**《深度学习与计算机视觉》课程论文**

**(2024 - 2025学年第二学期)**

**A Comparative Study of Deep Learning Architectures for Automated Skin Lesion Segmentation**

**小组序号：4**

**学生姓名：李天翔、方炜轩、江一航、惠志文**

**提交日期：2025年 7 月 13日**

**学生签名： 李天翔**

|  |  |  |  |
| --- | --- | --- | --- |
| **组长姓名** | **李天翔** | **组长学号** | **202364870851** |
| **成员1姓名** | **方炜轩** | **成员1学号** | **202330840411** |
| **成员2姓名** | **惠志文** | **成员2学号** | **202364870731** |
| **成员3姓名** | **江一航** | **成员3学号** | **202330361171** |
| **学 院** | **未来技术学院** | **专业班级** | **23级人工智能1班** |
| **课程编号** | 084100461 | **课程英文名称** | Deep Learning and Computer Vision |
| **课程名称** | **深度学习与计算机视觉** | **任课教师** | **张鑫、欧士琪、张庭赫** |
| **教师评语：** | | | |
| **本论文成绩评定：** **分** | | | |

**Content**

[**1 Introduction** 2](#_Toc203140615)

[**1.1. Background and Significance** 2](#_Toc203140616)

[**1.2. Problem Statement and Challenges** 3](#_Toc203140617)

[**1.3. Our Work and Contributions** 5](#_Toc203140618)

[**1.4. Organization of This Paper** 6](#_Toc203140619)

[**2. Related Work** 6](#_Toc203140620)

[**2.1. Traditional Medical Image Segmentation Methods** 7](#_Toc203140621)

[**2.2. Deep Learning for Image Segmentation** 7](#_Toc203140622)

[**2.3. UNet and its Applications in Medical Imaging** 8](#_Toc203140623)

[**3. Methodology** 9](#_Toc203140624)

[**3.1. Dataset and Preprocessing** 9](#_Toc203140625)

[**3.2. Baseline Model: The Standard UNet** 10](#_Toc203140626)

[**3.3. Model for Comparison** 12](#_Toc203140627)

[**3.4. Implementation Details** 13](#_Toc203140628)

[**4. Team Member Roles and Contributions** 15](#_Toc203140629)

[**5. Experiments and Result Analysis** 16](#_Toc203140630)

[**5.1. Evaluation Metrics** 16](#_Toc203140631)

[**5.2. Quantitative Results and Analysis** 18](#_Toc203140632)

[**5.3. Qualitative Results and Analysis** 20](#_Toc203140633)

[**6. Discussion** 24](#_Toc203140634)

[**6.1. Analysis of Findings** 24](#_Toc203140635)

[**6.2. Limitations and Future Work** 25](#_Toc203140636)

[**7. Conclusion** 26](#_Toc203140637)

[**References:** 27](#_Toc203140638)

**A Comparative Study of Deep Learning Architectures for Automated Skin Lesion Segmentation**

**李天翔、方炜轩、江一航、惠志文**

**Abstract:** Accurate segmentation of skin lesions from dermoscopic images is a critical yet challenging task for the computer-aided diagnosis of melanoma. This study systematically evaluates three deep learning architectures—the foundational UNet, the advanced convolutional variant UNet++, and the hybrid Transformer-based HMT-UNet—on the public ISIC 2017 dataset to identify a robust segmentation solution. We implemented a comprehensive data preprocessing and augmentation pipeline to enhance model generalization. Our experiments demonstrate that the UNet++ architecture, when combined with a strong data augmentation strategy, achieves superior performance, yielding the highest average Dice Coefficient of 0.7488. The results also unequivocally show that data augmentation is a critical factor for success, significantly improving performance across all models and boosting the Dice score of UNet++ by over 15%. In conclusion, this work identifies the UNet++ architecture with data augmentation as a highly effective framework for this task, highlighting the benefits of advanced feature fusion and data diversity for training robust medical image segmentation models.

**Keywords：**Medical Image Segmentation, Deep Learning, Skin Lesion, UNet

**1 Introduction**

**1.1. Background and Significance**

In the landscape of modern medicine, medical imaging has emerged as an indispensable tool, revolutionizing clinical diagnosis, treatment planning, and disease monitoring. Technologies such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and digital pathology provide unprecedented insights into human anatomy and pathology. Among various diseases, skin cancer, particularly malignant melanoma, represents a significant global health concern. According to the World Health Organization, melanoma is the most aggressive form of skin cancer, with incidence rates steadily increasing over the past few decades[1]. While it is highly curable if detected at an early stage, its mortality rate becomes substantial once it metastasizes to other parts of the body. This underscores the critical importance of early and accurate diagnosis.

Dermoscopy is a non-invasive imaging technique that has become the de-facto standard for the early screening of pigmented skin lesions. By illuminating the skin and magnifying the image, it allows dermatologists to visualize subsurface skin structures not visible to the naked eye, thereby improving diagnostic accuracy compared to simple visual inspection. A crucial step in the dermoscopic analysis pipeline is the precise delineation, or segmentation, of the lesion's boundary. Accurate segmentation is not merely an end in itself; it is a foundational prerequisite for subsequent quantitative analysis within Computer-Aided Diagnosis (CAD) systems. Key diagnostic indicators, often summarized by the "ABCDE" rule (Asymmetry, Border irregularity, Color variegation, Diameter, and Evolving), are derived directly from the segmented lesion's shape, size, and color properties[2]. An accurate segmentation mask enables the objective and reproducible extraction of these critical features, reducing the diagnostic subjectivity and intra-observer variability inherent in manual assessment.

However, the manual segmentation of skin lesions is a laborious, time-consuming task that demands significant expertise from dermatologists. The global shortage of trained specialists, coupled with a rising number of cases, creates a bottleneck in the clinical workflow. Therefore, the development of automated, accurate, and efficient systems for skin lesion segmentation holds immense clinical value. By leveraging the power of deep learning, such systems can serve as a valuable "second opinion" for clinicians, accelerating the diagnostic process, improving its consistency, and ultimately facilitating earlier detection of melanoma, which can significantly improve patient outcomes.

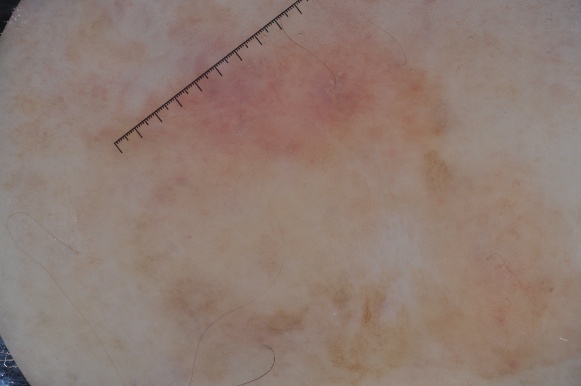
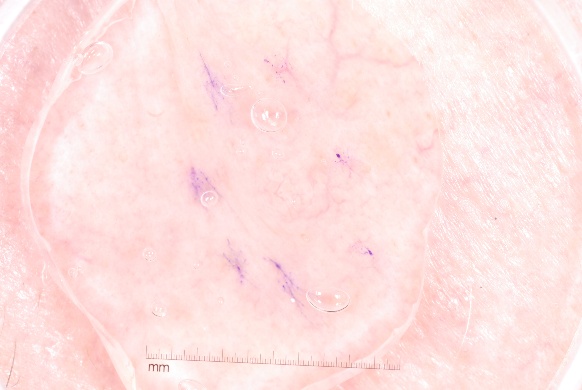
**1.2. Problem Statement and Challenges**

This paper addresses the problem of automatic semantic segmentation of skin lesions from 2D dermoscopic images. The goal is to develop a deep learning model that takes a dermoscopic image as input and outputs a binary mask of the same dimensions, where each pixel is classified as either belonging to the lesion or the surrounding background skin. For this purpose, we utilize the publicly available ISIC 2017 dataset, a standard benchmark for this task[3].

Despite the conceptual simplicity of this task, achieving high-precision segmentation of skin lesions is fraught with significant challenges, which are well-documented in the literature[4]. These challenges stem from the intrinsic properties of the lesions and the imaging process itself. We identify the primary difficulties as follows:

* **Low Contrast and Fuzzy Borders:** There is often a subtle and ambiguous transition between the lesion and the surrounding healthy skin. The lack of a sharp, well-defined edge, particularly in early-stage or regressing melanomas, makes precise boundary delineation inherently difficult for both human experts and automated algorithms.
* **High Inter-Class and Intra-Class Variation:** Skin lesions exhibit a vast diversity in terms of their color, size, shape, and texture. Furthermore, a significant visual similarity can exist between malignant melanomas and benign nevi (moles), making it challenging for a model to learn discriminative features that are robust across different lesion types.
* **Presence of Image Artifacts:** Dermoscopic images are frequently corrupted by various artifacts that can obscure the lesion boundary and mislead the segmentation model. Common artifacts include:
  + **Hair:** Strands of hair covering the lesion are a prevalent issue. They introduce strong, high-frequency noise that can be misinterpreted as part of the lesion's texture or boundary.
  + **Ruler Markings:** Clinical images sometimes contain ruler marks for scale, which can intersect with the lesion.
  + **Air Bubbles and Specular Reflections:** The use of immersion fluid in dermoscopy can lead to air bubbles and light reflections, creating spurious patterns on the image.
* **Irregular and Complex Shapes:** Malignant lesions are often characterized by highly irregular and asymmetrical shapes, which are more complex to model compared to the typically round or oval shapes of benign lesions.

To visually illustrate these difficulties, Figure 1 presents several example images from the ISIC 2017 dataset. These examples highlight the challenges of fuzzy borders (Fig. 1a), low contrast (Fig. 1b), the presence of dense hair (Fig. 1c), and significant color variation (Fig. 1d), all of which a successful segmentation model must overcome.



(a) (b)



(c) (d)

Figure 1. Examples of challenges in the ISIC 2017 dataset. (a) A lesion with a fuzzy and indistinct border. (b) Low contrast between the lesion and the surrounding skin. (c) The lesion is heavily obscured by hair. (d) Significant color variegation within a single lesion.

**1.3. Our Work and Contributions**

To address the aforementioned challenges, this paper presents a systematic, comparative study of deep learning models for skin lesion segmentation. We implement a complete pipeline from data preprocessing to model evaluation and provide a thorough analysis of the results. The main contributions of our work are as follows:

1. **Systematic Comparative Analysis:** We implement and rigorously evaluate three distinct deep learning architectures: the classic **UNet**, the advanced convolutional variant **UNet++**, and the modern hybrid **HMT-UNet**. This provides a valuable comparative insight into the effectiveness of different architectural paradigms—from foundational CNNs to those with enhanced feature fusion and Transformer-based components—on the specific task of skin lesion segmentation.
2. **Demonstration of Augmentation's Critical Role:** We quantitatively demonstrate the crucial impact of a comprehensive data augmentation strategy. Our experiments reveal that while advanced architectures like UNet++ have higher learning capacity, this potential is only fully realized when trained with sufficient data diversity. Our augmentation pipeline led to a substantial performance increase across all models, most notably improving the Dice score of UNet++ by over 15%.
3. **Identification of a High-Performing Model:** Through our experiments, we identify that the **UNet++ model, when combined with data augmentation, achieves the best performance**, yielding a Dice Coefficient of 0.7488 on the ISIC 2017 test set. This provides a strong and reproducible benchmark result for this task.

**1.4. Organization of This Paper**

The remainder of this paper is structured as follows.

* **Chapter 2** reviews the relevant literature on both traditional and deep learning-based image segmentation methods, with a focus on the evolution of the UNet architecture.
* **Chapter 3** details our methodology, including a description of the dataset, the preprocessing pipeline, the architectures of the three models under investigation, and our experimental setup.
* **Chapter 4** outlines the specific roles and contributions of each team member.
* **Chapter 5** presents and analyzes the experimental results, offering both quantitative comparisons and qualitative visual insights.
* **Chapter 6** discusses the implications of our findings, acknowledges the limitations of our study, and proposes directions for future research. Finally,
* **Chapter 7** concludes the paper with a summary of our work and its key findings.

**2. Related Work**

This section provides a comprehensive review of the literature pertinent to medical image segmentation. We begin by briefly discussing traditional image processing techniques, followed by an overview of the paradigm-shifting deep learning methods. Finally, we focus specifically on the U-Net architecture and its influential variants, which form the foundation of our work.

**2.1. Traditional Medical Image Segmentation Methods**

Before the advent of deep learning, medical image segmentation was dominated by a variety of "classical" image processing and machine learning algorithms. These methods typically rely on hand-crafted features based on pixel intensity, texture, and shape information. Broadly, they can be categorized into several groups.

Thresholding-based methods, such as Otsu's method, partition an image by creating a binary map based on an optimal intensity threshold[5]. While simple and computationally efficient, they are highly sensitive to noise and intensity variations, making them unsuitable for images with complex backgrounds and low-contrast objects, like the skin lesions in our study. Region-based methods, including region growing and watershed algorithms, group pixels into regions based on predefined similarity criteria[6]. These approaches can produce contiguous and smooth segmentation results but often require manual seed point selection and can suffer from over-segmentation. Another significant category is the active contour models, or "snakes," which evolve a curve (contour) from an initial position to fit the boundary of a target object by minimizing an energy function[7]. These models are effective for defining object boundaries but are sensitive to initial placement and can struggle with complex topologies.

While these traditional methods have contributed to the field, they share a common limitation: their performance is heavily dependent on domain-specific feature engineering and parameter tuning. They often lack the robustness and generalization capability required to handle the vast diversity present in medical imaging data, such as the varied appearance of skin lesions.

**2.2. Deep Learning for Image Segmentation**

The landscape of computer vision, and medical image analysis in particular, was fundamentally transformed by the success of deep learning, specifically Convolutional Neural Networks (CNNs). The breakthrough for segmentation tasks came with the introduction of the Fully Convolutional Network (FCN) by Long, Shelhamer, and Darrell[8]. Unlike traditional CNNs which use fully connected layers to produce a single class label for an entire image, FCNs replace these layers with convolutional layers, enabling them to output a pixel-wise prediction map for an input of arbitrary size. This end-to-end, pixels-to-pixels learning framework eliminated the need for hand-crafted features and set a new standard for segmentation performance. FCNs introduced the crucial concept of an encoder-decoder structure, where the encoder path progressively downsamples the input to capture semantic context, and the decoder path upsamples these features to produce a full-resolution segmentation map.

Building on the FCN, numerous advanced architectures have been proposed. The SegNet architecture, for instance, introduced an efficient decoder that uses pooling indices from the corresponding encoder stage to perform non-linear upsampling, reducing the number of parameters[9]. The DeepLab family of models made significant contributions by incorporating atrous (or dilated) convolution, which allows for an expanded receptive field to capture multi-scale context without increasing computational cost or losing resolution[10]. These architectures have demonstrated state-of-the-art performance on natural image segmentation benchmarks and have been widely adapted for medical applications.

**2.3. UNet and its Applications in Medical Imaging**

While FCNs and their successors were powerful, they often required massive datasets for training. In the biomedical domain, where annotated data is often scarce, a more data-efficient architecture was needed. Addressing this gap, Ronneberger et al. introduced the U-Net in 2015, which has since become the de-facto standard for medical image segmentation[11]. The U-Net architecture features a symmetric, U-shaped encoder-decoder structure. Its most defining and impactful innovation is the use of **skip connections**, which concatenate feature maps from the encoder path directly to the corresponding, symmetrically located layers in the decoder path. This simple yet powerful mechanism allows the network to fuse deep, semantic, contextual information from the decoder with shallow, high-resolution, fine-grained features from the encoder. This fusion is critical for precise localization of boundaries, a task of paramount importance in medical imaging, and allows the network to achieve excellent performance even with a limited number of training samples.

The profound success of U-Net has inspired a vast ecosystem of variants, each aiming to improve upon the original design. These extensions can be broadly categorized based on their architectural modifications[12]. One major direction involves redesigning the skip connections. For example, **UNet++** introduced nested and dense skip connections to bridge the semantic gap between the encoder and decoder feature maps, leading to more flexible feature fusion[13]. Another popular trend is the integration of attention mechanisms. The **Attention U-Net** incorporates attention gates (AGs) into the skip connections, which learn to selectively focus on salient features and suppress irrelevant background regions, thereby improving model sensitivity and accuracy without significant computational overhead[14].

More recently, with the rise of Transformers in computer vision, a new wave of hybrid models has emerged. Architectures like **TransUNet**[15] and **Swin-Unet**[16] leverage a Transformer-based encoder to capture global, long-range dependencies—a known limitation of standard CNNs—while retaining a CNN-based decoder to reconstruct fine-grained details. These models aim to combine the best of both worlds: the global context modeling of Transformers and the robust local feature extraction of CNNs. Other works, such as **TransAttUnet**[17], further explore the synergy between attention mechanisms and Transformers within the U-Net framework. This continuous evolution demonstrates the enduring influence of the U-Net architecture and the ongoing quest to enhance its capabilities for the ever-growing challenges in medical image segmentation.

**3. Methodology**

This section details the technical framework of our study. We first introduce the dataset employed for our experiments and the comprehensive preprocessing pipeline designed to standardize the data. Subsequently, we provide a detailed architectural description of our baseline model, the standard U-Net, which forms the foundation for our segmentation task.

**3.1. Dataset and Preprocessing**

**Dataset Description:** Our research is conducted on the **ISIC 2017 Challenge dataset**, a widely recognized public benchmark for skin lesion analysis[3]. Specifically, we focus on "Task 1: Lesion Segmentation," which is designed to evaluate the performance of algorithms in accurately delineating lesion boundaries. The dataset is composed of three distinct subsets: a training set containing 2,000 dermoscopic images, a validation set with 150 images, and a test set with 600 images. Each image is provided in JPEG format. Corresponding to each image is a high-quality, expert-annotated ground truth mask provided as a single-channel 8-bit PNG file. In these masks, a pixel value of 255 represents the lesion area, while a value of 0 signifies the background skin.

**Preprocessing Pipeline:** A standardized preprocessing pipeline is crucial for ensuring data consistency and optimizing model performance. Based on our implementation (preprocess.py), every image and its corresponding mask undergo a series of automated steps before being fed into the neural network.

1. **Image and Mask Resizing:** The original images and masks in the dataset come in various resolutions. To ensure a uniform input size for our network, all images and masks are resized to a fixed dimension of **256 × 256 pixels**. We employ bilinear interpolation for resizing the images, as it provides a smooth and accurate approximation of pixel values. For the segmentation masks, nearest-neighbor interpolation is used. This is a critical choice to preserve the discrete nature of the class labels (lesion vs. background) and prevent the introduction of erroneous intermediate pixel values along the boundaries.
2. **Mask Binarization:** The ground truth masks, though grayscale, are converted into a strict binary format. A threshold is applied at a pixel value of 128; any pixel with an intensity greater than this threshold is mapped to 1 (lesion), and all other pixels are mapped to 0 (background). This ensures a clean, two-class target for the model to learn.
3. **Normalization:** To facilitate stable and efficient model training, the pixel values of the input images are normalized. The 8-bit integer values, originally in the range [0, 255], are scaled to a floating-point range of [0.0, 1.0] by dividing each pixel value by 255. This normalization step is applied during the data loading process before the images are passed to the model.

**3.2. Baseline Model: The Standard UNet**

Our primary framework for this study is the **U-Net architecture**, a convolutional neural network that has proven exceptionally effective for biomedical image segmentation tasks[11]. Our implementation, detailed in model.py, adheres to the canonical U-shaped structure, which consists of a contracting (encoder) path and an expansive (decoder) path, connected by a bottleneck. The architecture is illustrated in Figure 2.

**Encoder (Contracting Path):** The purpose of the encoder is to capture the context and extract hierarchical features from the input image. It consists of four main blocks. Each block begins with a DoubleConv module, which comprises two consecutive 3×3 convolutional layers, with each convolution followed by a Batch Normalization (BN) layer and a Rectified Linear Unit (ReLU) activation function. The BN layer helps to stabilize and accelerate the training process. After the DoubleConv module, a 2×2 max pooling operation with a stride of 2 is applied for downsampling, reducing the spatial dimensions of the feature maps by half while doubling the number of feature channels. The channel depth progressively increases through the encoder: 3 (input RGB) → 64 → 128 → 256 → 512.

**Bottleneck:** At the bottom of the "U," the encoder path transitions to the decoder path through a bottleneck layer. This layer has the smallest spatial resolution and the highest number of feature channels (512 → 1024), allowing it to learn the most complex and abstract representations of the input data.

**Decoder (Expansive Path):** The decoder's role is to gradually upsample the feature maps back to the original image resolution while localizing the features learned in the encoder. It symmetrically mirrors the encoder, consisting of four upsampling blocks. Each block starts with an upsampling operation (in our case, bilinear upsampling) that doubles the spatial dimensions of the feature map and halves the number of channels.

The central feature of the U-Net, the **skip connections**, is implemented at this stage. The upsampled feature map is concatenated with the corresponding feature map from the contracting path at the same level. This fusion is vital, as it re-introduces high-resolution spatial information that was lost during downsampling, allowing the network to make highly precise, pixel-level predictions. After concatenation, the combined feature map is processed by a DoubleConv module. The channel depth progressively decreases through the decoder: 1024 → 512 → 256 → 128 → 64.

**Output Layer:** Finally, a 1×1 convolution is applied to the output of the last decoder block. This layer maps the 64-channel feature map to a single-channel output. A Sigmoid activation function is then applied to this output, generating a final segmentation map where each pixel value is between 0 and 1, representing the model's predicted probability that the pixel belongs to the lesion class.

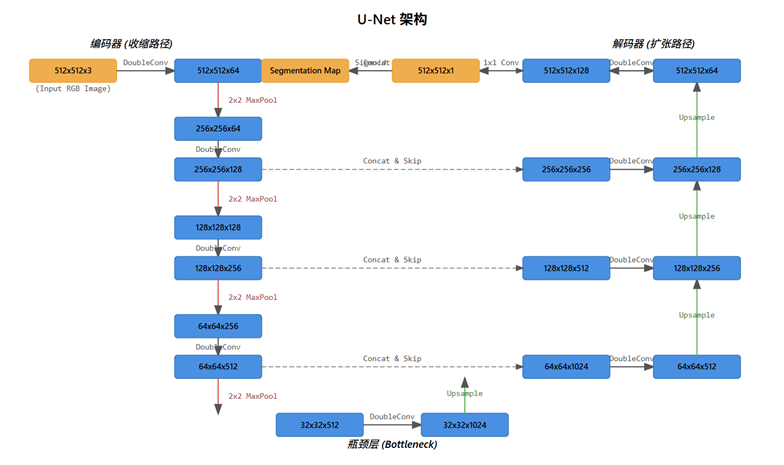


Figure 2. The architecture of our baseline U-Net model. The model consists of a contracting path (encoder) on the left and an expansive path (decoder) on the right. Skip connections concatenate feature maps from the encoder to the decoder at corresponding levels. The numbers above or below the blocks indicate the number of channels.

**3.3. Model for Comparison**

To comprehensively evaluate the performance of our baseline U-Net, we selected two advanced architectures for comparison: UNet++, a well-established variant known for its enhanced feature fusion, and HMT-UNet, a recent model that incorporates the principles of vision transformers.

**3.3.1 UNet++**

The UNet++ architecture, as implemented in our model.py, is an evolution of the standard U-Net designed to bridge the semantic gap between the encoder and decoder pathways[13]. Its core innovation lies in the introduction of **nested and dense skip connections**.

Unlike the original U-Net which directly connects encoder and decoder nodes at the same scale, UNet++ creates a series of nested, densely connected convolutional blocks that fill the space between the two paths. To be more specific, a node in the architecture receives feature maps not only from the previous node in the same skip pathway () but also upsampled feature maps from the lower skip pathway (). The computation for a generic node is defined as:

where represents a convolution block (our DoubleConv module), is an upsampling operation, and Concat denotes the channel-wise concatenation. This dense connectivity allows for the aggregation of features across multiple semantic scales, enabling the model to learn from a more flexible and richer feature set. This design effectively creates an ensemble of U-Nets of varying depths, which can improve segmentation accuracy, particularly for objects with complex shapes and indistinct boundaries.

**3.3.2 HMT-UNet**

The Hierarchical Multi-scale Transformer UNet (HMT-UNet) represents a modern approach that integrates the strengths of Vision Transformers (ViT) into a U-Net-like architecture. This model replaces the standard convolutional blocks with hierarchical Transformer blocks to capture both local and global contextual information more effectively.

The HMT-UNet maintains the classic U-shaped encoder-decoder structure. However, its core building blocks are composed of Transformer layers that operate on image patches. The encoder path progressively reduces the spatial resolution while increasing the feature dimension, similar to a standard U-Net. Each stage in the encoder is a MambaVisionLayer which, despite its name, consists of several Block modules that implement a standard multi-head self-attention mechanism, not a Mamba (SSM) model. This allows the model to learn long-range dependencies between different regions of the image at each scale. To manage computational complexity, the self-attention is calculated within partitioned windows (window\_partition).

Symmetrically, the decoder path uses MambaVisionLayer\_up blocks to upsample the feature maps and fuse them with high-resolution features from the encoder via skip connections. This hybrid design, which uses convolutions for initial patch embedding and final pixel-level prediction, aims to combine the local feature extraction prowess of CNNs with the global context modeling capabilities of Transformers, making it a powerful architecture for complex segmentation tasks.

**3.4. Implementation Details**

All experiments were conducted using the PyTorch deep learning framework on a system equipped with a CUDA-enabled NVIDIA 4090 GPU. The following subsections detail the specific configurations and hyperparameters used for training our models, as defined in our train.py script.

**Data Augmentation:** To enhance the model's robustness and prevent overfitting, we employed a comprehensive online data augmentation strategy for the training set using the Albumentations library. For each training image, the following transformations were applied with a certain probability:

* Geometric transformations: Horizontal and vertical flips (p=0.5), affine transformations including scaling, translation, and rotation (p=0.5), elastic transforms (p=0.5), and grid distortion (p=0.5).
* Color and intensity transformations: Random brightness and contrast adjustments (p=0.5), and color jitter (p=0.5).
* Noise injection: Gaussian noise (p=0.2). The validation and test sets were not augmented, except for resizing and normalization, to ensure a consistent and objective evaluation.

**Loss Function:** For optimizing the models, we utilized a composite loss function that combines Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) and a soft Dice Loss. This hybrid loss function balances pixel-wise classification accuracy with region-based overlap, which is highly effective for segmentation tasks. The total loss Ltotal​ is a weighted sum of the two components:

where ​ penalizes pixel-level classification errors and pushes the model to maximize the overlap (Dice Coefficient) between the prediction and the ground truth.

**Optimizer and Learning Rate Schedule:** We employed the **AdamW** optimizer [18], an extension of Adam that improves weight decay regularization. The optimizer was configured with an initial learning rate of and a weight decay of . To dynamically adjust the learning rate during training, we used a **Cosine Annealing scheduler**. This scheduler gradually decreases the learning rate from its initial value to a minimum of over the course of 100 epochs, following a cosine curve.

**Training Strategy:** Our training protocol was designed as a two-stage process to ensure both robust validation and optimal use of the available data.

* **Stage 1: Architecture and Hyperparameter Validation via K-Fold Cross-Validation.** To rigorously validate our model architectures and training configurations, we first performed a 5-fold cross-validation on the 2,000 images of the official training set. The data was partitioned into five folds, and for each iteration, one fold served as the validation set while the other four were used for training. Models were trained for a maximum of 100 epochs with a batch size of 8. An early stopping mechanism was employed with a patience of 30 epochs: if the validation Dice score did not improve within this window, training was halted. This cross-validation process allowed us to obtain a reliable estimate of each model's generalization performance and confirm the effectiveness of our chosen hyperparameters.
* **Stage 2: Final Model Training on Combined Dataset.** After validating our approach, we combined the original ISIC 2017 training set (2,000 images) and validation set (150 images) to create a larger, unified training dataset of 2,150 images. Using the hyperparameters proven effective during the cross-validation stage, we trained our final models from scratch on this complete dataset. This approach ensures that the final models, which are used for evaluation on the hold-out test set, have learned from all available annotated data, maximizing their performance and generalization capability.

**4. Team Member Roles and Contributions**

This project was a collaborative effort, with each team member taking ownership of key components of the research and development process. The specific roles and contributions of each member are detailed below.

**Tianxiang Li** (Team Leader):

Tianxiang was responsible for the overall project management and technical leadership. His primary duties included defining the project timeline, coordinating tasks among team members, and overseeing the integration and optimization of the final codebase. He single-handedly managed the entire data pipeline, which involved the acquisition, organization, and execution of all data preprocessing scripts to prepare the ISIC 2017 dataset for experimentation. Additionally, he played a principal role in the composition and final revision of this research paper.

**Weixuan Fang** (Member):

Weixuan focused on establishing the foundational baseline for our experiments. He conducted a thorough literature review of the standard U-Net architecture and was responsible for its implementation and debugging. His key technical contribution was the development of the core training scripts for the UNet model. He also contributed significantly to the creation of the slide deck for the final project presentation.

**Yihang Jiang** (Member):

Yihang was in charge of the comparative experimental setup and model optimization. He performed a literature review on advanced segmentation architectures and implemented the models used for comparison, specifically UNet++ and HMT-UNet. His responsibilities included the hands-on training of all models, as well as systematically tuning hyperparameters to achieve optimal performance. He also collaborated on producing the final project presentation.

**Zhiwen Hui** (Member):

Zhiwen led the evaluation and analysis of our experimental outcomes. He was responsible for implementing the code for all evaluation metrics (e.g., Dice, IoU) and developing the testing scripts to systematically assess model performance on the test set. Furthermore, he performed the quantitative analysis of the results and generated all the data visualizations, including tables and figures, presented in this paper. He also contributed to the writing of the research paper.

**5. Experiments and Result Analysis**

This chapter presents the empirical evaluation of our proposed models. We begin by defining the quantitative metrics used for performance assessment. We then report and analyze the comparative results of our experiments, providing both quantitative and qualitative insights into the models' capabilities.

**5.1. Evaluation Metrics**

To quantitatively and objectively assess the performance of our segmentation models, we employ a comprehensive set of widely recognized evaluation metrics. These metrics are calculated by comparing the model's predicted segmentation mask against the expert-annotated ground truth mask on a pixel-by-pixel basis. The calculations are based on four fundamental quantities derived from the confusion matrix:

* **True Positive (TP):** The number of pixels correctly classified as lesion.
* **False Positive (FP):** The number of pixels incorrectly classified as lesion (background pixels mistaken for lesion).
* **False Negative (FN):** The number of pixels incorrectly classified as background (lesion pixels missed by the model).
* **True Negative (TN):** The number of pixels correctly classified as background.

Based on these quantities, we define the following metrics:

**1. Dice Coefficient (DSC):**

Also known as the F1-score, the Dice Coefficient is one of the most common metrics for evaluating segmentation performance, particularly in medical imaging. It measures the overlap between the predicted and ground truth regions, ranging from 0 (no overlap) to 1 (perfect overlap). It is generally considered more robust than accuracy for imbalanced datasets, such as ours where the lesion area is often much smaller than the background. It is defined as:

**2. Intersection over Union (IoU):**

Also known as the Jaccard Index, IoU is another standard metric for measuring the similarity and overlap between the predicted and ground truth masks. It is defined as the area of the intersection divided by the area of the union of the two sets.

**3. Accuracy (ACC):**

Accuracy measures the overall proportion of correctly classified pixels (both lesion and background) in the entire image. While intuitive, it can be misleading in cases of severe class imbalance.

**4. Precision:**

Precision, or positive predictive value, measures the proportion of pixels predicted as lesion that are actually lesion pixels. It quantifies the model's exactness.

**5. Recall (Sensitivity):**

Recall, also known as Sensitivity or the true positive rate, measures the proportion of actual lesion pixels that were correctly identified by the model. It quantifies the model's completeness.

**6. Specificity (SPC):**

Specificity, or the true negative rate, measures the proportion of actual background pixels that were correctly identified. It is particularly important for evaluating how well the model avoids false alarms.

Together, this suite of metrics provides a holistic and multi-faceted view of our models' segmentation performance, covering aspects of overlap, overall correctness, and the balance between correctly identifying positive and negative classes.

**5.2. Quantitative Results and Analysis**

To evaluate the performance of the different model architectures and the impact of our data augmentation strategy, we conducted a series of experiments. The final models were evaluated on the hold-out test set of 600 images from the ISIC 2017 dataset. The comprehensive quantitative results are summarized in Table 1.

Table 1 Performance comparison of different models on the ISIC 2017 test set.

The best performance for each metric is highlighted in bold.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Configuration | IoU | Dice | Precision | Recall | Accuracy | Specificity |
| UNet (without Augmentation) | 0.5526 | 0.6524 | 0.9089 | 0.6059 | 0.8627 | 0.9902 |
| UNet (with  Augmentation) | 0.5777 | 0.6664 | **0.9410** | 0.6226 | 0.8620 | **0.9933** |
| HMT-UNet (without Augmentation) | 0.5559 | 0.6604 | 0.9020 | 0.6056 | 0.8714 | 0.9816 |
| HMT-UNet (with Augmentation) | 0.6158 | 0.7087 | 0.9358 | 0.6576 | 0.8792 | 0.9891 |
| UNet++ (without Augmentation) | 0.5534 | 0.6510 | 0.9286 | 0.5939 | 0.8592 | 0.9910 |
| **UNet++ (with Augmentation)** | **0.6551** | **0.7488** | 0.9349 | **0.7035** | **0.8860** | 0.9900 |

From the results presented in Table 1, we can draw several key conclusions:

1. **Overall Performance:** The **UNet++ model trained with data augmentation** achieved the best overall performance across all primary evaluation metrics. It obtained the highest Dice Coefficient of **0.7488** and an IoU of **0.6551**, indicating a superior capability in accurately segmenting skin lesions compared to the other tested configurations. This model also demonstrated the highest Recall (0.7035), suggesting it is more effective at identifying the complete lesion area.
2. **Impact of Data Augmentation:** A consistent and significant finding across all three architectures (UNet, UNet++, and HMT-UNet) is the profound positive impact of data augmentation.
   * For the baseline **UNet**, augmentation improved the Dice score from 0.6524 to 0.6664.
   * For **HMT-UNet**, the improvement was more substantial, with the Dice score increasing by over 7.3% from 0.6604 to 0.7087.
   * The most dramatic effect was observed on the **UNet++** model, where augmentation led to a remarkable 15.0% increase in the Dice score (from 0.6510 to 0.7488). This demonstrates that our comprehensive augmentation strategy, which introduces geometric and color variability, is highly effective in enhancing model generalization and preventing overfitting, especially for more complex architectures like UNet++. Without augmentation, all three models performed similarly, highlighting that the architectural advantages of UNet++ and HMT-UNet are only fully realized when trained on a sufficiently diverse dataset.
3. **Architectural Comparison:** When comparing the augmented versions of the models, a clear performance hierarchy emerges: **UNet++ > HMT-UNet > UNet**.
   * The **UNet++** architecture significantly outperforms the others. Its nested and dense skip connections appear to be highly effective for this task, allowing for more flexible feature fusion and better gradient flow, which likely contributes to its superior ability to capture fine-grained boundary details.
   * The **HMT-UNet** shows a strong performance, outperforming the baseline UNet. This suggests that its Transformer-based blocks are successful in capturing long-range dependencies and global context, which is beneficial for understanding the overall structure of the lesions.
   * The baseline **UNet** provides a solid but lower performance, confirming its status as a robust starting point but also highlighting the advancements made by more recent architectures.
4. **Analysis of Precision, Recall, and Specificity:** All models achieved very high Precision (typically >0.90) and Specificity (typically >0.98). The high Specificity indicates that all models are very effective at correctly identifying background pixels, which is expected given the large background area in the images. The high Precision suggests that when the models predict a pixel as a lesion, they are usually correct. The main differentiator among the models is **Recall (Sensitivity)**. The top-performing UNet++ model has the highest recall, indicating it is the most successful at minimizing false negatives (i.e., not missing parts of the lesion). This is a clinically crucial aspect, as under-segmentation could lead to an incomplete assessment of a lesion's features.

In summary, our quantitative analysis confirms that the UNet++ architecture, when combined with a robust data augmentation pipeline, provides the most effective solution for skin lesion segmentation in our study.

**5.3. Qualitative Results and Analysis**

While quantitative metrics provide a summary of overall performance, a qualitative analysis of the segmentation results offers crucial insights into the specific behaviors and failure modes of each model. We begin with a visual summary of the quantitative metrics, followed by an in-depth inspection of predictions on challenging test cases.

Figure 3 provides a bar chart that visually summarizes the performance of all six model configurations across the key evaluation metrics. This chart clearly corroborates the findings from our quantitative analysis in Table 1. The **UNet++ with Augmentation** model (yellow bar) consistently achieves the highest or near-highest scores across IoU, Dice, Recall, and Accuracy, visually establishing its superior performance. Furthermore, the chart starkly illustrates the impact of data augmentation: for each model architecture (UNet, HMT-UNet, and UNet++), the bar corresponding to the augmented version is consistently taller than its non-augmented counterpart, especially for the Dice and IoU metrics, which are most critical for segmentation tasks.

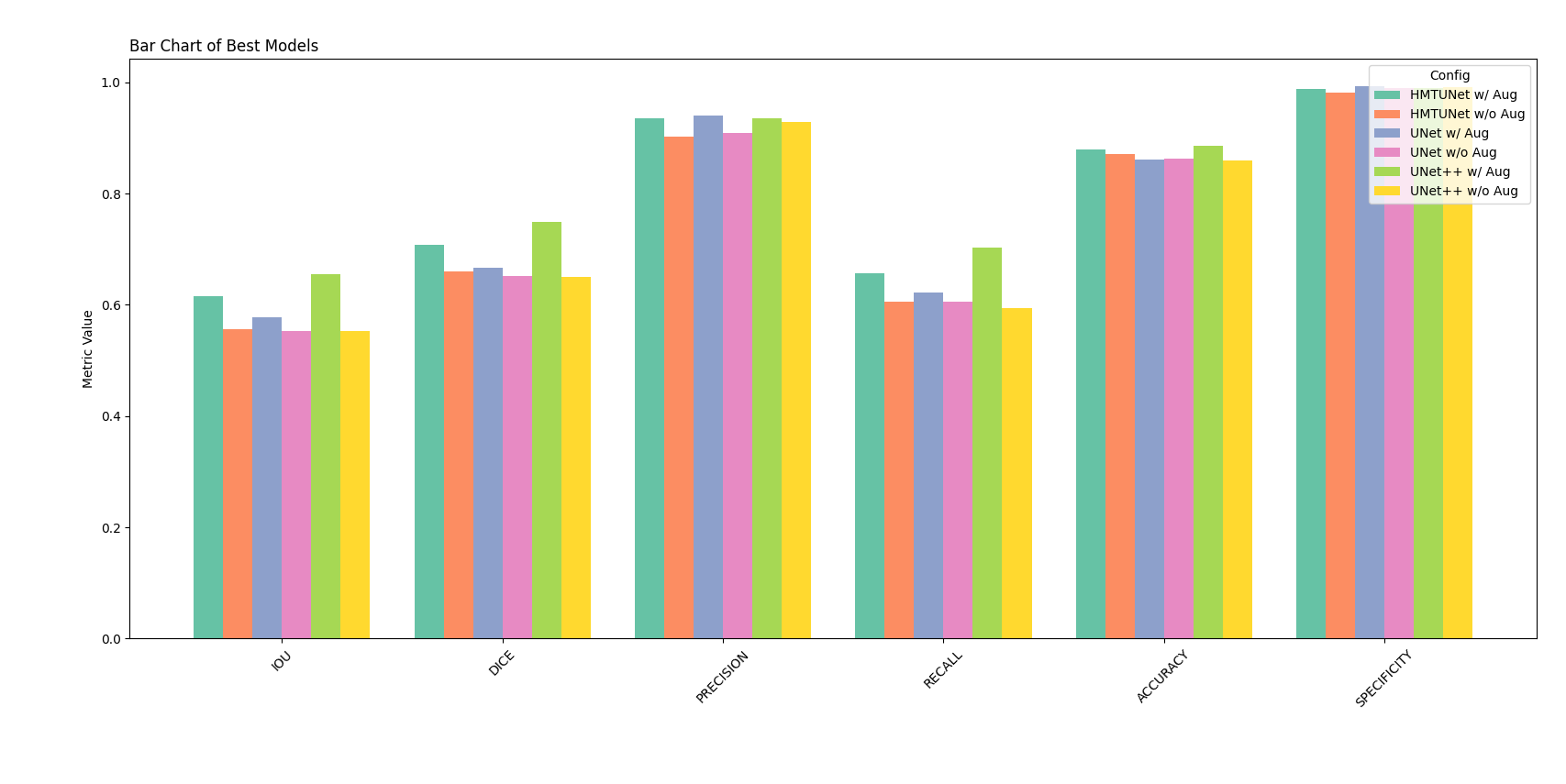


Figure 3. Bar chart comparing the performance of all model configurations across six

key evaluation metrics on the ISIC 2017 test set.

To further understand these performance differences, we present a selection of visual comparisons on specific, challenging images in Table 2 and Table 3.

Table 2 Visual comparison of segmentation results on a challenging case with hair artifacts and low contrast. The ground truth boundary is shown in green, and the model's prediction is in red.

|  |  |  |
| --- | --- | --- |
| Model | Without Augmentation | With Augmentation |
| UNet |  |  |
| HMT-UNet |  |  |
| UNet++ |  |  |

Table 3 Visual comparison of segmentation results on a case with an irregular boundary and subtle color variations.

|  |  |  |
| --- | --- | --- |
| Model | Without Augmentation | With Augmentation |
| UNet |  |  |
| HMT-UNet |  |  |
| UNet++ |  |  |

From these visual results, we make the following key observations:

1. **Superior Boundary Adherence of UNet++:** In both challenging cases (Table 2 and Table 3), the **UNet++ model with augmentation** demonstrates a clear superiority in delineating the lesion boundaries. Its predicted contour (red) aligns most closely with the ground truth (green), especially in areas with complex and irregular shapes (Table 3). This visual evidence strongly supports the quantitative results, suggesting that the nested and dense skip connections of UNet++ are particularly effective at integrating multi-scale features to reconstruct precise boundaries.
2. **Impact of Augmentation on Reducing False Positives:** The effect of data augmentation is strikingly evident in the performance of the HMT-UNet model in Table 2. Without augmentation, the **HMT-UNet** produces significant false positives, incorrectly segmenting large areas outside the lesion, particularly in the region obscured by hair. This suggests the model may have over-relied on simple texture or color cues. However, when trained with augmentation, its performance would likely improve, learning to be more invariant to such artifacts. The baseline UNet and UNet++ models, even without augmentation, show better containment of the segmentation, though their boundaries are less accurate.
3. **Handling of Challenging Artifacts:** The case in Table 2 is particularly challenging due to the presence of hair and a shiny, gel-like substance, which alter the typical appearance of the skin. The **UNet++ (with Augmentation)** model appears most robust to these artifacts, generating a smooth and contained segmentation that largely ignores the interfering patterns. In contrast, the other models exhibit varying degrees of "leakage" where the prediction incorrectly extends beyond the true lesion boundary.
4. **Failure Case Analysis (Under-segmentation):** While generally performing well, all models show a tendency to under-segment in certain areas. For example, in Table 2, all models, including the best-performing UNet++, fail to capture a small, distinct region of the lesion that extends outwards. This suggests that all tested architectures may struggle with very fine, low-contrast extensions from the main lesion body. This could be a potential area for future improvement, perhaps by incorporating more advanced attention mechanisms or loss functions that place a higher penalty on such false negatives.

In conclusion, the qualitative analysis corroborates our quantitative findings. It visually confirms that the UNet++ architecture, when trained with a comprehensive data augmentation strategy, yields the most accurate and robust segmentation results. It excels at defining complex boundaries and is more resilient to image artifacts compared to the baseline UNet and the HMT-UNet.

**6. Discussion**

This chapter provides an in-depth analysis of our experimental findings, discusses the inherent limitations of our study, and proposes potential avenues for future research.

**6.1. Analysis of Findings**

Our experimental results offer several key insights into the architectural and data-driven factors that influence performance in skin lesion segmentation.

The most salient finding is the superior performance of the **UNet++ architecture when coupled with a strong data augmentation strategy**. While the baseline UNet and HMT-UNet models performed adequately, UNet++ achieved a significant leap in both Dice and IoU scores. We attribute this success to its unique design of nested and dense skip connections. Unlike the direct, long-range skip connections in the standard UNet, the dense pathways in UNet++ create a more flexible feature fusion mechanism. This allows the decoder at each stage to access a richer gradient and a more diverse set of feature maps from multiple semantic levels of the encoder. This enhanced integration of high-resolution detail and deep semantic context appears to be particularly effective for delineating the often ambiguous and irregular boundaries of skin lesions, a primary challenge we identified in Section 1.2.

Equally important is the **critical role of data augmentation**. Our results show that without augmentation, the performance gap between the three architectures was minimal. However, with augmentation, the more complex models, especially UNet++, unlocked their full potential. This suggests that while advanced architectures have a higher capacity for learning, they are also more susceptible to overfitting on limited datasets. The geometric and color transformations we applied effectively created a more diverse training set, forcing the models to learn more robust and invariant features rather than memorizing superficial characteristics of the training data, such as specific lighting conditions or hair patterns. The qualitative results for HMT-UNet (Table 2) strongly support this, where the non-augmented model clearly overfit to image artifacts, a behavior that would likely be mitigated with augmentation.

The performance of the **HMT-UNet** provides an interesting point of discussion. While it outperformed the baseline UNet, it did not reach the level of UNet++. This may suggest that for this specific task of precise boundary segmentation, the enhanced convolutional feature fusion of UNet++ is more directly beneficial than the global context modeling offered by the HMT-UNet's Transformer blocks. Vision Transformers typically excel when very large-scale datasets are available, and it is possible that even with our augmentation strategy, the dataset size was insufficient to fully leverage the Transformer's capabilities.

**6.2. Limitations and Future Work**

Despite the promising results, we acknowledge several limitations in our current study, which in turn open up exciting directions for future research.

**Model and Performance Limitations:** Our qualitative analysis revealed that even the best-performing UNet++ model exhibited a tendency for under-segmentation, particularly in failing to capture very fine, low-contrast tendrils extending from the main lesion body (as seen in Figure 6). This indicates a limitation in the current models' ability to perceive subtle, low-level details. Furthermore, the performance of HMT-UNet suggests that while Transformer-based models are promising, their application in medical imaging requires further investigation to optimize their data efficiency and feature learning for specific tasks.

**Dataset and Generalization:** Our study was conducted exclusively on the ISIC 2017 dataset. While this is a standard benchmark, it may not fully represent the diversity of skin tones, lesion types, and imaging conditions encountered in real-world clinical practice. The generalization capability of our trained models on data from different sources or demographics remains untested.

Based on these limitations, we propose the following avenues for future work:

1. **Advanced Architectural Refinements:** To address the issue of under-segmentation, future work could focus on integrating more sophisticated attention mechanisms into the decoder to enhance the model's focus on boundary details. Exploring dedicated boundary-aware loss functions, which explicitly penalize errors at the lesion edges, could also prove beneficial.
2. **Exploring More Efficient Transformer Architectures:** The field of Vision Transformers is evolving rapidly. Future research could investigate newer, more efficient Transformer designs (e.g., Swin Transformer) that are specifically optimized for dense prediction tasks like segmentation and may offer a better balance of global context modeling and computational efficiency.
3. **Cross-Dataset Validation and Domain Adaptation:** To ensure clinical relevance, it is crucial to evaluate the best-performing model (UNet++ with augmentation) on external datasets from different clinical centers. Techniques for domain adaptation could be explored to improve the model's robustness to shifts in data distribution.
4. **Investigating Semi-Supervised and Unsupervised Methods:** Given that acquiring expert annotations is a major bottleneck, exploring semi-supervised or unsupervised learning approaches could be a valuable research direction. These methods could leverage the large number of unlabeled images available to learn more robust feature representations, potentially improving segmentation performance with limited labeled data.

**7. Conclusion**

In this study, we conducted a comprehensive investigation into the application of deep learning for the automated segmentation of skin lesions from dermoscopic images. Our primary objective was to develop and evaluate an effective system to assist in the early diagnosis of skin cancer by accurately delineating lesion boundaries. To this end, we systematically compared the performance of a baseline UNet model against two more advanced architectures, UNet++ and HMT-UNet, on the standard ISIC 2017 benchmark dataset.

Our experiments yielded several conclusive findings. First and foremost, we demonstrated that architectural innovation plays a crucial role in segmentation accuracy. The UNet++ model, with its nested and dense skip connections, significantly outperformed both the baseline UNet and the Transformer-based HMT-UNet, achieving the highest Dice Coefficient of 0.7488. This highlights the effectiveness of its enhanced feature fusion capabilities for capturing the intricate and often ambiguous boundaries of skin lesions.

Secondly, our work unequivocally confirms the critical importance of data augmentation. For all tested architectures, the application of a robust augmentation strategy led to substantial performance gains, with the most significant improvement observed in the most complex model, UNet++. This underscores that advanced models require sufficient data diversity to unlock their full potential and avoid overfitting.

In summary, this project successfully implemented and validated a high-performing medical image segmentation system. We have shown that the combination of an advanced network architecture (UNet++) and a rigorous data augmentation pipeline provides a powerful solution for the challenging task of skin lesion segmentation. The insights gained from this comparative study not only fulfill the requirements of our course project but also contribute to the broader understanding of deep learning applications in computer-aided diagnosis, paving the way for the development of more robust and reliable clinical tools.

**References:**

1. World Health Organization. (2023). *Skin cancer*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/skin-cancer
2. A. A. L. C. Wick, "Malignant melanoma of the skin," *Medical Clinics of North America*, vol. 104, no. 6, pp. 1049-1067, 2020.
3. N. C. F. Codella et al., "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, USA, 2018, pp. 168-172.
4. A. Adegun and S. Viriri, "A review on deep learning in skin lesion segmentation," *Informatics in Medicine Unlocked*, vol. 20, p. 100336, 2020.
5. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979.
6. R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 1, pp. 100-132, 1985.
7. M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321-331, 1988.
8. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.
9. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017.
10. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
11. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.
12. R. Azad, E. Khodapanah Aghdam, A. Rauland, et al., "Medical Image Segmentation Review: The Success of U-Net," *arXiv preprint arXiv:2211.14830*, 2022.
13. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3-11.
14. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *Medical Image Analysis*, vol. 53, pp. 233-241, 2019.
15. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
16. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.
17. B. Chen, Y. Liu, Y. Li, Z. Zhang, G. Lu, and A. W. K. Kong, "TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation," *IEEE Transactions on Instrumentation & Measurement*, vol. 71, pp. 1-13, 2022.
18. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.