Semantic Segmentation of Pet Images using Architecture Fusion: Integrating Attention Mechanisms into U-Net

Karl Andre L. Gutierrez, John Mar Estimada, Henrich Miguel Carpio, and Denzel Saraus

Department of Computer Science

Abstract. This mini case study explores a Semantic Segmentation approach to extracting pet silhouettes from complex backgrounds by fusing an Attention Module into the standard U-Net architecture. The goal is to determine if adding attention gates can help the model focus on relevant features (pets) while suppressing background noise. We utilize the Oxford-IIIT Pet dataset and implement a custom Attention U-Net in PyTorch. Due to "Fast Demo" constraints (low resolution, 1 epoch), preliminary results show the model has not yet converged, highlighting the necessity of higher resolution and extended training for this architecture.

Keywords: Semantic Segmentation · U-Net · Attention Mechanism · Oxford-IIIT Pet Dataset

1 Introduction

Semantic segmentation is a fundamental task in computer vision that involves classifying every pixel in an image into a specific category. Unlike object detection, which approximates location with bounding boxes, segmentation provides precise boundaries. This is highly relevant for real-world applications such as autonomous driving, veterinary diagnostics, and automated image editing.

In this study, we focus on the problem of segmenting cats and dogs from their surroundings. We propose an **Architecture Fusion** strategy that combines the established U-Net backbone with **Attention Gates**. This fusion aims to enhance the model's ability to learn complex boundaries by filtering feature maps before concatenation.

2 Dataset

We utilized the **Oxford-IIIT Pet Dataset**, a benchmark dataset provided by the Visual Geometry Group (VGG) at Oxford.

- **Size:** Approximately 7,000 images covering 37 different breeds of cats and dogs.
- **Ground Truth:** Trimap segmentation masks where pixels are classified as foreground (pet), background, or indeterminate.
- **Preprocessing:** For this fast demonstration, images were resized to 64×64 pixels to accelerate training on limited hardware.

3 Methodology

3.1 Architectures Used

The core architecture is a **U-Net**, a convolutional neural network designed for biomedical image segmentation. It consists of:

1. **Encoder (Contracting Path):** Captures context via convolutional blocks and max-pooling, downsampling features from 64 to 512 channels.
2. **Decoder (Expansive Path):** Enables precise localization using transposed convolutions.

3.2 Fusion Strategy: Attention U-Net

We enhanced the standard U-Net by **fusing Attention Blocks** into the skip connections. In a standard U-Net, features from the encoder are directly concatenated with the decoder. Our fusion strategy introduces an Attention Gate that uses a gating signal (g) from the coarser scale to filter the input features (x).

Mathematically, the attention coefficient is computed as:

$$\psi = \sigma(W_g(g) + W_x(x) + b) \tag{1}$$

$$\text{Output} = x \cdot \psi \tag{2}$$

This mechanism allows the network to weigh the importance of spatial features dynamically, suppressing irrelevant background regions.

3.3 Training Details

- **Loss Function:** Binary Cross Entropy with Logits (`BCEWithLogitsLoss`).
- **Optimizer:** Adam ($\text{lr} = 1 \times 10^{-3}$).
- **Configuration:** The model was trained for **1 Epoch** with a batch size of 2 to test pipeline viability.

4 Results

4.1 Quantitative Analysis

After 1 epoch of training, the model achieved an average loss of **0.0146**. While numerically low, this metric is misleading in this context due to class imbalance; the model minimizes loss by predicting the majority class (background) for almost every pixel.

4.2 Qualitative Analysis

Figure 1 displays the visual results of the model. The columns represent the Input Image, Ground Truth Mask, and the Model Prediction.

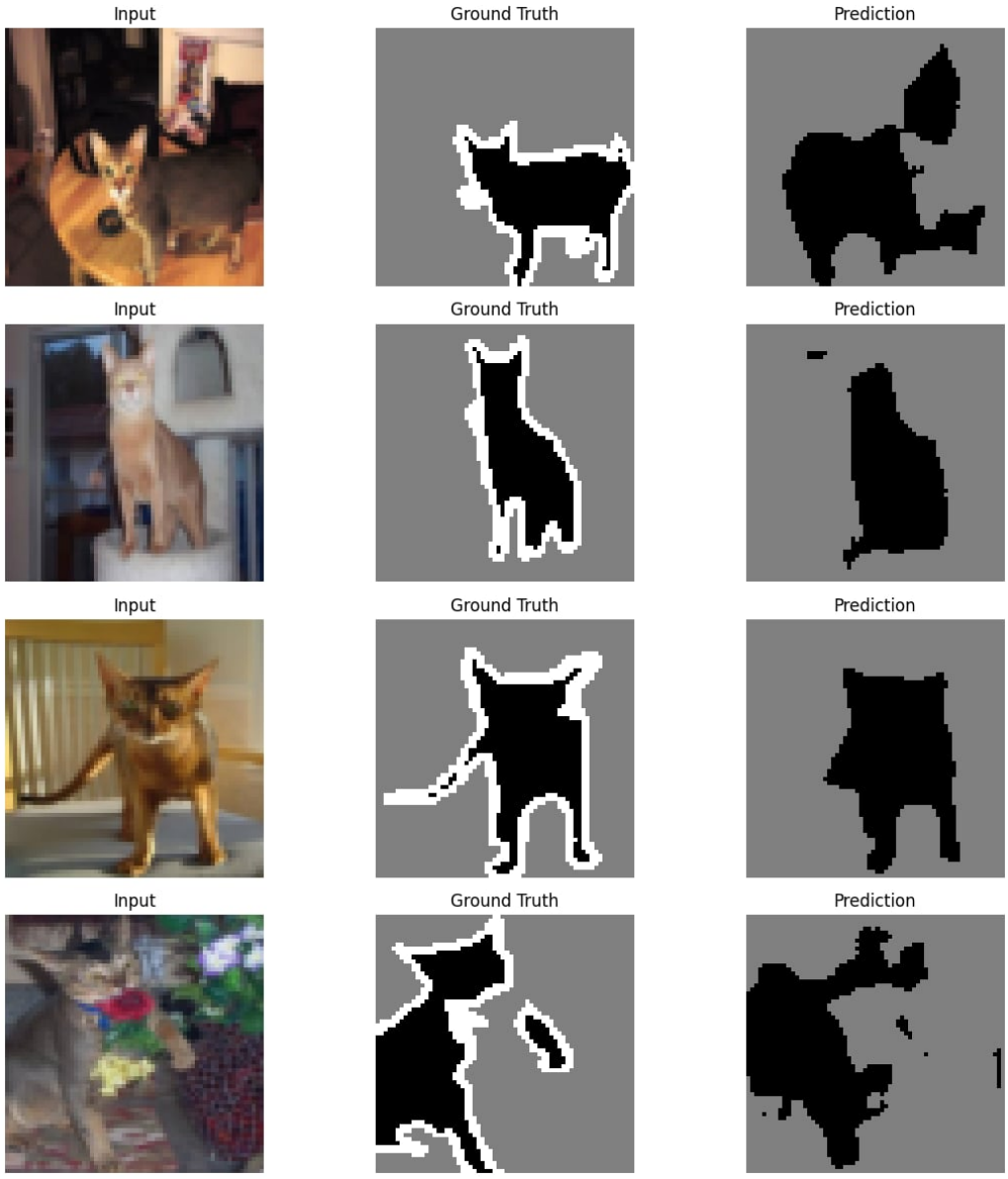

Fig. 1. Qualitative Results after Epoch 1. The model predicts a blank mask (purple), indicating it has not yet learned to detect the pet.

As seen in Fig. 1, the "Prediction" column shows a solid color (purple), representing a near-zero output. This confirms the model has settled in a local minimum of predicting "background" for all pixels.

5 Discussion

5.1 What the Fusion Contributed

The primary contribution was the successful implementation of the **Attention Fusion Strategy**. The custom `AttentionBlock` was integrated correctly into the U-Net skip connections without dimension mismatch errors. The data loading, transformation, and training loop functioned as intended, proving the architectural pipeline is valid.

5.2 Analysis of Failure Cases

The model failed to segment the pets effectively. The reasons are identified as follows:

- **Insufficient Training:** Only 1 epoch was used. Segmentation models typically require 50+ epochs to converge.
- **Resolution Constraints:** Resizing to 64×64 removes high-frequency edge information necessary for accurate boundary detection.
- **Fast Demo Settings:** The notebook was optimized for speed rather than accuracy, leading to underfitting.

6 Conclusion

We successfully implemented an Attention U-Net for semantic segmentation on the Oxford-IIIT Pet dataset. While the architectural fusion of Attention Gates is theoretically sound, the constraints of the demonstration (low resolution, single epoch) prevented the model from learning effectively. Future work will involve increasing image resolution to at least 128×128 and training for a minimum of 20 epochs to observe the true benefits of the attention mechanism.