This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Self-Supervised Denoising Reconstruction Outperforms Foundation Models for Capsule Endoscopy Abnormality Classification

Vamshi Krishna Kancharla[1], Dasari Naga Raju[2], and Kaveti Pavan Kumar[3]

[1] Indian Institute of Science, Bangalore, India
vamshi.kancharla461@gmail.com
[2] Independent Researcher, Andhra Pradesh, India
raajuuu1998@gmail.com
[3] Hokkaido University, Sapporo, Hokkaido, Japan
pavankumar.kaveti.c5@elms.hokudai.ac.jp

Abstract. Video Capsule Endoscopy has become an indispensable diagnostic tool for gastrointestinal disorders due to its non-invasive nature and ability to capture high-resolution images of the small intestine. However, the enormous volume of data generated during a single procedure makes manual inspection labor-intensive, time-consuming, and prone to inter-observer variability. In this work, we propose a novel framework that combines self-supervised learning through a U-Net-based masked autoencoder with supervised feature extraction using EfficientNet-B7 for multi-class abnormality classification in Endoscopy images. The U-Net model is first trained in a self-supervised manner using Gaussian noise removal and masked reconstruction to learn robust visual representations without requiring annotations. The learned encoder features are then fused with EfficientNet-B7 features to form a rich, discriminative representation for classification. We evaluate our approach on the Capsule Vision 2024 Challenge dataset consisting of ten abnormality classes and a dominant normal class. Experimental results demonstrate that the proposed fusion framework achieves a macro-F1 of **0.842** and balanced accuracy of **0.940**, outperforming frozen Self-Supervised Learning ViT baselines including EndoViT and DINOv2. The study highlights the effectiveness of task-specific self-supervised representation learning and feature fusion in addressing class imbalance and improving diagnostic accuracy in real-world medical imaging scenarios.

Keywords: Video Capsule Endoscopy · Autoencoders · Self-Supervised Learning · Masked Image Modeling · UNet

1 Introduction

Conditions such as Crohn's disease, celiac disease, gastrointestinal bleeding, ulcers, polyps, and intestinal cancers often require early detection for effective

treatment and improved patient outcomes. Traditional endoscopic procedures, although effective, are invasive and uncomfortable for patients. Video Capsule Endoscopy (VCE) has emerged as a revolutionary, patient-friendly alternative that allows non-invasive visualization of the entire small intestine. A swallowable capsule equipped with a miniature camera captures tens of thousands of images as it traverses the GI tract, providing clinicians with valuable diagnostic information. Despite its clinical advantages, VCE introduces new challenges in data management and interpretation. A single examination can generate between 57,000 and over 1,000,000 frames, depending on the duration of the recording and frame rate. Manual review of these frames by gastroenterologists is extremely time-consuming, often requiring two to three hours per patient. This process is not only labor-intensive but also subject to fatigue, human bias, and inter-observer variability, which can lead to missed lesions or false positives. Artifacts such as bubbles, food debris, intestinal folds, and motion blur further complicate accurate interpretation.

Early automation in VCE relied on manual feature extraction, such as color histograms and texture descriptors, paired with standard classifiers [1]. These methods, however, lack the generalization power needed for diverse patient data. CNNs have since replaced hand-crafted features, becoming the standard for medical image analysis [2]. Among these, EfficientNet-B7 [3] stands out for its balance of accuracy and computational efficiency, achieving state-of-the-art results in fine-grained classification tasks. Several recent studies confirm that fusing EfficientNet features with other architectures improves detection rates for bleeding and ulcers [4,5]. However, most existing approaches rely heavily on supervised learning, which requires large amounts of annotated data. In the medical domain, labeling is expensive, time-consuming, and requires expert knowledge, making it difficult to scale. Furthermore, VCE datasets are typically highly imbalanced, with normal frames vastly outnumbering abnormal ones, often leading to biased models that perform well on the majority class but poorly on rare yet clinically important abnormalities. Self-supervised learning has recently gained attention as a powerful paradigm for learning meaningful representations from unlabeled data. Masked autoencoders [7] have shown strong performance in learning scalable and generalizable representations. While Masked Image Modeling was popularized by Vision Transformers [6], it is equally effective for CNN-based architectures in medical imaging. The U-Net, originally designed for segmentation, is particularly well-suited for this reconstruction task due to its skip connections, which preserve spatial details essential for identifying small lesions. More recently, large-scale pretrained Vision Transformers such as EndoViT [8] and DINOv2 [11] have been explored as frozen feature extractors paired with lightweight MLP classifiers, representing strong SSL baselines. However, such frozen-encoder approaches are inherently constrained by their pretraining distribution and cannot adapt to the specific noise and artifact characteristics of VCE data.

In this paper, we propose a novel framework that integrates a U-Net-based masked autoencoder trained in a self-supervised manner using Gaussian noise removal with supervised feature extraction using EfficientNet-B7. The learned

encoder features are fused with EfficientNet-B7 features and passed through fully connected layers for multi-class classification of ten clinically relevant VCE abnormalities: angioectasia, bleeding, erosion, erythema, foreign body, lymphangiectasia, polyp, ulcer, worms, and normal. By combining task-specific self-supervised pretraining with discriminative CNN features, we aim to improve diagnostic accuracy, reduce dependence on labeled data, and address class imbalance particularly on minority classes where frozen ViT encoders fall short. We evaluate our approach on the Capsule Vision 2024 Challenge dataset [9] and demonstrate that the proposed fusion framework achieves a macro-F1 of **0.842** and a balanced accuracy of **0.940**, outperforming both standalone CNN models and strong frozen SSL ViT baselines.

2 Dataset and Preprocessing

All experiments are conducted on the Capsule Vision 2024 Challenge dataset [9], consisting of high-resolution VCE frames from multiple patients and diverse clinical settings, with each image annotated by expert gastroenterologists into one of ten categories: angioectasia, bleeding, erosion, erythema, foreign body, lymphangiectasia, polyp, ulcer, worms, and normal. The dataset exhibits severe class imbalance the validation set contains 16,132 samples with the Normal class alone accounting for over 12,000 images, while classes such as worms, erythema, and angioectasia are comparatively rare. We strictly adhere to the official challenge splits with no external data or annotations.

Table 1. Per-class distribution of the Capsule Vision 2024 dataset used in all experiments. The Worms class is under-represented by a factor of 181 relative to Normal.

Class	Train	Validation
Normal	28,663	12,287
Erosion	2,694	1,155
Polyp	1,162	500
Angioectasia	1,154	497
Bleeding	834	359
Lymphangiectasia	796	343
Foreign Body	792	340
Erythema	691	297
Ulcer	663	286
Worms	158	68
Total	**37,607**	**16,132**

All images are resized to EfficientNet-B7's input resolution and color-normalized to reduce inter-patient and inter-device variability. Standard augmentation flips, rotations, zooming, and brightness adjustments is applied during training. For self-supervised pretraining, images are used without annotations and Gaussian

noise is injected to create corrupted inputs, enabling the U-Net to learn denoising and reconstruction representations.

3 Proposed Methodology

The proposed framework leverages both self-supervised and supervised learning paradigms to achieve robust classification of gastrointestinal abnormalities in VCE images. The overall pipeline consists of two major components: (1) a self-supervised U-Net-based masked autoencoder for learning rich visual representations through image reconstruction, illustrated in Fig. 2, and (2) a supervised EfficientNet-B7 backbone for discriminative feature extraction. These two streams are integrated through feature fusion and followed by a fully connected classifier. The comprehensive workflow of our proposed dual-branch network is illustrated in Fig. 1.

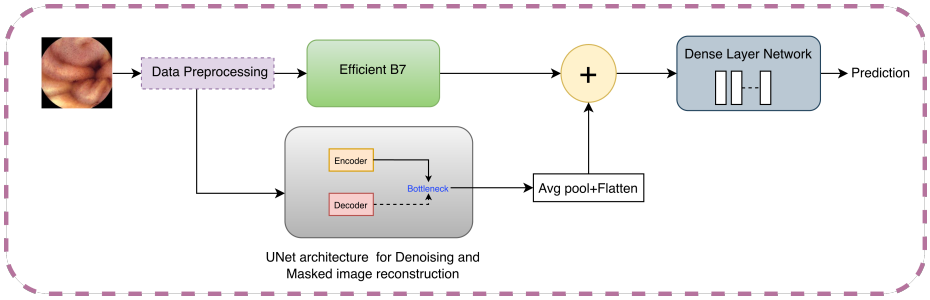


Fig. 1. Proposed block diagram combining EfficientNet-B7 and self-supervised U-Net for feature fusion and VCE abnormality classification.

3.1 Self-Supervised Reconstruction Module (U-Net)

The first stage of the framework focuses on learning meaningful representations from unlabeled VCE images. Inspired by masked autoencoders and denoising autoencoders, we design a U-Net architecture trained to reconstruct clean images from corrupted inputs. By forcing the network to recover missing or noisy regions, it must learn high-level semantic and structural information about the gastrointestinal anatomy and potential abnormalities. We explore two corruption strategies: random patch masking and Gaussian noise injection. Empirical evaluation showed that Gaussian noise removal produced superior reconstruction quality and more stable training, likely due to its ability to preserve global context while still challenging the network.

The U-Net architecture consists of an encoder with four convolutional blocks containing 32, 64, 128, and 256 channels, respectively, followed by a bottleneck

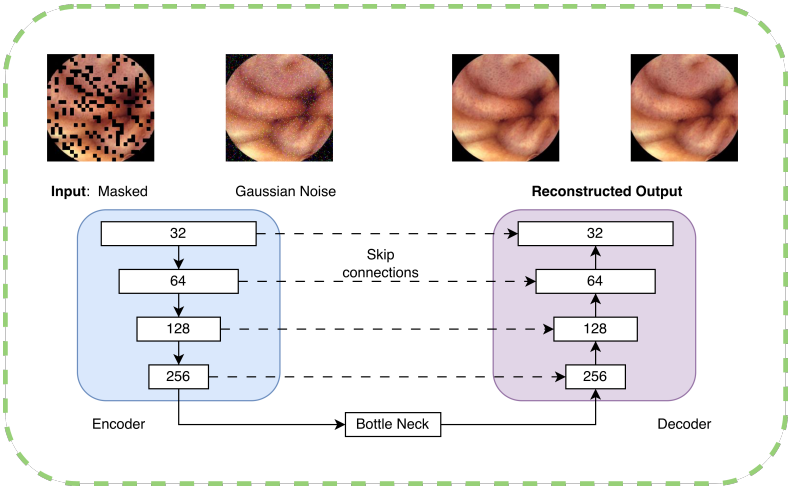


Fig. 2. U-Net architecture for denoising and masked image reconstruction used in the self-supervised pretraining stage.

layer with 512 channels. Each block includes convolution, batch normalization, and ReLU activation, along with max pooling for downsampling. The decoder mirrors the encoder structure with upsampling layers and skip connections that concatenate corresponding encoder features to preserve spatial information. The output layer uses a sigmoid activation to generate the reconstructed image. The model is trained using mean squared error between the output and the clean target image, enabling the encoder to capture high-level representations invariant to noise and artifacts common in VCE images.

3.2 Feature Fusion and Classification

In parallel to the self-supervised branch, we employ EfficientNet-B7 as a supervised feature extractor, chosen due to its strong performance and ability to capture both fine-grained details and global context. The network is initialized with ImageNet-pretrained weights and fine-tuned on the VCE dataset, with the global average pooling layer output used as the EfficientNet feature representation. After training the U-Net autoencoder, the encoder is used to extract features from input images, which are passed through average pooling and flattened to form a feature vector. This vector is then concatenated with the EfficientNet-B7 feature vector, resulting in a fused representation that combines reconstruction-based and discriminative features. The fused vector is fed into fully connected layers with ReLU activation and dropout for regularization, with the final layer using softmax activation to output class probabilities for the ten target classes. All models are trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 256, with categorical cross-entropy as the loss function. The self-

supervised U-Net is pretrained first, followed by supervised fine-tuning of the fusion model.

4 Experiments and Results

4.1 Comparison with SSL ViT Baselines

We compare against frozen SSL ViT baselines using two strong pretrained encoders: **EndoViT** [8], pretrained on a large endoscopic image corpus, and **DINOv2** [11], pretrained via self-distillation on diverse natural images. For each backbone we report both the MLP-only baseline and the best-performing variant found (hard class augmentation for EndoViT; Conditional DDPM for DINOv2), ensuring a fair and rigorous comparison against their strongest configurations. We report macro-averaged F1-score and balanced accuracy as primary metrics, as Macro-F1 treats all classes equally and is especially informative for minority-class performance, which is clinically critical in the presence of severe class imbalance.

Table 2. Comparison against frozen SSL ViT baselines on the Capsule Vision 2024 validation set.

Model	Macro-F1	Balanced Accuracy
EndoViT+ MLP	0.7769	0.7678
EndoViT+ Hard Augmentation	0.7740	0.7598
DINOv2+ MLP baseline	0.7871	0.7733
DINOv2+ Condition DDP	0.7945	0.7809
Proposed Fusion (Ours)	**0.842**	**0.940**

Table 2 reveals a consistent results across paradigms. The EndoViT MLP baseline achieves a macro-F1 of 0.7769 and balanced accuracy of 0.7678; even its best variant (hard class augmentation) reaches only 0.7740, with balanced accuracy slightly dropping to 0.7598. DINOv2 fares better overall, with the MLP baseline at 0.7871 and its best variant (Conditional DDPM) at 0.7945 and balanced accuracy of 0.7809. Despite these being strong, large-scale pretrained encoders, their frozen backbones cannot adapt to the specific noise, artifact, and class-imbalance characteristics of VCE data. The proposed Efficient Fusion U-Net surpasses all baselines with a macro-F1 of 0.842 and a balanced accuracy of 0.940, confirming that task-specific reconstruction pretraining combined with fine-tuned CNN features yields representations far more suited to this domain than frozen ViT encoders.

4.2 Class-Wise Performance Analysis

A detailed per-class analysis in Table 3 reveals important insights. The model achieves very high precision and recall for critical classes such as ulcer (F1 =

0.982), normal (F1 = 0.982), and worms (F1 = 0.971), while angioectasia, bleeding, and lymphangiectasia also show strong performance with F1-scores above 0.83. Performance on erythema and polyp is comparatively lower, with F1-scores of 0.572 and 0.721 respectively, likely due to subtle visual appearance and limited sample sizes. The macro-averaged F1-score of 0.842 demonstrates good balance across classes, while the weighted F1-score of 0.939 reflects strong overall performance, confirming that self-supervised feature learning and fusion help mitigate class imbalance.

Table 3. Detailed classification metrics for the proposed Efficient Fusion U-Net.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Angioectasia	0.865	0.813	0.838	497
Bleeding	0.840	0.822	0.831	359
Erosion	0.785	0.764	0.774	1,155
Erythema	0.614	0.535	0.572	297
Foreign Body	0.911	0.844	0.876	340
Lymphangiectasia	0.888	0.854	0.871	343
Normal	0.978	0.986	0.982	12,287
Polyp	0.692	0.752	0.721	500
Ulcer	0.989	0.976	0.982	286
Worms	0.944	1.000	0.971	68
Overall Accuracy			**0.940**	16,132
Macro Avg	0.851	0.835	0.842	16,132
Weighted Avg	0.939	0.940	0.939	16,132

5 Discussion

Our experiments reveal two complementary insights about self-supervised learning for VCE classification. The gap between frozen SSL ViT baselines and our proposed method is addressing EndoViT and DINOv2, despite strong pretraining on large datasets, achieve macro-F1 scores of only 0.777 and 0.787 with frozen backbones. This limitation arises because generic, and even endoscopy-domain, pretraining does not expose the encoder to the specific corruption patterns such as motion blur, specular reflections, and intestinal fluid artifacts that are prevalent in VCE.A frozen representation, regardless of its richness, cannot adapt to these domain-specific characteristics. In contrast, our U-Net is explicitly trained to denoise and reconstruct VCE images, compelling the encoder to internalize these characteristic degradation patterns through its reconstruction objective. Furthermore, among CNN-based self-supervised approaches, Gaussian noise-based reconstruction provides a more stable pretext task than masked patch modeling. VCE images exhibit strong spatial continuity, and patch masking introduces artificial discontinuities that allow the decoder to rely on local

interpolation or shortcut features rather than learning semantically meaningful representations. In contrast, Gaussian noise preserves the global spatial structure while still requiring the encoder to recover fine-grained details, resulting in more robust and transferable feature representations.

Regarding feature fusion, simple concatenation outperformed the attention-based(Foundation models) variant, suggesting that the features from both branches are already complementary and highly discriminative. Minority classes, such as erythema and polyps, remain challenging due to their subtle visual characteristics and limited number of training samples. Nevertheless, the improved recall achieved on these classes compared with all baseline methods indicates that VCE-specific self-supervised pretraining provides an effective regularization mechanism, enhancing the model's ability to learn discriminative representations even under severe class imbalance.

From a clinical perspective, the high detection performance achieved for critical abnormalities, including bleeding (F1 = 0.831) and ulcers (F1 = 0.982), highlights the potential of the proposed framework as a computer-aided triage system for prioritizing clinically significant VCE frames during review. Nevertheless, several limitations remain. First, the current framework processes each frame independently and does not exploit temporal information, despite the fact that gastrointestinal abnormalities in VCE frequently extend across consecutive frames. Incorporating temporal modeling through sequence-based architectures may further improve diagnostic consistency and robustness. Second, the computational complexity of the EfficientNet-B7 backbone may limit deployment in real-time or resource-constrained clinical settings. Future work will therefore investigate model compression and knowledge distillation techniques to transfer the learned representations to lightweight architectures while preserving diagnostic performance.

6 Conclusion

This paper presented a dual-branch framework integrating task-specific self-supervised U-Net pretraining with EfficientNet-B7 feature fusion for VCE abnormality classification. Against frozen SSL ViT baselines including EndoViT and DINOv2, our proposed method achieves a macro-F1 of **0.842** and balanced accuracy of **0.940**, this states that domain-adaptive reconstruction pretraining yields representations that large-scale frozen models cannot match despite their scale advantage. Looking ahead, we aim to incorporate temporal modeling to exploit frame-to-frame consistency and investigate class-balancing loss functions to further improve performance on rare pathologies such as erythema and polyps.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Okeke, S., et al.: Precision enhancement in wireless capsule endoscopy: A Review. Front. Artif. Intell. **8**, 1529814 (2025). https://doi.org/10.3389/frai.2025.1529814
2. Soffer, S., et al.: Recent technological advances in video capsule endoscopy: a comprehensive review. Clin. Endosc. **58**(1), 1–12 (2025). https://doi.org/10.5946/ce.2025.135
3. Togacar, M., Cömert, Z., Ergen, B.: A Novel Multi-Feature Fusion Method for Classification of Gastrointestinal Diseases Using Endoscopy Images. Diagnostics **12**(10), 2316 (2022). https://doi.org/10.3390/diagnostics12102316
4. Kundu, R., et al.: Classification of wireless capsule endoscopy images for bleeding using deep features fusion. Sci. Rep. **12**, 1458 (2022). https://doi.org/10.1109/ICECCME55909.2022.9987916
5. Khan, M.A., et al.: Deep Feature Fusion and Optimization-Based Approach for Stomach Disease Classification. Sensors **22**(7), 2801 (2022). https://doi.org/10.3390/s22072801
6. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF ICCV, pp. 9650–9660 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
7. He, K., et al.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF CVPR, pp. 16000–16009 (2022). https://doi.org/10.1109/CVPR52688.2022.01553
8. Batic, D., et al.: EndoViT: pretraining vision transformers on a large collection of endoscopic images. Int. J. Comput. Assist. Radiol. Surg. **19**, 13–21 (2024). https://doi.org/10.1007/s11548-024-03091-5
9. Handa, P., et al.: Capsule vision 2024 challenge: Multi-class abnormality classification for video capsule endoscopy. arXiv preprint arXiv:2408.04940 (2024). https://doi.org/10.48550/arXiv.2408.04940
10. Handa, P., et al.: Training and Validation Dataset of Capsule Vision 2024 Challenge. Figshare (2024). https://doi.org/10.6084/m9.figshare.26403469.v1
11. Oquab, M., et al.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024). https://arxiv.org/abs/2304.07193