Att-Next for Skin Lesion Segmentation with Topological Awareness

C. Katar[a,b], O.B. Eryilmaz[c], E.M. Eksioglu[b]

[a]*Electrics and Electronics Department, Turkish-German University, Istanbul, 34820, Turkey, cihan.katar@tau.edu.tr*
[b]*Electronics and Communication Engineering Department, Istanbul Technical University, Istanbul, 34469, Turkey, katar20@itu.edu.tr, eksioglue@itu.edu.tr*
[c]*School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK, obe851@student.bham.ac.uk*

Abstract

Skin lesion segmentation is crucial for the early detection and accurate diagnosis of dermatological conditions, as precise boundary delineation enables better identification of lesion features. While Convolutional Neural Networks (CNNs) and hybrid CNN-Attention models have achieved notable success in this task, they often struggle to segment fine-grained lesion boundaries and suppress irrelevant tumor-like artifacts. They also tend to neglect topological features, which are crucial for accurately identifying complex lesions. To address these limitations, we propose a novel hybrid model that integrates ConvNeXt blocks with self-attention mechanisms. The model is also enhanced by a topological loss combined with Binary Cross Entropy (BCE) loss. This approach enables the model to better capture both local and global context, accurately delineate lesion boundaries, and suppress irrelevant regions, all without relying on a pre-trained backbone. Our method is evaluated on four publicly available skin lesion datasets: ISIC 2016, ISIC 2018, HAM10000, and PH2. Performance is assessed using segmentation metrics such as the Dice coefficient and Jaccard index. Experimental results demonstrate that the proposed model outperforms state-of-the-art (SOTA) methods, including MISSFormer, Swin-UNet, LeViT-UNet FAT-Net, Att-UNet, DoubleU-Net, DeepLabV3 and TransUNet. Notably, the model achieves a Jaccard index of 0.8529 and a Dice coefficient of 0.913 on the ISIC 2018 dataset, surpassing the performance of given SOTA models in boundary delineation and tumor-like region suppression. These results highlight the potential of our hybrid ConvNeXt-Attention model with topological loss to improve lesion segmentation accuracy, which would lead to more effective and precise dermatological diagnoses.

Keywords: Deep Learning, ConvNeXt module, Multi Head Attention, Topological Loss, Segmentation, Skin Lesion

1. Introduction

Skin cancer, one of the most prevalent forms of cancer, continues to affect more people each year (Roky et al., 2025). Among the different types of skin cancer, melanoma is the most aggressive and accounts for 90% of the deaths associated with cutaneous tumors (Garbe et al., 2016). Early intervention in skin cancer, particularly melanoma, is crucial to ensure high survival rates in the growing number of cases. Melanoma survival exceeds 95% when detected early, while it decreases to 20% when detected late (Garrison et al., 2023). Identifying melanoma, especially in its early stages, can be challenging, even for highly experienced dermatologists. It is estimated that experienced dermatologists achieve around 70% sensitivity when diagnosing melanoma using only visual examination (Garbe et al., 2016). On the other hand, recent works have shown that machine learning (ML) based algorithms are more accurate and can better support dermatological clinical studies.

Medical image segmentation commonly relies on an Encoder-Decoder network with skip connections, a structure effectively represented by U-Net (Ronneberger et al., 2015). U-Net has become a staple in medical image segmentation, inspiring numerous variations in recent years. Since the introduction of Vision Transformers by Dosovitskiy et al. (2020), Transformer-based models have grown in popularity due to their strong ability to capture global context, which greatly benefits segmentation tasks. Initially developed for Natural Language Processing (NLP), Transformers have proven effective in semantic segmentation by employing attention mechanisms that selectively focus on crucial regions of the input. Moreover, recent developments in ConvNeXt, built upon traditional ConvNet principles, have demonstrated promising results in medical image segmentation (Liu et al., 2022).

Skin lesion segmentation remains a challenging task due to the complex nature of skin lesions and their variability in medical images. Figure 1 illustrates the challenges inherent in skin lesion segmentation, including external markings, hair occlusions, artifacts, and varying lesion shapes, colors, and textures. Ambiguous borders and low contrast between lesions and surrounding healthy skin further complicate accurate segmentation. Existing approaches, as discussed in the related works section, attempt to address these challenges but often face limitations such as increased model complexity, high computational costs, reliance on pretrained models, and a loss of global context while focusing on capturing local details like lesion boundaries. Moreover, skin lesions, such as melanoma, possess intrinsic topological structures that are reflected in their connected components and higher-order features. These approaches often fall short in effectively capturing both the global and local topological context of tumors, resulting in suboptimal segmentation performance. To address these challenges, we propose the topology-aware Att-Next model, a hybrid approach that integrates ConvNeXt modules with attention mechanisms in the deeper layers of the network. ConvNeXt and attention mechanisms effectively capture both spatial details by extracting fine-grained local features and the broader global context, without requiring any pretrained backbone. Moreover, the model gains topological awareness through the use of a topological loss function. This function is based on the persistent homology of cubical filtration, where H_0 and H_1 features are considered, and the Wasserstein distance is used to quantify the loss value. The inclusion of topological loss enables the model to capture the underlying structure of the lesions, making it easier to differentiate between true lesion boundaries and noise. To the best of our knowledge, our proposed work is the first to explicitly incorporate a topological loss function for skin lesion segmentation. In this paper, we propose an efficient hybrid ConvNeXt-Attention architecture, topology-aware Att-Next, for skin lesion segmentation. Our contributions are as follows:

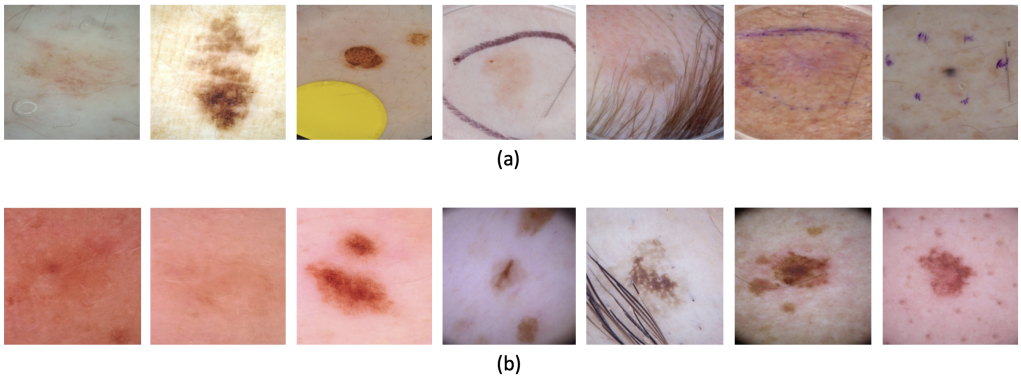

Figure 1: Challenges in skin lesion segmentation: (a) Images from the ISIC 2018 dataset highlight external markings, hair occlusions, artifacts, and varying lesion shapes, colors, and textures, which complicate accurate classification and boundary detection. (b) Images from the HAM10000 dataset illustrate ambiguous borders and low contrast between lesions and surrounding skin, increasing the difficulty of precise segmentation.

1) Att-Next combines modified ConvNeXt components and attention blocks in a U-shaped architecture to capture both short- and long-range dependencies.

2) We redesigned the ConvNeXt modules by reducing the kernel size, adding normalization layers, and replacing the Multi-Layer Perceptron (MLP) in the Transformer encoder with a ConvNeXt module. This redesign improves local detail capture, training stability, and feature representation for segmentation tasks.

3) We introduced a topological loss based on the Wasserstein distance, utilizing homological features H_0 and H_1, to effectively capture the data's topological structure and further improve segmentation accuracy.

The content for the rest of the paper is as follows. Chapter 2 discusses related works in the field. Chapter 3 introduces the proposed model architecture, emphasizing the use of topological loss functions to enhance segmentation performance. Chapter 4 describes the experimental setup, including dataset characteristics, preprocessing steps, and training strategies. Chapter 5 presents a comprehensive analysis of the results, incorporating both quantitative data and qualitative insights. The performance of the proposed model is compared with existing approaches, and the improvements achieved through ablation studies are highlighted. Additionally, this chapter discusses the model's generalization capabilities, main findings, limitations, and future recommendations. Chapter 6 concludes the paper.

2. Review of Related Works

2.1. Model Structures utilized in the Literature

In medical image segmentation tasks, capturing both global and local contextual information, as well as understanding the connectivity and structure of regions, is crucial. In recent years, numerous CNN models have been proposed, showcasing their capability in these tasks. For instance, Liu et al. (2022) introduced ConvNeXt modules, which utilize depthwise and pointwise lightweight separable convolutions (Chollet, 2017), enhanced with LayerNorm and GeLU activation functions. Similarly, Liu et al. (2022) proposed ConvUNeXt, which incorporates ConvNeXt modules within a U-shaped structure to reduce model complexity and parameter count.

Additionally, Zhang et al. (2023) developed the BCU-Net, a parallel U-Net-ConvNeXt architecture specifically designed for medical image segmentation tasks, further improving performance and efficiency. Han et al. (2022) demonstrated that, compared to the standard U-Net, their ConvNeXt-based U-shaped model reduces the number of parameters by 20%.

Although CNNs have achieved significant success, their limitations in capturing global feature information, due to the constraint of local receptive fields, have paved the way for the introduction of Transformer models. Unlike CNNs, Transformers extend the model's ability to focus on both image details and boundaries by capturing feature information across the entire image. However, Transformer-based models often lack the translational invariance and local correlation biases found in CNNs, meaning that they generally require large datasets to outperform CNNs. To tackle this challenge, hybrid models combining CNNs and Transformers have become increasingly prevalent in medical image segmentation. Att-UNet, introduced by (Oktay et al., 2018), incorporates attention mechanisms into the U-Net framework to improve the focus on salient features. TransUNET, developed by (Chen et al., 2024b), was the first to modify the Vision Transformer (ViT) into a U-Net architecture. UNETR, presented by (Hatamizadeh et al., 2022), employs a purely Transformer-based encoder for segmentation tasks. To address computational complexity, (Cao et al., 2023) introduced Swin-Unet, an architecture that uses self-attention within shifted windows to enhance efficiency. TransFuse, described by (Zhang et al., 2021), combines CNN and Transformer features to effectively integrate spatial and global context. LeViT-U, proposed by (Xu et al., 2024), incorporates LeViT transformers into a U-Net framework, providing efficient feature learning. AS-Net, presented by (Hu et al., 2022), improves discriminative power by utilizing both spatial and channel attention mechanisms. FAT-Net, described by (Wu et al., 2022), integrates a Transformer branch within a CNN-based encoder-decoder architecture and employs feature adaptation modules alongside a memory-efficient decoder to capture local and global contexts. MISSFormer, introduced by (Huang et al., 2023), employs hierarchical Transformer blocks to enhance segmentation performance. Additionally, TransAttUnet, a Transformer-based attention-guided U-Net, improves segmentation across various medical imaging tasks by leveraging self-aware attention modules and multi-scale skip connections (Chen et al., 2024a). Furthermore, D-TrAttUnet, a hybrid CNN-Transformer model with dual decoders, is designed to segment lesions and organs simultaneously (Bougourzi et al., 2024).

Hybrid models that combine ConvNeXt and attention mechanisms have gained attention in recent years due to their potential advantages, especially in classification tasks. FNeXter, a U-shaped network that integrates ConvNeXt and Transformer blocks, incorporates Region-Aware Spatial Attention (RASA) and a Self-Adaptive Multi-Scale Feature Fusion Attention (SMFFA) module (Niu et al., 2024). This design efficiently extracts local features via ConvNeXt and captures long-range dependencies through Transformers. However, the architectural complexity increases computational demands and training overhead. A two-stream network was introduced to fuse ConvNeXt and Swin Transformers using a simplified MIX-Block for feature fusion, reducing computational costs compared to more intricate hybrid models (Wang et al., 2023). Nevertheless, reliance on pre-trained backbones may limit adaptability to highly diverse or noisy datasets. To address artifacts, ConvNeXt-ST-AFF enhances hybrid architectures by combining ConvNeXt and Swin Transformers with Attentional Feature Fusion (AFF) modules and Efficient Channel Attention (ECA) (Hao et al., 2023). Recently, MedNeXt has been proposed as a fully Transformer-driven ConvNeXt segmentation model designed to tackle the limitations of hybrid architectures in medical imaging by incorporating residual inverted bottlenecks and compound scaling techniques (Roy et al.). However, its ability to accurately segment small and low-contrast

lesions remains a challenge. Similarly, Response Fusion Attention U-ConvNeXt (RFAU-CNxt) integrates ConvNeXt with novel attention mechanisms to enhance segmentation accuracy in fundus images and optic disc and cup segmentation. Despite its strong performance, the increased architectural complexity can lead to longer training times (Mallick et al., 2023).

Many approaches struggle to simultaneously capture both fine local details and broader contextual relationships while maintaining a smaller parameter count and reduced computational cost. Additionally, some methods rely on pretrained models, which may limit their adaptability to highly diverse or noisy datasets. To address these challenges, we propose a topology-aware hybrid ConvNeXt-Transformer approach for skin lesion segmentation.

2.2. Use of Topological Awareness

Recently, Topological Data Analysis (TDA), particularly persistent homology (PH), has been employed in deep learning segmentation tasks to enhance the topological similarity between segmented regions and the ground truth, which may otherwise be missed by traditional pixel-wise comparisons (Mosinska et al.). Various approaches have been proposed to incorporate persistent homology into deep learning models. These methods include extracting topological features from input images via PH and feeding them into a CNN(Hofer et al., 2017), or creating topologically aware layers (Hofer et al., 2020; Love et al., 2023) and networks (Papamarkou et al., 2024).

In line with our proposed work, some studies have integrated topological information into deep learning models through loss functions, such as those presented by (Hu et al., 2019; Clough et al., 2020; Gupta et al., 2022; Demir et al., 2023; Yang et al., 2021). Topological loss functions have demonstrated improved performance, particularly in detecting thin structures, which is essential for the accurate segmentation of certain subjects. These functions have shown effectiveness in segmenting vascular networks, road maps, and other data involving high-dimensional interactions, such as 3D fMRI data.

Specific to our dataset, the ground truth skin lesions are connected structures. Consequently, the model output can be constrained to maintain the same topological structures as the ground truth across different scales. Additionally, border irregularities are critical features for detecting malignant lesions (Marghoob et al., 2019). Vandaele et al. (2020) showed the effectiveness of a persistent homology-based approach in unsupervised skin lesion segmentation. While their approach utilized persistent homology, they did not incorporate a topological loss function based on persistent homology. By integrating this topological loss into our proposed hybrid ConvNeXt-Attention model, we enhance skin lesion segmentation, achieving more precise and topologically reliable outputs.

3. Proposed model with Topological Awareness

3.1. Model Structure

The methods mentioned earlier are not yet sufficient to fully address the challenges discussed in the introduction, and achieving precise segmentation of skin lesions in dermoscopic images remains a complex task. The proposed model integrates attention mechanisms and modified ConvNeXt blocks to efficiently capture both global and local contexts, as well as short- and long-range dependencies, while considering a reduction in model complexity, parameters, and FLOPs. Within a U-shaped architecture, we sequentially employ modified ConvNeXt and Multi-head self-attention blocks.

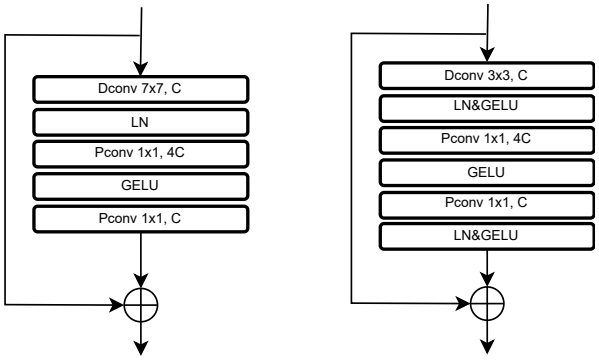

Figure 2: Traditional (left) and Modified (right) ConvNeXt Block

In medical imaging, such as skin lesion segmentation, capturing fine local details like edges and textures is crucial. To achieve this, we modified the ConvNeXt block by reducing the convolutional kernel size from 7×7 to 3×3, as illustrated in Figure 2. The input tensor is first processed through a depthwise convolution ($DConv_{3\times3}$) to extract spatial information. Layer Normalization (LN) and the Gaussian Error Linear Unit ($GELU$) activation function are then applied to introduce non-linearity. Next, pointwise convolutions ($PConv_{1\times1}$) are used to initially expand the channels to $4C$ and subsequently reduce them back to C. This combination of depthwise and pointwise convolutions captures multi-scale local features. By incorporating additional normalization and activation functions, we achieved more efficient feature extraction, improved model convergence, and overall enhanced performance on the skin lesion segmentation tasks. Experimental results verified that reducing the kernel size and increasing normalization improved stability, enabled more effective learning, and enhanced overall performance compared to using fewer normalization layers.

As part of the model structure, the first two stages incorporate double modified ConvNeXt modules, which are well-suited for capturing local spatial features, such as edges and textures, to extract feature maps. These modules utilize pointwise and depthwise separable convolutions, which are efficient in terms of model parameters and computational cost. In the third and fourth stages, we employ an attention mechanism following the modified ConvNeXt modules to capture the global context of the tumor regions.

In addition, we replace the standard MLP in the Transformer encoder with modified ConvNeXt modules. Unlike MLPs, which process each pixel independently of its neighbors, ConvNeXt modules preserve local spatial relationships, a critical factor in tasks like image segmentation. By substituting MLPs with ConvNeXt modules, we mitigate the limitations caused by the inability of the former to account for spatial location. Figure 3 illustrates the comparison between the traditional transformer encoder and the proposed attention mechanism integrated with ConvNeXt blocks.

The input is first embedded as a sequence of features (or called tokens). Let X represent the input embeddings or tokens. The token sequence $X \in \mathbb{R}^{N\times C}$, with length N and channel dimension C, is fed into the ConvNeXt blocks, one of which can be expressed as:

Modified ConvNeXt module can be formulated as follows

$$X' = MHSA(X) + X_{\text{in}} \tag{1}$$

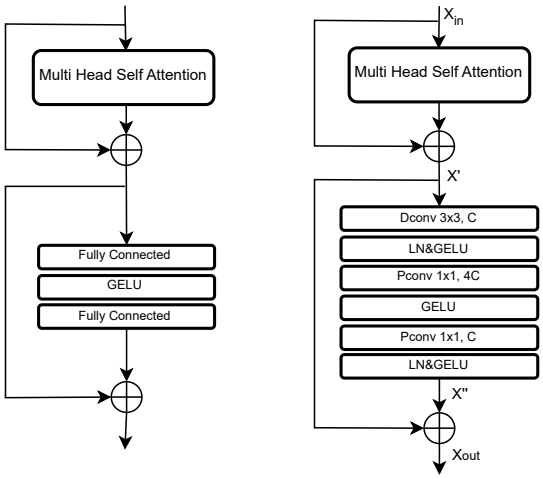

Figure 3: Comparison of Traditional Transformer Encoder (left) vs. Attention with Modified ConvNeXt Block (right)

$$X'' = \mathrm{LN}(\mathrm{PConv}_{1\times1}(\sigma(\mathrm{PConv}_{1\times1}(\sigma(\mathrm{LN}(\mathrm{DConv}_{3\times3}(X', C))), 4C)), C)) \tag{2}$$

$$X_{\text{out}} = X'' + X' \tag{3}$$

$\sigma(\cdot)$ denotes the activation function, which is the GELU. MHSA represents Multi-Head Self Attention, which captures global relationships in the input sequence. $\mathrm{DConv}_{3\times3}$ is the Depthwise Convolution operation with a 3×3 kernel, used to capture spatial information across individual channels, producing C output channels. $\mathrm{PConv}_{1\times1}$ is the Pointwise Convolution with a 1×1 kernel, used for channel mixing and operates either with $4C$ or C output channels to control dimensionality. LN is applied after convolutions to stabilize training and enhance gradient flow. Finally, Y represents the output tensor, which is obtained after combining these operations.

In the downsampling layers, we use 3×3 convolutions to increase the channel size and max-pooling with a stride of 2 to decrease spatial dimensions, which helps emphasize edges necessary for segmentation. For upsampling, 3×3 convolutions reduce the channel count, and nearest-neighbor upsampling increases the spatial size. Although transposed convolutions were considered, they did not improve validation scores and increased the parameter count.

Combining all these implementations, we constructed a U-shaped network for 2D image segmentation, as depicted in Figure 4. The input tensor is defined as $X \in \mathbb{R}^{B\times C\times H\times W}$, where $B = 8$ (batch size), $C = 3$ (input channels), $H = 256$, and $W = 256$. In the encoder, the channel dimensions progressively double from 3 up to 512, while the spatial dimensions are halved at each downsampling step. In the decoder, this process is reversed so that the channel dimensions decrease from 512 back to 1, and the spatial dimensions are incrementally restored to their original size.

We designed different model variants, starting with either attention mechanisms or ConvNeXt blocks. However, experimental results showed that using ConvNeXt blocks from the start led to higher Dice and Jaccard scores and lower computational costs.

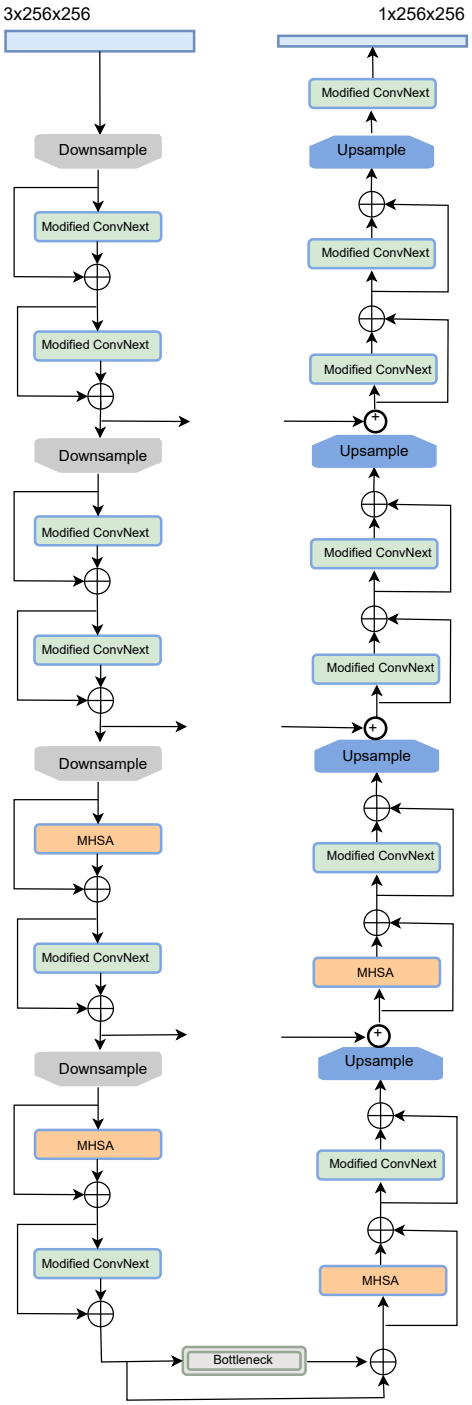

Figure 4: Overall Design of Att-Next Architecture

3.2. *Topological Data Analysis (TDA)*

TDA comprises a set of methods aimed at extracting meaningful insights from the topological structures of data embedded in a topological space (Carlsson, 2009). A topological space is a mathematical construct characterised by properties such as continuity, connectedness and convergence. TDA analyses topological features that are often represented in low-dimensional structures, such as simplicial or cubical complexes. In this section we will briefly introduce the PH and most common filtration methods. For a more detailed introduction to TDA, readers are referred to (Chazal and Michel, 2021; Coskunuzer and Akçora, 2024).

PH, the main tool in TDA, tracks the evolution of homologies, such as connected components, loops and voids, across different scales. The PH process begins by mapping the data into a topological space using a filtration method. Filtration methods, such as Vietoris-Rips and cubical filtration, incrementally increase a scale parameter that acts as a threshold for connecting data points. Essentially, PH addresses the thresholding problem by considering all possible thresholds. This threshold, or scale, can be defined in various ways, such as using Euclidean distance to connect points or using pixel values in digital images to determine which pixels are included. As the scale increases, new simplices or cubes (depending on the filtration type) are added to the complex, each associated with its corresponding scale value.

Formally, given a simplicial complex K, we create a sequence of subcomplexes, denoted as $K = \{K_p \mid 0 \leq p \leq m\}$, such that $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K$, by applying the chosen filtration method. This process reflects the incremental addition of simplices as the scale increases. A simplicial complex is constructed from simplices, which include vertices (0-simplices), edges (1-simplices), triangles (2-simplices) and their higher-dimensional analogues.

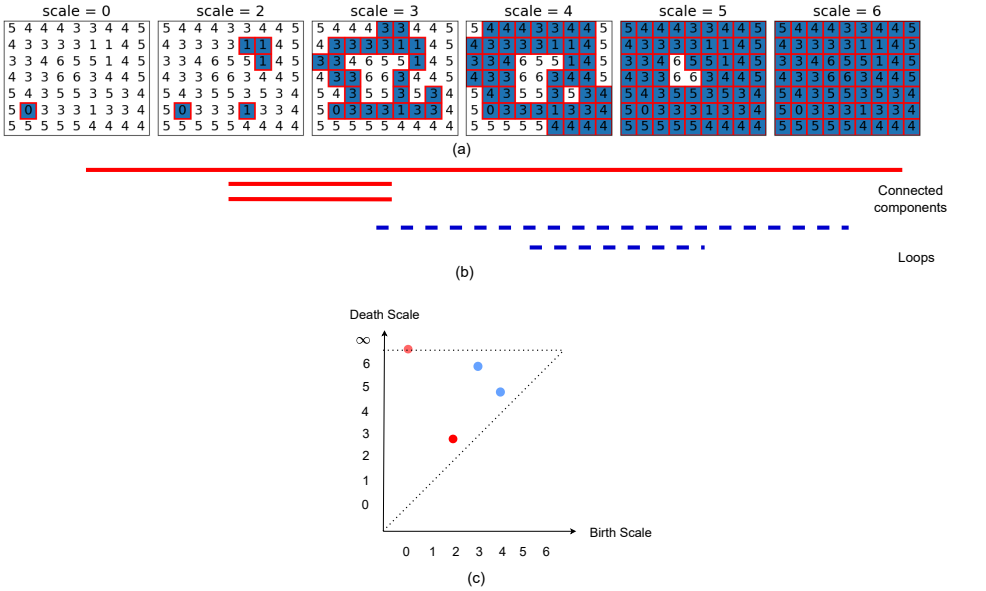

Figure 5: Illustration of Persistent Homology for Cubical Complex Analysis at Different Scales: (a) Filtration Process Across Scales, (b) Barcode Diagram of Connected Components and Loops, (c) Persistence Diagram Depicting Birth and Death of Topological Features

In contrast, cubical complexes use cubes, including vertices (0-dimensional cubes), edges (1-dimensional cubes), squares (2-dimensional cubes) and higher-dimensional cubes. Digital images are particularly well-suited for cubical filtration due to their grid-like structure, which naturally aligns with the cubical representation. In a cubical complex, digital images are represented as nested sequences of 2D images by considering the gray-scale values, $\gamma_{i,j} \in [0, 255]$, of each pixel, denoted by $\triangle_{i,j} \subset \mathbf{X}$. Given a sequence of thresholds ($0 \leq t_1 < t_2 < \cdots < t_N \leq 255$), we generate a series of 2D binary images, $X_1 \subset X_2 \subset \cdots \subset X_N$, where X_n represents the set of pixels $\triangle_{i,j} \subset \mathbf{X}$ that have values exceeding the threshold t_n. In other words, we start with an empty 2D image of size $n \times m$ and progressively activate pixels by coloring them black as their values exceed each threshold t_n. This process is illustrated in Figure 5.

This cubical complex is stored in a boundary matrix from which we can infer the persistence of the homologies after a Gaussian elimination process. For a given homology, σ, we have the scale at which it first appears, b and the scale at which it disappears, d. The difference between d and b is called the persistence of that homology. For each homology dimension, we have the (b, d) pairs. These pairs are usually summarized in barcodes or Persistence Diagrams (PDs). The persistence diagram for dimension k is then the collection of all such pairs, $PD_k = \{(b_\sigma, d_\sigma)\}$. This information allows us to identify significant homologies that persist across different scales.

3.3. Topological Loss

Accurately segmenting skin lesions is particularly challenging due to the presence of irrelevant tumor-like regions that are difficult for models to distinguish. To address this, we implemented a topological loss function to help the model capture and preserve the intrinsic topology of tumors. Incorporating topological loss during training enhances the model's ability to learn and retain crucial topological information for accurate tumor segmentation.

To achieve this, we compared the persistence diagrams (PDs) of the model's output and the ground truth using the Wasserstein distance ($p = 2$). The Wasserstein distance measures the minimal cost to transform one distribution into another, assessing the similarity between the PDs. A smaller distance indicates greater similarity between the topological features of the model's predictions and the ground truth, as illustrated in Figure 6.

Given two persistence diagrams $D_1 = \{(b_i, d_i)\}_{i=1}^{n}$ and $D_2 = \{(b'_j, d'_j)\}_{j=1}^{m}$, the 2-Wasserstein distance is defined as follows.

$$W_2(D_1, D_2) = \left(\inf_{\varphi: D_1 \to D_2} \sum_{i=1}^{n} \|x_i - \varphi(x_i)\|^2 \right)^{\frac{1}{2}} \tag{4}$$

- $x_i = (b_i, d_i)$ are points in D_1,
- $y_j = \varphi(x_i)$ are points in D_2,
- $\|\cdot\|$ is typically the L_2 norm in the Euclidean plane,
- $\varphi : D_1 \to D_2$ is a bijection (matching) between the points of the two diagrams, extended to include points on the diagonal (where birth equals death).

We also use BCE loss (L_ω^{BCE}) as defined below.

$$L_\omega^{BCE}(G, P) = -\frac{1}{N} \sum_{i=1}^{N} (G_i \log P_i + (1 - G_i) \log(1 - P_i)) \tag{5}$$

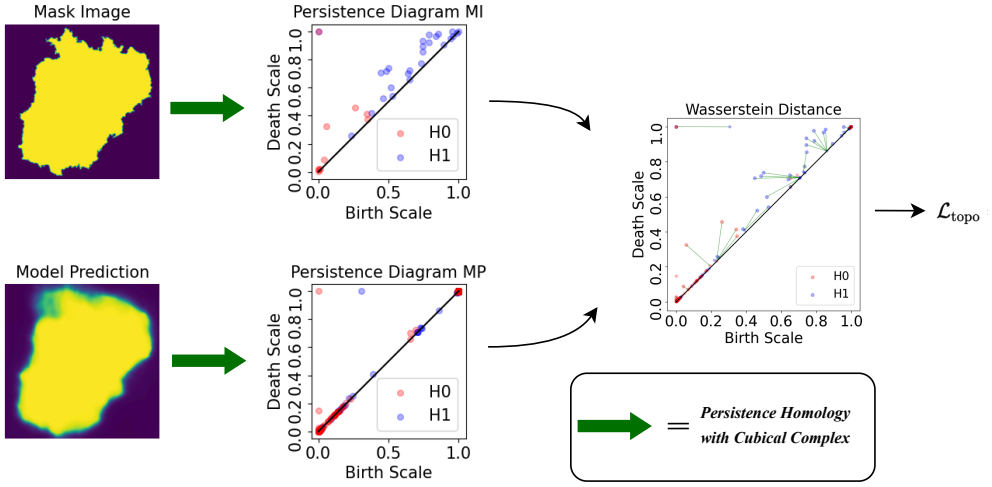

Figure 6: Topological Loss Integration for Tumor Segmentation Using Persistence Diagrams, Wasserstein Distance. Green Arrow representing calculation of Persistent Homology with a Cubical Complex

Here, G is the ground truth segmentation, P is the predicted segmentation, and N is the total number of pixels in the image.

The total loss used in our training is a combination of the BCE loss and the Wasserstein distance-based topological loss (L_{ω}^{Topo}). It is expressed as:

$$L_{total} = L_{\omega}^{BCE} + \lambda_1 W_2(D_1, D_2) \tag{6}$$

In this equation, λ_1 represents the weighting coefficient used to adjust the balance between the contributions of each loss term.

4. Experiments

4.1. Dataset

The ISIC dataset offers a comprehensive collection of dermoscopic images, crucial for segmentation tasks in computer vision research. The ISIC 2018 dataset comprises 2,694 images, with original dimensions ranging from 540×722 to 4500×6800 pixels. To ensure uniformity during model training, all images were resized to 256×256 pixels.

In addition to the ISIC 2018 dataset, we also utilized the HAM10000 dataset, which is another large collection of dermoscopic images containing 10,015 images of various common pigmented skin lesions. The images in the HAM10000 dataset have dimensions of 600×450 pixels and, similar to the ISIC 2018 dataset, were resized to 256×256 pixels for consistency and to match the model's input requirements. The HAM10000 dataset offers a diverse set of lesions, providing a rich source of training data to improve model generalization.

To test the robustness and generalization of our trained model, we evaluated it on the ISIC 2016 and PH2 datasets (Mendonça et al., 2013) using the proposed architecture and loss function. The ISIC 2016 dataset consists of 900 images, while the PH2 dataset includes 200 dermoscopic

images, providing an opportunity to demonstrate the model's generalization capability on relatively small datasets.

4.2. Implementation Details

We used 80% of the dataset for training, 10% for validation, and 10% for testing. The network was trained end-to-end using the AdamW optimizer, with an initial learning rate of 1×10^{-4}. A cosine annealing learning rate schedule was applied, reducing the learning rate by a factor of $1/10$ during training to gradually decrease the learning rate and encourage better convergence towards the end of training. The batch size was set to 8, and the model was trained for 400 epochs. All experiments were implemented in PyTorch and executed on an NVIDIA GeForce RTX 2080 Ti GPU. The source code will be made available upon publication of the paper.

To enhance network performance and improve convergence speed, we employed a combination of loss functions, including the BCE loss (L_{ω}^{BCE}) and a topological loss derived from the persistence information of the data. We set the weight of the topological loss (λ_1) to 0.1 in the combined BCE-Topoloss function.

Furthermore, to address computational complexity, we applied average pooling to reduce image size while preserving the lesion's structural features. Instead of using patchify methods, which can fragment lesion integrity and are unsuitable for skin lesion datasets, we opted for average pooling to reduce image size.

4.3. Augmentation

In our study, we applied various data augmentation techniques to enhance the robustness and generalization of our model. We utilized standard augmentations such as random horizontal and vertical flips, as well as random rotations. Additionally, we incorporated advanced methods like CutMix and CutOut, each applied with a probability of 0.5 during training. CutMix is an augmentation technique where patches are cut and pasted among training images, and the ground truth labels are mixed proportionally to the area of the patches. This encourages the model to learn from combined visual features and improves its robustness. CutOut, on the other hand, involves randomly masking out square regions of the input images during training, which forces the model to focus on less prominent features and prevents overfitting to specific image regions. The implementations of CutMix and CutOut augmentations are illustrated in Figure 7.

We used a 25×25-pixel cutout box to zero out randomly selected areas, which could sometimes cover the entire tumor region. In such cases, we skipped the augmentation to avoid losing critical tumor information.

We also experimented with other augmentation techniques, including color jitter, perspective transformations, and affine transformations. However, we observed that these additional augmentations did not improve the model's performance metrics and instead resulted in increased computational overhead.

4.4. Evaluation metrics

To assess our model's performance, we rely on five well-established metrics: Dice coefficient (DSC), Intersection over Union (IoU), Precision, Recall, and Accuracy. Dice and IoU are particularly common in segmentation tasks because they assess the overlap and consistency between the predicted segmentation and the ground truth. Specifically, Dice measures similarity by calculating twice the overlap area divided by the total number of pixels, providing a balanced

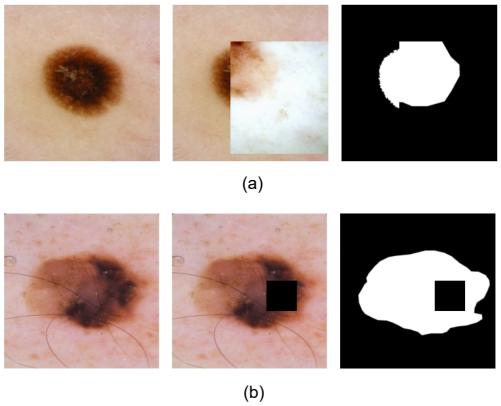

Figure 7: Implementation of Advanced Augmentation Techniques: (a) CutMix Augmentation, (b) CutOut Augmentation

assessment even when class distribution is uneven. On the other hand, IoU calculates the ratio of intersection to union between predicted and actual segmentation masks, offering an intuitive measure of overlap.

In addition to these metrics, Precision and Recall provide a more detailed understanding of model performance. Precision indicates how many of the predicted positive pixels are correct, highlighting the model's ability to minimize false positives. Recall, by contrast, measures how well the model identifies all relevant positive pixels, reflecting its ability to detect all areas of interest. By considering these metrics together, we gain a comprehensive evaluation of our model, ensuring it consistently produces accurate segmentation results while effectively identifying the regions of interest.

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{7}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

5. Results

In this section, we present the evaluation results of our proposed topology-aware Att-Next model, comparing its performance against several baseline models. The ISIC 2018 and HAM10000 datasets were used for training and testing, while the PH2 and ISIC 2016 datasets were also evaluated on testing to assess the model's generalization capability alongside the ISIC

Table 1: Comparison results on the ISIC 2018 dataset

Methods	IoU	DSC	Rec.	Prec.	Acc.	FLOPs (G)	Params (M)
TransUNet (Chen et al., 2024b)	0.7805	0.8741	0.8872	0.8707	0.9509	33.69	67.08
LeViT-U (Xu et al., 2024)	0.7817	0.8791	0.9251	0.8615	0.9501	18.90	19.89
U-Net (Ronneberger et al., 2015)	0.8001	0.8830	0.8940	0.8770	0.9523	65.56	37.66
Swin-Unet (Cao et al., 2023)	0.8012	0.8853	0.8875	0.8885	0.9537	8.00	27.17
MISSFormer (Huang et al., 2023)	0.8133	0.8953	0.8934	0.9045	0.9582	9.86	42.46
DeepLabV3 (Chen et al., 2017)	0.8178	0.8842	0.8556	0.8834	0.94780	43.40	42.00
Att-UNet (Oktay et al., 2018)	0.8199	0.8955	0.9117	0.8846	0.9525	66.69	34.88
TransAttUnet (Chen et al., 2024a)	0.8224	0.9000	0.9178	0.8883	0.9623	88.87	25.97
DoubleU-Net (Jha et al., 2020)	0.8259	0.9024	0.9291	0.8816	0.9610	53.89	29.29
MedNeXt (Roy et al.)	0.8278	0.9045	0.9034	0.9108	0.9605	15.55	12.49
FAT-Net (Wu et al., 2022)	0.8303	0.9061	0.9243	0.8935	0.9620	22.84	28.76
DTrAttUnet (Bougourzi et al., 2024)	0.8358	0.9092	0.9194	0.9025	0.9630	42.03	132.51
RFAU-CNxt (Mallick et al., 2023)	0.8484	0.9100	0.9252	0.9132	0.9673	18.08	100.41
Att-Next (Proposed Model)	**0.8529**	**0.9130**	**0.9279**	**0.9169**	**0.9675**	13.47	13.88

Table 2: Validation IoU and Dice scores of different models on the ISIC 2018

Model	Validation IoU	Validation Dice
U-Net (Ronneberger et al., 2015)	0.80645	0.88845
MISSFormer (Huang et al., 2023)	0.83403	0.90776
FAT-Net (Wu et al., 2022)	0.83919	0.91059
ATT-Next (Proposed Model)	**0.84888**	**0.91721**

2018 and HAM10000 datasets. To ensure a fair comparison, we utilized publicly available implementations of the baseline models and applied the same augmentation techniques across all experiments.

5.1. Comparison with competing methods

ISIC 2018 Dataset: As summarized in Table 1, topology-aware Att-Next was evaluated against recent models discussed in the related works section, using the ISIC 2018 dataset for comparison. Att-Next outperforms all these models, achieving the highest Intersection over Union of 0.8529, the highest Dice Similarity Coefficient of 0.913, and the highest recall, precision, and accuracy metrics, confirming its superiority in segmentation tasks on the ISIC 2018 dataset.

Table 1 demonstrates that Att-Next achieves an IoU of 0.8529, surpassing RFAU-CNxt by 0.53%, TransAttUnet by 3.71%, MedNeXt by 3.04%, FAT-Net by 2.73%, and DoubleU-Net by 3.27%. In terms of DSC, Att-Next attains 0.913, which is 0.30% higher than RFAU-CNxt, 1.44% higher than TransAttUnet, 0.94% higher than MedNeXt, and 0.76% higher than FAT-Net. Regarding recall, Att-Next achieves 0.9279, outperforming RFAU-CNxt by 0.29%, TransAttUnet

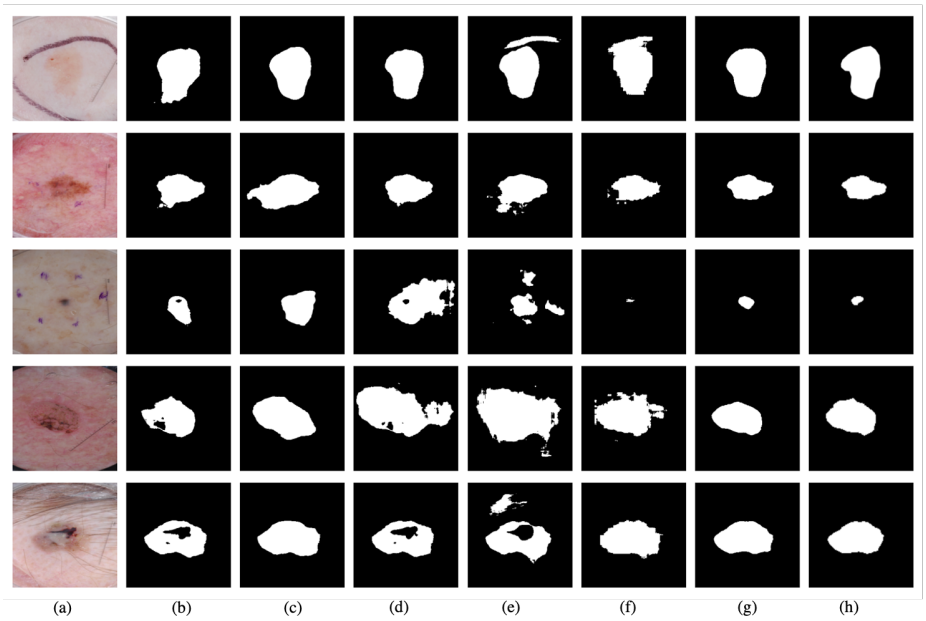

Figure 8: Qualitative Comparison of Lesion Segmentation Results on Challenging Images Across Multiple Models on ISIC 2018. From left to right: (a) Input Image, (b) U-Net, (c) DoubleU-Net, (d) Attention U-Net, (e) TransUNet, (f) Swin-Unet, (g) Proposed Att-Next and (h) Ground Truth Mask.

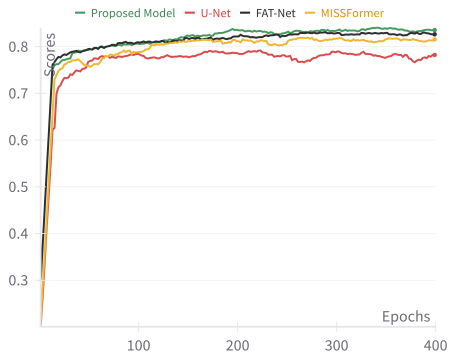

Figure 9: Validation IoU Score Comparison on Tthe ISIC 2018

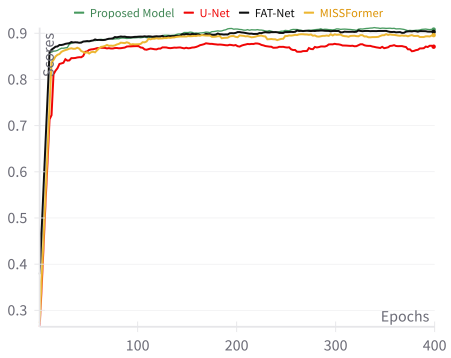

Figure 10: Validation Dice Score Comparison on the ISIC 2018

by 1.10%, MedNeXt by 2.73%, and FAT-Net by 0.39%. For precision, Att-Next attains 0.9169, which is 0.41% higher than RFAU-CNxt, 3.22% higher than TransAttUnet, 0.67% higher than MedNeXt, and 2.62% higher than FAT-Net. Furthermore, Att-Next achieves the highest accuracy of 0.9675, which is 0.02% higher than RFAU-CNxt, 0.53% higher than TransAttUnet, 0.73% higher than MedNeXt, and 0.68% higher than FAT-Net.

Regarding computational complexity, Att-Next achieves impressive performance with an IoU of 0.8529, DSC of 0.913, and Acc. of 0.9675, requiring only 13.466 GFLOPs and 13.876 mil-

Table 3: Comparison results on the HAM10000 dataset.

Methods	IoU	DSC	Rec.	Prec.	Acc.	FLOPs(G)	Params(M)
U-Net (Ronneberger et al., 2015)	0.8607	0.9197	0.9111	0.9310	0.9516	65.56	37.66
Swin-Unet (Cao et al., 2023)	0.8684	0.9282	0.9287	0.9312	0.9608	8.00	27.17
Att-UNet (Oktay et al., 2018)	0.8896	0.9406	0.9287	**0.9551**	0.9677	66.69	34.88
DoubleU-Net (Jha et al., 2020)	0.8915	0.9418	0.9449	0.9409	0.9679	53.89	29.29
Att-Next (Proposed Model)	**0.902**	**0.9442**	**0.9435**	0.9469	**0.9701**	13.47	13.88

lion parameters. In comparison, models like RFAU-CNxt (100.41 million parameters, 18.08 GFLOPs), DTrAttUnet (132.51 million parameters, 42.03 GFLOPs), and FAT-Net (28.76 million parameters, 22.84 GFLOPs) exhibit much higher computational costs.

Figure 8 presents a qualitative comparison of lesion segmentation performance on challenging images using different models. Notably, the proposed Att-Next architecture provides more accurate segmentation of lesion boundaries compared to other models, particularly in cases with ambiguous or noisy backgrounds, as evident in the examples shown.

Figure 9 and Figure 10 illustrate the validation performance of different models. These figures show that the proposed topology-aware Att-Next model consistently achieves the highest IoU and Dice scores across epochs. Table 2 summarizes the quantitative comparison of IoU and Dice scores achieved in validation.

HAM1000 Dataset: Table 3 shows that Att-Next outperforms four models on the HAM10000 dataset, achieving a Dice Similarity Coefficient of 0.9442, which is 0.3% higher than DoubleU-Net and 0.4% higher than Att-UNet. Att-Next also achieves a IoU of 0.902, which is 1.1% higher than DoubleU-Net and 1.2% higher than Att-UNet. In terms of recall, Att-Next achieves 0.9435, which is slightly lower than DoubleU-Net (0.9449) but still competitive. For precision, Att-Next attains 0.9469, surpassing DoubleU-Net by 0.7% and U-Net by 1.7%. Additionally, Att-Next reaches the highest accuracy of 0.9701, which is 0.4% higher than DoubleU-Net and 1.8% higher than U-Net. Overall, Att-Next consistently outperforms the other models across most evaluation metrics in lesion segmentation tasks on the HAM10000 dataset.

5.2. Generalization Capabilities

We evaluated the generalization capability of our proposed model by training it on two datasets, ISIC 2018 and HAM10000, and testing it on four datasets: ISIC 2018, HAM10000, ISIC 2016, and PH2. Tables 4 and 5 present the performance comparison of our method, Att-Next, with SOTA techniques, including MISSFormer (Huang et al., 2023), DoubleU-Net (Jha et al., 2020), FAT-Net (Wu et al., 2022), and Att-UNet (Oktay et al., 2018).

Table 4 presents the results of training the Att-Next model on the ISIC 2018 dataset and testing it on three different datasets: ISIC 2016, HAM10000, and PH2. The results demonstrate that Att-Next achieves the highest Jaccard, F1 score, and Accuracy on ISIC 2016, outperforming other models, including DoubleU-Net, which records the best Recall. On the HAM10000 dataset, Att-Next leads with the highest Jaccard, F1 score, and Recall, while FAT-Net performs best in Precision. For the PH2 dataset, Att-Next achieves the top Jaccard, F1 score, and Precision, whereas DoubleU-Net records the highest Recall and Accuracy.

Similarly, Table 5 presents the results of training the Att-Next model on the HAM10000 dataset and testing it on three different datasets: ISIC 2018, ISIC 2016, and PH2. The results

Table 4: Generalization results of models trained on ISIC 2018 and tested on different datasets.

Model	Dataset	Jaccard (IoU)	F1 (Dice)	Recall	Precision	Acc.
MISSFormer (Huang et al., 2023)	ISIC 2016	0.8467	0.9153	0.9216	0.9132	0.9557
	HAM	0.8191	0.8980	0.8858	0.9174	0.9455
	PH2	0.8081	0.8938	0.9118	0.8803	0.9229
DoubleU-Net (Jha et al., 2020)	ISIC 2016	0.8894	0.9412	**0.9469**	0.9359	0.9684
	HAM	0.8302	0.9001	0.9089	0.9103	**0.9529**
	PH2	0.8592	0.9212	**0.9921**	0.8821	0.9568
FAT-Net (Wu et al., 2022)	ISIC 2016	0.8883	0.9408	0.9453	0.9375	0.9674
	HAM	0.8155	0.8961	0.8777	**0.9218**	0.9479
	PH2	0.8538	0.9207	0.9926	0.8598	0.9467
Att-Next (Proposed Model)	ISIC 2016	**0.8909**	**0.9420**	0.9341	**0.9511**	**0.9698**
	HAM	**0.8329**	**0.9072**	**0.9105**	0.9096	0.9501
	PH2	**0.8635**	**0.9263**	0.9752	**0.8839**	0.9483

Table 5: Generalization results of models trained on HAM10000 and tested on different datasets.

Model	Dataset	Jaccard (IoU)	F1 (Dice)	Recall	Precision	Acc.
Att-UNet (Oktay et al., 2018)	ISIC 2018	0.8025	0.8883	0.9170	**0.8677**	0.9534
	ISIC 2016	0.8519	0.9194	0.9114	0.9324	0.9583
	PH2	0.8506	0.9191	0.9368	0.9044	0.9450
DoubleU-Net (Jha et al., 2020)	ISIC 2018	0.7979	0.8855	0.9279	0.8546	0.9545
	ISIC 2016	0.8593	0.9238	**0.9235**	0.9272	0.9588
	PH2	0.8426	0.9139	0.9349	**0.8960**	0.9402
ATT-Next (Proposed Model)	ISIC 2018	**0.8127**	**0.8938**	**0.9335**	0.8651	**0.9562**
	ISIC 2016	**0.8702**	**0.9302**	0.9206	**0.9419**	**0.9608**
	PH2	**0.8594**	**0.9240**	**0.9882**	0.8688	**0.9509**

further highlight the superior generalization capabilities of Att-Next. On ISIC 2018, Att-Next achieves the highest Jaccard and F1 score, outperforming DoubleU-Net and Att-UNet. For ISIC 2016, Att-Next again records the best Jaccard and F1 score, while DoubleU-Net achieves competitive Recall. On the PH2 dataset, Att-Next achieves the highest Jaccard, F1 score, and Recall, effectively capturing the diverse and complex lesion characteristics.

Our experimental results confirm that topology-aware Att-Next consistently achieves the best performance across all datasets in terms of Jaccard (IoU), F1 (Dice) score, and Accuracy, demonstrating its ability to generalize robustly across different datasets and lesion types.

5.3. Ablation Study

In our experiments, we integrated both a traditional transformer encoder with an MLP and a modified ConvNeXt architecture incorporating self-attention. Initially, we trained the model

with both ConvNeXt modules. Subsequently, we introduced the attention mechanism into the deeper layers, incorporating both the traditional MLP, Convnext and Modified ConvNeXt modules. We also trained and compared the models under three loss scenarios for each of the different architectures: the traditional combined loss Dice-BCE, a triple loss that combines Dice-BCE and Topological loss (Dice-BCE-Topo), and a combined loss focused on BCE and Topological loss (BCE-Topo).

ISIC2018 Dataset: Table 6 presents the effect of different model structures on segmentation performance for the ISIC 2018 dataset. The table compares various configurations, including ConvNeXt and its modifications with attention mechanisms and different loss functions. Metrics such as IoU, Dice Score, Precision, Recall, and Accuracy are reported for each model structure and loss combination. The highlighted values indicate the best performance within each model structure across different evaluation metrics.

In Table 6 a gradual improvement in model performance is observed as the architecture evolves from the basic ConvNeXt to more complex hybrid structures with attention mechanisms. Starting with ConvNeXt, the use of BCE-Topo loss yields the best performance, suggesting that adding topological information enhances segmentation quality. Incorporating Attention-ConvNeXt enhances the ConvNeXt structure, with BCE-Topo loss consistently delivering the best results, highlighting the advantage of combining local and global context and topological consistency. Introducing Modified ConvNeXt blocks and a traditional transformer encoder further improves performance, especially with BCE-Topo loss, though some limitations are noted with Dice-BCE-Topo loss. The highest performance across all metrics is achieved by combining attention mechanisms with the modified ConvNeXt block, resulting in the most robust and effective model. This progression demonstrates that increasing architectural complexity through hybrid and modified structures can enhance segmentation quality.

Figure 11 shows the visual output of model predictions using various architectures with the BCE-Topo loss function. The ConvNext for all stages provides reasonable segmentation but

Table 6: Effect of Model Architectures and Loss Functions on Skin Lesion Segmentation for the ISIC 2018 Dataset

Model structure	**Loss Function**	**IoU**	**Dice**	**Prec.**	**Rec.**	**Acc**
ConvNexT	Dice-BCE	0.8105	0.8928	0.8831	0.9101	0.9624
	Dice-BCE-Topo	0.8287	0.9043	**0.9081**	0.9073	0.9623
	BCE-Topo	**0.8317**	**0.9061**	0.8999	**0.9167**	**0.9637**
M.ConvNexT ; Attention-MLP	Dice-BCE	0.8315	0.9022	**0.9046**	0.9026	0.9542
	Dice-BCE-Topo	0.8251	0.8965	0.9000	0.8978	0.9583
	BCE-Topo	**0.8389**	**0.9085**	0.8941	**0.9110**	**0.9607**
ConvNexT ; Attention-ConvNexT	Dice-BCE	0.8352	0.9051	**0.9135**	0.9012	0.9574
	Dice-BCE-Topo	0.8326	0.9072	0.9079	0.9125	0.9649
	BCE-Topo	**0.8385**	**0.9097**	0.9134	**0.9127**	**0.9659**
M. ConvNexT ; Attention-M. ConvNexT	Dice-BCE	0.8474	0.9164	0.9098	**0.9256**	0.9656
	Dice-BCE-Topo	0.8340	0.9080	0.9078	0.9126	0.9614
	BCE-Topo	**0.8551**	**0.9208**	**0.9279**	0.9169	**0.9675**

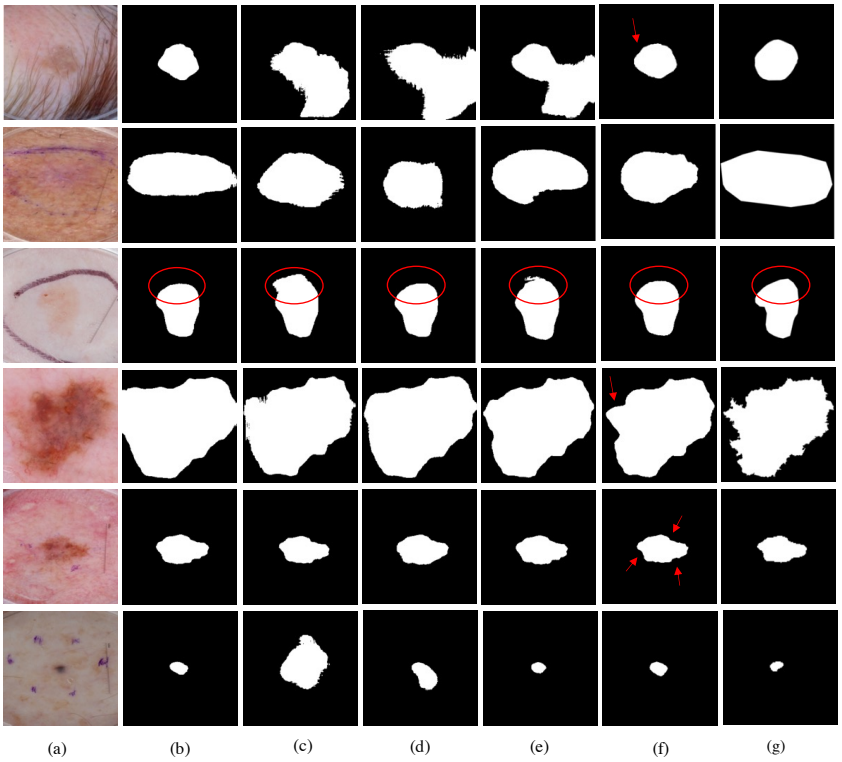

Figure 11: Skin Lesion Segmentation with Diverse Architectures on ISIC 2018: (a) Input Images, (b) ConvNext for All Stages, (c) Traditional Transformer Encoder for the Last Two Stages, (d) MLP Replaced with ConvNext in the Last Two Stages from (c), (e) ConvNext Replaced with Modified ConvNext in the Last Two Stages from (d), (f) Topology-Aware Att-Next (g) Ground Truth.

tends to over-smooth lesion boundaries, reducing precision. The Traditional Transformer Encoder for the last two stages captures more global features, improving context but sometimes causing less accurate boundary details. Replacing the MLP with ConvNext in the last two stages enhances boundary detection, resulting in more precise segmentation by retaining local feature details, yet some segmentation errors remain. The best performance is observed when Modified ConvNext is used for all stages, along with the MLP in the transformer encoder, effectively balancing local feature extraction with global dependencies. Red arrows indicate areas where segmentation was successfully captured, while red circles highlight improved consistency in feature segmentation across lesions. The Att-Next model with topological awareness demonstrates the most accurate visual segmentation results, producing clearer boundaries and fewer misclassified regions compared to other architectures.

Figure 12 shows the segmentation results of skin lesions on the ISIC 2018 dataset using different loss functions. Among all models, the one trained with BCE-TopoLoss demonstrates the most precise segmentation, accurately capturing lesion boundaries and minimizing misclassifications.

HAM10000 Dataset: Table 7 presents the effect of different loss functions on the segmentation performance for the HAM10000 dataset, evaluated using the best-performing model ar-

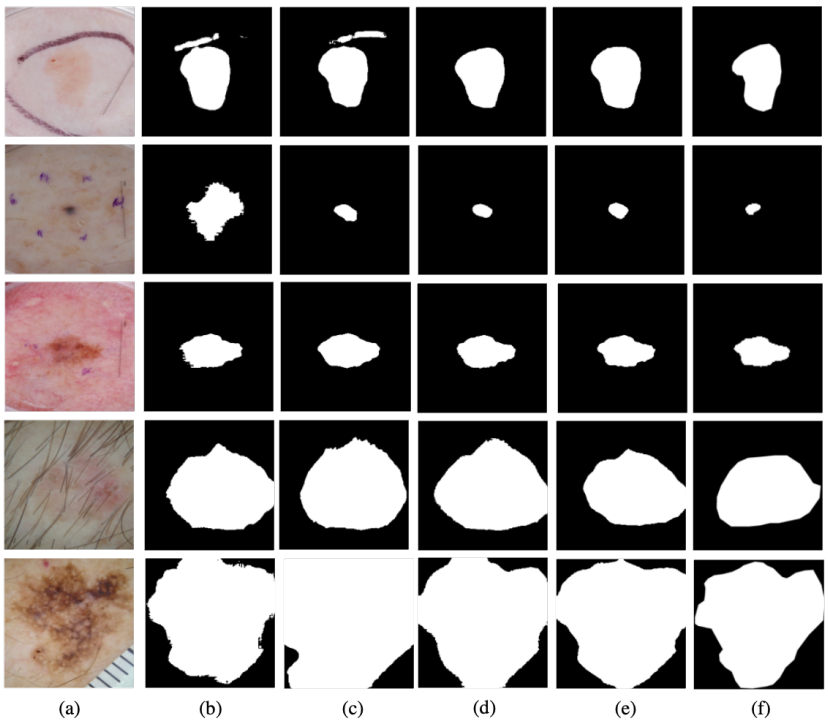

Figure 12: Skin Lesion Segmentation with Varied Loss Functions on ISIC 2018 Dataset. (a) Input Images, (b) Model trained with Dice&TopoLoss, (c) Model trained with Dice&BCE Loss, (d) Model trained with Dice&BCE&TopoLoss, (e) Model trained with BCE&TopoLoss, (f) Ground Truth

chitecture from the ISIC 2018 dataset. The BCE-Topo loss function consistently achieves the highest scores across most evaluation metrics, enhancing the model's ability to accurately capture the essential features of skin lesions. The inclusion of topological loss along with BCE seems to aid in achieving more accurate segmentation, as it improves the model's understanding of lesion structures. The Dice-BCE-Topo loss function also performs well, particularly in terms of boundary detection, suggesting that combining Dice loss with topological awareness can enhance segmentation consistency. However, it does not outperform the BCE-Topo combination. The Dice-BCE loss, while effective in providing reasonable segmentation results, shows lower performance compared to BCE-Topo, highlighting the importance of the topological component in improving segmentation accuracy and robustness.

Table 7: Impact of Various Loss Functions on Skin Lesion Segmentation Performance for the HAM10000 Dataset

Loss Function	**IoU**	**Dice**	**Prec.**	**Rec.**	**Acc**
Dice-BCE-Topo	0.8867	0.9388	0.9422	0.9383	0.9667
Dice-BCE	0.8702	0.9255	0.9346	0.9183	0.9508
BCE-Topo	**0.902**	**0.9442**	**0.9435**	**0.9469**	**0.9701**

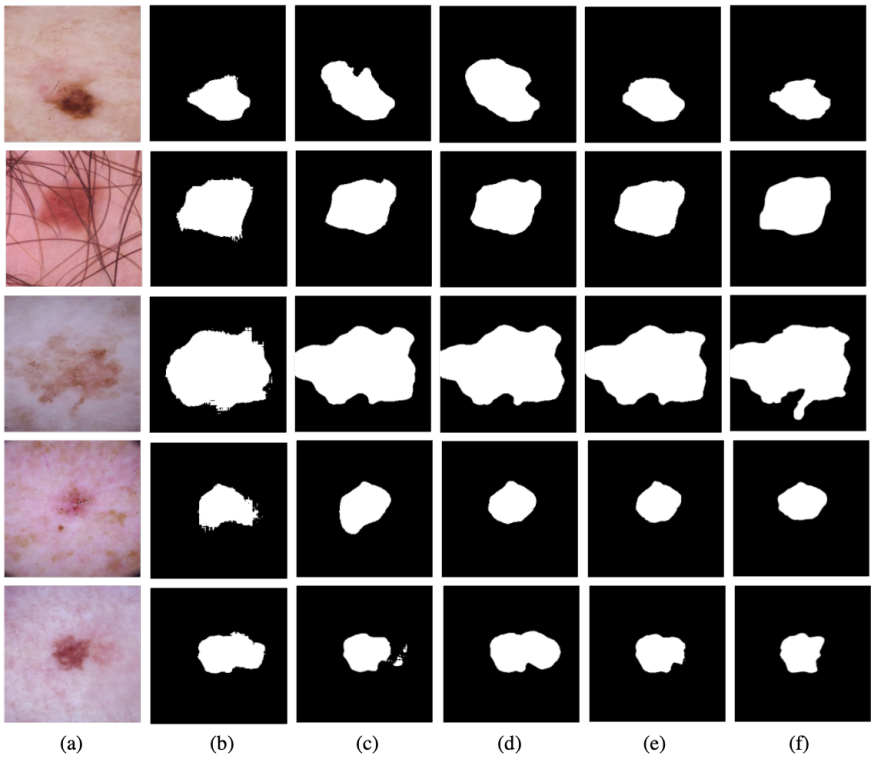

Figure 13: Skin Lesion Segmentation with Varied Loss Functions on HAM 10000 dataset (a) Input Images, (b) Model trained with Dice&TopoLoss, (c) Model trained with Dice&BCE Loss, (d) Model trained with Dice&BCE&TopoLoss, (e) Model trained with BCE&TopoLoss, (f) Ground Truth

As depicted in Figure 13, the model trained with BCE-TopoLoss similarly exhibits the best segmentation results among all evaluated loss functions. The outputs demonstrate superior boundary accuracy and reduced noise compared to other loss functions, effectively capturing the diverse and complex patterns.

5.4. Main Findings

Our proposed topology-aware Att-Next model demonstrates improved segmentation accuracy and generalization performance across various datasets. On the ISIC 2018 dataset, the Att-Next model achieved an IoU of 0.8529 and an DSC of 0.913, surpassing existing approaches. When tested on unseen datasets after training on ISIC 2018, the model exhibited robust performance. On ISIC 2016, it achieved an IoU of 0.8909, a Dice score of 0.9420, and an accuracy of 0.9698. On the HAM10000 dataset, it recorded an IoU of 0.8329, a Dice score of 0.9072, and a recall of 0.9105. Similarly, on the PH2 dataset, the Att-Next model achieved an IoU of 0.8635, a Dice score of 0.9263, and a precision of 0.8839. On the HAM10000 dataset, it achieved an IoU of 0.902 and an DSC of 0.9442, consistently outperforming existing methods.

To further evaluate generalization, we tested it on three unseen datasets: ISIC 2018, ISIC 2016, and PH2, as shown in Table 5. The proposed Att-Next model again demonstrated superior performance compared to given SOTA approaches. On the ISIC 2018 dataset, Att-Next achieved

an IoU of 0.8127 and a Dice score of 0.8938. For ISIC 2016, the model attained an IoU of 0.8702 and a Dice score of 0.9302, outperforming the other methods. On the PH2 dataset, Att-Next achieved the highest IoU of 0.8594 and a Dice score of 0.9240, while also recording the best recall of 0.9882.

Our model design was compared to various architectural configurations presented in the ablation study. The findings, as illustrated in Figure 11, demonstrate that incorporating modified ConvNeXt modules with attention mechanisms captures local details and global structures. As the model architecture evolves, we observe progressively better segmentation outputs, particularly noticeable in the improved delineation of lesion boundaries.

In addition to architectural modifications, we evaluated the model with different loss combinations. While Dice-BCE loss is effective for general segmentation tasks, it exhibits limitations in handling complex topological structures like skin lesions. It focuses on pixel-wise overlap and probability alignment, often failing to accurately delineate boundaries or penalize irrelevant tumor-like regions. This approach struggles with small or subtle lesion areas due to class imbalance and lacks the ability to enforce global or local structural consistency.

In contrast, Topo-BCE loss leverages persistent homology to capture the lesion's topological features, such as connected components (H_0) and loops (H_1), ensuring structural integrity. By penalizing discrepancies in these features using the Wasserstein distance, Topo-BCE loss improves boundary accuracy, suppresses irrelevant regions, and enhances robustness to noise and artifacts, making it better suited for tasks requiring precise and topologically accurate segmentation.

5.5. Limitations and Future Recommendations

We employ modified ConvNeXt modules and attention in the deeper layers to mitigate the heavy computational cost, particularly during matrix multiplication operations in multi-head self-attention. Future work could explore integrating attention into the initial layers or fusion modules.

While our proposed topology-aware Att-Next model achieves robust segmentation performance, we observe that the computational cost associated with training using topological loss remains a challenge. The complexity of computing topological features from a cubical filtration is typically $O(n^w)$, where n is the input size and w is the matrix multiplication exponent, currently estimated to be approximately 2.376. This limitation highlights the need for more efficient methods to compute these topological features (Wagner et al.).

We also observed that combining Dice and Topological loss can lead to gradient explosion, necessitating careful tuning of the loss coefficients to ensure stable training. This combination remains an open area for further research, as it requires addressing gradient instability and identifying optimal loss coefficient strategies for improved performance.

6. Conclusions

In this work, we introduced a novel hybrid model, topology-aware Att-Next, which combines ConvNeXt and Transformer architectures to address key challenges in lesion segmentation. Our model was trained on the ISIC 2018 and HAM10000 datasets, and its effectiveness was validated on four public datasets: ISIC 2018, HAM10000, ISIC 2016, and PH2.

To enhance feature extraction, the ConvNeXt modules were modified by reducing the kernel size and increasing the number of normalization layers and activation functions. These modifications enable the model to capture finer local details and improve training stability. Additionally,

the MLP in the Transformer encoder was replaced with a modified ConvNeXt module, allowing the model to better capture local spatial relationships and improve overall performance.

The integration of topological loss in our model represents a key contribution to skin lesion segmentation. This approach enables the model to capture the underlying topological structure of the data, enhancing its capacity to extract critical topological information. Ablation studies underscore the significant impact of this design choice on model performance, particularly in improving segmentation precision by effectively removing irrelevant tumor regions.

By incorporating topological loss, Att-Next demonstrates superior performance compared to several recently published models, as shown in Tables 1 and 3. The model achieves the highest DSC and IoU values among all compared methods, particularly excelling in handling imbalanced and noisy datasets.

Overall, our findings demonstrate that the topology-aware Att-Next is a highly effective solution for lesion segmentation. These results offer promising insights for future advancements in medical image analysis.

References

Bougourzi, F., Dornaika, F., Distante, C., Taleb-Ahmed, A., 2024. D-trattunet: Toward hybrid cnn-transformer architecture for generic and subtle segmentation in medical images. Computers in Biology and Medicine 176, 108590. doi:10.1016/j.compbiomed.2024.108590.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2023. Swin-unet: Unet-like pure transformer for medical image segmentation, in: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.), Computer Vision – ECCV 2022 Workshops, Springer Nature Switzerland, Cham. pp. 205–218.

Carlsson, G., 2009. Topology and data. Bulletin of the American Mathematical Society 46, 255–308.

Chazal, F., Michel, B., 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in artificial intelligence 4, 667963. doi:10.3389/frai.2021.667963.

Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K., 2024a. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Transactions on Emerging Topics in Computational Intelligence 8, 55–68. doi:10.1109/TETCI.2023.3309626.

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A., Zhou, Y., 2024b. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis 97, 103280. doi:10.1016/j.media.2024.103280.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. ArXiv abs/1706.05587.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. doi:10.1109/CVPR.2017.195.

Clough, J.R., Oksuz, I., Byrne, N., Aboagye, E.O., Montana, G., King, A.P., Schnabel, J.A., 2020. Topological loss functions for deep learning based image segmentation using persistent homology. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 1555–1566. doi:10.1109/TPAMI.2019.2958821.

Coskunuzer, B., Akçora, C.G., 2024. Topological methods in machine learning: A tutorial for practitioners. `arXiv:5833867`.

Demir, A., Massaad, E., Kiziltan, B., 2023. Topology-aware focal loss for 3d image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 580–589.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Garbe, C., Peris, K., Hauschild, A., Saiag, P., Middleton, M., Bastholt, L., Grob, J.J., Malvehy, J., Newton-Bishop, J., Stratigos, A.J., Pehamberger, H., Eggermont, A.M., 2016. Diagnosis and treatment of melanoma. european consensus-based interdisciplinary guideline – update 2016. European Journal of Cancer 63, 201–217. doi:10.1016/j.ejca.2016.05.005.

Garrison, Z.R., Hall, C.M., Fey, R.M., Clister, T., Khan, N., Nichols, R., Kulkarni, R.P., 2023. Advances in early detection of melanoma and the future of at-home testing. Life 13. doi:10.3390/life13040974.

Gupta, S., Hu, X., Kaan, J., Jin, M., Mpoy, M., Chung, K., Singh, G., Saltz, M., Kurc, T., Saltz, J., Tassiopoulos, A., Prasanna, P., Chen, C., 2022. Learning Topological Interactions for Multi-Class Medical Image Segmentation, in:

Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham. pp. 701–718. doi:10.1007/978-3-031-19818-2˙40.

Han, Z., Jian, M., Wang, G.G., 2022. Convunext: An efficient convolution neural network for medical image segmentation. Knowledge-Based Systems 253. doi:10.1016/j.knosys.2022.109512.

Hao, S., Zhang, L., Jiang, Y., Wang, J., Ji, Z., Zhao, L., Ganchev, I., 2023. Convnext-st-aff: A novel skin disease classification model based on fusion of convnext and swin transformer. IEEE Access 11, 117460–117473. doi:10.1109/ACCESS.2023.3324042.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. UNETR: Transformers for 3d medical image segmentation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 1748–1758. doi:10.1109/WACV51458.2022.00181.

Hofer, C., Graf, F., Niethammer, M., Kwitt, R., 2020. Topologically densified distributions, in: International Conference on Machine Learning, PMLR. pp. 4304–4313.

Hofer, C., Kwitt, R., Niethammer, M., Uhl, A., 2017. Deep learning with topological signatures. Advances in neural information processing systems 30.

Hu, K., Lu, J., Lee, D., Xiong, D., Chen, Z., 2022. As-net: Attention synergy network for skin lesion segmentation. Expert Systems with Applications 201, 117112. doi:10.1016/j.eswa.2022.117112.

Hu, X., Chen, Y., Guibas, L.J., Nie, Q., Bao, F., 2019. Topology-preserving deep image segmentation using discrete morse theory, in: Advances in Neural Information Processing Systems, pp. 3384–3395.

Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y., 2023. Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging 42, 1484–1494. doi:10.1109/TMI.2022.3230943.

Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D., 2020. DoubleU-Net: A deep convolutional neural network for medical image segmentation, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems, Institute of Electrical and Electronics Engineers Inc.. pp. 558–564. doi:10.1109/CBMS49503.2020.00111.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

Love, E.R., Filippenko, B., Maroulas, V., Carlsson, G., 2023. Topological convolutional layers for deep learning. Journal of Machine Learning Research 24, 1–35.

Mallick, S., Paul, J., Sil, J., 2023. Response fusion attention u-convnext for accurate segmentation of optic disc and optic cup. Neurocomputing 559, 126798. doi:10.1016/j.neucom.2023.126798.

Marghoob, N.G., Liopyris, K., Jaimes, N., 2019. Dermoscopy: a review of the structures that facilitate melanoma detection. Journal of Osteopathic Medicine 119, 380–390.

Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J., 2013. Ph2 - a dermoscopic image database for research and benchmarking, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437–5440. doi:10.1109/EMBC.2013.6610779.

Mosinska, A., Marquez-Neila, P., Kozinski, M., Fua, P., . Beyond the Pixel-Wise Loss for Topology-Aware Delineation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3136–3145. doi:10.1109/CVPR.2018.00331.

Niu, Z., Deng, Z., Gao, W., Bai, S., Gong, Z., Chen, C., Rong, F., Li, F., Ma, L., 2024. Fnexter: A multi-scale feature fusion network based on convnext and transformer for retinal oct fluid segmentation. Sensors 24, 2425. doi:10.3390/s24082425.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

Papamarkou, T., Birdal, T., Bronstein, M., Carlsson, G., Curry, J., Gao, Y., Hajij, M., Kwitt, R., Liò, P., Di Lorenzo, P., et al., 2024. Position paper: Challenges and opportunities in topological deep learning. arXiv preprint arXiv:2402.08871 .

Roky, A.H., Islam, M.M., Ahasan, A.M.F., Mostaq, M.S., Mahmud, M.Z., Amin, M.N., Mahmud, M.A., 2025. Overview of skin cancer types and prevalence rates across continents. Cancer Pathogenesis and Therapy 3, 89–100. doi:10.1016/j.cpt.2024.08.002.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241. doi:10.1007/978-3-319-24574-4˙28.

Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jäger, P.F., Maier-Hein, K.H., . MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation, in: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland. pp. 405–415. doi:10.1007/978-3-031-43901-8˙39.

Vandaele, R., Nervo, G.A., Gevaert, O., 2020. Topological image modification for object detection and topological image

processing of skin lesions. Scientific Reports 10, 21061. doi:10.1038/s41598-020-77933-y.

Wagner, H., Chen, C., Vuçini, E., . Efficient Computation of Persistent Homology for Cubical Data, in: Peikert, R., Hauser, H., Carr, H., Fuchs, R. (Eds.), Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications. Springer, pp. 91–106. doi:10.1007/978-3-642-23175-9˙7.

Wang, C., Zhang, J., He, J., Luo, W., Yuan, X., Gu, L., 2023. A two-stream network with complementary feature fusion for pest image classification. Engineering Applications of Artificial Intelligence 124, 106563. doi:10.1016/j.engappai.2023.106563.

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z., 2022. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Medical Image Analysis 76, 102327. doi:10.1016/j.media.2021.102327.

Xu, G., Zhang, X., He, X., Wu, X., 2024. LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation, in: Liu, Q., Wang, H., Ma, Z., Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R. (Eds.), Pattern Recognition and Computer Vision, Springer Nature, Singapore. pp. 42–53. doi:10.1007/978-981-99-8543-2˙4.

Yang, J., Hu, X., Chen, C., Tsai, C., 2021. A Topological-Attention ConvLSTM Network and Its Application to EM Images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 217–228. doi:10.1007/978-3-030-87193-2˙21.

Zhang, H., Zhong, X., Li, G., Liu, W., Liu, J., Ji, D., Li, X., Wu, J., 2023. Bcu-net: Bridging convnext and u-net for medical image segmentation. Computers in Biology and Medicine 159, 106960. doi:10.1016/j.compbiomed.2023.106960.

Zhang, Y., Liu, H., Hu, Q., 2021. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 14–24. doi:10.1007/978-3-030-87193-2_2.