Optimizing R2U-Net for Skin Cancer Segmentation with Attention and Hybrid Loss

Sreejita Das

Advisor: **Dr. Kaustuv Nag**

Department of Computer Science and Engineering
Indian Institute of Information Technology Guwahati

This report is submitted for the course of
CS300 : Project-I

IIIT Guwahati March 2025

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This report is my own work and contains nothing that is the outcome of work done in collaboration with others except as specified in the text and Acknowledgements.

Sreejita Das
Roll: 2201201,
3rd year, Bachelors of Technology,
Department of Computer Science and Engineering,
Indian Institute of Information Technology Guwahati.

Acknowledgements

I express my sincere gratitude to Dr. Kaustuv Nag for their expert guidance and steadfast support throughout this project. I am also thankful to the faculty of the Department of Computer Science and Engineering at the Indian Institute of Information Technology Guwahati for their valuable mentorship. Additionally, I appreciate the support of my peers during this academic endeavor. Lastly, I extend my gratitude to my parents and family for their consistent encouragement and support throughout my studies.

Abstract

Accurate segmentation of skin cancer lesions is crucial for early diagnosis but remains challenging due to class imbalance and the need for precise boundary delineation. This project advances the R2U-Net model, a neural network designed for medical image segmentation, by integrating attention gates and a hybrid loss function combining Dice and weighted Binary Cross-Entropy (BCE) losses. These enhancements sharpen the model's focus on lesion regions and optimize its performance in balancing overlap accuracy with pixel-level precision. Evaluations on subsets of the ISIC 2017 dataset reveal substantial improvements in segmentation accuracy, underscoring the approach's potential for real-world impact. This work delivers a practical solution for enhancing medical image segmentation and establishes a robust foundation for future advancements in the field.

Table of Contents

List of Figures

Chapter 1

Introduction

Medical image segmentation is a critical process in healthcare, enabling the precise identification of anatomical structures in medical images. U-Net [5] revolutionized medical image segmentation with its encoder-decoder architecture. Building on U-Net, RU-Net [1] introduced recurrent convolutional units to enhance feature accumulation over time. R2U-Net [1] further advanced this by combining recurrent and residual units, boosting feature extraction and training stability.

1.0.1 Problem Statement

Accurate segmentation of skin cancer lesions is essential for early diagnosis and treatment, yet it remains a challenge in medical imaging. Two obstacles hinder this task: **class imbalance**, and the need for **precise boundary detection** of irregular lesion edges. While R2U-Net excels in feature extraction and training stability, it struggles to capture fine details—like subtle textures—and maintain accuracy across imbalanced classes, often causing over-segmentation or missed boundaries in complex cases

1.0.2 Solution

We propose an enhanced R2U-Net framework by incorporating **attention gates** and a **hybrid loss function** that blends Dice and weighted Binary Cross-Entropy (BCE) losses. Attention gates [**?**], refine skip connections to prioritize lesion regions, sharpening the model's focus on subtle features. Simultaneously, the hybrid loss improves overlap accuracy and mitigates class imbalance more effectively than traditional loss functions.

1.1 Objective

This study aims to enhance the R2U-Net model, a neural network designed for medical image segmentation, to improve the accuracy and reliability of skin cancer lesion segmentation. The investigation focuses on two primary objectives: **first**, to integrate attention gates and a hybrid loss function—combining Dice and weighted Binary Cross-Entropy (BCE) losses—into the R2U-Net architecture, addressing challenges such as class imbalance and precise boundary detection; and **second**, to evaluate the effectiveness of these enhancements in improving segmentation performance, particularly in terms of overlap accuracy and pixel-level precision. By analyzing the impact of attention gates and the hybrid loss on R2U-Net's ability to segment skin cancer lesions, this research seeks to contribute to the development of more accurate diagnostic tools and offer insights into optimizing neural network components for medical imaging.

1.2 Motivation

The timely identification of skin cancer, notably melanoma, plays a pivotal role in enhancing patient prognosis, as early intervention markedly elevates treatment success rates. Traditional diagnostic approaches, however, depend extensively on manual segmentation of dermoscopic images, a process that is both labor-intensive and subject to inconsistencies. These challenges underscore the pressing demand for automated and precise segmentation techniques to support efficient diagnosis. This project is driven by the potential to refine the R2U-Net model—a cutting-edge convolutional neural network tailored for medical image segmentation—by overcoming its limitations in managing class imbalance and delineating intricate lesion boundaries. Through the incorporation of attention gates and a hybrid loss function, this work seeks to develop an advanced model that enhances segmentation accuracy while offering a practical, scalable tool for clinical deployment.

Chapter 2

Related Work

2.1 Literature Survey

2.1.1 U-Net

Ronneberger et al. [5] introduced U-Net in 2015, a pioneering model for medical image segmentation that uses an encoder-decoder architecture with skip connections to preserve spatial details, making it effective for skin cancer lesion segmentation [5].

2.1.2 3D U-Net

Çiçek et al. [6] presented 3D U-Net in 2016, extending U-Net for volumetric medical image segmentation by learning from sparsely annotated 3D data, applicable to tasks requiring spatial depth [6].

2.1.3 V-Net

Also in 2016, Milletari et al. [3] introduced V-Net, a 3D fully convolutional network with residual connections for volumetric segmentation, incorporating a Dice loss to address class imbalance in medical imaging [3].

2.1.4 RU-Net and R2U-Net

Alom et al. [1] proposed RU-Net and R2U-Net in 2018, enhancing U-Net with recurrent convolutional layers (RU-Net) and recurrent-residual units (R2U-Net) for better feature accumulation and training stability in skin cancer segmentation [1].

2.2 Primary Model : R2UNet

Recurrent U-Net (RU-Net) and Recurrent Residual U-Net (R2U-Net), introduced by Alom et al. [1], are advanced extensions of U-Net, designed to improve feature extraction for tasks like skin cancer lesion segmentation on the ISIC 2017 dataset [1]. RU-Net incorporates Recurrent Convolutional Layers (RCLs) into its encoder-decoder framework, enabling feature accumulation over time steps (e.g., $t = 2$ or $t = 3$) to capture fine details in dermoscopic images[1].

R2U-Net builds on RU-Net by integrating RCLs with residual connections, forming Recurrent Residual Convolutional Units (RRCUs), which enhance training stability and robustness while maintaining the same number of parameters as U-Net [1]. The RRCUs combine recurrent feature accumulation with residual learning, mitigating the vanishing gradient problem and improving performance on fine lesion details, though class imbalance remains a challenge [1].

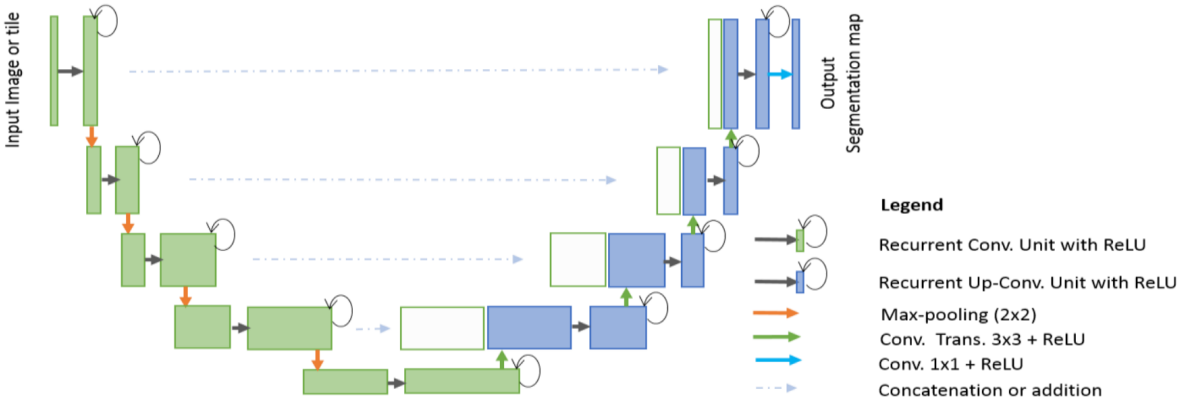

Fig. 2.1 R2U-Net architecture using recurrent residual convolutional units

2.3 Working Principle

R2U-Net processes input dermoscopic images through its encoder-decoder architecture to produce segmentation maps, leveraging Recurrent Residual Convolutional Units (RRCUs) in each convolutional block [1]. As shown in Fig. 2.2, the encoder downsamples the input image using max-pooling, while the decoder upsamples feature maps via up-convolution, with skip connections preserving spatial details.

Within each RRCU, the Recurrent Convolutional Layers (RCLs) first accumulate features over time steps (e.g., $t = 2$ or $t = 3$). For an input x_l at the l-th layer, the RCL output at time step t for a pixel at position (i, j) on the k-th feature map is:

$$O_{ijk}^{l}(t) = (w_k^f)^T * x_l^{f(i,j)}(t) + (w_k^r)^T * x_l^{r(i,j)}(t-1) + b_k,$$

where $x_l^{f(i,j)}(t)$ and $x_l^{r(i,j)}(t-1)$ are inputs to the standard and recurrent convolutions, w_k^f and w_k^r are their respective weights, and b_k is the bias [1]. This output is passed through a ReLU activation, $\mathcal{F}(x_l, w_l) = \max(0, O_{ijk}^{l}(t))$. The RRCU then applies a residual connection, computing the block's output as:

$$x_{l+1} = x_l + \mathcal{F}(x_l, w_l),$$

where x_l is the input to the RRCU [1]. The recurrent operation enhances feature accumulation, while the residual connection stabilizes training by allowing gradients to flow directly through the network, enabling R2U-Net to capture fine lesion details.

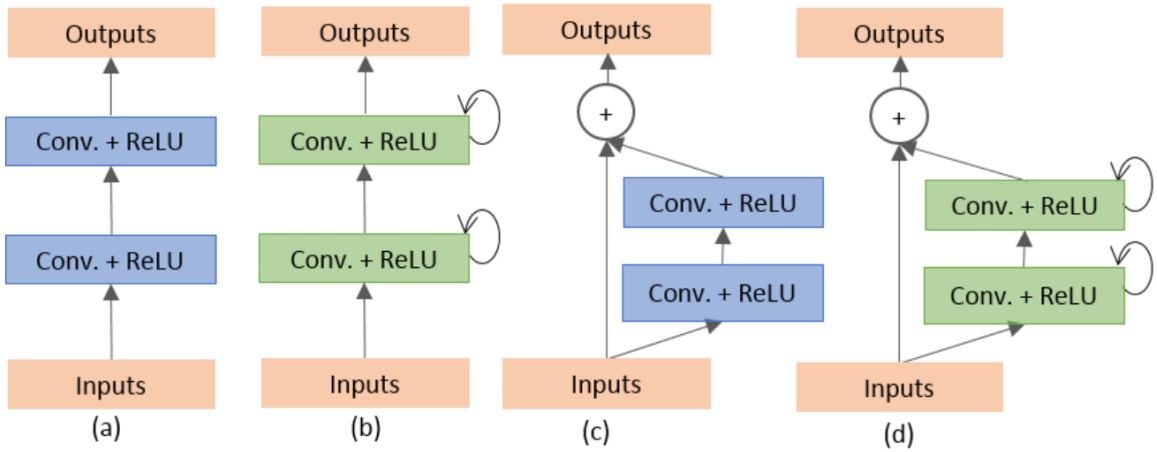

Fig. 2.2 Different variant of convolutional and recurrent convolutional units (a) Forward convolutional units, (b) Recurrent convolutional block (c) Residual convolutional unit, and (d) Recurrent Residual convolutional units (RRCU).

2.4 Notations

In the working principle of R2U-Net, the Recurrent Convolutional Layer (RCL) output at time step t for a pixel at position (i, j) on the k-th feature map in the l-th

layer is denoted as $O_{ijk}^{l}(t)$. The input to the RCL at the l-th layer is represented as x_l, where $x_l^{f(i,j)}(t)$ and $x_l^{r(i,j)}(t-1)$ are the inputs to the standard and recurrent convolutions, respectively. The weights of the standard convolutional layer and the RCL for the k-th feature map are denoted as w_k^f and w_k^r, respectively, and b_k is the bias term. The output of the RCL after ReLU activation is expressed as $\mathcal{F}(x_l, w_l)$. For the Recurrent Residual Convolutional Unit (RRCU), the output at the $(l+1)$-th layer is denoted as x_{l+1}, which is computed by adding the input x_l to the RCL output $\mathcal{F}(x_l, w_l)$.

2.5 Dataset

The dataset used in this project is the International Skin Imaging Collaboration (ISIC) 2017 dataset, which is a widely recognized benchmark for skin lesion segmentation tasks [2]. It consists of dermoscopic images of skin lesions, annotated with ground truth masks for segmentation. The dataset is divided into three subsets: training, validation, and test sets, as summarized in Table 2.1.

Table 2.1 Distribution of the ISIC 2017 Dataset

Data	**Sample Images**	**Ground Truth**
Training	2000	2000
Validation	150	150
Test	600	600

A sample image from the ISIC 2017 dataset, along with its ground truth mask, is shown in Fig. 2.3. This dataset is particularly relevant to this project as it provides a standardized benchmark for evaluating the effectiveness of R2U-Net and its enhancements in skin cancer lesion segmentation.

2.6 Evaluation Metrics

To assess the performance of R2U-Net for medical image segmentation, several standard evaluation metrics are employed, as outlined in Alom et al. [1]. These metrics are particularly relevant for evaluating segmentation models on the ISIC 2017 dataset, where precise delineation of skin cancer lesions is critical. The metrics used, along with their formulas, are as follows:

Fig. 2.3 A sample dermoscopic image from the ISIC 2017 dataset (bottom) and its corresponding ground truth segmentation mask (top).

- **Accuracy (AC)**: Represents the proportion of correctly classified pixels (both lesion and background) in the image. It is calculated as:

$$AC = \frac{TP + TN}{TP + TN + FP + FN},$$

  where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

- **Sensitivity (SE)**: Also known as the true positive rate, this metric evaluates the model's ability to correctly identify lesion pixels. It is given by:

$$SE = \frac{TP}{TP + FN},$$

- **Specificity (SP)**: Also referred to as the true negative rate, this metric assesses the model's ability to correctly identify background pixels, helping to minimize false positives in segmentation. It is expressed as:

$$SP = \frac{TN}{TN + FP},$$

- **Dice Coefficient (DC)**: Measures the overlap between the predicted segmentation and the ground truth, emphasizing the similarity between the two regions.

The DC is defined as:

$$DC = 2\frac{|GT \cap SR|}{|GT| + |SR|},$$

where GT refers to the ground truth and SR refers to the segmentation result.

- **Jaccard Similarity (JS)**: Also known as the Intersection over Union (IoU), this metric quantifies the ratio of the intersection of the predicted and ground truth regions to their union, providing a measure of spatial overlap. It is calculated as:

$$JS = \frac{|GT \cap SR|}{|GT \cup SR|},$$

2.7 Notations

In the evaluation metrics for R2U-Net, TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively. The ground truth is denoted as GT, and the segmentation result as SR. The intersection and union of GT and SR are represented as $|GT \cap SR|$ and $|GT \cup SR|$, respectively, while $|GT|$ and $|SR|$ denote their respective cardinalities.

2.8 Loss Function

The R2U-Net model is trained using the Binary Cross-Entropy (BCE) loss function to optimize its performance for medical image segmentation tasks, such as skin cancer lesion segmentation on the ISIC 2017 dataset [1]. The BCE loss measures the difference between the predicted probabilities and the ground truth labels, making it suitable for binary segmentation tasks. The BCE loss is defined as:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

This loss function ensures that the model learns to accurately classify each pixel as either part of the lesion or the background, addressing challenges like class imbalance in dermoscopic images.

2.9 Notations

In the BCE loss function for R2U-Net, BCE represents the Binary Cross-Entropy loss. The total number of pixels in the image is denoted as N. For the i-th pixel, y_i is the ground truth label (0 or 1, indicating background or lesion), and $\hat{y}_i$ is the predicted probability of the pixel belonging to the lesion class, ranging between 0 and 1.

2.10 Training Algorithm

Algorithm 1 Training R2U-Net for Lesion Segmentation

Init R2U-Net, Adam optimizer
for epoch = 1 to `num_epochs` **do**
  for batch $(images, GT)$ in `train_loader` **do**
    $SR = \sigma(\text{R2U-Net}(images))$
    BCE: $-\frac{1}{N}\sum_{i=1}^{N}[GT_i \log(SR_i) + (1 - GT_i)\log(1 - SR_i)]$
    Update params via gradient descent
  end for
end for

Trains R2U-Net using BCE loss to segment lesions from background over epochs.

2.11 Evaluation Algorithm

Algorithm 2 Evaluation Metrics for Segmentation Performance

Input: SR, GT, threshold = 0.5
Binarize: $SR = (SR > 0.5)$, $GT = (GT > 0.5)$
$TP = \sum(SR = 1 \wedge GT = 1)$, $TN = \sum(SR = 0 \wedge GT = 0)$
$FP = \sum(SR = 1 \wedge GT = 0)$, $FN = \sum(SR = 0 \wedge GT = 1)$
AC: $\frac{TP+TN}{TP+TN+FP+FN}$
SE: $\frac{TP}{TP+FN}$, SP: $\frac{TN}{TN+FP}$
PC: $\frac{TP}{TP+FP}$, F1: $2 \cdot \frac{\text{SE} \cdot \text{PC}}{\text{SE}+\text{PC}}$
JS: $\frac{|SR \cap GT|}{|SR \cup GT|}$, DC: $\frac{2 \cdot |SR \cap GT|}{|SR| + |GT|}$

Computes metrics by comparing binarized segmentation with ground truth.

2.12 Data Preprocessing Algorithm

Algorithm 3 Preprocessing ISIC 2017 Data for R2U-Net

Load ISIC 2017 images, masks
Resize to 256×256
if mode = `train` **then**
    Augment (flips, rotations), $p = 0.8$
end if
Normalize to [0, 1]
Convert to tensors
Batch (`batch_size`)

Preprocesses images and masks for training or inference.

2.13 Inference Algorithm

Algorithm 4 Inference for Lesion Segmentation with R2U-Net

Load trained R2U-Net
Eval mode
Input: image
Preprocess: resize, normalize, tensor
$SR = \sigma(\text{R2U-Net}(image))$
Binarize: $SR = (SR > 0.5)$

Segments a new image into lesion and background.

Detailed Working of the Algorithms

The algorithms collectively form the workflow for segmenting skin cancer lesions using R2U-Net on the ISIC 2017 dataset, separating lesions (foreground) from the background. Here's a step-by-step explanation of each algorithm:

Data Preprocessing: This algorithm prepares the ISIC 2017 dataset for R2U-Net. It starts by loading dermoscopic images and their corresponding ground truth masks, which label each pixel as lesion (1) or background (0). The images and masks are resized to a uniform 256×256 resolution to ensure consistency. During training, data augmentation (e.g., random flips and rotations) is applied with a probability of

0.8 to increase dataset variety and improve model robustness. The pixel values are then normalized to the range [0, 1] for numerical stability. Finally, the images and masks are converted into tensors and grouped into batches for efficient processing during training or inference.

Training: The training algorithm optimizes R2U-Net to segment lesions accurately. It begins by initializing the R2U-Net model and the Adam optimizer. For each epoch, the algorithm iterates over batches of preprocessed images and ground truth masks. R2U-Net processes each batch through its encoder-decoder architecture, where the encoder extracts features (e.g., edges, textures) via downsampling, and the decoder reconstructs a segmentation map via upsampling, with recurrent residual units enhancing feature capture. The output logits are passed through a sigmoid activation to produce a probability map (SR), where each pixel's value (between 0 and 1) indicates the likelihood of it being a lesion. The Binary Cross-Entropy (BCE) loss is computed by comparing SR with the ground truth, measuring the difference between predicted probabilities and actual labels. The model parameters are updated using gradient descent to minimize this loss.

Evaluation Metrics Computation: It takes the segmentation result (SR) and ground truth (GT) as inputs, binarizing both using a threshold of 0.5 (values above 0.5 are classified as lesion, below as background). It then calculates True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) by comparing the binarized SR and GT. Using these, it computes standard metrics: Accuracy (AC), Sensitivity (SE), Specificity (SP), Precision (PC), F1-score, Jaccard Similarity (JS) and Dice Coefficient (DC).

Inference: The inference algorithm applies the trained R2U-Net to segment a new dermoscopic image. It starts by loading the trained model and setting it to evaluation mode to disable training-specific operations. The input image is preprocessed by resizing to 256×256, normalizing, and converting to a tensor. R2U-Net processes the image to produce a probability map (SR) via a sigmoid activation, where each pixel's value indicates the likelihood of it being a lesion. The map is binarized using a threshold of 0.5 to create a segmentation mask, labeling each pixel as lesion (1) or background (0).

2.14 Notations

The following notations are used in the algorithms: R2U-Net denotes the R2U-Net model; *images* and *image* represent input dermoscopic images (batch and single,

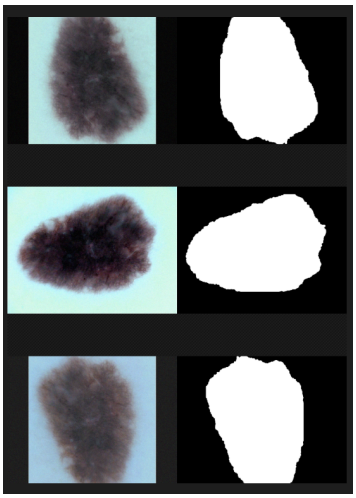

Fig. 2.4 An example of an augmented sample from ISIC-2017 with random flipping, color jittering and rotating.

respectively); GT is the ground truth mask (0 for background, 1 for lesion); SR is the segmentation result (probability map); σ is the sigmoid activation; N is the number of pixels; TP, TN, FP, and FN are True Positives, True Negatives, False Positives, and False Negatives; AC, SE, SP, PC, F1, JS, and DC represent Accuracy, Sensitivity, Specificity, Precision, F1-score, Jaccard Similarity, and Dice Coefficient, respectively; p is the augmentation probability; and `batch_size` is the batch size for data loading.

Chapter 3

Proposed Model

The proposed model incorporates two novel modifications to the standard R2U-Net architecture to enhance its performance in skin cancer lesion segmentation: a combined BCE-Dice loss function and attention gates within skip connections.

3.1 Proposed improvments to R2UNet

3.1.1 Hybrid Loss Function

The first approach is the integration of a combined BCE-Dice loss function, achieved by adding the Binary Cross-Entropy (BCE) and Dice losses with weighted contributions to address class imbalance. The combined loss is defined as $L = \alpha \cdot \text{BCE} + \beta \cdot (1 - \text{Dice})$, where α and β are hyperparameters that balance the contributions of BCE, which focuses on pixel-wise classification accuracy, and Dice loss, which emphasizes the overlap between the predicted and ground truth masks. Dice loss is added because it directly optimizes the overlap between the predicted and ground truth masks, emphasizing the intersection-over-union metric, which is particularly effective for imbalanced datasets as it gives equal importance to both classes regardless of their pixel counts.

3.1.2 Attention Gates

The second approach involves the incorporation of attention gates within the skip connections of R2U-Net. Attention gates [4] in image segmentation are mechanisms used in deep learning models, particularly in convolutional neural networks (CNNs), to enhance the focus on relevant regions of an image while suppressing irrelevant

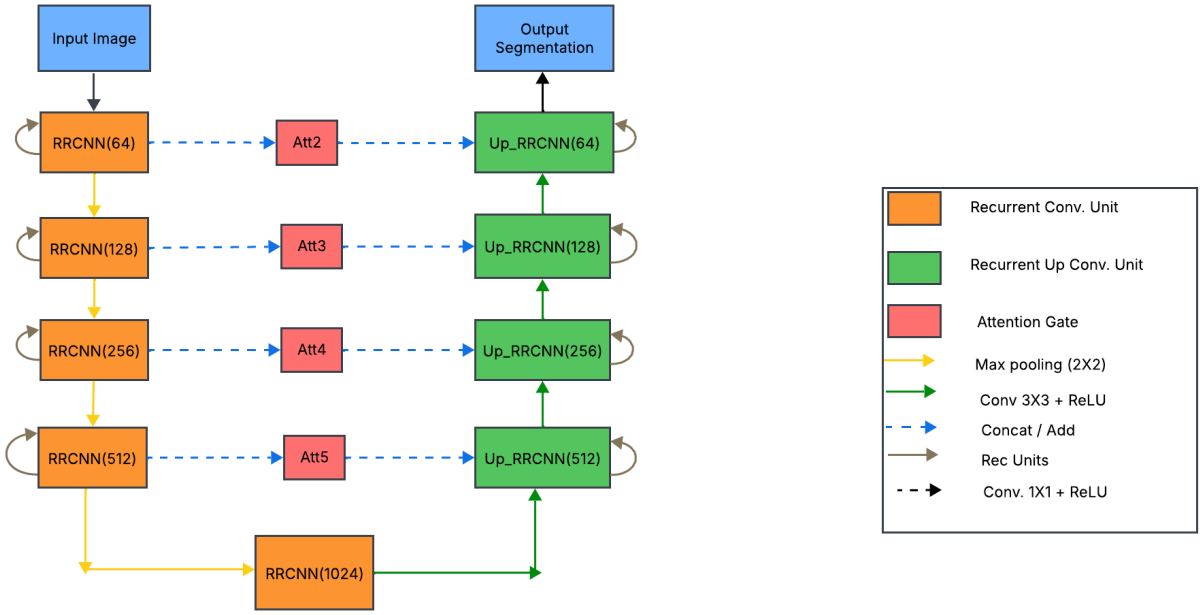

Fig. 3.1 Block diagram of proposed attention-gated R2UNet model.

or less important areas. The gate computes an attention score, or coefficients α_l for each spatial location in the feature map. This score is usually a value between 0 and 1, where:

- A score close to 1 indicates high relevance (the region is preserved or amplified).
- A score close to 0 indicates low relevance (the region is suppressed).

The final feature map is obtained by the attended feature computation $x'_l = \alpha_l \cdot x_l$. This mechanism prioritizes encoder features most relevant to lesion regions, suppressing irrelevant background features, and allows the model to focus on critical areas of the dermoscopic image.

3.2 Key Findings

References

[1] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*. 1, 3, 4, 5, 6, 8

[2] ISIC Archive (2017). ISIC 2017: Skin Lesion Analysis Towards Melanoma Detection. https://challenge.isic-archive.com/landing/2017. 6

[3] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. 3

[4] 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*. 13

[5] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. 1, 3

[6] Çiçek, , Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432. 3