Abnormality Detection in PET Imaging Using Foundation Models: A Comparative Study and Analysis

Nathan Vandekerckhove, Ide Van den Borre, Simon De Keyser, Maarten Larmuseau, and Aleksandra Pižurica

***Abstract*—The escalating clinical demand for positron emission tomography (PET) is accompanied by a heterogeneity in scanners, tracers and acquisition protocols, making PET data complex to interpret, especially when the lesions of interest are subtle or minute. Motivated by this challenge, we investigate abnormality detection on PET scans through the lens of large-scale foundation models and other pre-trained vision encoders, whose broad applicability has recently energised deep-learning research. We present the first systematic 2-D study that contrasts such pre-trained backbones with traditional, task-specific architectures in both binary classification and semantic segmentation. For classification, embeddings were extracted from frozen encoders (DINOv2, RAD-DINO, ConvNeXt, etc.) and fed to a lightweight fully-connected head, while a traditional CNN provided the baseline. For segmentation, frozen pretrained encoders–both off-the-shelf and custom-built–are grafted onto the nnU-Net pipeline, exploring information fusion at the bottleneck and across multiple feature levels. Experiments demonstrate that DINOv2-based embeddings achieve the most promising results in classification, surpassing the CNN baseline across all metrics while requiring only lightweight training once embeddings are harvested. For segmentation, plain nnU-Net remains superior, indicating that PET-specific fine-tuning of general-purpose encoders may be necessary. These findings chart the promise and current limits of foundation models for clinical PET analysis.**

***Index Terms*—Positron emission tomography, Foundation models, Anomaly detection, Transfer Learning, Classification, Semantic segmentation.**

I. Introduction

Positron emission tomography (PET) is a crucial diagnostic imaging modality, particularly in oncology [1]. PET scans offer valuable insights into cellular metabolic activities, making them highly effective for detecting and managing various types of cancer. PET imaging has demonstrated particular value in identifying and characterizing abnormalities such as tumors, lesions, and other anomalies that may not be clearly visible with traditional anatomical imaging modalities. Development of reliable automatic analysis techniques for PET images is desirable as a supportive tool for radiologists. Yet, despite significant advances in image processing and deep learning, fully automatising PET image analysis and interpretation remains very challenging, mainly because current deep-learning methods still require large amounts of labelled data. Accurate labeling of anomalies requires specialized training, is inherently subjective, and consumes substantial time and resources. The complexity of PET images, characterized by varying signal intensities, radiotracer variability and the frequent presence of subtle, small-scale abnormalities, further compounds these challenges. Additionally, the scarcity of qualified medical imaging specialists exacerbates this issue, resulting in limited availability of accurately labeled datasets [2]. Consequently, there is a growing interest in automating abnormality detection using advanced computational methods, aiming to overcome these limitations by enhancing diagnostic consistency, reducing reliance on human annotations, and facilitating rapid and reliable interpretation of large imaging datasets. Automated abnormality detection can significantly streamline clinical workflows, enable timely diagnosis, and ultimately improve patient management and outcomes [3].

Concurrent with these developments in medical imaging, the field of deep learning has experienced significant advancements, notably through the emergence of foundation models [4]. Foundation models are large-scale neural networks pre-trained on extensive and diverse datasets, capable of generalizing well across various downstream tasks. These models have demonstrated remarkable efficacy in fields ranging from natural language processing to computer vision, often outperforming traditional, task-specific models by leveraging their extensive prior knowledge. Given the limited availability of labeled PET datasets, foundation models present a promising avenue for enhancing automated abnormality [5].

This thesis seeks to explore the applicability and effectiveness of these foundation models, alongside other prominent pre-trained architectures, in detecting abnormalities within PET images. Building on this aim, we deliver the first systematic head-to-head evaluation of these models–spanning both off-the-shelf and newly crafted architectures, from lightweight frozen CNNs to large vision transformers–and analyse their performance-versus-complexity trade-offs across two key tasks: slice-wise binary classification of normal versus abnormal images and voxel-wise segmentation that precisely delineates abnormal regions.

The paper is structured as follows: Section II surveys related work. Section III explores the datasets used and explains the main methodology for both classification and segmentation approaches. Section IV presents and analyzes the results, while Section V concludes the study and outlines directions for future research.

II. Related Work

Early work on automated PET abnormality detection relied on modest, task-specific CNNs trained from scratch on single-centre datasets [6]. By contrast, large-scale 3-D pre-training on

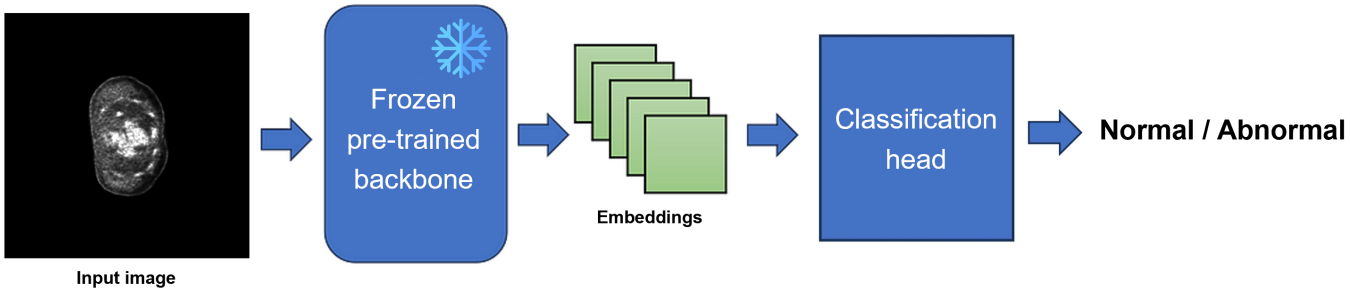

Fig. 1. General pipeline of the classification process using pre-trained models.

aggregated CT and MRI archives, exemplified by MedicalNet and Med3D, has repeatedly boosted performance when annotations are scarce [7]. Motivated by these gains, a handful of exploratory studies have begun porting natural-image vision transformers such as DINOv2 to organ-specific PET tasks, though the evidence remains anecdotal and confined to narrow applications [8]. Promptable segmentation frameworks like SAM have likewise inspired medical variants (MedSAM, MobileSAM), yet early reports reveal sharp performance drops on the low-contrast metabolic boundaries typical of PET lesions [9]. Consequently, the field still lacks a systematic comparison of frozen foundation encoders–spanning both CNN and ViT families–against task-specific baselines for PET abnormality analysis. This work closes that gap by providing the first systematic evaluation of foundation and other pre-trained vision models on PET imaging–a modality still far less explored than CT or MRI in deep-learning research–quantifying performance-versus-complexity trade-offs in both slice-level classification and voxel-level segmentation.

III. Methodology

A. Datasets

We conduct all experiments on AutoPET III, the MICCAI 2024 benchmark for whole-body tumor segmentation in PET/CT [10], [11]. The dataset contains 1 611 anonymised scans acquired at two centres and spans two complementary tracers: FDG (1 014 studies from 900 cancer patients plus 513 negative controls) and PSMA (597 studies from 378 prostate-cancer patients). Each case provides co-registered PET and diagnostic-quality CT volumes, expert 3-D lesion masks verified by nuclear-medicine physicians, and basic patient metadata. Both raw counts and SUV-scaled PET intensities are released in DICOM, NIfTI, and HDF5 formats, while the 597 lesion-free scans enable stringent false-positive assessment–making AutoPET III a demanding yet realistic test bed for generalisable segmentation algorithms across scanners and tracers.

B. Classification Architecture

We converted the 1 611 AutoPET III whole-body scans into a uniformly formatted 2-D dataset by resampling each PET volume to 400 × 400 × 300 voxels and slicing it axially into 300 grayscale images, paired with aligned masks to yield 483 300 labelled slices ($\approx 20\%$ abnormal). Patient-wise splits (75%/15%/10%) prevented leakage, and every 224 × 224 slice was intensity-scaled with a normalisation that divides the image by its own maximum SUV before linear mapping. This single preprocessing pipeline was applied to all baseline and transfer-learning models, ensuring that performance differences reflect the learning algorithms rather than data handling.

To anchor subsequent experiments we train a compact three-block convolutional neural network that follows a conventional pattern of convolution, batch normalisation, ReLU activation and max pooling, culminating in three fully connected layers and a sigmoid output. Weighted sampling is applied during optimisation so that every minibatch contains equal numbers of normal and abnormal slices, which prevents the network from collapsing to the majority class and makes the baseline a fair yardstick against which more sophisticated approaches can be measured.

The core of the classification study is a two-stage transfer-learning pipeline, visualized in Figure 1 that exploits powerful encoders trained on large-scale natural or medical image collections. In the first stage each preprocessed slice is passed through a frozen backbone–ConvNeXt-Base, ResNet-50, DINOv2-Small, DINOv2-Base or RAD-DINO–and the resulting embedding vector is extracted without any gradient updates to the encoder itself [9], [12]–[14]. For convolutional networks the embedding is taken from the penultimate pooling layer, whereas for vision transformers it is the representation attached to the global classification token. The dimensionalities range from 384 for DINOv2-Small to 2 048 for ResNet-50, but because all feature maps are written once to disk this computationally intensive step is performed a single time per model and the same immutable embeddings are reused across all subsequent classification experiments. Storing them also eliminates the need to load entire networks when only the downstream head is under investigation. In the second stage a lightweight classifier consisting of a dropout layer and a single linear transformation converts each embedding into a raw logit. Training this head resembles traditional logistic regression on fixed features: the only learnable parameters are the weights and bias of the final linear layer, so optimisation converges quickly and can be repeated with varied hyper-parameters,

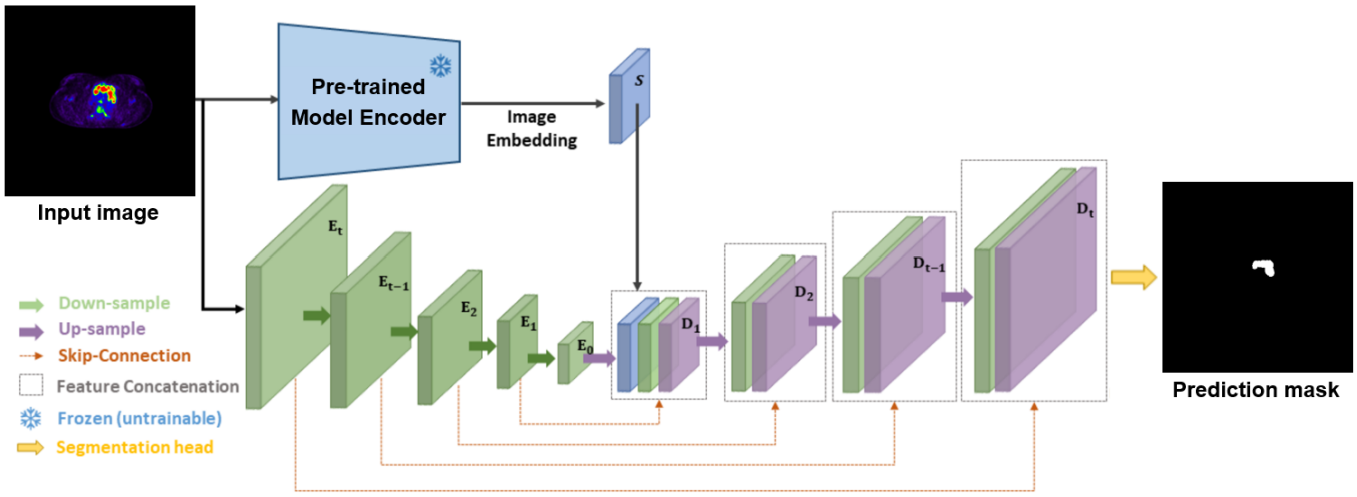

Fig. 2. General pipeline of the nnU-Net variants extended with an additional encoder of pre-trained models.

class-balance techniques or validation criteria at negligible cost. Because the backbone is frozen, performance differences isolate the representational power of solely the encoder.

Each model–baseline CNN or encoder-plus-head–is evaluated on the test set using accuracy, precision, recall, specificity, F1, ROC-AUC and average precision. Confusion, ROC and precision–recall curves supplement these scores

By freezing rich encoders and restricting training to a minimal classification head the proposed approach delivers several advantages: reproducibility is guaranteed because feature extraction is deterministic once completed; limited hardware resources are sufficient for experimentation because only a handful of parameters are updated; and researchers can iterate rapidly over imbalance remedies, calibration methods or voting ensembles without rerunning costly backbone passes. The baseline CNN anchors these findings in a familiar fully supervised framework, yet the transfer-learning results demonstrate how much more effectively large-scale pretraining captures the subtle patterns that distinguish normal from pathological uptake in PET slices.

C. Segmentation Architecture

Within the segmentation arm of our methodology we begin with a conventional two-dimensional nnU-Net baseline and then ask how much benefit can be harvested by injecting powerful, frozen feature extractors that were pre-trained on vast external corpora [15]. All experiments draw on the AutoPET III challenge data: 1 611 whole-body PET/CT studies of which 1 461 scans supply training folds while 150 are withheld for final testing. In every setting the PET channel alone is used, each volume is resampled to a common voxel spacing, and nnU-Net's built-in planner selects a patch size of 384 × 384 and an automated intensity normalisation (sample-specific z-score normalization). Training proceeds on random axial patches; inference re-assembles slice-wise predictions into full three-dimensional masks so that voxel-level overlap with the expert annotations can be computed.

The reference model is a standard 2D nnU-Net whose encoder and decoder are learned from scratch on the PET data. It is optimised with a composite Dice + binary-cross-entropy loss, uses class-balanced sampling to combat the extreme foreground scarcity, and is trained on progressively smaller fractions of the data (100%, 50%, 10%, 5%, 1%) so that later comparisons can reveal how quickly performance degrades once annotations become scarce.

To explore transfer learning three variants keep the nnU-Net decoder intact but augment the encoding path with a parallel, frozen backbone. In every case only the nnU-Net parameters are updated; gradients are never propagated into the external encoder, isolating the value of its pretrained representation.

The first variant, nnSAM, is an already existing extension of nnU-Net that couples the standard encoder with MobileSAM, a distilled version of the Segment-Anything Model that trades capacity for efficiency while preserving the rich spatial semantics learned from the SA-1B dataset [16], [17]. PET slices are resized to 1 024 × 1 024 to meet the SAM input contract, then passed through the frozen encoder. The resulting 64 × 64 feature map is interpolated to the spatial dimensions of nnU-Net's bottleneck and concatenated channel-wise with the native nnU-Net features right before the decoding stage begins. All subsequent layers, including the skip connections and decoder, are identical to the baseline. By supervising the combined representation with exactly the same Dice + BCE objective we ensure that any gain can be attributed to the pretrained semantics rather than to auxiliary losses or multitask heads. To gauge data-efficiency the full range of training fractions (100% to 1%) is repeated for nnSAM.

The second family, nnConvNeXt, two nnU-Net extensions newly developed for this study, investigates whether modern convolutional backbones can provide a stronger inductive bias than SAM's transformer-based design. In nnConvNeXt1 the

TABLE I
CLASSIFICATION PERFORMANCE COMPARISON OF ALL CONFIGURATIONS.

Model	Acc	AP	ROC-AUC	F1	Prec	Rec
DINOv2 Small	0.90	0.77	0.92	0.69	0.70	0.68
DINOv2 Base	**0.91**	**0.79**	**0.92**	**0.71**	0.71	**0.71**
RAD-DINO	0.89	0.72	0.90	0.66	0.66	0.66
ConvNeXt	0.90	0.77	0.91	0.69	0.70	0.68
ResNet	0.89	0.75	0.91	0.68	0.66	0.71
Baseline CNN	0.86	0.74	0.90	0.68	**0.73**	0.63

TABLE II
AVERAGE SEGMENTATION METRICS FOR ALL MODEL CONFIGURATIONS TRAINED ON 10% OF THE TRAINING DATASET.

Model	Dice	IoU	FPV (ml)	FNV (ml)
nnU-Net	**0.47**	**0.35**	8.61	**12.98**
nnSAM	0.45	0.34	8.83	14.26
nnConvNeXt1	0.46	0.34	9.84	15.80
nnConvNeXt2	0.44	0.33	10.71	14.88
nnDINO	0.45	0.34	**7.15**	14.48

ConvNeXt-Base encoder simply replaces MobileSAM in the bottleneck fusion scheme described above: slices are up-scaled to 1 024 × 1 024, pushed through ConvNeXt, and the deepest feature map is merged with nnU-Net at the bottleneck. In nnConvNeXt2 a more ambitious multi-level fusion is explored: ConvNeXt activations are harvested at four stratified stages (spatial strides 4, 8, 16, 32), each map is size-matched to its counterpart in nnU-Net, and the pairs are concatenated before the decoder consumes them. This design lets the network exploit pretrained features at native resolutions rather than relying solely on the bottleneck encoding. For both nnConvNeXt configurations, training was performed only once using the 10% subset of AutoPET III.

The final variant, nnDINO, also custom-designed, evaluates transformer features learned through self-supervised knowledge distillation. A frozen DINOv2-Base encoder ingests each PET slice after it has been resized to 224 × 224 and replicated across three channels. The classification token is discarded, leaving a grid of 14 × 14 patch embeddings (dimension 768) that is reshaped into a dense feature map. This map is interpolated to match bottleneck level of nnU-Net and concatenated there. As with the ConvNeXt experiments, the DINO variant is trained on 10% of the scans to highlight data-efficiency without incurring excessive computational overhead.

All five architectures–baseline, nnSAM, nnConvNeXt1, nnConvNeXt2, and nnDINO–are benchmarked on the held-out test cohort with the same quartet of overlap metrics: Dice, Intersection-over-Union (IoU), false-positive volume (FPV) and false-negative volume (FNV). While Dice and IoU reveal proportional agreement between prediction and truth, FP_{vol} and FN_{vol} translate those ratios into clinically meaningful units by summing the cubic centimetres of tissue that are respectively over-segmented or missed outright. Reporting both volumes is essential because two models can share an identical Dice yet differ drastically in the absolute amount of spurious uptake they hallucinate, a crucial distinction when false positives trigger costly follow-up scans or invasive biopsies. Finally, because Dice and IoU degenerate on "normal" scans whose ground-truth masks are empty, we exclude such cases from their computation. Presenting all four overlap metrics together yields a balanced evaluation that captures both proportional performance and clinically relevant error volumes.

Together these experiments chart a clear trajectory: start from a robust nnU-Net baseline, introduce frozen encoders that inject external knowledge at carefully chosen depths, and quantify gains not only in fractional overlap but also in the real-world volumes of tissue added or missed. By varying both dataset size and backbone design, the study shows when transfer learning truly pays off and which representations deliver the greatest boost per annotation.

IV. RESULTS AND DISCUSSION

A. Classification

The classification study used a two-step design that isolates backbone effects from preprocessing. After fixing a single, common preprocessing pipeline, we evaluated six architectures–ConvNeXt, ResNet-50, DINOv2-Small, DINOv2-Base, RAD-DINO and a custom CNN–across accuracy, average precision, ROC-AUC, F1, precision and recall. This setup lets the analysis focus squarely on how backbone choice shapes performance. Table I summarises the headline numbers. DINOv2 Base leads in every score, with DINOv2 Small and ConvNeXt essentially tied for second place, ResNet-50 a small step behind, and RAD-DINO trailing the transformer group despite its medical fine-tuning. The lightweight CNN baseline finishes last but still reaches a respectable 0.86 accuracy, 0.90 ROC-AUC and 0.74 AP, confirming that domain-specific convolutional features can capture a large fraction of the signal even without external data. Full ROC and precision–recall curves in the appendix reinforce

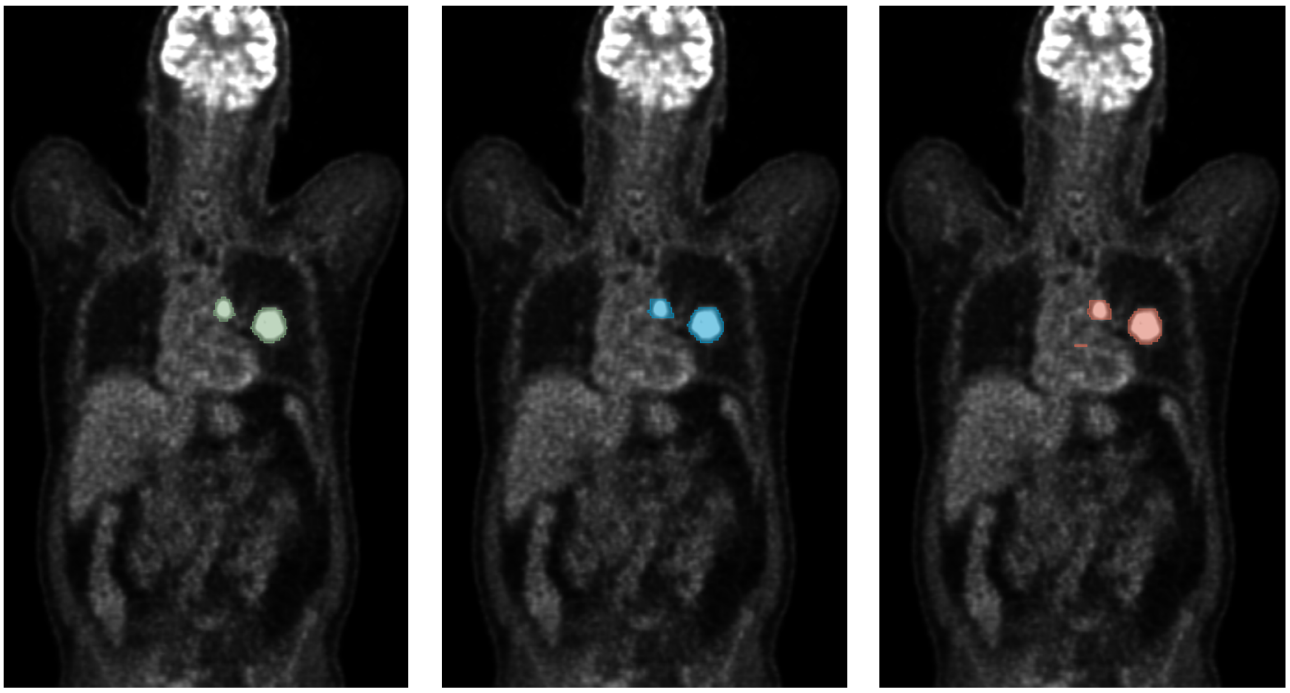

Fig. 3. Representative PET slice with ground-truth mask (green), nnSAM prediction (blue) and nnU-Net prediction (red).

the numerical hierarchy and reveal wider margins at clinically relevant false-positive operating points.

The transformer advantage aligns with the notion that PET lesions may appear anywhere in the slice and often manifest as subtle, distributed uptake; global self-attention therefore provides a favourable inductive bias. ConvNeXt narrows the gap by importing transformer-style design decisions–large kernels, LayerNorm, inverted bottlenecks–while retaining the efficiency of convolutions. ResNet-50, representing a more traditional CNN, cannot fully bridge that context deficit. RAD-DINO's unexpectedly weak showing highlights a modality mismatch: features tuned on grayscale chest X-rays do not transfer neatly to radionuclide distributions in PET and occasionally suppress cues that generic visual pre-training preserves. This suggests that domain adaptation helps only when the source and target share low-level statistics, motivating future work on a PET-specific "DINOv2-PET" corpus.

Computational considerations favour the frozen-backbone pipeline. Extracting embeddings is the only heavy step and must be done just once; afterwards, training a one-layer classifier completes in seconds and can be repeated for extensive hyper-parameter sweeps. In contrast, the from-scratch CNN needs full back-prop through all pixels each time a design variant is tested, turning even modest architecture searches into multi-hour endeavours. For deployment the situation reverses: the CNN is a compact monolith, whereas transformer pipelines require a separate embedding cache unless the encoder is bundled into the inference graph. Nonetheless, slice-wise inference with a frozen DINOv2 Base remains comfortably sub-second on modern GPUs, making the trade-off attractive for most clinical workflows.

Implementation overhead is similarly asymmetric. Pre-trained models reduce code to a handful of lines–load backbone from timm, disable gradients, forward pass, attach linear layer–whereas the custom CNN demanded numerous design choices on kernel sizes, paddings, activation placement and regularisation. The gain in simplicity is therefore commensurate with the gain in accuracy.

Summarising the league table: DINOv2 Base clearly dominates; DINOv2 Small and ConvNeXt form a strong second tier; ResNet-50 is competent but dated; RAD-DINO underperforms due to modality drift; the baseline CNN sets a credible floor. Together these results show that generic vision transformers trained on diverse natural images already transfer well to PET slice classification, whereas medical-specific fine-tuning must match the target modality to be effective. Future work will explore selective fine-tuning of upper transformer blocks and the acquisition of PET-centric pre-training data to push beyond the current state-of-the-art.

B. Segmentation

Segmentation performance was first compared between nnU-Net and its SAM-augmented variant (nnSAM) across a series of training-set sizes ranging from 1% to 100%. Results show a consistent trend: nnU-Net outperforms nnSAM on the core overlap metrics, Dice and IoU, at every training fraction. The performance gap is widest at the extremes: when trained on just 1% of the data, nnU-Net scores 0.29 on Dice versus 0.25 for nnSAM; at 100%, the scores are 0.57 and 0.55, respectively. This suggests that the added encoder in nnSAM, while rich in generic semantic features, introduces redundancy

or noise that complicates learning under limited supervision. The volumetric error metrics tell a more nuanced story: nnSAM occasionally achieves lower false-negative volumes, implying a slight bias toward more liberal inclusion of tumor voxels, but this often comes with increased false positives. A representative slice in Fig. 3 shows the trade-off: nnU-Net captures the primary tumor cleanly, whereas nnSAM adds an extra false-positive focus. Overall, nnU-Net remains the more balanced and reliable option across dataset sizes.

Focusing next on the 10% setting, the central scope of this analysis, we benchmarked five architectures: baseline nnU-Net, nnSAM, two ConvNeXt-based hybrids, and a transformer-enhanced nnDINO model. As summarised in Table II, the unmodified nnU-Net again leads in Dice, IoU, and false-negative volume, highlighting its strong inductive bias and effective use of limited training data. Among the hybrids, nnConvNeXt1 came closest, matching the IoU of nnU-Net and trailing by just 0.01 Dice points, though with higher missed volume. nnDINO and nnSAM offer comparable Dice and IoU, but differ in their error profiles. nnDINO reduces false positives more effectively, while nnSAM catches slightly more lesions. The second ConvNeXt variant (nnConvNeXt2), which employs multi-scale fusion, underperforms across all metrics, indicating that injecting pre-trained features at multiple encoder levels can overwhelm the decoder rather than enhance its capacity.

Despite differences in encoder origin–foundation models like SAM and DINOv2 versus conventional supervised ConvNeXt–all pre-trained variants cluster within a narrow Dice range of 0.44–0.46, and nearly identical IoU. This convergence suggests that frozen encoders, when used without fine-tuning, offer limited practical advantage. The decoder appears to dominate performance, and unless the added features are adaptively integrated or aligned with domain-specific priors, their utility remains constrained.

In sum, while encoder-augmented variants show potential, the fully trainable nnU-Net continues to outperform them in low-data PET segmentation. Further gains will likely depend not on more pre-trained features, but on smarter integration and domain-aware adaptation.

V. Conclusion

This work benchmarked modern deep-learning pipelines for slice-wise classification and voxel-wise segmentation of PET images. A comparison of frozen encoders with a single linear head showed that DINOv2-Base–a large-scale self-supervised vision transformer–performed best in accuracy, AP, ROC-AUC and F1, whereas RAD-DINO, tuned on X-rays, lagged, underscoring the need for modality-matched pre-training. For segmentation, a fully trainable nnU-Net remained dominant; adding frozen MobileSAM, ConvNeXt or DINOv2 features increased complexity yet delivered only marginal, mostly negative, gains, confirming that naïve encoder fusion cannot outdo nnU-Net's carefully engineered pipeline. Looking ahead, PET-centric foundation models, selective encoder fine-tuning, smarter multi-scale fusion, stronger domain generalisation and self-supervised pre-training on unlabelled scans are promising paths to close the remaining gap between research prototypes and routine clinical deployment.

Acknowledgment

This work was supported by the Special Research Fund of Ghent University under Grant BOF/IOP/2022/038, and by the Flanders AI Research Programme under Grant 174B09119.

References

[1] M. Fisher, "PET scan volumes continue to grow new IMV report shows," 2024. [Online]. Available: https://www.dotmed.com/news/story/62663

[2] Success Pitchers, "Medical image annotation challenges to overcome for better healthcare AI," 2023. [Online]. Available: https://successpitchers.com/medical-image-annotation-challenges-to-overcome-for-better-healthcare-ai/

[3] European Enterprise Network, "AI in nuclear medicine for automatic abnormalities detection on PET/CT FDG," 2023. [Online]. Available: https://een.ec.europa.eu/partnering-opportunities/ai-nuclear-medicine-automatic-abnormalities-detection-pet-ct-fdg

[4] F. Hashimoto and LastName, "Deep learning–based pet image denoising and reconstruction," 2024. [Online]. Available: https://link.springer.com/article/10.1007/s12194-024-00780-3

[5] F. Shamshad *et al.*, "Advances in medical image analysis with vision transformers," 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1361841523002608

[6] K. Kawauchi, S. Furuya, K. Hirata, C. Katoh, O. Manabe, K. Kobayashi, S. Watanabe, and T. Shiga, "A Convolutional Neural Network-Based System to Classify Patients Using FDG PET/CT Examinations," *BMC Cancer*, vol. 20, no. 227, pp. 1–11, 2020.

[7] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer Learning for 3D Medical Image Analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[8] J. Sun, Q. Zhao, and H. Li, "Adapting DINO for Self-Supervised Liver-Lesion Classification in PET Imaging," in *Proc. IEEE Int. Symp. Biomedical Imaging*, 2023.

[9] J. Ma, B. Wang, W. Kuo, X. Zhang, and Y. Xie, "Segment Anything in Medical Images," 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[10] S. Gatidis, T. Hepp, M. Fruh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Kustner, C. Cyran, and D. Rubin, "A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions," *Scientific Data*, vol. 9, no. 1, p. 601, October 4 2022.

[11] M. Ingrisch, J. Dexl, K. Jeblick, C. Cyran, S. Gatidis, and T. Kuestner, "Automated Lesion Segmentation in Whole-Body PET/CT." https://autopet-iii.grand-challenge.org/, 2024.

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, available at: https://arxiv.org/abs/2201.03545.

[13] H. Liu, Z. Zhang, H. Zeng, Y. Wang, J. Wang, Y. Wang, J. Fu, and S. Lu, "Fast Segment Anything," 2024, available at: https://arxiv.org/abs/2401.10815.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, available at: https://arxiv.org/abs/1512.03385.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, available at: https://arxiv.org/abs/1505.04597.

[16] Y. Zhang, H. Dong, J. Du, X. Yu, Z. Zhou, L. Xie, and Y. Xu, "Segment Anything in Medical Images," 2023, available at: https://arxiv.org/abs/2309.16967.

[17] X. Deng, J. Zhang, Z. Wu, P. Nie, Y. Tang, Z. Wang, P. Gao, Y. Bai, Y. Wang, C. Dong, Y. Shi, X. Cao, and X. Yang, "Segment Anything Meets Medical Images," 2023, available at: https://arxiv.org/abs/2306.14289.