Article | Received 14 May 2026; Revised 18 August 2026; Accepted 26 August 2026; Published XX September 2026
https://doi.org/10.55092/rl20260025

Perturbation-aware distillation for few-shot skin lesion segmentation

Yang Wang[1], Kan Yu[1], Yuanyuan Liu[2], Zijing Chen[1] and Zhe Chen[1,3,*]

[1] Department of Computer Science and Information Technology, La Trobe University, Flora Hill, Australia
[2] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China
[3] Australian Centre for AI in Medical Innovation (ACAMI), La Trobe University, Melbourne, Australia

* Correspondence author; E-mail: zhe.chen@latrobe.edu.au.

Highlights:

- Perturbation-aware distillation learns robust lesion features under prompt noise.
- MePoDIN weights reliable teacher channels under prompt uncertainty.
- Training uses perturbed GT boxes; inference uses prompt approximation.
- Robustness, ablation, and efficiency studies support the method design.
- Strong few-shot ISIC results with good transfer to PH2.

Abstract: Skin lesion segmentation supports computer-aided dermatology, but blurred boundaries, low contrast, and marked variation in dermoscopic appearance make reliable delineation difficult. Foundation models such as the Segment Anything Model (SAM) offer strong visual priors, although their sensitivity to prompt quality limits their reliability in few-shot medical settings. We present PADS, a novel perturbation-aware distillation framework for robust skin lesion segmentation under prompt variability. During training, we perturb bounding-box prompts derived from ground-truth masks to simulate realistic prompt noise. A lightweight relevance module then estimates channel-wise feature importance under these perturbations and guides selective distillation from MedSAM into a compact UNet. This training improves robustness to imperfect or noisy prompts, allowing pseudo-prompts to support reliable segmentation. At inference, the UNet either produces the segmentation directly or supplies a pseudo-prompt for optional MedSAM refinement. Experiments on ISIC 2018 and cross-dataset evaluation on PH2 show improved few-shot segmentation accuracy and slower performance degradation as prompt perturbation increases, compared with standard distillation. PADS adds only limited computational overhead relative to MedSAM. The experimental results support perturbation-aware distillation as a practical approach to robust few-shot medical image segmentation.

Keywords: few-shot segmentation; skin lesion segmentation; knowledge distillation; foundation model adaptation; prompt robustness

Copyright © 2026 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Skin cancer is a major global health concern, and its incidence has been rising steadily over recent decades. The World Health Organization (WHO) emphasizes the importance of early detection and treatment for improving outcomes (https://www.who.int/news-room/fact-sheets/detail/cancer). To diagnose skin cancer, clinicians use dermoscopic imaging as a non-invasive way to view skin lesions, but tracing their boundaries accurately is usually difficult and time-consuming, even for experienced clinicians. For machine-based skin lesion analysis, fuzzy edges, low contrast, shadows, and hair occlusion make this clinically important task particularly challenging. Since dermoscopic images often lack reliable structural information to distinguish lesions from the surrounding skin, even the most advanced algorithms struggle to provide reliable results. In CT or MRI scans, by comparison, anatomical structures typically have clearer boundaries and stronger shape priors. As a result, developing reliable automated skin lesion segmentation is important for supporting scalable and consistent computer-aided diagnosis [1,2].

To improve skin lesion diagnosis, researchers have turned to deep learning, which has brought substantial improvements to medical image segmentation over the past decade. UNet [3], for example, performs well using an encoder–decoder architecture with skip connections. Still, these models need large numbers of pixel-wise annotations to generalize well. In medical imaging, these annotations are costly and labor-intensive to obtain, and privacy and ethical considerations impose further constraints [4–6]. The shortage of labelled data remains a barrier to reliable clinical deployment. Few-shot learning (FSL) aims to address this problem by learning to generalize from a handful of labelled examples [7]. Existing few-shot segmentation methods [8–12] typically use support–query paradigms, prototype decomposition or matching, and contrastive representation learning, though these approaches may be difficult to use in clinical workflows. Few-shot models also struggle with irregular lesion shapes, high intra-class variation, and ambiguous boundaries in dermoscopic images. Methods such as MPBA-Net [13] and BDFormer [14] use boundary-aware or attention-enhanced mechanisms to address parts of this problem, but they still need extensive labelled datasets and comprehensive training.

More recently, large-scale foundation models have provided another approach to segmentation. The Segment Anything Model (SAM) [15] uses user-provided prompts to guide its predictions and can segment objects across diverse visual domains. It can effectively generalize to unseen objects without substantial task-specific retraining. For medical segmentation, this offers a way to use SAM's rich visual priors instead of training models entirely from scratch. However, prompt quality remains a major limitation, making it difficult to apply SAM to skin lesion segmentation. In particular, ambiguous lesion boundaries and subtle spatial cues in dermatological images make precise prompts hard to specify without expert input. Inaccurate prompts can reduce segmentation quality, and the limited supervision available in few-shot settings does little to correct these errors. MedSAM [16] and SAM-Med2D [17] use fine-tuning or adapter-based strategies to address the gap between general-domain and medical segmentation. These adaptations can be effective, but they still rely on large labelled datasets or carefully hand-annotated prompts. This limits their practical use in few-shot clinical settings. Throughout this paper, we use SAM to denote the general Segment Anything architecture or model family. Here, vanilla SAM denotes the original general-domain model [15], and MedSAM denotes the medical-domain model used as the frozen

teacher and optional refinement model in all perturbation-aware distillation (PADS) experiments.

Instead of eliminating prompts or retraining foundation models, we aim to learn lesion representations that remain robust under prompt uncertainty in a lightweight segmentation model. MedSAM encodes rich medical visual priors, yet not all internal features are equally relevant for lesion segmentation, and its predictions remain conditioned on specific prompt inputs. If reliable MedSAM representations are selectively distilled into a compact model under controlled prompt variability, the student can acquire stable lesion-aware priors with minimal supervision. Figure 1 compares the annotation requirements and reliance on manual prompts of PADS, fully supervised UNet segmentation, conventional few-shot segmentation, and prompt-based zero-shot SAM evaluation. PADS combines perturbation-aware distillation during few-shot training with a UNet-generated pseudo-prompt for optional MedSAM refinement at inference, aiming to improve robustness to prompt variation and reduce dependence on precise manual prompts.

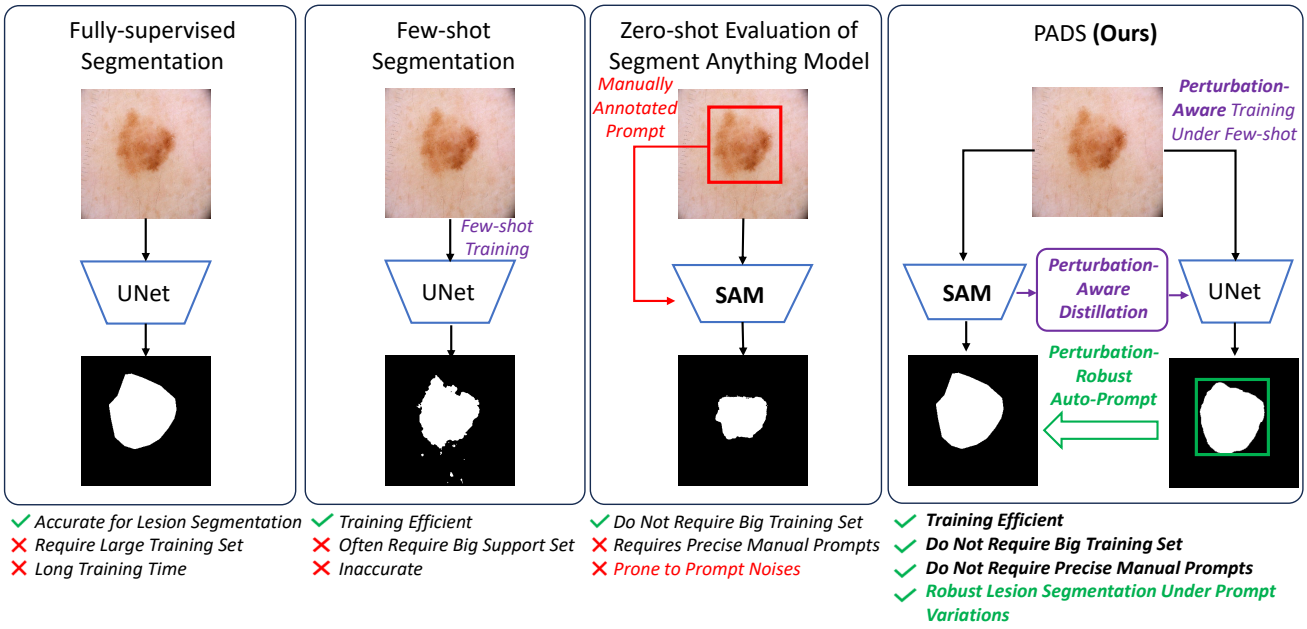

Figure 1. Conceptual overview of PADS for few-shot skin lesion segmentation. During training, PADS explicitly introduces prompt perturbations and learns which features of the frozen MedSAM teacher remain reliable under varying prompt conditions. This perturbation-aware distillation could enable the student network to learn stable lesion representations. At inference, the framework could generate promising pseudo-prompt rather than a manually specified prompt, improving robustness when prompt quality is imperfect.

Based on this insight, we propose PADS, a perturbation-aware distillation framework that transfers lesion-relevant knowledge from a frozen MedSAM teacher into a compact UNet under limited supervision. Whereas conventional perturbation-based training typically treats noise as augmentation and conventional feature distillation transfers teacher representations as fixed targets, PADS uses controlled translation and scale perturbations of ground-truth-derived boxes to create informative variation in the prompt-conditioned MedSAM response. The Meta-Prompt Distillation Instruction Network (MePoDIN) integrates these varying responses with MedSAM features and lesion supervision to estimate which teacher channels

remain reliable under prompt uncertainty, and the resulting relevance scores directly weight teacher–student feature alignment. Thus, perturbation is not merely used to expose the model to noisy prompts. It determines what knowledge should be transferred, emphasizing stable lesion evidence while suppressing prompt-sensitive features. To the best of our knowledge, we present the first skin lesion segmentation study to explicitly link prompt uncertainty, reliability estimation, and selective feature distillation. The novelty is in how these components work together, rather than in perturbation, attention, or distillation alone. This design reduces prompt sensitivity without fine-tuning MedSAM or optimizing prompts. At inference, the trained student can either segment independently or generate a *UNet-generated lesion-informed pseudo-prompt*, defined as a bounding box extracted from its predicted lesion mask, for optional MedSAM refinement. Hereafter, we use the shorter term *pseudo-prompt*.

We evaluate PADS on ISIC 2018 [18,19] and PH2 [20] under standard few-shot settings. The experiments are organized to answer four practical questions: whether the method improves segmentation accuracy against strong baselines, whether perturbation-aware relevance weighting is necessary, whether robustness is maintained under degraded prompts, and what computational overhead is introduced. In practice, PADS can be useful to learning-based robotic and automated dermatological perception pipelines, in which lesion segmentation may depend on approximate locations produced by upstream detection, camera positioning, scanning, or other visual modules. Errors in these automatically generated locations can degrade prompt-sensitive MedSAM segmentation. By using prompt-variable MedSAM supervision, PADS trains a compact perception model that is less dependent on precise localisation. The student can operate independently or provide a pseudo-prompt for optional MedSAM refinement.

Our main contributions are as follows:

- We formulate prompt variability as a representation learning problem and propose a perturbation-aware distillation mechanism that transfers stable lesion-relevant knowledge from MedSAM to a compact UNet.
- We introduce MePoDIN, a lightweight feature relevance module that estimates channel-wise importance under prompt uncertainty and enables relevance-weighted distillation beyond standard feature matching or generic channel attention.
- We demonstrate a practical application pathway in which the trained UNet provides a pseudo-prompt for optional MedSAM refinement, while controlled experiments on ISIC 2018 and PH2 validate accuracy, robustness, ablation behavior, and efficiency.

2. Related work

2.1. Skin lesion segmentation

Skin lesion segmentation is an important task in computer-aided diagnosis, enabling precise analysis of lesion attributes such as shape, size, and colour—key factors in melanoma assessment [1,2,21,22]. Clinical systems widely apply rule-based heuristics such as the ABCD rule [23]. Early methods for automated skin lesion diagnosis used handcrafted features [21,24] alongside classical techniques like color thresholding, region growing, and SVMs [25]. These methods did not generalise well because they were quite sensitive to contrast, hair artefacts, and background noise [26]. Furthermore, they generally

required heavy image pre-processing, such as hair removal and contrast enhancement [27], which could be both unreliable and inefficient.

Deep learning, especially Convolutional Neural Networks (CNNs), has substantially advanced the field since its introduction. For example, FCNs [28] introduced end-to-end pixel-level segmentation. U-Net [3] later became the standard model because of the simplicity and effectiveness of its encoder-decoder architecture and skip connections. Many variants followed the original implementation, especially in medical contexts. Yuan *et al.* [29] enabled full-image inference to improve the efficiency of UNet, while Feng *et al.* [30] added pyramid modules to better capture multi-scale context. Li *et al.* [31] introduced Dense Deconvolutional Networks to improve upsampling fidelity, achieving more precise modeling results. These models have performed well on ISIC [19] and PH2 [20], and some have reached dermatologist-level accuracy [32]. Although UNet is effective and popular, it still suffers from data shortages and vulnerability in local predictions. Some other studies tried hybrid approaches. For example, Bissoto *et al.* [33] focused on dataset curation and showed that training only on challenge-specific data gives more reliable results than including heterogeneous datasets. Ali *et al.* [34] combined CNNs with Gaussian Mixture Models. Qi *et al.* [35] proposed CQENet, using confidence estimation and statistical attention to better delineate boundaries. More recent lesion-specific hybrids include enhanced U-Net variants such as EIU-Net [36], attention-based networks like ACCPG-Net [37], transformer-guided designs [38], and boundary-aware CNN–Transformer variants [13,14,39]. These methods improve contour quality, but robustness and annotation efficiency remain challenges in routine clinical settings.

Despite these architectural improvements, CNN-based methods remain data-hungry and typically require extensive manual annotations to generalize well. Their effectiveness typically scales with the quantity and quality of pixel-level annotations, which are expensive and time-consuming to acquire in medical settings [4]. In contrast, our method focuses on few-shot segmentation by distilling knowledge from a foundation model, which is a novel research direction in this area.

2.2. Large foundation models in medical image analysis and medical image segmentation

The availability of large-scale datasets has enabled the rise of foundation models, particularly Transformer-based architectures [40], which scale performance with model size and data volume. Initially developed for natural language tasks powered by models like GPT-3 [41], Gemini [42], and LLaMA [43], Transformers have since been extended to vision through the Vision Transformer (ViT) [44], which models long-range dependencies by treating image patches as token sequences. ViT and its variants have since shown outstanding capabilities, achieving highly accurate visual modeling results in classification, detection and segmentation.

To take advantage of Transformers, researchers have also applied this architecture to medical imaging tasks, including disease diagnosis and organ segmentation [45]. For example, hybrid models like TransUNet [46] combine CNN-based encoders with Transformer decoders to address the inductive biases that ViT lacks (*e.g.,* locality, translation invariance). However, adapting these models to a specific clinical domain still usually takes large amounts of labelled data and extensive fine-tuning. Domain shifts, hallucination, and poor explainability also make them difficult to deploy in clinical practice [47].

For segmentation, SAM [15] offers a powerful foundation model trained on over a billion masks on

generic visual data. It uses ViT to encode images and can perform impressive zero-shot segmentation based on different types of prompts, such as points, boxes, and masks. In medical imaging, however, it requires either high-quality prompts or rich fine-tuning data to perform favorably due to a lack of domain-specific priors. More specifically, early evaluations found that vanilla SAM struggles to capture fine-grained medical boundaries without domain adaptation [48,49]. Several methods have tried to address these limitations by introducing an extra fine-tuning stage. MedSAM [16], for example, fine-tunes SAM on over 20 medical datasets with bounding box prompts. This improves accuracy, but both manual labels and accurate prompts are still required. AutoSAM [50] replaces prompts with a learned mask prediction network, while SAM-Med2D [17] uses lightweight adapters for interactive refinement. More recently, Medical SAM Adapter [51] has improved transferability to some extent. Even so, these approaches need extra tuning and data, which can make them less practical in low-resource clinical settings.

Our approach takes an alternative route. We follow a few-shot learning setting to avoid the need for excessive manually specified prompts and propose to distill lesion-relevant knowledge from MedSAM into a compact UNet. We do not modify or fine-tune MedSAM itself. Instead, we use selective feature distillation to draw on its medical visual priors, allowing data-efficient segmentation that could be practical for clinical use without relying on handcrafted prompts.

2.3. Data-efficient segmentation in skin lesion segmentation

A major challenge in medical image segmentation is the scarcity of annotated data, as pixel-wise labelling is labor-intensive and requires expert input [52]. Few-shot learning (FSL) addresses this by enabling generalisation from just a few labelled samples [9], making it particularly valuable in dermatology.

Most few-shot segmentation methods fall into two categories: prototype-based learning and meta-learning. Prototype methods, such as CGNet [53] and Siamese architectures [54], use feature similarity between support and query images but often struggle with lesion-specific challenges like low contrast and boundary ambiguity. Meta-learning approaches like CRNet [9] and iMAML [55] simulate episodic tasks to improve adaptability, while data-efficient architectures for small-scale medical datasets (e.g., PATrans [56]) and recent hierarchical prototype learning in Biomedical Signal Processing and Control [57] further highlight the need for robust boundary modelling under limited annotation budgets. Related object detection and tracking studies have also explored pixel-level fully convolutional tracking, recursive and condensed context modelling, recurrent Transformer decoding, and general-specific embeddings for few-shot detection [58–62]. Related progress in few-shot 3D point-cloud semantic segmentation provides more perspective on protocol design and the use of complementary priors. An *et al.* [63] identify foreground-leakage and sparse-sampling artefacts in established benchmarks, introduce a standardised evaluation setting, and optimise support–query correlations through COSeg. Their later MM-FSS method [64] combines point clouds with textual category labels and available 2D image information. Their work illustrates how multimodal cues can help a model generalise to novel classes. In a more generalised setting, GFS-VL [65] integrates sparse but accurate few-shot annotations with dense but noisy pseudo-labels from a 3D vision–language model, while still retaining base-class segmentation. These studies focus on episodic 3D scene segmentation, not fixed-subset 2D dermoscopic segmentation, so they are not directly comparable to PADS experimentally. Even so, they provide three useful lessons for our

setting: (1) evaluation protocols need to be clearly stated, (2) pretrained or multimodal knowledge can help when labels are sparse, and (3) noisy auxiliary guidance should be filtered based on its reliability. PADS follows these principles, using controlled prompt perturbations and MePoDIN-weighted MedSAM distillation to transfer reliable lesion information into a compact student for medical image segmentation. Recently, researchers have also attempted to adapt SAM to low-data settings. For example, PerSAM and PerSAM-F [66] use prompt tuning and lightweight finetuning on SAM, with promising results on natural images. However, their performance drops obviously on medical images, where lesion boundaries are subtle and annotations are scarce [67]. Instead, AutoSAM tries to learn prompts or reduce the need for them, but still requires heavy finetuning. For more information, Pachetti *et al.* [68] give a comprehensive review of FSL in medical imaging. They discuss the main technical approaches and gaps that remain, including limited task diversity, a lack of methodological standardization, and poor robustness in real-world deployment. For segmentation task, in particular, methods like RobustEMD [69] have been developed to address these limitations by improving domain robustness across modalities and institutions. These efforts still point to the need for prompt or feature learning strategies that suit the task and domain.

Recent foundation-model-based methods have also tackled low-data segmentation through automatic prompting, parameter-efficient adaptation, and reference-conditioned inference. ProtoSAM [70], for example, uses DINOv2 features for prototype-based coarse localisation, then refines point and box prompts with SAM for automated one-shot medical segmentation. SAMed [71] adapts SAM to medical semantic segmentation by applying LoRA-based parameter-efficient fine-tuning to the image encoder and optimising the prompt encoder and mask decoder. More recently, SANSA [72] uses SAM2's semantic representations for few-shot segmentation, with annotated reference examples provided as masks, boxes, points, or scribbles. These recent methods are relevant alternatives to our distillation framework. However, they use labelled examples in different ways, such as for model optimisation, support-conditioned inference, or automatic prompt generation. The term "k-shot" does not mean exactly the same procedure in all these methods. PADS trains its student for the task using a subset of k labelled images, and needs no labelled support examples during normal inference. Once trained, the student either predicts directly or produces a pseudo-prompt for optional MedSAM refinement. By contrast, ProtoSAM and SANSA still need labelled support or reference examples during inference, while SAMed uses the labelled subset to adapt model parameters. We therefore describe SANSA as using target-domain parameter-training-free, support-conditioned inference. It performs no ISIC- or PH2-specific parameter optimisation, but still receives labelled support masks during inference. Its reported standard deviation reflects how sensitive it is to the choice of support set. Earlier methods such as OSLSM [8] use episodic support–query meta-learning. These methods pursue related low-data segmentation goals, but use different training and inference paradigms.

Other recent studies have looked at limited supervision in related domains. AdaptSAM [73] adapts SAM for cross-domain few-shot medical image segmentation, aiming to improve transfer across heterogeneous medical imaging domains. This makes it the most directly related of the reviewer-suggested works. Meta-Transfer Derm-Diagnosis [74] studies episodic learning, supervised transfer learning, and self-supervised pretraining for long-tailed few-shot skin-disease classification. Its task is classification rather than pixel-wise segmentation, but it shows why transfer strategy and class imbalance matter when

labelled dermatological data are scarce. UniMF [75] introduces a unified multimodal framework for industrial anomaly detection, trained with only a few normal samples. Both its industrial multimodal setting and its anomaly-detection objective are quite different from dermoscopic lesion segmentation. Still, its use of shared representations with very limited normal data shows that data-efficient representation learning is useful beyond our task. We discuss these studies as complementary evidence from cross-domain medical segmentation, long-tailed dermatological classification, and industrial anomaly detection, but the latter two are not directly comparable segmentation baselines.

Knowledge distillation offers another option. Xiao *et al.* [67] propose region-aware distillation to suppress irrelevant background and report better efficiency than traditional matching-based FSL. Flamingo [76], a vision-language model, shows the potential of multimodal few-shot transfer, although its large scale limits clinical deployment. Even with these advances, most methods still need substantial computation or data, or depend too heavily on handcrafted prompts [45,77]. Our method reduces the need for manual prompt design by transferring robust lesion priors learned under prompt uncertainty into a compact student model. We also support "Bi-Coop." inference: the UNet provides a pseudo-prompt to guide MedSAM, and the two probability maps are averaged to obtain the final prediction.

3. Methodology

3.1. Overview of the PADS framework

We propose PADS, a framework that reduces sensitivity to prompt quality by transferring lesion-relevant representations from a frozen MedSAM teacher into a compact UNet using only a few labelled samples. MedSAM is the teacher instantiated in all PADS training and refinement experiments; SAM refers only to the broader architecture or model family. The framework identifies MedSAM feature channels that remain reliable when the prompt is imperfect. Figure 2 provides a high-level overview.

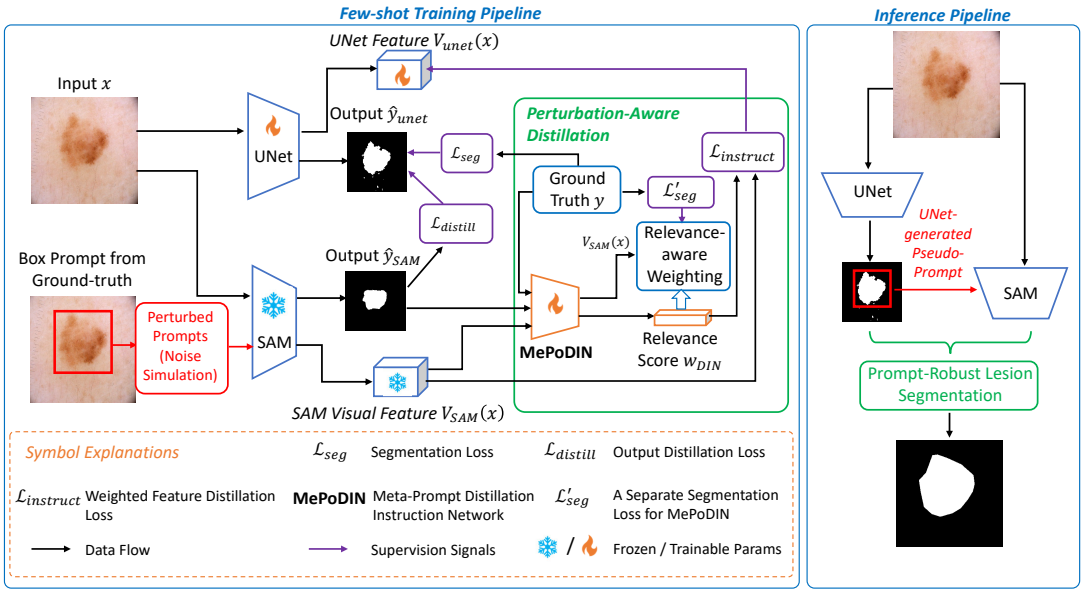

Figure 2. Architecture and workflow of PADS. left: few-shot training with perturbation-aware, relevance-weighted distillation from frozen MedSAM to UNet. right: Bi-Coop. inference, where the UNet supplies a pseudo-prompt to MedSAM and the two probability maps are averaged for the final prediction.

Technically, PADS converts prompt uncertainty into a feature-transfer control signal. During training, controlled translations and scale changes are applied to a ground-truth-derived box while the MedSAM teacher remains frozen. The resulting prompt-conditioned MedSAM predictions, together with the MedSAM feature map and lesion annotation, are provided to MePoDIN to estimate channel-wise reliability. These scores then weight the teacher–student feature-distillation loss. Consequently, ordinary perturbation training would use the noisy-prompt cases only as additional robustness examples, whereas PADS uses the MedSAM responses to determine how strongly each teacher channel should supervise the student.

As shown in the figure, the framework has four operational stages: (1) the student UNet predicts an initial lesion mask; (2) during training, a bounding box derived from the ground-truth mask is perturbed and used to guide MedSAM so that the teacher is observed under realistic prompt noise; (3) MePoDIN estimates the relevance of MedSAM feature channels with respect to lesion structures under these prompt variations; and (4) relevance-weighted distillation, at both mask and feature levels, transfers robust lesion information to the student model. Perturbation generation, reliability estimation, and weighted feature transfer thus form a coupled learning mechanism rather than independent augmentation, attention, and distillation components.

Formally, given a few-shot training set $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^{H\times W\times 3}$ and $y_i \in \{0,1\}^{H\times W}$, the goal is to learn a segmentation function $f : \mathbb{R}^{H\times W\times 3} \to \{0,1\}^{H\times W}$ that generalizes well to unseen lesion types. The instantiated teacher is MedSAM, denoted as $\mathscr{F}_{\text{MedSAM}}$, with visual encoder V_{MedSAM} and prompt encoder P_{MedSAM}. Given an image x and prompt p, MedSAM produces a segmentation mask via:

$$\hat{y}_{\text{MedSAM}} = \mathscr{F}_{\text{MedSAM}}(x,p) = V_{\text{MedSAM}}(x) \otimes P_{\text{MedSAM}}(p)$$

where $\otimes$ denotes a fusion operation. MedSAM is adapted from the SAM architecture for medical image segmentation and remains dependent on the quality of its bounding-box prompt [16]. Prompt-reducing SAM-family variants [78] still require considerable fine-tuning and supervision.

PADS uses MedSAM's medical visual priors without modifying the teacher. A student model $\mathscr{F}_{\text{unet}}$, guided by MePoDIN, learns to internalize lesion-relevant knowledge through relevance-weighted feature distillation. After training, this student can both generate segmentation masks and supply pseudo-prompts to MedSAM for enhanced refinement. The pseudo-prompt-guided MedSAM prediction used in Bi-Coop. is:

$$\hat{y}^{\text{pseudo}}_{\text{MedSAM}} = \mathscr{F}_{\text{MedSAM}}\Big(x, \mathscr{B}(\mathscr{F}_{\text{unet}}(x))\Big), \tag{1}$$

where $\mathscr{B}(\cdot)$ extracts a bounding box from the UNet's predicted mask. This Equation describes the MedSAM branch rather than the final ensemble; the output-fusion operation is described in Section 3.5. The rest of the section details the architecture and training process of PADS. In the following, we describe in detail how PADS is trained.

3.2. UNet student model

To enable lightweight segmentation without requiring precise prompts in low-data regimes, we adopt a compact UNet as the student network $\mathscr{F}_{\text{unet}}$. It is trained to predict lesion masks using a few labelled samples, while its intermediate features are aligned with MedSAM's via our perturbation-aware relevance-guided

distillation strategy. Given an input image x, the UNet produces a segmentation prediction:

$$\hat{y}_{\text{unet}} = \mathscr{F}_{\text{unet}}(x;\theta_s), \tag{2}$$

where θ_s are learnable parameters of UNet. For feature-level distillation, we extract a deep semantic representation $V_{\text{unet}}(x)$ from the deepest encoder layer:

$$V_{\text{unet}}(x) = \text{Interp}\left(\text{Conv}_{1\times1}\left(\text{Enc}^{(5)}_{\text{unet}}(x)\right)\right), \tag{3}$$

where $\text{Enc}^{(5)}_{\text{unet}}$ is the deepest encoder output, $\text{Conv}_{1\times1}$ is a projection layer to match MedSAM's channel dimension, and $\text{Interp}(\cdot)$ performs bilinear interpolation to align spatial size with $V_{\text{MedSAM}}(x)$. In practice, the UNet consists of 5 encoder-decoder levels adapted for 320×320 input images. Each encoder uses a convolutional block followed by downsampling, and each decoder uses upsampling and convolution. Group normalisation [79] is applied after each block to improve stability under small-batch training. This architecture could balance expressiveness and efficiency. The deep encoder captures high-level lesion semantics, while the projected feature ensures compatibility with MedSAM for effective cross-model distillation.

3.3. MePoDIN feature relevance module

To reduce reliance on manual prompt specification while still using MedSAM's medical visual priors, we introduce MePoDIN, a lightweight feature relevance module. MePoDIN explicitly models prompt variability during training and estimates the reliability of MedSAM feature channels under different prompt conditions. This allows the framework to selectively distill stable, lesion-relevant representations while suppressing prompt-sensitive activations. Here, "meta-prompt" refers to assessing prompt-conditioned teacher reliability to guide distillation, rather than an explicit meta-learning procedure.

MePoDIN takes three inputs: the MedSAM feature map $V_{\text{MedSAM}}(x)$, the MedSAM-predicted mask $\hat{y}_{\text{MedSAM}}$, and the ground-truth mask y. All mask inputs are spatially downsampled (e.g., 64×64) and encoded using lightweight CNNs with group normalisation [79] and ReLU activation. Two parallel branches, V_{DIN} and P_{DIN}, are designed to estimate feature relevance and prompt reliability, respectively.

The overall per-channel relevance weights $w_{\text{DIN}} \in \mathbb{R}^C$ are computed as

$$w_{\text{DIN}} = \tilde{w}_V \cdot \tilde{w}_p, \tag{4}$$

where $\tilde{w}_V$ measures channel-wise feature alignment with the lesion structure, and $\tilde{w}_p$ reflects the reliability of the prompt under which the MedSAM prediction is generated.

Feature relevance estimation. To evaluate how strongly each MedSAM feature channel contributes to lesion segmentation, we compute a feature importance vector $\tilde{w}_V \in \mathbb{R}^C$ as

$$\tilde{w}_V = V_{\text{DIN}}(V_{\text{MedSAM}}(x), y) = \text{MLP}\Big(\text{CNN}(V_{\text{MedSAM}}(x) \oplus y)\Big), \tag{5}$$

where $\oplus$ denotes channel-wise concatenation. This branch learns to measure semantic alignment between individual feature channels and the ground-truth lesion mask, producing per-channel relevance scores.

Prompt reliability estimation. In practice, prompts can vary quite a bit in quality. To simulate this during training, we generate a reference bounding box $\mathscr{B}(y)$ from the ground-truth mask and jitter its centre

location and side lengths using bounded random offsets. This gives us prompts ranging from nearly correct to moderately imprecise, similar to the coarse initialisation that can result from imperfect user input or automatic localisation. We then pass these perturbed prompts to MedSAM to obtain $\hat{y}_{\text{MedSAM}}$. Since MedSAM's prediction depends on the prompt, differences between $\hat{y}_{\text{MedSAM}}$ and y give an indication of prompt reliability. We use this information to estimate a scalar prompt relevance score $\tilde{w}_p \in \mathbb{R}^1$ via

$$\tilde{w}_p = P_{\text{DIN}}(\hat{y}_{\text{MedSAM}}, y) = \text{MLP}\Big(\text{CNN}(\hat{y}_{\text{MedSAM}} \oplus y)\Big), \tag{6}$$

where this branch uses its own parameters, separate from those of the feature branch. A higher $\tilde{w}_p$ means that the prompt produces a prediction that agrees well with the ground truth. This suggests that the corresponding MedSAM features are more reliable for distillation.

MePoDIN integrates channel-wise feature alignment with prompt reliability to give more weight to stable lesion-related representations and less weight to features that are sensitive to prompt perturbations. In other words, it learns not only whether a MedSAM channel is active, but also whether that channel remains useful when the prompt is less accurate. With these perturbation-aware weights, the student can learn representations that are robust to prompt variation rather than rely on a specific prompt configuration. In addition, MePoDIN uses low-resolution features and compact CNN encoders, so it stays lightweight and computationally efficient. This makes it suitable for few-shot training and settings where resources are limited.

MePoDIN is therefore distinct from generic attention mechanisms. SENet [80] generates channel weights from globally aggregated feature statistics, while CBAM [81] derives sequential channel and spatial attention from feature activations. In contrast, MePoDIN jointly uses the teacher feature map, the prompt-conditioned teacher mask, and the lesion ground truth. Its weights consequently estimate task relevance and prompt reliability for cross-model distillation rather than general activation importance. MePoDIN is not presented as a new form of attention in isolation. Its contribution is to use lesion- and prompt-conditioned reliability to regulate which teacher features are transferred to the student.

3.4. Overall training losses

To train PADS effectively under few-shot supervision, we design a composite loss that integrates three components: (1) supervised segmentation loss, (2) prediction-level distillation loss, and (3) feature-level distillation modulated by MePoDIN.

(1) Segmentation Loss—We use binary cross-entropy (BCE) to directly supervise the student UNet's predictions:

$$\mathscr{L}_{\text{seg}} = \text{BCE}(\hat{y}_{\text{unet}}, y), \tag{7}$$

where $\hat{y}_{\text{unet}} = \mathscr{F}_{\text{unet}}(x)$ and y is the ground truth. This loss ensures that the student learns to directly predict accurate lesion boundaries.

(2) Output Distillation Loss—To encourage the student to mimic MedSAM's confidence map, we minimize the mean squared error between the sigmoid-normalized logits of both models:

$$\mathscr{L}_{\text{distill}} = \text{MSE}\left(\sigma(\hat{y}_{\text{MedSAM}}),\ \sigma(\hat{y}_{\text{unet}})\right), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This guides the student using MedSAM's soft outputs, which

encode richer spatial cues than hard labels. This loss facilitates soft-label supervision by encouraging the student model to approximate the teacher's output, which implicitly encodes finer information.

(3) Instructed Feature Distillation Loss—We align the deep features of the student and teacher by computing a KL-divergence loss over each feature channel, weighted by MePoDIN's relevance scores:

$$\mathscr{L}_{\text{instruct}} = \sum_{c=1}^{C} (w_{\text{DIN}}^{(c)})^{\gamma} \cdot \text{KL}\left(V_{\text{MedSAM}}^{(c)}(x) \,\|\, V_{\text{unet}}^{(c)}(x)\right), \tag{9}$$

where $w_{DIN}^{(c)} \in w_{DIN}$ is the importance score for the channel from MePoDIN, and γ controls weighting sharpness. Higher γ values amplify the contribution of informative channels, while smaller values distribute focus more uniformly. We find $\gamma = 2$ offers a good trade-off in practice. We will discuss this in experiments.

Overall Loss—The overall training objective is:

$$\mathscr{L}_{\text{PADS}} = \mathscr{L}_{\text{seg}} + \lambda_1 \mathscr{L}_{\text{distill}} + \lambda_2 \mathscr{L}_{\text{instruct}}, \tag{10}$$

with λ_1 and λ_2 controlling the relative importance of distillation signals. This setup ensures that the student learns both from ground truth and the internal and external knowledge of the teacher, modulated by MePoDIN. The choice of λ_1 and λ_2 will be discussed in the following Experiments section.

MePoDIN Loss—To learn meaningful relevance weights w_{DIN}, MePoDIN is trained using an auxiliary segmentation objective over the MedSAM feature map:

$$\mathscr{L}'_{\text{seg}} = \text{BCE}\left(\text{Conv}_{1\times1}(w_{\text{DIN}} \odot V_{\text{MedSAM}}(x)),\, y\right), \tag{11}$$

where $\odot$ denotes channel-wise multiplication and $\text{Conv}_{1\times1}$ predicts a mask from the reweighted MedSAM features. Although MePoDIN receives the ground truth y during training, the learned vector w_{DIN} is low-dimensional and incapable of directly encoding spatial mask of y, preventing label leakage. This loss allows MePoDIN to identify and emphasize channels most relevant to lesion segmentation. We optimize MePoDIN separately from the student UNet to prevent gradient interference. Since MePoDIN focuses on evaluating feature relevance independent of the student's architecture or training dynamics, separate optimisation ensures better stability and modularity.

3.5. Bidirectional cooperation during inference

During inference, the trained UNet first produces a lesion probability map and may operate independently when lightweight deployment is preferred. In Bi-Coop. mode, its predicted mask is converted into the bounding-box pseudo-prompt used in Equation (1), allowing MedSAM to generate a complementary lesion probability map. The two probability maps are then averaged pixel-wise, and the averaged map is binarized using the same decision rule adopted for evaluation to obtain the final segmentation.

The term "bidirectional cooperation" therefore denotes two complementary contributions: the UNet guides MedSAM through the pseudo-prompt, while both the UNet and MedSAM predictions contribute to the final ensemble. It does not imply iterative feedback, parameter updating, or propagation of the MedSAM output back through the UNet. This design supports both lightweight UNet-only inference and the higher-accuracy Bi-Coop. mode evaluated in the following ablations.

4. Experiments

4.1. Experimental setup

Implementation Details. All experiments were conducted using PyTorch on a single NVIDIA RTX 4090 GPU. For the three independent trials, we used random seeds 42, 1024 and 2025. In each trial, the same seed controlled labelled-subset sampling and stochastic training and was applied to Python, NumPy, and PyTorch; deterministic execution was enabled where applicable. Models were trained for up to 100 epochs with early stopping based on validation Dice score. The model checkpoint with the best validation performance was selected and evaluated once on the official test set. All test results reported in the main comparison tables correspond to this fixed checkpoint and are not used for hyperparameter tuning. Unless otherwise specified, hyperparameters are set to $\gamma = 2.0$, $\lambda_1 = 0.5$, and $\lambda_2 = 1.0$ (as determined from validation experiments), and bidirectional cooperation (Section 3.5.) is used for evaluation. For the fully supervised UNet–PADS comparison, the applicable model-selection settings are matched: both use the fixed validation split, a maximum of 100 epochs, early stopping based on validation Dice, selection of the best validation checkpoint, and evaluation on the same fixed test set.

Dataset. We evaluate PADS on two public dermoscopic datasets. ISIC 2018 (Task 1) [19] provides 2594 labelled images in the official training partition and 1000 images in the official test partition. Following our protocol, 100 images are held out from the official training partition for validation, and few-shot subsets are sampled from the remaining training images. We focus exclusively on Task 1 (lesion boundary segmentation). PH2 [20] contains 200 high-resolution dermoscopic images annotated by dermatologists. PH2 is used only for cross-dataset generalisation evaluation without fine-tuning. ISIC 2018 is an established public benchmark for skin-lesion segmentation because it includes substantial variation in lesion appearance, size, colour, contrast, boundary ambiguity, and imaging artefacts, while its official partition supports reproducible comparisons. PH2 provides a complementary evaluation using an independent dataset and acquisition source; direct testing without PH2-specific fine-tuning therefore provides evidence of cross-dataset generalisation. Nevertheless, ISIC 2018 and PH2 are curated public dermoscopic datasets and cannot fully represent the diversity of clinical institutions, imaging devices, patient populations, skin tones, or real-world deployment conditions. Broader multi-centre clinical validation therefore remains necessary. The split is fixed across all experiments, and validation data are used exclusively for hyperparameter selection and ablation.

Preprocessing. All images are resized to 320×320 pixels. We apply standard normalisation and data augmentation including: random horizontal/vertical flipping, random rotation ($\pm 15^\circ$), random scaling and cropping, elastic deformation, color jitter, and Gaussian blur. These augmentations are applied consistently across all compared methods. In particular, the fully supervised UNet and PADS use the same input resolution, normalisation, augmentation pipeline, validation split, test preprocessing, and evaluation metrics.

Few-Shot Setting and Sampling Protocol. In this work, few-shot segmentation refers to task-specific model training using $k \in \{5, 10, 20\}$ labelled dermoscopic images randomly selected from the ISIC 2018 training partition. For each k, we use three independently sampled subsets and random-seed runs (42, 1024 and 2025) unless otherwise stated, and report trainable-model results as the mean and standard deviation. PADS uses each subset for supervised segmentation and distillation. After training,

inference requires no labelled support images because the student predicts directly or generates its own pseudo-prompt for optional MedSAM refinement. This fixed-subset paradigm differs from episodic support–query meta-learning, such as OSLSM [8], and from methods that consume labelled support examples during inference. Its annotation budgets are aligned with low-data settings reported for methods such as iMAML [55], although PADS itself is not an episodic meta-learning method. Across all trials, the validation and test sets remain fixed, the checkpoint selected according to validation performance is evaluated on the test set, and PH2 is used only for direct cross-domain evaluation without fine-tuning. The standard deviations reported in the tables quantify sensitivity to the particular few-shot samples and random training splits. Smaller standard deviations therefore indicate more stable learning under limited supervision, whereas larger values are expected in the 5-shot setting because each sampled image has a stronger influence on the learned lesion representation. Standard deviations are not applicable to every comparison entry. Fixed-prompt zero-shot vanilla SAM and MedSAM are deterministic under the specified model, image preprocessing, and fixed centre-box prompt; because these baselines involve no training, random support sampling, or stochastic prompt generation, there is no across-run variability to report. For earlier comparison methods evaluated under sufficiently comparable settings, we reproduce the values reported in the corresponding papers and do not add a standard deviation when the original publication provides only a point estimate. By contrast, the recent SAM-based methods marked † are reproduced using their released implementations, and their standard deviations are reported under the sampling procedures described above.

Training Strategy The student UNet is trained using the composite loss defined in Equation (10), which integrates: Supervised segmentation loss ($\mathscr{L}_{seg}$); Output-level distillation loss ($\mathscr{L}_{distill}$); MePoDIN-guided feature distillation loss ($\mathscr{L}_{instruct}$). MePoDIN is optimized separately using the auxiliary loss in Equation (11) to avoid gradient interference with the student network. We use the Adam optimizer with an initial learning rate of 5×10^{-3}. Learning rate decay is applied if validation performance plateaus. The supervision and objective are necessarily method-specific. The fully supervised UNet is trained only with the segmentation loss using the full labelled training pool, whereas PADS uses a k-image subset and additionally applies output-level and MePoDIN-weighted feature distillation. Thus, all applicable data processing, model-selection, and evaluation settings are matched, but the number of training images and loss functions are intentionally different. The Adam configuration and learning-rate schedule stated above describe PADS; we do not claim that method-specific optimisation hyperparameters are identical unless explicitly reported.

Prompt Perturbation During Training. Bounding-box prompts are generated from ground-truth masks and then perturbed using random translation and scaling. This makes the training condition explicit: the teacher always receives GT-derived boxes, but these boxes are intentionally degraded to simulate realistic prompt noise. As a result, MePoDIN learns prompt-quality-aware relevance weighting from controlled perturbations rather than uncontrolled prompt errors.

Computational Efficiency. We report inference time measured on RTX 4090: UNet inference: ~ 6 ms per image; MedSAM inference: ~ 72 ms per image; Full PADS pipeline: ~ 78 ms per image.

Inference. During deployment, the trained UNet is the default model and can be used independently for lightweight inference. When higher accuracy is needed, the UNet prediction is converted into a pseudo-prompt for optional MedSAM refinement.

Evaluation Metrics. We evaluate segmentation performance using: Dice Similarity Coefficient (DSC): $\text{DSC} = \frac{2|P \cap G|}{|P|+|G|}$; Intersection over Union (IoU): $\text{IoU} = \frac{|P \cap G|}{|P \cup G|}$, where P denotes the predicted mask and G the ground truth. Dice is used as the primary metric due to its robustness to class imbalance, which is common in lesion segmentation [8,55,82]. IoU is reported as a complementary measure for completeness.

4.2. Compared methods

Table 1 compares PADS with methods covering several low-data segmentation paradigms. The fully supervised UNet provides an upper-reference model trained with the complete ISIC training set. Vanilla SAM and MedSAM are evaluated without target-domain training using a fixed centre-box prompt. The established few-shot and data-efficient methods include prior-mask-guided Few-shot Segmentation [67], OSLSM [8], CANet [83], PANet [84], CRNet [9], the semi-supervised method of Feyjie *et al.* [82], and the implicit-gradient meta-learning method iMAML [55]. The recent foundation-model comparisons comprise ProtoSAM [70], which performs one-shot support-conditioned automatic prompting, SAMed [71], which applies LoRA-based SAM adaptation, and SANSA [72], which performs target-domain parameter-training-free, support-conditioned inference with SAM2. ProtoSAM, SAMed, and SANSA were not originally developed for this task. We adapted their publicly released implementations to skin-lesion segmentation and reproduced the reported comparison results under the protocols described above. Lastly, our PADS uses fixed-k training to distil prompt-robust knowledge from a frozen MedSAM teacher into a compact UNet student.

For methods whose published ISIC results use a sufficiently comparable evaluation setting, including iMAML and the earlier few-shot baselines, we report the values provided in the corresponding publications. For trainable methods reproduced under our protocol, we follow the same sampling strategy used for PADS: models are trained on matched labelled subsets where applicable, and mean and standard deviation are calculated across independent data splits and random seeds. For support-conditioned methods without target-domain parameter training, including ProtoSAM and SANSA, no parameter optimisation is performed on ISIC or PH2 during evaluation, labelled support examples are nevertheless supplied at inference, and mean and standard deviation are calculated across independently sampled support splits. For SANSA specifically, these support examples are supplied as labelled masks, so its deviation measures sensitivity to support-set selection rather than training-seed variation. Dice and IoU are calculated per image and averaged across the fixed test set. PH2 is used only for direct cross-domain evaluation without PH2-specific training, fine-tuning, support selection, or hyperparameter tuning.

4.3. Training Strategy

We jointly train the student UNet with segmentation supervision and perturbation-aware distillation from MedSAM via MePoDIN. The full objective includes segmentation, prediction-level, and feature-level losses (Equation (10)). MePoDIN is trained independently to learn a set of relevance weights w, which, when applied to MedSAM features, enable segmentation through a shallow MLP head. In practice, the student is encouraged to follow teacher evidence that remains useful across perturbed prompts, rather than matching every teacher activation equally. We use the Adam optimizer with a learning rate of 5×10^{-3}. Data augmentations include rescaling, cropping, saturation, and rotation to mitigate overfitting on limited samples.

4.4. Evaluation on ISIC 2018

4.4.1. Overall comparison

To assess the effectiveness of PADS, we compare it against strong baselines on the ISIC 2018 test set, including a fully supervised UNet, zero-shot SAM variants, established few-shot methods, and our staged distillation variants. The best-performing model on the validation set is used for evaluation. Table 1 reports ISIC 2018 Dice scores together with each method's evaluation setting. The subsequent experiments then examine controlled prompt robustness, ablations of perturbation and feature weighting, and computational efficiency so that the method is evaluated as a learning framework rather than only as a pipeline. In Table 1, the fixed-centre-box vanilla SAM and MedSAM rows are deterministic point estimates and therefore have no standard deviation. Earlier methods shown without a deviation are reported directly from their publications, whereas the reproduced methods marked † and our trainable models include uncertainty estimates under their stated sampling protocols.

Table 1. Overall comparison of Dice scores on ISIC 2018 under the reported low-data evaluation settings. Methods marked † are reproduced by the authors using publicly released implementations. For SANSA, ‡ denotes target-domain parameter-training-free, support-conditioned inference. Bold values indicate the best performance.

Method	Training data	ISIC Dice
Fully supervised reference		
Fully-supervised UNet	2594 labelled images	0.858 ± 0.004
Zero-shot SAM variants (fixed centre box as prompt)		
Vanilla SAM	0	0.426
MedSAM	0	0.580
Few-shot/data-efficient methods		
Few-shot Segm [67]	5-shot	0.386
OSLSM [8]	5-shot	0.589
CANet [83]	5-shot	0.615
Semi-sup Few-shot Segm [82] (with extra data)	5-shot	0.613
	10-shot	0.614
	20-shot	0.608
PANet [84]	5-shot	0.627
CRNet [9]	5-shot	0.623
iMAML [55] (with extra data)	20-shot	0.832
Low-data SAM-based methods (reproduced)		
ProtoSAM† [70]	1-shot	0.440 ± 0.027
SANSA† [72]	20-shot‡	0.684 ± 0.070
SAMed† [71]	20-shot	0.798 ± 0.006
PADS (Ours)	5-shot	**0.774** ± 0.035
	10-shot	**0.805** ± 0.028
	20-shot	**0.849** ± 0.007

Comparison with Fully Supervised UNet. A conventional UNet trained on the full ISIC 2018 training partition achieves 0.858 $\pm$ 0.004 Dice on the test set. In comparison, with only 20 labelled images, PADS reaches a Dice score of 0.849. This is close to the fully supervised reference, despite using far fewer labelled samples. This result suggests that transferring robust lesion priors into a lightweight student through perturbation-aware distillation can alleviate the need for annotations.

Comparison with Zero-shot Vanilla SAM and MedSAM: To evaluate zero-shot vanilla SAM and MedSAM, we use centred bounding boxes (70% of the image size). These prompts are intentionally coarse because accurate prompts are unavailable at test time. MedSAM performs better than vanilla SAM with these rough prompts, but its DSC (0.580) is still well below PADS. The gap suggests that relying directly on prompt-based predictions is not reliable enough when lesion boundaries are complex.

Comparison with Few-shot and Data-efficient Approaches: PADS also compares favorably with few-shot and data-efficient approaches. With 5-shot training, PADS reaches a DSC of 0.774, higher than PANet (0.627), CRNet (0.623), and recent semi-supervised baselines [82]. The few-shot result indicates that PADS can extract useful lesion priors even when only a small labelled subset is available.

Competing with Meta-Learning: iMAML [55] benefits from episodic meta-training and auxiliary data, reaching a promising Dice score of 0.832 at 20-shot. PADS still performs better (0.849), without either meta-training procedures or additional external datasets. The comparison shows that the proposed distillation strategy remains competitive without episodic meta-training or external data.

Comparison with Recent Low-Shot SAM-Based Methods. Table 1 also includes three recent methods with available code, which we reproduced for dermoscopic segmentation. ProtoSAM achieves 0.440 $\pm$ 0.027 Dice with its one-shot support-conditioned protocol. The low Dice score shows the difficulty of transferring lesion correspondence from a single annotated example across highly variable dermoscopic images. SAMed is the strongest reproduced baseline, with 0.798 $\pm$ 0.006 Dice after LoRA-based adaptation on 20 labelled images. PADS reaches 0.849 $\pm$ 0.007 Dice at 20-shot, a gain of 0.051 in mean Dice, outperforming these reproduced baselines while keeping the foundation-model teacher frozen and using a compact student for deployment. SANSA reaches 0.684 $\pm$ 0.070 Dice with target-domain parameter-training-free, support-conditioned inference. It performs no ISIC-specific parameter optimisation, but still uses labelled support masks during inference. Its larger standard deviation suggests that the choice of support set has a considerable effect on its results. These methods do not all follow the same low-shot protocol. ProtoSAM and SANSA use labelled support examples during inference, SAMed adapts SAM parameters on ISIC, and PADS trains a prompt-free student on a fixed labelled subset. We therefore keep each method's core protocol, but use the same dermoscopic preprocessing, test partitions, thresholding, and metric implementations.

Summary. Overall, PADS comes close to the fully supervised UNet reference (0.858 $\pm$ 0.004 Dice) and gives the strongest result among the low-data methods evaluated on ISIC 2018. Our evaluation includes both early few-shot baselines and recent methods based on automatic prompting, parameter-efficient adaptation, and SAM2-based support-conditioned inference, while making their different learning protocols clear. The standard deviation also decreases from 5-shot to 20-shot, suggesting that PADS gives more repeatable results as more labelled images become available.

4.5. Ablation study

To better understand the design choices behind PADS and isolate the contribution of each component, we conduct six groups of ablations: (1) Loss Effects: the impact of segmentation supervision, output-level distillation, and MePoDIN-guided feature distillation in Equation (10); (2) Training Perturbation Analysis: the effect of removing bounding-box perturbation during training; (3) Prompting Strategy: the effect of different inference prompts, including fixed centre boxes, random boxes, and pseudo-prompts derived from the trained UNet; (4) Robustness to Prompt Perturbation: controlled performance degradation under increasing bounding-box noise; (5) MePoDIN *vs.* Conventional Attention Mechanisms: whether perturbation-aware relevance modelling offers benefits beyond generic attention mechanisms; and (6) Hyperparameter Effects: sensitivity to the channel sharpening factor γ in Equation (9) and the loss weights λ_1 and λ_2.

Unless otherwise specified, ablation experiments are conducted on the ISIC 2018 validation set rather than the test set. The validation set enables efficient experimentation while exhibiting trends consistent with the final test results. Where applicable, we primarily report the 20-shot setting because it provides more stable estimates than the lower-shot settings.

Throughout Section 4.5, we use five fixed definitions. Few-shot UNet uses segmentation supervision only. Standard Distillation adds output-level distillation but has no feature-level alignment. Unweighted Feature Distillation additionally aligns teacher and student features with equal channel weighting. PADS without Perturbation uses MePoDIN relevance weighting without training-time prompt perturbation. Full PADS uses MePoDIN weighting together with training-time prompt perturbation. All variants follow the same few-shot sampling protocol unless otherwise stated.

4.5.1. Ablation study on loss effects

We assess the contribution of each supervision component in Equation (10) through a staged ablation. We begin with supervised segmentation only ($\mathscr{L}_{seg}$), denoted as *Few-shot UNet*. We then add prediction-level distillation from MedSAM ($\mathscr{L}_{distill}$), referred to as *Standard Distillation*. Finally, we incorporate the proposed MePoDIN-guided feature distillation term ($\mathscr{L}_{instruct}$), forming *Full PADS*. To ensure reliable evaluation under few-shot conditions, where performance is sensitive to support sample selection, we conduct experiments across three independent random splits for each k-shot setting ($k = 5, 10, 20$) and report the mean and standard deviation.

Results are shown in Table 2. It shows that a UNet trained solely with $\mathscr{L}_{seg}$ struggles under limited supervision. For example, it achieves a Dice score of 0.587 ± 0.100 in the 5-shot setting and 0.704 ± 0.008 in the 20-shot setting, with relatively high standard deviation in low-shot regimes. Introducing output-level distillation improves performance consistently, raising Dice to 0.622 ± 0.040 (5-shot) and 0.729 ± 0.004 (20-shot). This indicates that soft supervision from MedSAM provides useful global priors.

Table 2. Loss-component ablation on the ISIC 2018 validation set across multiple random splits. Few-shot UNet uses segmentation supervision only; standard distillation adds output-level distillation; full PADS further uses MePoDIN-weighted feature distillation and training-time prompt perturbation. Bold values indicate the best performance.

Learning Schemes	Metric	5-shot	10-shot	20-shot
Few-shot UNet	Jaccard index (IoU)	0.424 ± 0.060	0.560 ± 0.009	0.590 ± 0.008
	Dice score	0.587 ± 0.100	0.668 ± 0.010	0.704 ± 0.008
Standard Distillation	Jaccard index (IoU)	0.496 ± 0.032	0.587 ± 0.021	0.640 ± 0.008
	Dice score	0.622 ± 0.041	0.679 ± 0.010	0.729 ± 0.004
Full PADS (Ours)	Jaccard index (IoU)	**0.665** ± 0.019	**0.694** ± 0.011	**0.764** ± 0.007
	Dice score	**0.755** ± 0.040	**0.792** ± 0.020	**0.856** ± 0.008

In contrast, Full PADS achieves the best results across all shot settings. Dice improves to 0.755 ± 0.040 (5-shot), 0.792 ± 0.020 (10-shot), and 0.856 ± 0.008 (20-shot), with corresponding gains in IoU. The performance gap widens as more supervision becomes available, suggesting that MePoDIN-guided feature alignment makes more effective use of limited data. Table 2 also shows that variability is greatest with only five labelled samples and decreases in the 20-shot setting.

The loss ablations demonstrate that segmentation supervision, output-level distillation, and relevance-weighted feature distillation play complementary roles.While output distillation transfers coarse structural knowledge from MedSAM, the proposed feature-level guidance selectively emphasizes lesion-relevant representations, leading to improved accuracy and reduced standard deviation under few-shot learning.

4.5.2. Effect of training perturbation

We further investigate the role of prompt perturbation during training. In this variant, we disable bounding-box perturbation and always use the ground-truth box to prompt MedSAM. All other components remain unchanged. Table 3 reports results under the 20-shot setting on the ISIC 2018 validation set.

Table 3. Feature-distillation and perturbation ablation on the ISIC 2018 validation set (20-shot). Standard Distillation has no feature alignment; unweighted Feature Distillation uses equal channel weights; PADS without Perturbation uses MePoDIN without prompt perturbation; full PADS includes both MePoDIN and prompt perturbation. Bold values indicate the best performance.

Method	IoU	Dice
Standard Distillation	0.640 ± 0.008	0.729 ± 0.004
Unweighted Feature Distillation	0.670 ± 0.007	0.799 ± 0.009
PADS without Perturbation	0.732 ± 0.006	0.812 ± 0.008
Full PADS	**0.764** ± 0.007	**0.856** ± 0.008

The results show a progressive improvement across the fixed variants. Unweighted Feature Distillation improves Dice from 0.729 for Standard Distillation to 0.799, demonstrating the benefit of feature alignment

beyond output supervision. PADS without Perturbation further increases Dice to 0.812 through MePoDIN relevance weighting. Full PADS reaches 0.856 Dice when training-time prompt perturbation is added, a 4.4-point gain over PADS without Perturbation. The gain indicates that controlled prompt variability helps MePoDIN identify stable lesion-relevant channels.

4.5.3. Prompting strategies

The fixed centre-box result is deterministic for the specified model and preprocessing and is therefore reported without a standard deviation. In contrast, the random-box baseline samples prompts independently, so its mean and standard deviation quantify variability across random prompt draws. The learned UNet and PADS prompt rows retain the uncertainty associated with their few-shot training runs.

To analyze the effect of prompting strategies, we evaluate both standalone UNet inference and MedSAM inference under different prompt sources, as summarized in Table 4.

Table 4. Ablation study on prompting strategies on the ISIC 2018 validation set (20-shot). Standard distillation uses output-level distillation without feature alignment, whereas full PADS includes MePoDIN-weighted feature distillation and training-time prompt perturbation. Bidirectional cooperation (Bi-Coop.) averages the full PADS student and pseudo-prompt-guided MedSAM probability maps. Bold values indicate the best performance.

Prompt Type	**IoU**	**Dice**
UNet Inference (No MedSAM)		
Standard Distillation	0.640 ± 0.008	0.729 ± 0.004
Full PADS (UNet-only)	0.750 ± 0.006	0.834 ± 0.008
MedSAM Inference		
Fixed centre box (70%)	0.427	0.552
Random Box (70%)	0.403 ± 0.005	0.509 ± 0.009
Pseudo-prompt from Standard Distillation	0.692 ± 0.009	0.776 ± 0.010
Pseudo-prompt from Full PADS	0.752 ± 0.008	0.845 ± 0.008
Full PADS + Bidirectional cooperation (Ours)	**0.764** ± 0.007	**0.856** ± 0.008

Effect of Representation Learning. Without using MedSAM at inference, the Full PADS student achieves 0.834 Dice, well above Standard Distillation (0.729 Dice). Relevance-aware feature distillation therefore improves the student even without prompt-based refinement. The UNet's strong standalone performance indicates that the main improvement comes from better representation learning, rather than reliance on the foundation model at inference.

Effect of Prompt Quality. MedSAM performs poorly with naive prompts. A fixed centre box yields 0.552 Dice, and a random box lowers this to 0.509 Dice. These generic spatial priors do not provide enough information to locate lesions reliably, and the results reflect MedSAM's sensitivity to prompt quality. Pseudo-prompts substantially improve performance: the pseudo-prompt from Standard Distillation raises the Dice score to 0.776, while the pseudo-prompt from Full PADS reaches 0.845. Relevance-aware training produces more spatially precise lesion predictions with less noise, accounting for this gain.

Effect of Bidirectional Cooperation. The Full PADS student achieves 0.834 Dice, and MedSAM guided by its pseudo-prompt achieves 0.845 Dice. Averaging their probability maps in Bi-Coop. increases Dice to 0.856. The two predictions complement each other, but the main gain still comes from the learned student representation.

The prompting ablations demonstrate that: (1) relevance-aware distillation (*MePoDIN*) significantly enhances the student model, (2) accurate pseudo-prompts support MedSAM refinement, and (3) Bi-Coop. provides a modest ensemble gain. Together, these components enable robust few-shot lesion segmentation without requiring precisely specified prompts at test time, with primary gains driven by improved representation learning in the student model.

Additional Comparison with Vanilla SAM and MedSAM. We also report results using both vanilla SAM and MedSAM prompted by our PADS-trained UNet. PADS with vanilla SAM refinement achieves a Dice score of 0.816, which is lower than PADS with MedSAM refinement (0.856 Dice). This comparison confirms that medical-domain adaptation improves segmentation quality in dermoscopic images. With a fixed centre-box prompt, vanilla SAM achieves approximately 0.51 Dice; using the PADS pseudo-prompt substantially improves its result. MedSAM remains the teacher and default refinement model in all other PADS experiments.

4.5.4. Robustness to prompt perturbation

A core motivation of PADS is to reduce the student's sensitivity to prompt quality by explicitly modelling prompt-induced feature relevance during training. To validate whether MePoDIN improves robustness rather than merely acting as standard feature distillation, we conduct a controlled prompt perturbation study. In particular, during inference, we progressively corrupt the bounding-box prompts supplied to MedSAM by adding spatial perturbations of increasing magnitude (0%, 10%, 20%, 30%, 50%). We emphasize that the perturbation study is conducted under a controlled experimental setting. Specifically, bounding boxes are derived from *ground-truth masks* and then artificially perturbed to simulate varying levels of prompt inaccuracy. This differs from the normal inference pipeline, where prompts are generated from the UNet prediction, and ground-truth annotations are not available.

The purpose of this design is to isolate and evaluate the sensitivity of different training strategies to prompt quality under identical and controlled perturbation magnitudes. By using ground-truth-derived boxes as a reference, we ensure that the noise level is precisely quantified and comparable across methods. This allows us to fairly assess robustness to prompt degradation without conflating it with upstream UNet prediction errors.

Three fixed variants are evaluated: (1) Few-shot UNet, trained with segmentation supervision only; (2) Standard Distillation, trained with segmentation and output-level distillation but no feature alignment; and (3) Full PADS, which additionally uses MePoDIN-weighted feature distillation and training-time prompt perturbation ($\gamma = 2$). For this controlled experiment, the GT-derived box serves only as the reference prompt and replaces the normal UNet-generated pseudo-prompt. At each noise level, the same perturbation is applied to Standard Distillation and Full PADS. MedSAM receives this externally controlled box, and Bi-Coop. averages its probability map with the corresponding student output. Few-shot UNet consumes no prompt and is therefore unchanged across perturbation levels. Its

prompt-independent Dice and IoU values are shown once using merged cells because the same fixed checkpoint is evaluated at every noise level. No prompt-induced standard deviation is reported; this does not imply zero variability across independently trained checkpoints. Dice and IoU are reported.

Results and Analysis. Table 5 is a controlled sensitivity test rather than a report of normal deployment performance. Its 0% condition uses an unperturbed GT-derived box, which is substantially more accurate than the automatically generated pseudo-prompt used in Table 4. This explains why the 0% values in Table 5 are higher than the corresponding results under standard inference in Table 4. The relevant comparison is therefore the degradation trend under the same increasing perturbation, rather than the absolute 0% value relative to normal inference.

Table 5. Controlled prompt perturbation study on the ISIC 2018 validation set (20-shot). A GT-derived reference box replaces the normal UNet-generated pseudo-prompt, and the same perturbation is applied to standard distillation and full PADS. Bi-Coop. averages the respective student output with the MedSAM output; few-shot UNet receives no prompt, so its prompt-independent results are displayed once using merged cells.

Noise (%)	Few-shot UNet		Standard Distillation		Full PADS (Ours)	
	Dice	IoU	Dice	IoU	Dice	IoU
0	0.786	0.686	0.925 ± 0.006	0.868 ± 0.004	0.942 ± 0.007	0.896 ± 0.008
10			0.896 ± 0.007	0.826 ± 0.006	0.918 ± 0.008	0.858 ± 0.009
20			0.850 ± 0.008	0.757 ± 0.007	0.874 ± 0.009	0.788 ± 0.009
30			0.793 ± 0.014	0.680 ± 0.010	0.820 ± 0.011	0.713 ± 0.010
50			0.660 ± 0.015	0.527 ± 0.012	0.700 ± 0.011	0.569 ± 0.011

(1) Standard Distillation is much more prompt-sensitive. When no perturbation is applied (0%), Standard Distillation achieves strong performance (Dice = 0.925). However, as perturbation increases to 50%, performance drops sharply to 0.660 (a decrease of 26.5 Dice points). This indicates that conventional distillation transfers representations that remain vulnerable to degraded prompt quality.

(2) Full PADS exhibits consistently improved robustness. Across all perturbation levels, Full PADS outperforms Standard Distillation. More importantly, the performance gap widens as noise increases. At 0% noise, the gain is modest (+1.7 Dice), while at 50% noise the gain increases to +4.0 Dice. This trend suggests that MePoDIN effectively suppresses prompt-sensitive feature channels and emphasizes lesion-relevant representations that remain stable under degraded prompting conditions. To further quantify robustness, we compute a relative retention ratio defined as Dice(50%) / Dice(0%). The retention ratio is 0.714 for Standard Distillation and 0.743 for Full PADS, further confirming that Full PADS degrades more gracefully under increasing prompt perturbation.

(3) Few-shot UNet is stable but limited. The Few-shot UNet remains constant across perturbation levels, confirming that it is independent of prompt quality. However, its overall performance ceiling (Dice = 0.7860) is substantially lower than the proposed approach under moderate perturbation levels. This highlights the advantage of selectively distilling MedSAM priors while mitigating prompt dependency.

In summary, this experiment directly validates the core hypothesis of PADS: perturbation-aware

relevance modelling is essential for reducing prompt sensitivity in foundation-model distillation. If MePoDIN were equivalent to standard channel attention or naive distillation, performance degradation trends would closely match Standard Distillation. Instead, Full PADS demonstrates consistently slower degradation under increasing perturbation, indicating that the learned relevance weights encode prompt robustness during training. Therefore, the proposed mechanism changes how teacher representations are transferred by selectively filtering prompt-sensitive noise.

4.5.5. Comparison with conventional attention mechanisms

The purpose of this experiment is to determine whether MePoDIN's benefit can be reproduced by generic attention mechanisms. We replace MePoDIN with SENet [80] or CBAM [81] while keeping all other components unchanged, including the UNet student, frozen MedSAM teacher, 20-shot training subset, training-time prompt perturbation, output-level and feature-level distillation losses, optimisation settings, validation set, and evaluation metrics. The only changed component is the feature-weighting mechanism.

Unweighted Feature Distillation assigns equal importance to all teacher feature channels. SENet generates channel weights from globally pooled teacher-feature statistics. CBAM applies conventional attention derived from feature activations in the feature-distillation pathway. MePoDIN estimates relevance using the MedSAM feature map, the prompt-conditioned MedSAM prediction, and the ground-truth lesion mask, thereby incorporating lesion relevance and prompt reliability during training. SENet and CBAM are attention-based weighting baselines rather than independent segmentation networks.

Table 6 shows that SENet improves over Unweighted Feature Distillation by 0.038 IoU and 0.005 Dice, while CBAM improves by 0.036 IoU and 0.002 Dice. Compared with the stronger attention baseline, SENet, Full PADS with MePoDIN further improves IoU by 0.056 and Dice by 0.052. Because the student, frozen teacher, training subset, prompt perturbations, distillation objectives, optimisation settings, and evaluation protocol are held fixed, the weighting module is the only controlled replacement in this experiment. The limited gains from SENet and CBAM indicate that the improvement cannot be attributed merely to inserting a generic attention mechanism. Instead, the substantially stronger result obtained by MePoDIN provides evidence that conditioning feature weights on lesion supervision and prompt reliability supports more effective teacher–student transfer than weighting based only on activation statistics. This controlled comparison supports, rather than by itself proves, the proposed methodological distinction.

Table 6. Comparison of feature-weighting mechanisms on the ISIC 2018 validation set under the 20-shot setting. Unweighted feature distillation uses equal channel weights, while full PADS uses MePoDIN weighting. All methods otherwise use the same student, frozen MedSAM teacher, prompt perturbation strategy, distillation losses, and training protocol. Bold values indicate the best performance.

Weighting Mechanism	**IoU**	**Dice**
Unweighted Feature Distillation	0.670 ± 0.007	0.799 ± 0.009
SENet	0.708 ± 0.003	0.804 ± 0.002
CBAM	0.706 ± 0.008	0.801 ± 0.006
Full PADS (MePoDIN, Ours)	**0.764** ± 0.007	**0.856** ± 0.008

4.5.6. Effects of hyperparameters

PADS includes several hyperparameters that influence performance: the exponent γ in Equation (9), which controls the sharpness of MePoDIN's relevance weighting, and the loss weights λ_1 and λ_2, which balance output-level and feature-level distillation, respectively. We analyze each component under the 20-shot setting on the ISIC 2018 validation set.

Effect of γ. Table 7 shows that performance improves steadily as γ increases from 0 to 2.0. When $\gamma = 0$, feature-level distillation is applied uniformly across channels (*i.e.*, without relevance weighting), yielding 0.799 Dice. Increasing γ progressively sharpens channel importance, allowing the model to emphasize lesion-relevant features while suppressing noisy activations. Performance peaks at $\gamma = 2.0$ (Dice 0.856, IoU 0.764), suggesting that moderate sharpening achieves the best balance between selectivity and stability. When γ is further increased to 3.0, performance slightly decreases (Dice 0.842), indicating that overly sharp weighting may overemphasize a subset of channels and reduce useful contextual information. These results confirm that relevance-aware weighting is beneficial and that $\gamma = 2.0$ provides an effective default configuration.

Table 7. Effect of γ (channel importance sharpening factor in Equation (9)) on segmentation performance. The $\gamma = 0$ setting is unweighted feature distillation, whereas $\gamma = 2$ is the full PADS default. Bold values indicate the best performance.

γ	0	0.5	1.0	2.0 (default)	3.0
IoU	0.670 ± 0.007	0.710 ± 0.006	0.735 ± 0.008	**0.764** ± 0.007	0.744 ± 0.008
Dice	0.799 ± 0.009	0.814 ± 0.007	0.839 ± 0.009	**0.856** ± 0.008	0.842 ± 0.008

Clarification of the $\gamma = 0$ Setting. The $\gamma = 0$ configuration is Unweighted Feature Distillation, not Standard Distillation. Standard Distillation uses segmentation and output-level supervision only, whereas $\gamma = 0$ retains feature-level alignment with equal channel weighting. This distinction explains why $\gamma = 0$ performs better than Standard Distillation.

Effect of Loss Weights λ_1 and λ_2. Table 8 evaluates the influence of output-level distillation weight λ_1 and feature-level distillation weight λ_2. For λ_1, increasing its value from 0.0 to 0.5 improves performance (Dice 0.735 to 0.856), confirming that soft supervision from MedSAM provides meaningful structural guidance beyond ground-truth labels.

Table 8. Effect of output distillation loss weight λ_1 and feature distillation loss weight λ_2 on performance. Default values are $\lambda_1 = 0.5$, $\lambda_2 = 1.0$. Bold values indicate the best performance.

λ_1 (**with** $\lambda_2 = 1.0$)	0.0	0.1	0.5 (default)	1.0	2.0
IoU	0.675 ± 0.003	0.734 ± 0.005	**0.764** ± 0.007	0.722 ± 0.008	0.723 ± 0.008
Dice	0.735 ± 0.005	0.801 ± 0.006	**0.856** ± 0.008	0.824 ± 0.009	0.823 ± 0.008
λ_2 (**with** $\lambda_1 = 0.5$)	0.0	0.1	0.5	1.0 (default)	2.0
IoU	0.693 ± 0.004	0.725 ± 0.003	0.729 ± 0.008	**0.764** ± 0.007	0.707 ± 0.005
Dice	0.766 ± 0.006	0.805 ± 0.007	0.818 ± 0.006	**0.856** ± 0.008	0.795 ± 0.009

However, further increasing λ_1 (1.0 or 2.0) leads to performance degradation, likely due to over-regularization that constrains the student model excessively. A similar trend is observed for λ_2. Performance improves as λ_2 increases from 0.0 to 1.0 (Dice 0.766 to 0.856), demonstrating the importance of feature-level alignment. Beyond this point, performance decreases (Dice 0.795 at $\lambda_2 = 2.0$), suggesting diminishing returns or over-constrained feature matching.

Overall, these experiments show that output-level and feature-level distillation are complementary. Moderate weighting, specifically $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$, achieves the best trade-off between guidance and adaptability and is therefore adopted as the default configuration.

4.6. Computational efficiency

To assess the practical deployment cost of PADS, we evaluate its computational efficiency in terms of the number of parameters (Params), floating-point operations (FLOPs), inference latency, and peak GPU memory usage. All measurements are conducted on a single NVIDIA RTX 4090 GPU with batch size 1 and input resolution 320×320. We report the average latency per forward pass and the peak additional GPU memory allocated relative to a fixed baseline (CUDA context and model weights loaded). All evaluations are performed under (torch.inference_mode()) to reflect inference-time cost.

Table 9 reports costs for two deployment modes. UNet-only inference requires 144 GFLOPs, 5.68 ms, and 599 MiB per image. Full Bi-Coop. additionally invokes MedSAM and averages both probability maps, requiring 1131 GFLOPs, 75.96 ms, and 2404 MiB for the higher-accuracy output.

Table 9. Computational efficiency comparison (batch size = 1). Peak memory reports the GPU allocation during inference. Bi-Coop. includes UNet-guided MedSAM inference and probability-map averaging.

Method	Params (M)	FLOPs (GFLOPs)	Latency (ms)	Peak Mem (MiB)
UNet	28.51	144	5.68	599
MedSAM	122.25	977	65.08	2402
MePoDIN†	3.22	20.56	10.67	40
PADS (ours, full with Bi-Coop.)	150.76	1131	75.96	2404

† MePoDIN is a training-only component evaluated independently for reference; it is not executed during testing and does not contribute to inference latency.

The appropriate cost comparison depends on the deployment mode. Relative to standalone MedSAM, the full PADS refinement pipeline adds 10.88 ms of latency (75.96 ms *vs.* 65.08 ms) and almost no additional peak memory (2404 MiB *vs.* 2402 MiB); the dominant cost remains the MedSAM backbone. Relative to the standalone student, however, invoking MedSAM refinement is substantially more expensive. MePoDIN is used only during training to learn relevance weights and is not executed in either inference mode; its independently listed footprint is therefore not included in test-time latency.

PADS offers a choice between efficiency and accuracy at inference: lightweight UNet-only prediction when speed and memory are priorities, or optional full MedSAM refinement when improved segmentation accuracy justifies the additional cost. The description of limited additional overhead applies specifically to full PADS relative to MedSAM, not relative to the standalone UNet student.

4.7. Generalisation on PH2

Table 10 provides the complete numerical cross-domain comparison on PH2, including the supervised reference, published baselines, and recently reproduced SAM-based methods. Figure 3 complements the table by visualising the Dice trends of the three directly matched configurations, including Few-shot UNet, Standard Distillation (labelled "UNet + MedSAM Distillation" in the figure), and PADS, across the 5-shot, 10-shot, and 20-shot settings. PADS consistently achieves the highest Dice score, while all three configurations generally improve as the labelled ISIC training subset increases. In Table 10, entries without a standard deviation are point estimates reported directly by the corresponding publications. The recent methods marked † are reproduced, and their standard deviations are calculated under the stated ISIC sampling protocol before direct PH2 evaluation. Although this direct PH2 evaluation strengthens the evidence beyond a single dataset, it does not constitute full clinical validation, and broader evaluation across institutions, devices, patient populations, and skin tones remains necessary. For the few-shot models on PH2, the reported standard deviations reflect both few-shot sampling variability on ISIC and the additional uncertainty introduced by cross-domain transfer. Variability is greater in the 5-shot setting and decreases at 20-shot, indicating that PADS learns more stable transferable lesion cues as the training subset becomes more representative.

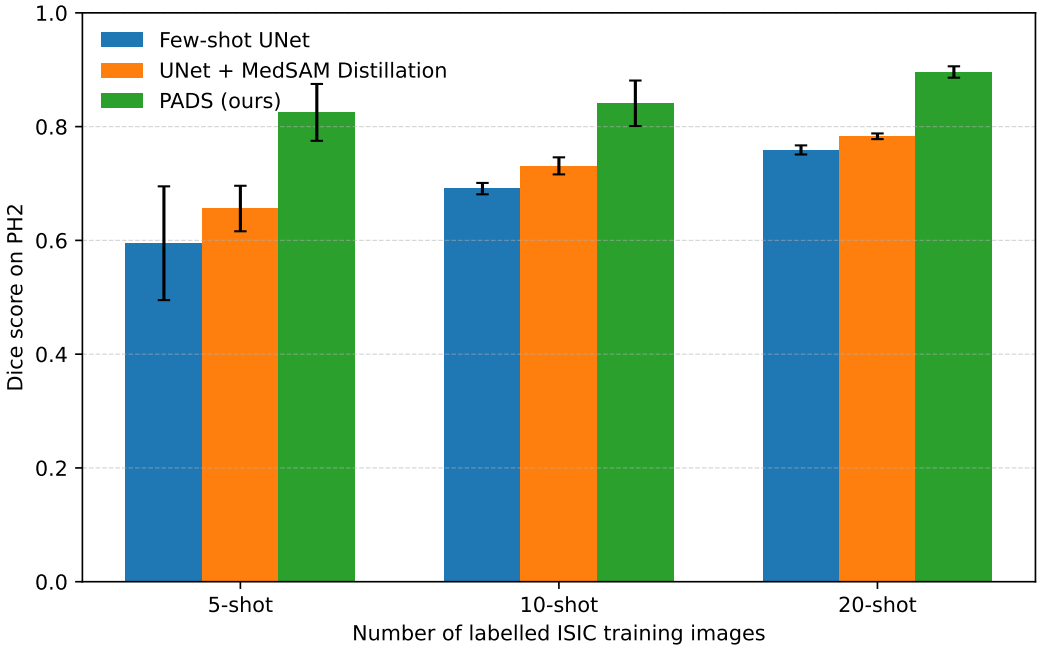

Figure 3. Cross-domain Dice performance on PH2 under different few-shot settings. All models are trained exclusively on ISIC 2018 and evaluated directly on PH2 without fine-tuning. Error bars indicate the standard deviation across three independent few-shot samplings and random seeds.

Table 10. Cross-domain generalisation performance on PH2. Methods marked † are reproduced by the authors using publicly released implementations. No PH2-specific training, fine-tuning, support selection, or hyperparameter tuning is performed. For SANSA, ‡ denotes target-domain parameter-training-free, support-conditioned inference. Bold values indicate the best performance.

Method	Evaluation Setting	IoU	Dice Score
Fully Supervised UNet	2594 Labelled Images	0.831 ± 0.007	0.901 ± 0.009
Semi-sup Few-shot Segmentation [82] (with extra data)	5-shot	–	0.741
	10-shot	–	0.747
	20-shot	–	0.748
Few-shot UNet	5-shot	0.418 ± 0.050	0.595 ± 0.100
	10-shot	0.576 ± 0.011	0.691 ± 0.010
	20-shot	0.638 ± 0.009	0.759 ± 0.008
Standard Distillation	5-shot	0.518 ± 0.035	0.656 ± 0.040
	10-shot	0.645 ± 0.015	0.731 ± 0.015
	20-shot	0.695 ± 0.006	0.783 ± 0.005
Low-shot SAM-based methods (reproduced)			
ProtoSAM† [70]	1-shot	0.343 ± 0.013	0.470 ± 0.015
SANSA† [72]	20-shot‡	0.684 ± 0.069	0.714 ± 0.108
SAMed† [71]	20-shot	0.752 ± 0.014	0.837 ± 0.009
PADS (Ours)	5-shot	**0.731** ± 0.065	**0.825** ± 0.050
	10-shot	**0.758** ± 0.047	**0.841** ± 0.040
	20-shot	**0.823** ± 0.009	**0.896** ± 0.010

PADS consistently achieves superior performance across all k-shot settings. In the 20-shot case, it reaches a Dice Score of 0.896 $\pm$ 0.010, outperforming all few-shot baselines and approaching the fully supervised UNet reference on PH2 (Dice: 0.901, IoU: 0.831). The cross-domain result shows PADS's strong ability to learn transferable lesion representations from limited supervision. By contrast, the few-shot UNet baseline struggles to generalize, especially in the 5-shot setting (Dice: 0.595). MedSAM distillation improves upon this (Dice: 0.656 at 5-shot; 0.783 at 20-shot), confirming the value of foundation model priors, but still falls short of PADS. A semi-supervised method from [82], which uses additional external data, achieves a maximum Dice of 0.748. Yet PADS surpasses this across all settings without any external data. Among the recent reproduced baselines, SAMed achieves the strongest PH2 result with 0.752 $\pm$ 0.014 IoU and 0.837 $\pm$ 0.009 Dice, followed by SANSA with 0.684 $\pm$ 0.069 IoU and 0.714 $\pm$ 0.108 Dice and ProtoSAM with 0.343 $\pm$ 0.013 IoU and 0.470 $\pm$ 0.015 Dice. PADS therefore improves the mean Dice by 0.059 over SAMed. The protocols differ in this comparison: SANSA performs no ISIC- or PH2-specific parameter optimisation but uses labelled ISIC support masks during inference, and its deviation reflects support-set selection; ProtoSAM also uses ISIC support examples during inference, SAMed adapts SAM parameters on ISIC, and PADS transfers knowledge from a frozen teacher into a compact student before direct PH2 evaluation. The cross-domain results show the effectiveness of our perturbation-aware relevance learning strategy.

In summary, PADS excels on the source domain (ISIC 2018) and generalises robustly to unseen data (PH2), approaching the supervised reference and outperforming few-shot baselines in real-world low-resource scenarios.

4.8. Qualitative evaluation

Figure 4 shows qualitative comparisons of segmentation results on three representative ISIC 2018 test cases. We visualise predictions from zero-shot vanilla SAM (using a fixed centre-box prompt), Few-shot UNet, and our proposed PADS alongside ground-truth annotations. In these examples, zero-shot vanilla SAM often produces incomplete or misaligned masks, particularly in the first and third cases, where large portions of the lesion are missed. The incomplete or misaligned masks show vanilla SAM's strong dependence on prompt quality and its limited robustness without medical-domain adaptation.

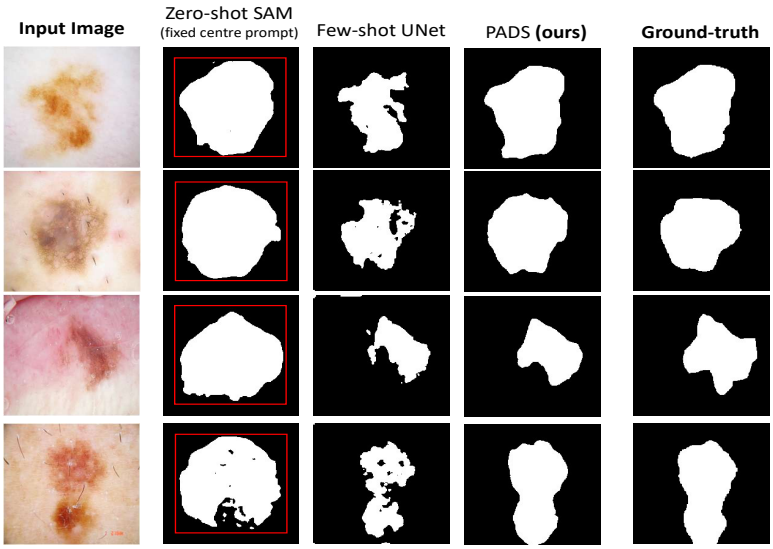

Figure 4. Qualitative Comparison on ISIC dataset.

Few-shot UNet improves lesion coverage but tends to over-segment, spilling into surrounding skin and resulting in more false positives. Its lack of prior knowledge makes it prone to imprecise boundaries, especially under low supervision. In contrast, PADS delivers the most accurate and consistent results. Its outputs exhibit high overlap with ground truth, better preserve lesion shape, and maintain sharp, clean boundaries, even in cases with fuzzy or irregular contours.

These visual comparisons reinforce PADS's key strengths: by learning feature relevance under prompt perturbation to guide selective distillation from MedSAM, and by using a pseudo-prompt for MedSAM refinement, it localizes clinically relevant regions more effectively. This leads to more precise, reliable segmentation, which is important for diagnosis and treatment planning in dermatological imaging.

5. Conclusion

We present PADS, a perturbation-aware distillation framework that reduces sensitivity to prompt quality by transferring knowledge from a frozen MedSAM teacher to a compact UNet. Rather than fine-tuning MedSAM or relying on carefully specified prompts, PADS learns robust lesion representations through relevance-guided distillation. MePoDIN estimates MedSAM feature reliability under varying prompt conditions and enables selective transfer of stable, lesion-relevant representations.

Experiments on ISIC 2018 and PH2 show that PADS approaches the fully supervised UNet references on both datasets (ISIC: 0.858 DSC; PH2: 0.901 DSC) and achieves the strongest performance among the evaluated methods under the reported dermoscopic low-data settings. The framework maintains good cross-domain generalisation and high segmentation quality with only 5 to 20 annotated samples. Ablation and robustness studies indicate that perturbation modelling and relevance-guided feature weighting are both needed to improve stability under prompt variation and limited supervision. Comparisons with recently released ProtoSAM, SAMed, and SANSA implementations further show that perturbation-aware distillation remains effective relative to automatic prompting, parameter-efficient foundation-model adaptation, and support-conditioned inference under dermoscopic low-data settings.

By coupling perturbation-aware distillation with Bi-Coop, PADS preserves MedSAM's medical priors while reducing reliance on manual prompts. At inference, the UNet supplies a pseudo-prompt to MedSAM and their probability maps are averaged; the UNet can also operate independently when efficiency is prioritised. This strategy offers a practical pathway for deploying foundation-model-based segmentation in data-scarce clinical settings.

Future work will validate PADS on broader multi-institutional and clinically acquired dermoscopic datasets, evaluate its robustness under naturally occurring perception and localisation errors, and test its integration within automated or robotic dermatological imaging pipelines. We will also investigate semi-supervised and unsupervised extensions to further reduce annotation requirements. These directions remain future work, and the current results should not be interpreted as validation of autonomous robotic or surgical operation.

Data availability statement

The datasets analyzed in this study are publicly available: ISIC 2018 and PH2.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used OpenAI ChatGPT, including GPT-5.5 and GPT-5.6 models, to assist with language editing and manuscript presentation. The tool was used to improve grammatical accuracy, clarity and readability, suggest alternative wording, refine sentence structure, and improve the flow and organisation of author-prepared text. All AI-assisted content was critically reviewed, verified and edited by the authors. The authors take full responsibility for the scientific content, interpretations, conclusions and final version of the manuscript.

Acknowledgments

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contribution

Conceptualization, Y.W.; methodology, Y.W. and K.Y.; software, Y.L.; validation, Y.W. and Z.J.C.; formal analysis, K.Y.; investigation, Y.W., Z.J.C. and Z.C.; data curation, Y.W. and Z.J.C.; writing—original draft preparation, Y.W.; writing—review and editing, K.Y., Y.L., Z.J.C. and Z.C.; visualization, Y.W.; supervision, K.Y. and Z.C.; project administration, Z.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

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

References

[1] Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, *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 *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, USA, April 4–7, 2018, pp. 168–172.

[2] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* 2018, 5(1):1–9.

[3] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, October 5–9, 2015, pp. 234–241.

[4] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* 2017, 42:60–88.

[5] Wang J, Zhu H, Wang SH, Zhang YD. A review of deep learning on medical image analysis. *Mobile Networks Appl.* 2021, 26(1):351–380.

[6] Xia Q, Zheng H, Zou H, Luo D, Tang H, *et al.* A comprehensive review of deep learning for medical image segmentation. *Neurocomputing* 2025, 613:128740.

[7] Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* 2020, 53(3):1–34.

[8] Shaban A, Bansal S, Liu Z, Essa I, Boots B. One-shot learning for semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, September 4–7, 2017, pp. 167.1–167.13.

[9] Liu W, Zhang C, Lin G, Liu F. CRNet: cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, June 13–19, 2020, pp. 4165–4173.

[10] Teng P, Liu W, Wang X, Wu D, Yuan C, *et al.* Beyond singular prototype: a prototype splitting strategy for few-shot medical image segmentation. *Neurocomputing* 2024, 597:127990.

[11] Wang S, Yu X, Chi J, Wu C, Gao X. PONet: prototype optimization network for few-shot medical image segmentation. *Neurocomputing* 2025, 652:131113.

[12] Zou M, Yuan W, Liu Y, Zeng D, Zhao G. Few-shot medical image segmentation via CLIP-driven dual contrastive learning. *Neurocomputing* 2026, 673:132815.

[13] Huang X, Ma Y, Mei X, Sun M, She Q, *et al.* Lesion boundary detection for skin lesion segmentation based on boundary sensing and CNN-transformer fusion networks. *Artif. Intell. Med.* 2025, p. 103190.

[14] Ji Z, Ye Y, Ma X. BDFormer: boundary-aware dual-decoder transformer for skin lesion segmentation. *Artif. Intell. Med.* 2025, 162:103079.

[15] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, October 1–6, 2023, pp. 4015–4026.

[16] Ma J, He Y, Li F, Han L, You C, *et al.* Segment anything in medical images. *Nat. Commun.* 2024, 15(1):654.

[17] Cheng J, Ye J, Deng Z, Chen J, Li T, *et al.* SAM-Med2D. *arXiv* 2023, arXiv:2308.16184.

[18] Cassidy B, Kendrick C, Brodzicki A, Jaworek-Korjakowska J, Yap MH. Analysis of the ISIC image datasets: usage, benchmarks and recommendations. *Med. Image Anal.* 2022, 75:102305.

[19] Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, *et al.* Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). *arXiv* 2019, arXiv:1902.03368.

[20] Mendonça T, Ferreira PM, Marques JS, Marcal AR, Rozeira J. PH 2-A dermoscopic image database for research and benchmarking. In *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, July 3–7, 2013, pp. 5437–5440.

[21] Celebi ME, Kingravi HA, Uddin B, Iyatomi H, Aslandogan YA, *et al.* A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graphics* 2007, 31(6):362–373.

[22] Cao W, Yuan G, Liu Q, Peng C, Xie J, *et al.* ICL-Net: global and local inter-pixel correlations learning network for skin lesion segmentation. *IEEE J. Biomed. Health. Inf.* 2022, 27(1):145–156.

[23] Ali AR, Li J, O'Shea SJ. Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images. *PLoS One* 2020, 15(6):e0234352.

[24] Schaefer G, Krawczyk B, Celebi ME, Iyatomi H. An ensemble classification approach for melanoma diagnosis. *Memet. Comput.* 2014, 6:233–240.

[25] Celebi ME, Mendonca T, Marques JS. *Dermoscopy Image Analysis*, 1st ed. Boca Raton: CRC Press, 2015.

[26] Yu L, Chen H, Dou Q, Qin J, Heng P. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med. Imaging* 2016, 36(4):994–1004.

[27] Ballerini L, Fisher RB, Aldridge B, Rees J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. *Color Med. Image Anal.* 2013, pp. 63–86.

[28] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, June 8–10, 2015, pp. 3431–3440.

[29] Yuan Y, Chao M, Lo YC. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Trans. Med. Imaging* 2017, 36(9):1876–1886.

[30] Feng S, Zhao H, Shi F, Cheng X, Wang M, *et al.* CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med. Imaging* 2020, 39(10):3008–3018.

[31] Li H, He X, Zhou F, Yu Z, Ni D, *et al.* Dense deconvolutional network for skin lesion segmentation. *IEEE J. Biomed. Health. Inf.* 2018, 23(2):527–537.

[32] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017, 542(7639):115–118.

[33] Bissoto A, Perez F, Valle E, Avila S. Skin lesion synthesis with generative adversarial networks. In *Proceedings of the Held in conjunction with the 21st International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018)*, Granada, Spain, September 16–20, 2018, pp. 294–302.

[34] Ali R, Hardie RC, De Silva MS, Kebede TM. Skin lesion segmentation and classification for ISIC 2018 by combining deep CNN and handcrafted features. *arXiv* 2019, arXiv:1908.05730.

[35] Qi Y, Wei L, Yang J, Xu J, Wang H, *et al.* CQENet: a segmentation model for nasopharyngeal carcinoma based on confidence quantitative evaluation. *Comput. Med. Imaging Graphics* 2025, 123:102525.

[36] Yu Z, Yu L, Zheng W, Wang S. EIU-Net: enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation. *Comput. Biol. Med.* 2023, 162:107081.

[37] Zhang W, Lu F, Zhao W, Hu Y, Su H, *et al.* ACCPG-Net: a skin lesion segmentation network with adaptive channel-context-aware pyramid attention and global feature fusion. *Comput. Biol. Med.* 2023, 154:106580.

[38] Xin C, Liu Z, Ma Y, Wang D, Zhang J, *et al.* Transformer guided self-adaptive network for multi-scale skin lesion image segmentation. *Comput. Biol. Med.* 2024, 169:107846.

[39] Liu S, Wang P, Lin Y, Zhou B. IDSNet: unifying local and global context for skin lesion image segmentation. *Biomed. Signal Process. Control* 2025, 108:107961.

[40] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Long Beach, USA, December 4–9, 2017, pp. 5998–6008.

[41] Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach.* 2020, 30:681–694.

[42] Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, *et al.* Gemini: a family of highly capable multimodal models. *arXiv* 2023, arXiv:2312.11805.

[43] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, *et al.* LlaMA: open and efficient foundation language models. *arXiv* 2023, arXiv:2302.13971.

[44] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, *et al.* An image is worth 16×16 words: transformers for image recognition at scale. *arXiv* 2021, arXiv:2010.11929.

[45] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, *et al.* Transformers in medical imaging: a survey. *Med. Image Anal.* 2023, 88:102802.

[46] Chen J, Lu Y, Yu Q, Luo X, Adeli E, *et al.* Transunet: transformers make strong encoders for medical image segmentation. *arXiv* 2021, arXiv:2102.04306.

[47] Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. *Med. Image Anal.* 2024, 91:102996.

[48] Mazurowski MA, Dong H, Gu H, Yang J, Konz N, *et al.* Segment anything model for medical image analysis: an experimental study. *Med. Image Anal.* 2023, 89:102918.

[49] Huang Y, Yang X, Liu L, Zhou H, Chang A, *et al.* Segment anything model for medical images? *Med. Image Anal.* 2024, 92:103061.

[50] Hu X, Xu X, Shi Y. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv* 2023, arXiv:2306.13731.

[51] Wu J, Wang Z, Hong M, Ji W, Fu H, *et al.* Medical sam adapter: adapting segment anything model for medical image segmentation. *Med. Image Anal.* 2025, p. 103547.

[52] Mirikharaji Z, Abhishek K, Bissoto A, Barata C, Avila S, *et al.* A survey on deep learning for skin lesion segmentation. *Med. Image Anal.* 2023, 88:102863.

[53] Gong W, Luo Y, Yang F, Zhou H, Lin Z, *et al.* CGNet: few-shot learning for intracranial hemorrhage segmentation. *Comput. Med. Imaging Graphics* 2025, p. 102505.

[54] Irfan M, Haq IU, Malik KM, Muhammad K. One-Shot learning for generalization in medical image classification across modalities. *Comput. Med. Imaging Graphics* 2025, p. 102507.

[55] Khadka R, Jha D, Hicks S, Thambawita V, Riegler MA, *et al.* Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. *Comput. Biol. Med.* 2022, 143:105227.

[56] Hu H, Zhang J, Yang T, Hu Q, Yu Y, *et al.* PATrans: pixel-adaptive transformer for edge segmentation of cervical nuclei on small-scale datasets. *Comput. Biol. Med.* 2024, 168:107823.

[57] Guo B, Huang W, Wang X. ABE-Mamba: few-shot medical image segmentation via adversarial bidirectional enhanced Mamba. *Expert Syst. Appl.* 2026, 298:129897.

[58] Chen Z, Li J, Chen Z, You X. Generic pixel level object tracker using bi-channel fully convolutional network. In *Proceedings of the 24th International Conference on Neural Information Processing (ICONIP)*, Guangzhou, China, November 14–18, 2017, pp. 666–676.

[59] Chen Z, Zhang J, Tao D. Recursive context routing for object detection. *Int. J. Comput. Vision* 2021, 129(1):142–160.

[60] Chen Z, Zhang J, Xu Y, Tao D. Transformer-based context condensation for boosting feature pyramids in object detection. *Int. J. Comput. Vision* 2023, 131(10):2738–2756.

[61] Chen Z, Zhang J, Tao D. Recurrent glimpse-based decoder for detection with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, June 19–24, 2022, pp. 5260–5269.

[62] Zhang X, Chen Z, Zhang J, Liu T, Tao D. Learning general and specific embedding with transformer for few-shot object detection. *Int. J. Comput. Vision* 2025, 133(2):968–984.

[63] An Z, Sun G, Liu Y, Liu F, Wu Z, *et al.* Rethinking few-shot 3D point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 17–21, 2024, pp. 3996–4006.

[64] An Z, Sun G, Liu Y, Li R, Wu M, *et al.* Multimodality helps few-shot 3D point cloud semantic segmentation. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, Singapore, April 24–28, 2025, pp. 50163–50184.

[65] An Z, Sun G, Liu Y, Li R, Han J, *et al.* Generalized few-shot 3D point cloud segmentation with vision–language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, June 11–15, 2025, pp. 16997–17007.

[66] Zhang R, Jiang Z, Guo Z, Yan S, Pan J, *et al.* Personalize segment anything model with one shot. *arXiv* 2023, arXiv:2305.03048.

[67] Xiao J, Xu H, Zhao W, Cheng C, Gao H. A prior-mask-guided few-shot learning for skin lesion segmentation. *Computing* 2023, pp. 1–23.

[68] Pachetti E, Colantonio S. A systematic review of few-shot learning in medical imaging. *Artif. Intell. Med.* 2024, 156:102949.

[69] Zhu Y, Li M, Ye Q, Wang S, Xin T, *et al.* RobustEMD: domain robust matching for cross-domain few-shot medical image segmentation. *Artif. Intell. Med.* 2025, p. 103197.

[70] Ayzenberg L, Giryes R, Greenspan H. ProtoSAM for automated one shot medical image segmentation using foundational models. *Sci. Rep.* 2025, 15:41482.

[71] Zhang K, Liu D. Customized segment anything model for medical image segmentation. *arXiv* 2023, arXiv:2304.13785.

[72] Cuttano C, Trivigno G, Averta G, Masone C. SANSA: unleashing the hidden semantics in SAM2 for few-shot segmentation. *arXiv* 2025, arXiv:2505.21795.

[73] Zhou W, Guan G, Yan Q, Zhao Y. AdaptSAM: adaptive SAM for cross-domain few-shot medical image segmentation. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kuala Lumpur, Malaysia, December 1–4, 2025, pp. 3389–3396.

[74] Özdemir Z, Keleş HY, Tanrıöver ÖÖ. Meta-transfer derm-diagnosis: exploring few-shot learning and transfer learning for skin disease classification in long-tail distribution. *IEEE J. Biomed. Health. Inf.* 2025, pp. 3132–3145.

[75] Jiang S, Ma Y, Zhou J, Wang Y, Liu M. Unified multimodal industrial anomaly detection via few normal samples. *IEEE Trans. Ind. Inform.* 2026, 22(8):6823–6834.

[76] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, *et al.* Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* 2022, 35:23716–23736.

[77] Yang G, Luo S, Greer P. A novel vision transformer model for skin cancer classification. *Neural Process. Lett.* 2023, 55(7):9335–9351.

[78] Shaharabany T, Dahan A, Giryes R, Wolf L. Autosam: adapting sam to medical images by overloading the prompt encoder. *arXiv* 2023, arXiv:2306.06370.

[79] Wu Y, He K. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, September 8–14, 2018, pp. 3–19.

[80] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, June 18–22, 2018, pp. 7132–7141.

[81] Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, Munich, Germany, September 8–14, 2018, pp. 3–19.

[82] Feyjie AR, Azad R, Pedersoli M, Kauffman C, Ayed IB, *et al.* Semi-supervised few-shot learning for medical image segmentation. *arXiv* 2020, arXiv:2003.08462.

[83] Zhang C, Lin G, Liu F, Yao R, Shen C. Canet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, June 16–20, 2019, pp. 5217–5226.

[84] Wang K, Liew JH, Zou Y, Zhou D, Feng J. Panet: few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, October 27–November 2, 2019, pp. 9197–9206.