DINOv2-UNet: Deployment-Friendly Colonoscopy Polyp Segmentation with Self-Supervised ViT Priors

Chengfei Cai[1,*], Jia Yu[1], Zehao Liu[1], Zhenyu Song[1], Zenan Lu[1], Lixing Tan[1], Guodong Hu[1]

[1]*Jiangsu Key Laboratory of Intelligent Drug Screening and Repositioning, College of Information Engineering, Taizhou University*, Taizhou, China

[*]Corresponding author: chengfeicai@tzu.edu.cn

***Abstract*—Accurate polyp segmentation is crucial for colorectal cancer screening but remains challenging due to low tissue contrast, imaging artefacts, and cross-domain variability. We propose DINOv2-UNet, a deployment-friendly framework that transfers large self-supervised vision transformers to colonoscopic segmentation through a representation-first design. The model couples a DINOv2 ViT-B/14 encoder with a lightweight U-shaped decoder, avoiding complex fusion modules to preserve efficiency. To ensure stable adaptation on label-scarce medical datasets, we introduce a principled fine-tuning recipe that involves partial optimisation of deep transformer blocks, differential learning rates, hybrid BCE–Dice supervision, and cosine scheduling with warm-up. Experiments on four public datasets, including Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-LaribPolypDB, achieve Dice scores of 0.927, 0.948, 0.925, and 0.931, respectively, consistently outperforming CNN- and transformer-based baselines. Despite the high-capacity backbone, the model runs at 53.9 FPS on an RTX 4050 GPU and over 23 FPS on a GTX 1080 Ti. These results demonstrate that well-adapted self-supervised priors can replace heavy decoders, enabling accurate and real-time clinical deployment.**

***Index Terms*—Polyp segmentation, colonoscopy, vision transformer, self-supervised learning, medical image segmentation**

I. Introduction

Accurate polyp segmentation plays a central role in computer-aided colorectal cancer screening, which is crucial for effective early detection [1]. However, automated segmentation remains challenging in practice due to low contrast between lesions and surrounding mucosa, significant appearance variations, frequent imaging artefacts, and pronounced domain shifts across acquisition devices and clinical centers [2].

Although recent advances in deep learning, ranging from standard CNN architectures [3]–[5] to modern attention mechanisms [6], have substantially improved segmentation performance, robustness and deployment efficiency remain critical concerns for real-world clinical systems. Foundation models based on large-scale Vision Transformers (ViTs) [7], [8] show strong representation learning and high potential for label-scarce medical tasks [9]–[11].

However, directly transferring such models to dense medical segmentation remains non-trivial. Straightforward fine-tuning of large ViTs often leads to unstable optimization or overfitting on small datasets, while freezing the encoder typically degrades task-specific performance [10]. To compensate, many existing approaches introduce sophisticated multi-scale fusion modules or heavy decoding heads [12]–[14], which increase architectural complexity, memory footprint, and inference latency, thereby limiting practical deployment in time-sensitive clinical workflows.

In this work, we revisit this trade-off from a different perspective and advocate a *representation-first* paradigm that prioritizes reliable feature transfer over decoder complexity. We propose **DINOv2-UNet**, a streamlined segmentation framework that couples a self-supervised DINOv2 [15] ViT-B/14 encoder with a lightweight U-shaped decoder. Rather than relying on elaborate attention-based fusion mechanisms, we focus on stabilizing the foundation encoder's adaptation to the medical domain via a principled, transfer-oriented optimization strategy. Specifically, we employ partial fine-tuning of deep transformer blocks while freezing shallow layers, together with differential learning rates and a warm-up cosine schedule, to balance representation plasticity and training stability under limited supervision. The main contributions of this work are summarized as follows:

- We introduce DINOv2-UNet, a deployment-oriented segmentation framework that combines a self-supervised Vision Transformer encoder with a lightweight U-shaped decoder, emphasizing architectural simplicity and practical feasibility.
- We propose a robust fine-tuning recipe for transferring large foundation models to label-scarce medical imaging tasks, based on partial encoder adaptation, differential learning rates, and warm-up cosine scheduling.
- We conduct comprehensive evaluations across four public datasets, including quantitative analyses and qualitative studies on representation visualization and training dynamics, to assess robustness and transferability.
- We demonstrate that the proposed model meets real-time clinical constraints, achieving high throughput without sacrificing segmentation accuracy, thereby supporting its applicability to online colonoscopic screening.

II. Related Work

A. Clinical and Data Generalizability Context

Generalizability across centers remains a key challenge for computer-aided endoscopy. Multi-center datasets and benchmarks consistently report notable performance degradation

when models are evaluated across different hospitals and devices, reflecting substantial domain shifts in imaging conditions and lesion appearance [16], [17]. These discrepancies limit the reliability of models trained on single-source data. Clinical studies further indicate that, beyond segmentation accuracy, inference latency and system stability strongly affect the integration of CADe and CADx systems into routine colonoscopy workflows [2]. As a result, both robustness and deployment efficiency are practical requirements for segmentation models in real-world settings.

B. Segmentation and Transfer Learning Trends

Self-supervised learning and foundation models have recently enabled effective representation learning from large-scale unlabeled data, offering improved label efficiency and transferability for medical imaging tasks [9], [10]. In colorectal lesion segmentation, recent studies have explored stronger backbones and enhanced feature modeling to improve robustness under challenging visual conditions [6], [14]. Foundation models are increasingly discussed as a potential means to improve cross-domain performance [11], yet their stable and efficient adaptation to dense prediction tasks remains an open issue.

C. Polyp Segmentation Architectures

Most existing methods are based on U-shaped convolutional architectures with skip connections, later extended with deeper backbones, multi-scale aggregation, and attention mechanisms [3], [5], [18]. While these designs improve benchmark performance, they often increase model complexity and computational cost. Transformer-based encoders have recently attracted interest due to their global receptive fields, but typically require heavy decoders to recover spatial detail, which can hinder real-time deployment [12], [13].

D. Self-supervised and Foundation-Model Transfer

Current transfer strategies range from linear probing and full fine-tuning to the introduction of task-specific adaptation modules and complex decoding heads. Although effective, these approaches may suffer from unstable optimization, increased parameter counts, or higher inference latency. In contrast, our work focuses on a lightweight adaptation strategy that combines partial fine-tuning of deep transformer layers with a compact decoder, aiming to balance segmentation accuracy, training stability, and deployment efficiency [15].

III. Methodology

Fig. 1 illustrates the proposed DINOv2-UNet framework. Our design highlights a partially frozen Vision Transformer encoder coupled with a streamlined U-shaped decoder. In this section, we first formulate the polyp segmentation task, then detail the encoder adaptation strategy, the decoding architecture, and finally the objective function and implementation recipe.

A. Problem Formulation

We formulate colorectal polyp segmentation as a pixel-wise binary classification task. Given an input RGB endoscopic image $x \in \mathbb{R}^{H\times W\times 3}$, the goal is to learn a mapping function f_θ that predicts a probabilistic map $p \in [0,1]^{H\times W}$ using a sigmoid activation, where 1 represents the polyp region and 0 represents the background mucosa.

$$p = \sigma\big(f_\theta(x)\big), \quad p \in [0,1]^{H\times W}. \tag{1}$$

where $x \in \mathbb{R}^{H\times W\times 3}$ is the endoscopic image, $f_\theta(\cdot)$ denotes the segmentation network parameterized by θ, and $\sigma(\cdot)$ is the sigmoid function that maps logits to probabilities.

B. Encoder with Partial Fine-tuning

The backbone of our framework is the DINOv2 ViT-B/14 encoder [15]. This model partitions the input image into non-overlapping patches of size 14×14 pixels. Each patch is linearly projected into an embedding space and processed through a stack of 12 transformer blocks.

To adapt this foundation model to the medical domain while preventing overfitting on small datasets, we employ a partial fine-tuning strategy. Specifically, we freeze the parameters of the shallow transformer blocks, defined as blocks 0 through 5. These layers capture generic low-level visual features such as textures and edges, which are highly transferable across domains. Conversely, we fine-tune the deeper blocks, defined as blocks 6 through 11. This allows the model to adapt high-level semantic representations to the specific characteristics of polyps, such as their unique shapes and lighting patterns. This strategy preserves the robust priors learned from large-scale self-supervision while ensuring the model is sensitive to clinical targets. The distinct separation between frozen and trainable layers stabilizes the optimization process and reduces the risk of catastrophic forgetting.

C. Streamlined Projection and U-shaped Decoding

The decoding architecture is deliberately designed to minimize computational overhead, thereby ensuring high inference throughput suitable for clinical deployment. We construct a lightweight U-shaped decoder that progressively recovers spatial details by leveraging multi-scale feature representations extracted from four specific transformer blocks. This hierarchical feature extraction strategy is inspired by established transformer-based segmentation networks [13].

1) Feature Projection Module: The transition from the ViT encoder to the decoder is mediated by a Feature Projection Module (FPM) that bridges the domain gap between 1D token sequences and 2D spatial maps. For each selected feature level l, the raw output from the transformer is first processed to discard the classification (CLS) token. The remaining spatial tokens are subsequently reshaped into a 2D feature grid $F^{(l)}$. Given the standard input resolution of 448×448 and a patch size of 14, these reshaped feature maps possess spatial dimensions of 32×32. To standardize the channel capacity across different scales for efficient processing, we apply a learnable 1×1 convolution that projects the high-dimensional

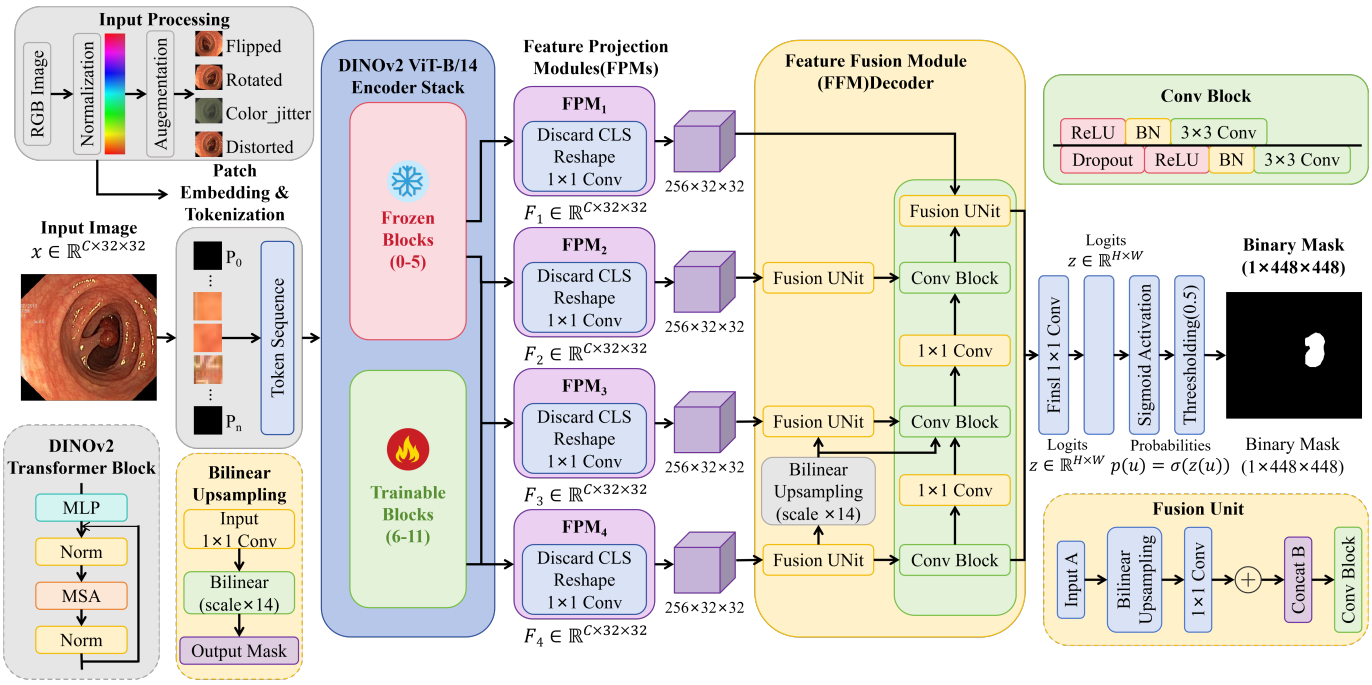


Fig. 1. Overview of DINOv2-UNet. The architecture consists of a DINOv2 encoder with partially frozen shallow blocks and a lightweight U-shaped decoder, designed for efficient and accurate polyp segmentation.

transformer embeddings to a unified channel width of $C = 256$.

2) Feature Fusion and Refining: We adopt a U-shaped reconstruction path to generate the high-resolution segmentation mask. The decoding process involves progressive bilinear upsampling of the deep feature maps to match the resolution of the corresponding projected features from the shallower encoder levels. Following spatial alignment, the features from both pathways are concatenated along the channel dimension to fuse semantic context with structural details.

$$\hat{P}_i = \mathrm{Concat}\big(\mathcal{U}(P_{i+1}),\, \phi(F_i)\big). \tag{2}$$

where i denotes the decoder stage, $\mathcal{U}(\cdot)$ is bilinear upsampling, $\phi(\cdot)$ is a 1×1 convolution for alignment, and $\mathrm{Concat}(\cdot,\cdot)$ concatenates features along channels. To refine these aggregated representations, we introduce a Regularized Dual-Conv Block at each decoding stage. This component comprises two sequential 3×3 convolutional layers, where each convolution is followed by Batch Normalization and a ReLU activation function. A critical design choice in this block is the injection of a Dropout layer between the two convolutional units. This regularization strategy helps mitigate overfitting on small medical datasets without relying on computationally expensive attention-based fusion mechanisms often found in recent literature, effectively balancing segmentation accuracy with model efficiency.

3) Prediction Head: The final segmentation probability map is derived from the output of the last fusion stage. The refined feature map is first passed through a 1×1 convolution to compress the channel dimension to a single output class. To restore the prediction to the original input resolution, the resulting logit map is upsampled by a factor of 14 using bilinear interpolation. Finally, a sigmoid activation function is applied to normalize the logits, producing a dense probability map $p \in [0, 1]^{H\times W}$ that represents the pixel-wise likelihood of polyp presence.

$$Z = \mathcal{U}_{\times 14}\left(\mathrm{Conv}_{1\times1}(P_1)\right) \in \mathbb{R}^{H\times W}. \tag{3}$$

where P_1 is the final fused feature map at the patch-grid resolution, $\mathrm{Conv}_{1\times1}(\cdot)$ projects features to a single-channel logit map, $\mathcal{U}_{\times14}(\cdot)$ upsamples by a factor of 14 to recover the input resolution, and Z is the full-resolution logit map.

D. Hybrid Loss Function

Polyp segmentation often suffers from severe class imbalance, as polyps occupy a small part of the image. To address this, we utilize a hybrid objective function $\mathcal{L}$ that combines the Binary Cross-Entropy (BCE) loss and the Dice loss:

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{BCE}} + (1 - \lambda)\mathcal{L}_{\mathrm{Dice}}. \tag{4}$$

where $\mathcal{L}_{\mathrm{BCE}}$ and $\mathcal{L}_{\mathrm{Dice}}$ denote the binary cross-entropy loss and Dice loss, respectively, and $\lambda \in [0, 1]$ controls their trade-off. In our experiments, we set the balancing coefficient λ to 0.5. The BCE term supervises pixel-level classification accuracy, while the Dice term optimizes the region-level overlap between the prediction and the ground truth, ensuring precise boundary delineation.

E. Training and Inference Recipe

1) Data Preprocessing and Augmentation: To ensure a unified and fair evaluation, we apply the same preprocessing

and splitting protocol across all datasets. All images are resized to a fixed resolution of 448×448 and normalized using the mean and standard deviation statistics from ImageNet. During training, we apply strong data augmentation to improve generalization. This includes random horizontal and vertical flipping, random rotation, random scaling and cropping, color jittering and Gaussian blur.

2) Optimization Strategy: The model is optimized using the AdamW optimizer. We implement a differential learning rate strategy to further stabilize the transfer learning process. The trainable encoder blocks are optimized with a lower learning rate of 1×10^{-5}, while the decoder layers are trained with a higher learning rate of 1×10^{-3}. We use a cosine annealing scheduler with a linear warm-up phase to adjust the learning rates dynamically over the course of training. The training runs for 80 epochs with a batch size of 8. Automatic Mixed Precision (AMP) is enabled to reduce memory usage and accelerate computation.

3) Inference Settings: For inference, we generate predictions using a single forward pass without test-time augmentation (TTA) by default to measure the raw speed and accuracy. However, for optional performance boosting, we also evaluate a horizontal flip TTA setting by averaging predictions from the original and flipped inputs (results are reported separately when used).

IV. Experiments

A. Datasets and Evaluation Protocol

We evaluate on four public datasets: **Kvasir-SEG** [19] (1,000 images), **CVC-ClinicDB** [20] (612 images), **CVC-ColonDB** [21] (380 images), and the challenging **ETIS-LaribPolypDB** [22] (196 images). To guarantee fair comparison, we implement a unified data processing protocol [16]. For each dataset, the model is trained and evaluated independently using a standard 80%/10%/10% random split for training, validation, and testing. We adhere to the in-domain evaluation setting to ensure performance gains are attributed to architectural improvements rather than data partitioning.

B. Implementation Details

Our pipeline is implemented in PyTorch. Images are resized to 448×448 and normalized using ImageNet statistics. We train the network for 80 epochs using the AdamW optimizer (weight decay 0.01) with a batch size of 8. To stabilize transfer learning, we employ a differential learning rate strategy [10], [11]: the decoder is optimized at 1×10^{-3}, whereas trainable encoder blocks use 1×10^{-5}. We apply a 5-epoch linear warm-up followed by cosine decay. Automatic Mixed Precision (AMP) is enabled, test-time augmentation is disabled by default, and the random seed is fixed to 42.

C. Metrics

We report Dice and IoU as primary overlap metrics. Given a predicted binary mask P and ground truth G, Dice and IoU are defined as

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}, \quad \text{IoU} = \frac{|P \cap G|}{|P \cup G|}. \tag{5}$$

TABLE I
Main segmentation results of DINOv2-UNet on four datasets.

Dataset	Dice ↑	IoU ↑
Kvasir-SEG	0.927	0.880
CVC-ClinicDB	0.948	0.904
CVC-ColonDB	0.925	0.864
ETIS-LaribPolypDB	0.931	0.875

Following standard practice, probabilities are binarized using a fixed threshold of 0.5 across all datasets. Unless otherwise stated, Dice and IoU are computed over the entire test set. For ablation studies where per-image averaging is more stable on small test sets, we additionally report mDice and mIoU, defined as the mean Dice and IoU across all test images. MAE is also reported when applicable.

V. Results

A. Overall Quantitative Results

Table I reports Dice/IoU on four datatsets. Across the four datatsets, performance remains consistently high despite notable dataset heterogeneity. These datasets differ in acquisition devices, illumination patterns, and polyp appearance, suggesting that SSL ViT priors can offer robust representations under diverse in-domain conditions.

B. Deployment Efficiency

For runtime evaluation, test-time augmentation is disabled, and FP32 inference is adopted unless otherwise specified. Latency is measured on both an NVIDIA GTX 1080 Ti and an NVIDIA RTX 4050 Laptop GPU to reflect practical deployment across different hardware generations. Throughput is evaluated with a batch size of 1 and an input resolution of 448×448, excluding data loading. The reported FPS is averaged over 300 forward passes after 50 warm-up iterations.

Two inference configurations are considered: a default setting and a speed-optimized setting, both using identical model weights and input resolution to ensure deployment consistency. The difference lies solely in inference-time implementation choices, including warm-up strategies and backend kernel optimizations. Accordingly, the FLOPs reported for the speed-optimized setting should be interpreted as effective estimates under the corresponding operator kernels, rather than as changes to the model architecture. Under FP32 inference, DINOv2-UNet achieves 53.92 FPS on the RTX 4050, demonstrating its feasibility for real-time clinical deployment. On a GTX 1080 Ti, the default setting requires 94.34 GFLOPs and 93.24M parameters, achieving 23.67 FPS. A speed-optimized inference implementation further improves throughput to 32.11 FPS (72.36 effective GFLOPs).

C. Ablation on Fine-tuning Strategy

We compare the proposed partial fine-tuning strategy with a frozen encoder (linear probing) and full fine-tuning of all ViT blocks, with results summarized in Table III. Freezing the encoder leads to clear performance degradation across all

TABLE II

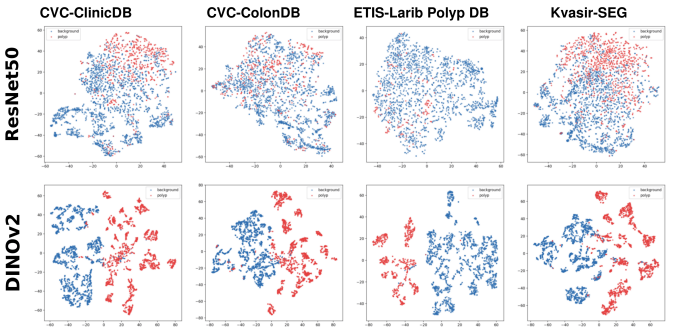
COMPARISON OF COMPUTATIONAL EFFICIENCY AND SEGMENTATION PERFORMANCE. **PARAMS**: NUMBER OF PARAMETERS (M). **FLOPS**: G. **FPS**: MEASURED ON AN NVIDIA GTX 1080 TI WITH BATCH SIZE 1 AND INPUT RESOLUTION 448×448 (DATA LOADING EXCLUDED). BASELINE FPS IS NOT REPORTED HERE DUE TO NON-UNIFIED HARDWARE/SETTINGS ACROSS SOURCES.

Model	Backbone	Params (M)	FLOPs (G)	FPS (1080Ti)	Dice				
					Kvasir	Clinic	Colon	ETIS	Avg.
U-Net	–	34.52	123.87	–	0.818	0.823	0.512	0.398	0.638
U-Net++	–	36.63	326.16	–	0.821	0.794	0.483	0.401	0.625
PraNet	ResNet-50	30.50	13.15	–	0.898	0.899	0.712	0.628	0.784
Polyp-PVT	PVT-v2	25.11	10.02	–	0.917	0.937	0.808	0.787	0.862
DINOv2-UNet (Ours)	ViT-B/14	93.24	94.34	**23.67**	0.927	**0.948**	**0.925**	**0.931**	**0.933**

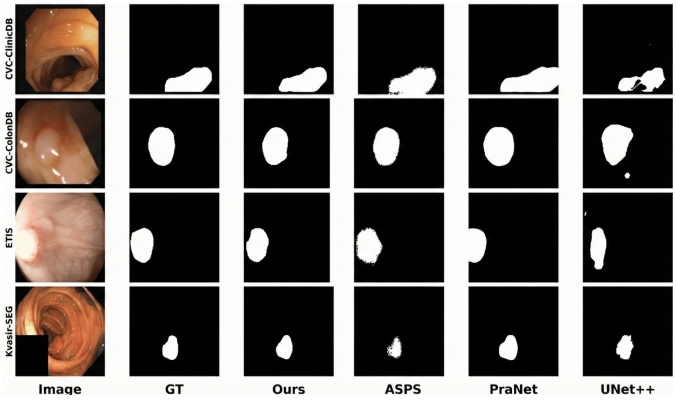
TABLE III

FINE-TUNING STRATEGY ABLATION ON ALL DATATSETS.

Dataset	Strategy	mDice ↑	mIoU ↑	MAE ↓
Kvasir-SEG	Frozen Encoder	0.899	0.831	0.043
Kvasir-SEG	Full Fine-tune	0.931	0.882	**0.029**
Kvasir-SEG	Partial FT	**0.936**	**0.887**	0.031
CVC-ClinicDB	Frozen Encoder	0.916	0.855	0.018
CVC-ClinicDB	Full Fine-tune	**0.948**	**0.904**	**0.011**
CVC-ClinicDB	Partial FT	0.946	0.900	0.012
CVC-ColonDB	Frozen Encoder	0.839	0.764	0.018
CVC-ColonDB	Full Fine-tune	**0.919**	**0.856**	**0.010**
CVC-ColonDB	Partial FT	0.917	0.851	0.010
ETIS	Frozen Encoder	0.913	0.846	0.008
ETIS	Full Fine-tune	**0.927**	**0.867**	**0.006**
ETIS	Partial FT	0.918	0.852	0.012



(a) t-SNE embedding across datasets.



(b) Qualitative comparison grid.

Fig. 2. Additional analysis of partial fine-tuning. The t-SNE embedding and qualitative grid provide complementary evidence on representation adaptation and segmentation behavior.

datasets. Full fine-tuning achieves comparable accuracy but exhibits noticeable instability on smaller datasets. In contrast, partial fine-tuning consistently delivers superior or comparable accuracy with improved stability. Qualitative results further confirm more coherent polyp delineation and fewer spurious fragments under challenging illumination and low-contrast boundaries. Visualization grids illustrate that these gains stem from effective adaptation of deep representations while preserving transferable shallow priors.

D. Ablation on Pretraining: Supervised vs. Self-Supervised Priors

To verify that our gains are driven by self-supervised priors rather than architecture alone, we compare a supervised ImageNet-pretrained ViT encoder against DINOv2 self-supervised pretraining, while keeping the same decoder and partial fine-tuning recipe. As shown in Table IV, DINOv2 consistently improves mDice and mIoU across all datasets, especially on the more challenging ETIS dataset, supporting the claim that SSL priors offer stronger transferability under limited annotations.

E. Decoder Design Trade-off: Complex vs. Simple Decoder

A common intuition is that ViT-based segmentation needs heavy fusion modules to recover spatial detail. We test this assumption by comparing a heavier complex decoder against

TABLE IV

PRETRAINING ABLATION (PARTIAL FINE-TUNING + SAME DECODER).

Dataset	Pretraining	mDice ↑	mIoU ↑	MAE ↓
Kvasir-SEG	ImageNet-Sup	0.921	0.862	0.032
	DINOv2-SSL	**0.936**	**0.890**	**0.024**
CVC-ClinicDB	ImageNet-Sup	0.913	0.857	0.016
	DINOv2-SSL	**0.946**	**0.905**	**0.009**
CVC-ColonDB	ImageNet-Sup	0.875	0.803	0.012
	DINOv2-SSL	**0.906**	**0.835**	**0.011**
ETIS	ImageNet-Sup	0.824	0.734	**0.012**
	DINOv2-SSL	**0.908**	**0.835**	0.013

TABLE V
DECODER ABLATION WITH DINOV2-SSL + PARTIAL FINE-TUNING.

Dataset	Decoder	mDice ↑	mIoU ↑	MAE ↓
Kvasir-SEG	Complex	0.911	0.854	0.040
	Simple	**0.938**	**0.891**	**0.026**
CVC-ClinicDB	Complex	**0.951**	**0.910**	**0.009**
	Simple	0.948	0.903	0.010
CVC-ColonDB	Complex	0.920	0.857	0.010
	Simple	**0.923**	**0.861**	**0.009**
ETIS	Complex	0.891	0.814	0.013
	Simple	**0.912**	**0.845**	**0.009**

Fig. 3. Training loss curves across datasets.

our deliberately simplified U-shaped decoder, while keeping the same encoder (DINOv2) and partial fine-tuning strategy. Table V shows that the simple decoder is competitive or better on most datasets, supporting our design choice for deployment-friendly inference.

F. Convergence and Training Stability

Fig. 3 reports the training loss curves on all datasets. The loss decreases rapidly in the early stage and converges smoothly thereafter, without evident oscillation or divergence. Similar convergence patterns are observed across datasets with different sizes and characteristics, indicating that the proposed training strategy remains stable when adapting the pretrained backbone to different data distributions.

VI. CONCLUSION

This paper presents DINOv2-UNet, combining a self-supervised DINOv2 encoder with a lightweight U-shaped decoder and a partial fine-tuning strategy for colonoscopy polyp segmentation. Extensive experiments demonstrate that our streamlined design achieves robust segmentation accuracy and real-time inference across diverse datasets. These findings prove that properly adapted foundation models can serve as effective medical segmentation backbones without relying on complex, heavy decoders. Future work will explore parameter-efficient adaptation and extensions to multi-class volumetric segmentation.

GENERATIVE AI DISCLOSURE

We used a large language model (LLM) to assist with English polishing and clarity of presentation. All technical content, experimental design, results, and conclusions were produced and verified by the authors.

REFERENCES

[1] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: A Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, 2024.

[2] M. H. J. Maas *et al.*, "A computer-aided polyp detection system in screening and surveillance colonoscopy: an international, multicentre, randomised, tandem trial," *Lancet Digit. Health*, vol. 6, no. 3, pp. e157–e165, 2024.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[4] Z. Zhou, M. M. R. Siddiquee *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[5] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.

[6] D.-P. Fan, G.-P. Ji *et al.*, "PraNet: Parallel reverse attention network for polyp segmentation," in *MICCAI*, 2020, pp. 263–273.

[7] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[8] Z. Liu, Y. Lin *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, 2021, pp. 10 012–10 022.

[9] K. He, X. Chen *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. CVPR*, 2022, pp. 16 000–16 009.

[10] R. Azad *et al.*, "Dive into the details of self-supervised learning for medical image analysis: In-depth study," *Med. Image Anal.*, vol. 86, p. 103000, 2023.

[11] M. Moor, O. Banerjee *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[12] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *MICCAI*, 2021, pp. 14–24.

[13] A. Hatamizadeh, V. Nath *et al.*, "UNETR: Transformers for 3D medical image segmentation," in *Proc. WACV*, 2022, pp. 1748–1758.

[14] B. Xiao, J. Hu *et al.*, "CTNet: Contrastive transformer network for polyp segmentation," *IEEE Trans. Cybern.*, vol. 54, no. 9, pp. 5040–5053, 2024.

[15] M. Oquab, T. Darcet *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[16] S. Ali *et al.*, "A multi-centre polyp detection and segmentation dataset for generalisability assessment," *Sci. Data*, vol. 10, p. 75, 2023.

[17] D. Jha *et al.*, "Validating polyp and instrument segmentation methods in colonoscopy through Medico 2020 and MedAI 2021 challenges," *Med. Image Anal.*, vol. 99, p. 103307, 2025.

[18] F. Isensee, P. F. Jaeger *et al.*, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[19] D. Jha, P. H. Smedsrud *et al.*, "Kvasir-SEG: A segmented polyp dataset," in *MultiMedia Modeling*, 2020, pp. 451–462.

[20] J. Bernal, F. J. Sánchez *et al.*, "WM-DOJO: A benchmark for polyp detection in colonoscopy videos," in *MICCAI*, 2015, pp. 1–9.

[21] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[22] J. Silva, A. Histace *et al.*, "Toward embedded detection of polyps in colonoscopy videos," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2666–2673, 2014.