upon SAMs, GleSAM inherits prompt-based segmentation and performs well on images of any quality.

timestep. The denoised output $\hat{z}_H$ is expected to be closer to the features extracted from clear images z_H. This single-step process significantly reduces computational overhead, making it more efficient when applied to segmentation models. After that, with $\hat{z}_H$ as input, the mask decoder can predict more precise masks, as: $m_p = \mathcal{D}_\theta(\hat{z}_H)$.

3.2. Feature Distribution Alignment

We employ the pre-trained U-Net in LDM as the denoising backbone. However, a significant challenge arises due to the substantial difference between the latent spaces in the original LDM (encoded by VAE) and segmentation models, leading to several technical issues for our application.

Firstly, there is a distribution gap between the two spaces and directly feeding segmentation features into U-Net may prevent it from fully exerting its denoising capabilities, as shown in the right part of Fig. 3. To address this gap, we introduce a Feature Distribution Alignment (FDA) technique. Specifically, we add an adaptation weight γ to scale the segmentation features, adjusting their variance to align more closely with the VAE's latent space. This adjustment ensures that the features are compatible with U-Net's optimal input space, improving the robustness and accuracy of the semantic interpretation and enhancing the denoising capability. The LQ feature denoising process in Eq. 3 can be updated as:

$$\hat{z}_H = \mathrm{GLE}(z_L) = \frac{\gamma z_L - \sqrt{1-\bar{\alpha}_T}\epsilon_\theta(\gamma z_L; T)}{\gamma\sqrt{\bar{\alpha}_T}}, \tag{4}$$

where we divide by γ to restore its original distribution. We experimentally verified in Sec. 5.4 that this simple operation effectively improves U-Net's denoising performance when applied to segmentation features.

3.3. Channel Expansion for Head-tail Layers

Another technical issue arises from the channel mismatch of the head and tail layers between the pre-trained U-Net and the segmentation features. The U-Net in LDMs is designed for 4-channel input and output ($h \times w \times 4$), which does not match the dimension of SAM's latent space ($h \times w \times 256$). We explore various methods to solve this problem (in Sec. 5.4) and empirically find that fine-tuning new head and tail layers or an encoder-decoder for segmentation features is ineffective. This is likely due to difficulties in aligning with the pre-trained model's parameters while preserving its generalization ability. To address this, we propose a Channel Replication and Expansion (CRE) method that replicates and concatenates the pre-trained weights of head and tail layers to match the required channel dimension. During training, the parameters of the head and tail layers remain frozen, while learnable LoRA layers are added to adapt to segmentation features. This approach effectively preserves the pre-trained generalization capacity while minimizing the number of learnable parameters.

3.4. Training Method

We employ a two-step fine-tuning process. In the first step, we fine-tune the denoising U-Net to reconstruct high-quality features. In the second step, we fine-tune the decoder with the restored features to further align the feature space for more accurate segmentation.

U-Net finetuning. To adapt the pre-trained U-Net to the segmentation framework while preserving its inherent generalization ability, we employ the LoRA [21] scheme to fine-tune all the attention layers in the U-Net. During this step, we freeze the pre-trained image encoder and U-Net layers and only fine-tune the added LoRA layers. The estimated feature is compared with the corresponding HQ feature z_H by a reconstruction loss, as:

$$\mathcal{L}_{\text{Rec}} = \mathcal{L}_{\text{MSE}}(\text{GLE}(z_L), z_H). \tag{5}$$

This step significantly enhances performance without fine-tuning SAM's parameters.

Decoder finetuning. Next, we use the reconstructed high-quality features to fine-tune the mask decoder for more precise segmentation. Our experiments demonstrate that fine-tuning either the entire decoder or only the output tokens with these features further improves segmentation accuracy while maintaining generalization on clear images. Focal Loss and Dice Loss are employed as segmentation loss functions, as:

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{Dice}}(m_p, m_g) + \mathcal{L}_{\text{Focal}}(m_p, m_g), \tag{6}$$

where m_p and m_g indicate predicted and ground-truth masks respectively.

4. Low-Quality Image Segmentation Dataset

We construct a comprehensive low-quality image segmentation dataset dubbed ***LQSeg*** that encompasses more complex and multi-level degradations, rather than relying on a single type of degradation for each image. The dataset is composed of images from several existing datasets with our synthesized degradations. In this section, we first model a multi-level degradation process of low-quality images and then detail the dataset composition.

4.1. Multi-level Degradation Modeling

To model a more practical and complex degradation process, inspired by the previous work in image reconstruction [54, 63], we utilize a mixed degradation method. Specifically, the degraded process is modeled as the random combination of the four common degradation models, including *Blur*, *Random Resize*, *Noise*, and *JPEG Compression*. Each degradation model encompasses various types, such as Gaussian and Poisson noise for *Noise*, ensuring the diversity of the degradation process.

To enrich the granularity of degradation, we employ multi-level degradation by adjusting the downsampling rates. We employed three different resize rates, *i.e.*, [1, 2, 4], which correspond to three degradation levels from slight to severe: LQ-1, LQ-2, and LQ-3. More implementation details are illustrated in *Supplementary Material*.

4.2. Dataset Composition

Based on the above multi-level degradation model, we construct ***LQSeg*** to train our model and evaluate the segmentation performance on different levels of low-quality images. The images in LQSeg are sourced from several well-known existing datasets in the community with our synthesized degradation. In detail, for the ***training set***, we utilize the entire training sets of LVIS [17], ThinObject-5K [31], and MSRA10K [9] as the source data and procedurally synthesize corresponding low-quality images. The ***evaluation set*** is sourced from four subsets, *i.e.*, 1) the whole test sets of ThinObject-5K and LVIS (seen sets), and 2) ECSSD [49] and COCO-val [32] (unseen sets). For each source image, We systematically generate three levels of degraded images to thoroughly assess the model's robustness.

5. Experiment

We conduct extensive experiments to verify our method across images of varying quality. All proposed techniques can be applied to SAM and SAM2, referred to as GleSAM and GleSAM2. In practice, our models perform well on low-quality (Tab. 1, 2) and clear images (Tab. 5) and they generalize effectively to unseen degradations (Tab. 3).

5.1. Experimental Setup

Implement Details. Our model is trained using the AdamW optimizer with the learning rate of 1×10^{-4} and batch size of 4. The pre-trained U-Net in Stable Diffusion (SD) 2.1-base [45] is adopted as the denoising backbone. We use 3 clicks as SAM's prompts by default. Our approach can be efficiently trained on 4× A100 GPUs within approximately 30 hours, during which we fine-tune the U-Net for 100K iterations and the decoder for only 20K iterations.

Evaluation Metrics. We employ three metrics to assess our model's performance, including Intersection over Union (IoU), Dice Coefficient (Dice), and Pixel Accuracy (PA).

5.2. Performance Comparisons

In this experiment, we evaluate the performance on the test set of our LQSeg, including seen-set (Tab. 1) and unseen-set (Tab. 2) evaluations. We compare our method with a set of comparison baselines to quantify the performance gains. For SAM-based comparisons, besides the original SAM [27], we also compare with the RobustSAM [8], which has improved robustness on the degraded dataset. Additionally, we compare with two-stage methods, *i.e.*, reconstructing images first with image reconstruction (IR) networks and passing the restored clear images to the SAM and SAM2. We use two state-of-the-art IR networks for comparison: PromptIR [41], and diffusion-based DiffBIR [33].

Method	ThinObject-5K						LVIS					
	LQ-3		LQ-2		LQ-1		LQ-3		LQ-2		LQ-1	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
SAM	0.6285	0.7286	0.7054	0.7939	0.7527	0.8343	0.4041	0.5005	0.4886	0.5838	0.5325	0.6282
RobustSAM	0.7015	0.7965	0.7648	0.8463	0.7922	0.8658	0.4517	0.5670	0.4962	0.6079	0.5262	0.6356
PromptIR-SAM	0.6380	0.7374	0.7146	0.8018	0.7434	0.8248	0.4020	0.4978	0.4705	0.5677	0.5222	0.6187
DiffBIR-SAM	0.7055	0.7917	0.7772	0.8531	0.7927	0.8652	0.5316	0.6307	0.5812	0.6913	0.6021	0.7090
GleSAM (Ours)	**0.7594**	**0.8413**	**0.8033**	**0.8738**	**0.8277**	**0.8920**	**0.5535**	**0.6756**	**0.5916**	**0.7066**	**0.6131**	**0.7244**
SAM2	0.7174	0.8000	0.7636	0.8373	0.7839	0.8536	0.5118	0.6174	0.5634	0.6633	0.6024	0.7004
PromptIR-SAM2	0.7119	0.7945	0.7753	0.8477	0.7801	0.8487	0.5017	0.6084	0.5529	0.6546	0.5875	0.6865
DiffBIR-SAM2	0.7348	0.8117	0.7832	0.8505	0.7974	0.8604	0.5651	0.6664	0.6004	0.7032	0.6278	0.7380
GleSAM2 (Ours)	**0.7947**	**0.8653**	**0.8300**	**0.8896**	**0.8527**	**0.9072**	**0.5738**	**0.6887**	**0.6082**	**0.7161**	**0.6361**	**0.7402**

Table 1. Performance comparison on the test set of Thinobject-5K [31] and LVIS [17] datasets (seen datasets) with different levels of degradation. From LQ-1 to LQ-3, the degree of degradation increases progressively. We report IoU and Dice for comparison. Our GleSAM and GleSAM2 consistently outperform other competitors, especially on the most challenging LQ-3 version. The words with boldface indicate the best results and those underlined indicate the second-best results.

Method	ECSSD						COCO					
	LQ-3		LQ-2		LQ-1		LQ-3		LQ-2		LQ-1	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
SAM	0.5219	0.6383	0.6156	0.7202	0.6778	0.7735	0.4048	0.5014	0.4848	0.5810	0.5392	0.6350
RobustSAM	0.6535	0.7659	0.7271	0.8240	0.7758	0.8606	0.4541	0.5711	0.4956	0.6087	0.5304	0.6401
PromptIR-SAM	0.5277	0.6442	0.6109	0.7188	0.6752	0.7726	0.4006	0.4969	0.4784	0.5753	0.5257	0.6219
DiffBIR-SAM	0.6675	0.7652	0.7408	0.8244	0.7737	0.8501	0.5260	0.6263	0.5742	0.6918	0.6112	0.7080
GleSAM (Ours)	**0.7332**	**0.8282**	**0.7944**	**0.8719**	**0.8257**	**0.8931**	**0.5495**	**0.6705**	**0.5900**	**0.7055**	**0.6158**	**0.7257**
SAM2	0.6534	0.7540	0.7292	0.8157	0.7625	0.8417	0.5176	0.6230	0.5681	0.6692	0.6018	0.7000
PromptIR-SAM2	0.6490	0.7526	0.7175	0.8082	0.7519	0.8351	0.5088	0.6153	0.5579	0.6600	0.5905	0.6914
DiffBIR-SAM2	0.7158	0.8026	0.7853	0.8588	0.8070	0.8745	0.5680	0.6711	0.6109	0.7190	0.6245	0.7348
GleSAM2 (Ours)	**0.7335**	**0.8258**	**0.7976**	**0.8726**	**0.8175**	**0.8853**	**0.5724**	**0.6867**	**0.6165**	**0.7246**	**0.6337**	**0.7371**

Table 2. Zero-shot performance comparison on the ECSSD [49] and COCO [32] datasets (unseen datasets) with different levels of degradation. These results indicate that GleSAMs possesses significant robustness in zero-shot segmentation across different levels of degradations.

Performance Comparison on Seen Datasets. In Tab. 1, we evaluate the performance of our GleSAM and GleSAM2 on two seen datasets: ThinObject-5K [31] and LVIS [17]. Each dataset contains three levels of degradation. Our method demonstrates superior performance across all degradation levels, effectively handling low-quality images. As the degradation level increases (from LQ-1 to LQ-3), GleSAMs show increasingly significant performance improvements compared to the baseline, highlighting its robustness against challenging degradations.

Performance Comparison on Un-seen Datasets. In Tab. 2, we evaluate the zero-shot segmentation performance of GleSAMs on two unseen datasets: ECSSD [49] and COCO [32], all of which are synthesized with different levels of degradations. GleSAM consistently outperforms other methods, particularly on the most challenging LQ-3 version, underscoring its strong zero-shot generalization capabilities and potential for real-world applications. To further assess segmentation quality, we plot the density distribution maps of image quality and segmentation IoU in Fig. 4. Compared to the baselines, our method achieves overall improvement and more stable performance on low-quality images, especially for those of inferior quality.

Validation with Other Degradations. To validate the model's generalization on other unseen degradations, we evaluated GleSAM and GleSAM2 on RobustSeg-style degradations [8], with the results presented in Tab. 3. These degradations were not used during training. Our method consistently outperforms SAM and SAM2 and even surpasses RobustSAM which is specifically trained on RobustSeg. This demonstrates the strong generalization of our method across diverse degradations.

5.3. Ablation Study

We conduct ablation experiments in Tab. 4, 5 to further understand the impact of each component of our framework.

Effect of Each Component. In Tab. 4, we validated the effectiveness of each proposed module in GleSAM. "Gle" indicates our framework with generative latent space enhancement. The results show that each introduced method significantly improves the performance. We also make a qualitative visualization in Fig. 5, which shows that the

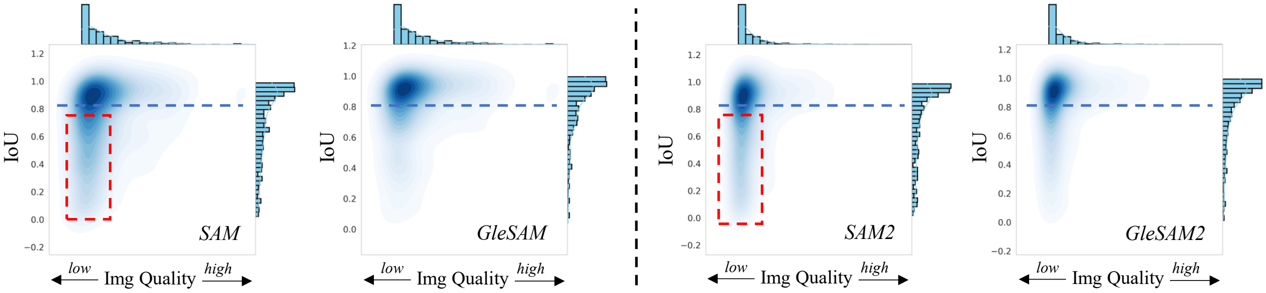


Figure 4. Density distribution maps about IoU and image quality across different methods, including SAM, GleSAM, SAM2, and GleSAM2. The image quality is calculated using the Laplacian operator in OpenCV. The red dashed box highlights the area where our method demonstrates improved segmentation performance compared to SAM, particularly in lower-quality images.

Method	ECSSD		COCO	
	IoU	Dice	IoU	Dice
SAM [27]	0.7833	0.8573	0.6512	0.7457
RobustSAM [8]	0.8427	0.9044	0.6543	0.7518
GleSAM (Ours)	**0.8568**	**0.9119**	**0.6678**	**0.7700**
SAM2	0.8211	0.8817	0.6817	0.7753
GleSAM2 (Ours)	**0.8611**	**0.9133**	**0.6869**	**0.7836**

Table 3. Zero-shot performance comparison on RobustSeg-style [8] degradations. Performance is tested on the unseen ECSSD and COCO datasets. Note that our methods are not trained on such degradations. The superior performance of our method demonstrates robustness against various forms of degradation.

Method	ECSSD		COCO	
	IoU	Dice	IoU	Dice
Baseline	0.6054	0.7107	0.4763	0.5725
+ Gle & CRE	0.6567	0.7657	0.4958	0.5963
+ Gle & CRE & FDA	**0.7104**	**0.8020**	**0.5166**	**0.6174**

Table 4. Ablation study of each component in the proposed method, evaluated on the unseen ECSSD and COCO datasets. Each additional component positively affects the performance, demonstrating the effectiveness of the proposed methods.

clearest latent feature is obtained when combining all modules. Additionally, in Tab. 5, without fine-tuning the decoder, our method improves the results on LQ images by about 7 points, while also preserving the robustness of SAM on clear conditions, indicating the robustness of our method for both degraded and clear images.

Effect of Fine-tuning SAM. We explore two common configurations to fine-tune SAM: fine-tuning the entire SAM's decoder and the output mask token. Our experiments reveal that directly fine-tuning SAM's decoder or output mask token on degraded images leads to a significant drop in zero-shot performance on clear images, with the IoU decreasing by nearly ***20 points***. In contrast, our method further improves performance on both low-quality and clear images by fine-tuning the entire decoder or mask token, enabling the segmentation of images with any quality.

Method	LQ		Clear		Average	
	IoU	Dice	IoU	Dice	IoU	Dice
w/o Fine-tuning SAM:						
SAM	0.5407	0.6416	0.7830	0.8554	0.6619	0.7485
Ours	**0.6135**	**0.7097**	**0.7846**	**0.8567**	**0.6991**	**0.7832**
Fine-tuning SAM:						
SAM-FT-T	0.6305	0.7374	0.5847	0.7029	0.6076	0.7202
SAM-FT-D	0.6327	0.7385	0.6071	0.7242	0.6199	0.7314
Ours-FT-T	0.6759	0.7751	**0.8061**	**0.8747**	0.7410	0.8249
Ours-FT-D	**0.6848**	**0.7825**	0.8022	0.8704	**0.7435**	**0.8265**

Table 5. Effect of Fine-tuning SAM. The performance is evaluated on the unseen ECSSD and COCO datasets. "FT-T" and "FT-D" indicate fine-tuning the SAM's mask token and decoder respectively. "LQ" indicates the mean performance on three levels of degraded data, and "Clear" indicates the results on the original clear images.

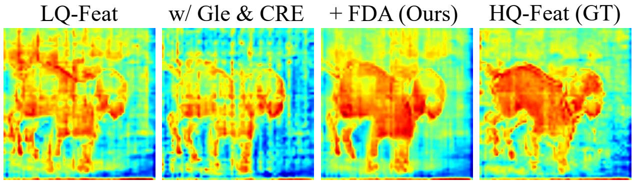
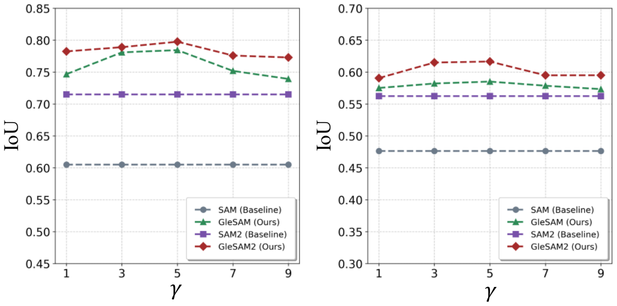


Figure 5. Qualitative visualization of the enhanced latent features. The clearest feature is obtained when combining all modules.

5.4. Detailed Analysis

Analysis of channel expansion methods. To address the issue of channel mismatch between SAM and the pre-trained U-Net, we explored various strategies including (a) using two simple convolutional layers to reduce and expand the channels of segmentation features as needed, (b) fine-tuning new head and tail layers from scratch, and (c) our CRE method. Our results (in Tab. 6 and Fig. 7) show that strategies (a) and (b) are ineffective, likely because the new layers couldn't leverage the pre-trained knowledge. In contrast, our method effectively resolves this issue by replicating pre-trained parameters and fine-tuning with LoRA.

Analysis of hyperparameter γ. We use an adaption weight



(a) ECSSD (b) COCO

Figure 6. Ablation study of adaption weight γ.

Method	IoU	Dice	PA
(a) Additional encoder and decoder	0.4544	0.5842	0.6106
(b) New head and tail layers	0.6014	0.7077	0.7782
(c) Replicate and Expansion (Ours)	**0.6567**	**0.7657**	**0.8400**

Table 6. Analysis of the proposed CRE. It significantly outperforms alternative approaches, achieving higher scores.

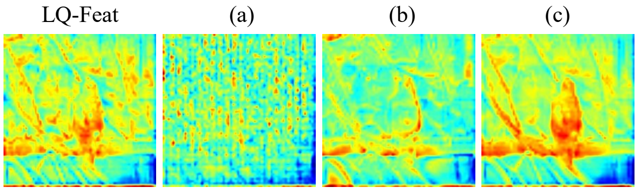


Figure 7. Qualitative visualization of the enhanced latent features generated by different channel expansion methods. Our proposed CRE method (c) produces more salient features.

γ to align the distribution of the latent space between LDM and SAM. To determine the optimal value of γ, we empirically test five different values on the ECSSD and COCO datasets. The results, shown in Fig. 6, suggest that $\gamma = 5$ is the most effective setting, providing strong generalization across all models and datasets. Therefore, we adopt $\gamma = 5$ as the default value in our experiments.

Analysis of LoRA ranks. We incorporate learnable LoRA layers to fine-tune the pre-trained denoising U-Net. Here, we evaluate the impact of different LoRA ranks on segmentation performance. The results are shown in Tab. 8. We tested the results and the corresponding number of learnable parameters for setting a rank to 4, 8, and 16. We found that setting the rank to 8 can obtain good results while maintaining an acceptable number of parameters.

Analysis of degradation levels. To investigate the necessity of training on all three degradation levels together, we train GleSAM on LQ datasets for each degradation level (LQ-1, LQ-2, LQ-3) individually. We evaluate the models' performance on ECSSD and COCO datasets using three degraded levels and RobustSeg-style degradations. The results in Tab. 7 show that training on all three degradation levels together contributes to more robust performance across varying levels of image degradation.

Training	Ours			LQ-RS	AVG
Data	LQ-1	LQ-2	LQ-3		
LQ-1	0.6981	0.6663	0.6076	0.7587	0.6827
LQ-2	0.7168	0.6843	0.6318	0.7621	0.6988
LQ-3	0.7185	0.6910	0.6403	0.7619	0.7030
LQ-1,2,3	**0.7208**	**0.6922**	**0.6414**	**0.7623**	**0.7042**

Table 7. Ablation study for degradation levels during training. LQ-RS indicates the RobustSeg-style [8] degradation. AVG indicates the average performance. We report the IoU for comparison. The results on the unseen ECSSD and COCO datasets reveal that multi-level degradation contributes to more robust performance.

Rank	LQ			Learnable
	IoU	Dice	PA	Params
4	0.7760	0.8597	0.9434	16.25M
8	**0.7844**	**0.8644**	**0.9456**	32.49M
16	0.7697	0.8543	0.9418	64.99M

Table 8. Analysis of LoRA ranks in U-Net.

Method	Learnable Params	Num. GPUs	Training Time (h)	Inference Speed (s)
SAM	1250 MB	256	N/A	0.32
GleSAM	47 MB	4	30	0.38 (+0.06)

Table 9. Computational requirements of GleSAM *vs* SAM.

Analysis of computational requirements. In Tab. 9, we report detailed training and inference comparisons between our GleSAM and SAM. Although GleSAM demonstrates significantly improved robustness, it introduces only marginal learnable parameters and incurs a slight trade-off in inference speed. The additional parameters can be efficiently optimized in 30 hours on four A100 GPUs.

6. Conclusion

We present GleSAM, a solution to enhance Segment Anything Models (SAM and SAM2) for robust segmentation across images of any quality, particularly those with severe degradation. We incorporate the generative ability of pre-trained diffusion models into the latent space of SAMs to enhance low-quality features, thus promoting more robust segmentation. Our approach is further supported by the LQSeg dataset, which includes diverse degradation types and levels, allowing for more comprehensive model training and evaluation. Extensive experiments demonstrate that GleSAM achieves superior performance on degraded images while maintaining generalization to clear ones, and it performs well on unseen degradations, highlighting its robustness and versatility. This work extends SAM capabilities, offering a practical approach for degraded scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62372382.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 3

[3] Huafeng Chen, Dian Shao, Guangqian Guo, and Shan Gao. Just a hint: Point-supervised camouflaged object detection. In *European Conference on Computer Vision*, pages 332–348, 2024. 1

[4] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *European Conference on Computer Vision*, pages 315–331, 2024. 1

[5] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[6] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023. 3

[7] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 909–919, 2023. 3

[8] Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Robustsam: Segment anything robustly on degraded images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4081–4091, 2024. 1, 2, 3, 5, 6, 7, 8

[9] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 2, 5

[10] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[11] Kazuki Endo, Masayuki Tanaka, and Masatoshi Okutomi. Semantic segmentation of degraded images using layer-wise feature adjustor. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3205–3213, 2023. 1, 2

[12] Shan Gao, Guangqian Guo, Hanqiao Huang, and CL Philip Chen. Go deep or broad? exploit hybrid network architecture for weakly supervised object classification and localization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[13] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv preprint arXiv:2307.02138*, 2023. 3

[14] Dazhou Guo, Yanting Pei, Kang Zheng, Hongkai Yu, Yuhang Lu, and Song Wang. Degraded image semantic segmentation with dense-gram networks. *IEEE Transactions on Image Processing*, 29:782–795, 2019. 2

[15] Guangqian Guo, Pengfei Chen, Xuehui Yu, Zhenjun Han, Qixiang Ye, and Shan Gao. Save the tiny, save the all: hierarchical activation network for tiny object detection. *IEEE transactions on circuits and systems for video technology*, 34 (1):221–234, 2023. 1

[16] Guangqian Guo, Dian Shao, Chenguang Zhu, Sha Meng, Xuan Wang, and Shan Gao. P2p: Transforming from point supervision to explicit visual prompt for object detection and segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. 1

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2, 5, 6

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[19] Glenn E Healey and Raghava Kondepudy. Radiometric ccd camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994. 1, 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, pages 6840–6851, 2020. 2, 3

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5

[22] Yihao Huang, Yue Cao, Tianlin Li, Felix Juefei-Xu, Di Lin, Ivor W Tsang, Yang Liu, and Qing Guo. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*, 2023. 1, 3

[23] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21741–21752, 2023. 3

[24] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8828–8838, 2020. 2

[25] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[26] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 9404–9413, 2019. 1

[27] 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. 1, 2, 3, 5, 7

[28] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13883–13893, 2024. 3

[29] Sohyun Lee, Taeyoung Son, and Suha Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18911–18921, 2022. 2

[30] Bo Li, Haoke Xiao, and Lv Tang. Asam: Boosting segment anything model with adversarial tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3699–3710, 2024. 3

[31] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 2, 5, 6

[32] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5, 6

[33] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3, 5

[34] Ce Liu, Richard Szeliski, Sing Bing Kang, C Lawrence Zitnick, and William T Freeman. Automatic estimation and removal of noise from a single image. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):299–314, 2007. 1, 3

[35] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2): 205–214, 2006. 1, 3

[36] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1, 3

[37] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 1

[38] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020. 2

[39] Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[40] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4157–4168, 2023. 3

[41] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[42] Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, Chenshuang Zhang, and Choong Seon Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023. 1, 3

[43] AN Rajagopalan et al. Improving robustness of semantic segmentation to motion-blur using class-centric augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10470–10479, 2023. 2

[44] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3

[45] 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. 2, 3, 5

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[47] Xinru Shan and Chaoning Zhang. Robustness of segment anything model (sam) for autonomous driving in adverse weather conditions. *arXiv preprint arXiv:2306.13290*, 2023. 1, 3

[48] Chuyun Shen, Wenhao Li, Yuhang Shi, and Xiangfeng Wang. Interactive 3d medical image segmentation with sam 2. *arXiv preprint arXiv:2408.02635*, 2024. 3

[49] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4): 717–729, 2015. 2, 5, 6

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3

[51] Chaowei Wang, Guangqian Guo, Chang Liu, Dian Shao, and Shan Gao. Effective rotate: Learning rotation-robust prototype for aerial object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[52] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything

model. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[53] Mengyu Wang, Henghui Ding, Jun Hao Liew, Jiajun Liu, Yao Zhao, and Yunchao Wei. Segrefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. *Advances in Neural Information Processing Systems*, 36:79761–79780, 2023. 3

[54] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2, 5

[55] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 1

[56] Yuqing Wang, Yun Zhao, and Linda Petzold. An empirical study on the robustness of the segment anything model (sam). *Pattern Recognition*, page 110685, 2024. 1, 2, 3

[57] Zhaoyang Wei, Pengfei Chen, Xuehui Yu, Guorong Li, Jianbin Jiao, and Zhenjun Han. Semantic-aware sam for point-prompted instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2024. 3

[58] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024. 3

[59] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6030–6038, 2024. 3

[60] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 3

[61] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2025. 3

[62] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[63] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2, 5

[64] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3

[65] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024. 3