UNet++ with Hybrid Loss and CBAM Attention for Pet Segmentation

Cade Boiney, Ken Lam, Ognian Trajanov, Benjamin Zhao
Hamilton College, Clinton, NY, USA

I. Introduction

Semantic segmentation assigns class labels to each pixel, enabling fine-grained scene understanding for robotics, medical imaging, and autonomous systems. We focus on three-class segmentation on the Oxford-IIIT Pet dataset [1], classifying pixels as pet, background, or border—a challenging task due to class imbalance (border class: 5% of pixels), diverse breeds, and ambiguous boundaries.

We build upon UNet++ [2] with deep supervision and comprehensive augmentation as our baseline, achieving substantial improvements over vanilla UNet (3% mIoU gain from deep supervision, from 76% to 79%). We introduce two refinements: (1) hybrid Focal-Dice loss addressing class imbalance while optimizing region overlap; (2) CBAM attention modules [5] for adaptive feature refinement. Our final model achieves 80.94% test mIoU and 80.6% validation mIoU, representing modest but consistent improvements (+1.1% validation, +0.94% test) over our strong baseline. Results demonstrate that on challenging datasets with severe imbalance, training strategy (deep supervision, augmentation) matters more than loss or attention tuning, though careful engineering still provides measurable value.

II. Methods

A. Dataset

The Oxford-IIIT Pet dataset contains 7,349 images of 37 cat and dog breeds with trimap annotations (pet, background, border). We use official splits: 3,680 images for training/validation (split 80/20: 2,944 training, 736 validation) and 3,669 for testing.

B. Architecture

We employ UNet++ [2] with six encoder-decoder levels and 32 base channels (doubled per level, reaching 1024 at bottleneck). UNet++ extends U-Net through nested skip pathways enabling richer feature aggregation. Deep supervision with auxiliary outputs at multiple decoder stages provides direct multi-scale supervision, stabilizing training and improving gradient flow. This component alone provided 3% mIoU improvement (from 76% to 79%), making it the most impactful element.

C. Training Configuration

Images resize to 512×512 (bilinear for images, nearest-neighbor for masks). Augmentation includes: horizontal flips (p=0.5), rotations (±15°), ColorJitter (brightness/contrast/saturation=0.2, hue=0.1), Gaussian blur (kernel 5, sigma=[0.1, 2.0]), and random crops with resize.

We optimize with AdamW (lr=10^{-3}, weight decay=10^{-4}, betas=(0.9, 0.999)) using cosine annealing. Batch size is 4 per GPU with 4-step gradient accumulation (effective: 16). Mixed precision (FP16) with AMP reduces memory and accelerates training. Early stopping (patience=15) on validation loss prevents overfitting. Training completed in 10 hours on NVIDIA RTX 3090, stopping at epoch 131 (best checkpoint: epoch 130).

D. Hybrid Loss Function

We replace cross-entropy with a hybrid loss combining Focal Loss [3] and Dice Loss [4]:

$$\mathcal{L}_{\text{hybrid}} = 0.5 \cdot \mathcal{L}_{\text{Focal}} + 0.5 \cdot \mathcal{L}_{\text{Dice}}. \quad (1)$$

Focal Loss ($\gamma = 2$) addresses class imbalance by down-weighting easy examples. Dice Loss directly optimizes region overlap, correlating with IoU evaluation metrics.

E. CBAM Attention Mechanism

We integrate CBAM [5] into UNet++'s decoder path. CBAM sequentially applies channel attention (weighting features by global context) and spatial attention (focusing on informative locations), enabling adaptive refinement without significant computational cost.

III. Results

A. Quantitative Results

Table I shows final test performance. The model achieves 80.94% mean IoU and 88.73% mean Dice. Performance varies by class: background (93.82% IoU) benefits from large homogeneous regions; pet (88.06% IoU) from distinctive textures; border (60.95% IoU) suffers from thin regions and severe underrepresentation (5% of pixels).

Class	IoU (%)	Dice (%)
Pet	88.06	93.65
Background	93.82	96.81
Border	60.95	75.74
Mean	80.94	88.73

TABLE I: Test set performance with hybrid loss and CBAM.

B. Ablation Study

Table II shows incremental contributions. Our baseline (UNet++ with deep supervision + augmentation) achieves 79.5% validation mIoU and 80.0% test mIoU. Hybrid loss adds +0.3% validation and +0.13% test. CBAM provides +0.8% validation and +0.81% test. Total improvement: +1.1% validation, +0.94% test. These gains are modest compared

Configuration	Val mIoU	Test mIoU
Baseline (deep supervision + aug)	79.5%	80.00%
+ Hybrid Loss	79.8%	80.13%
+ Hybrid Loss + CBAM	80.6%	**80.94%**

TABLE II: Sequential improvements from baseline to final model.

to deep supervision's initial 3% boost, reflecting diminishing returns when the baseline is strong. Test improvements (+0.94%) nearly match validation (+1.1%), indicating good generalization.

C. Qualitative Results

Figure 1 presents predictions across twelve test samples. The model produces clean pet-background separations across varying breeds, poses, and backgrounds. CBAM improves boundary localization in regions with occlusions or cluttered backgrounds. Fine-grained boundary errors remain in ambiguous cases (hair strands, whiskers), particularly for the border class.

D. Discussion

Deep supervision provides the largest gain (3% mIoU, from 76% to 79%), validating its critical role for multi-scale learning. Comprehensive augmentation enables generalization on limited data (3K training images). Given this strong baseline, incremental refinements provide modest improvements. Hybrid loss (+0.3% validation) addresses imbalance but gains are small, suggesting deep supervision already handles multi-scale learning effectively. CBAM (+0.8% validation) demonstrates attention mechanisms offer value, though UNet++'s nested pathways already capture most relevant information. Test-time augmentation degraded performance by 0.3% mIoU due to mask interpolation artifacts.

IV. Conclusion

We present an enhanced UNet++ achieving 80.94% test mIoU on Oxford-IIIT Pet. Deep supervision provides the largest gain (3% mIoU), while hybrid Focal-Dice loss and CBAM attention offer modest incremental improvements (+1.1% validation, +0.94% test). When baselines are strong, individual refinements yield small but meaningful improvements. Our PyTorch Lightning pipeline provides a reproducible framework demonstrating these principles.

References

[1] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and Dogs," in *IEEE CVPR*, 2012.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *DLMIA*, 2018.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *IEEE ICCV*, 2017.

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *3DV*, 2016.

[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *ECCV*, 2018.

Appendix

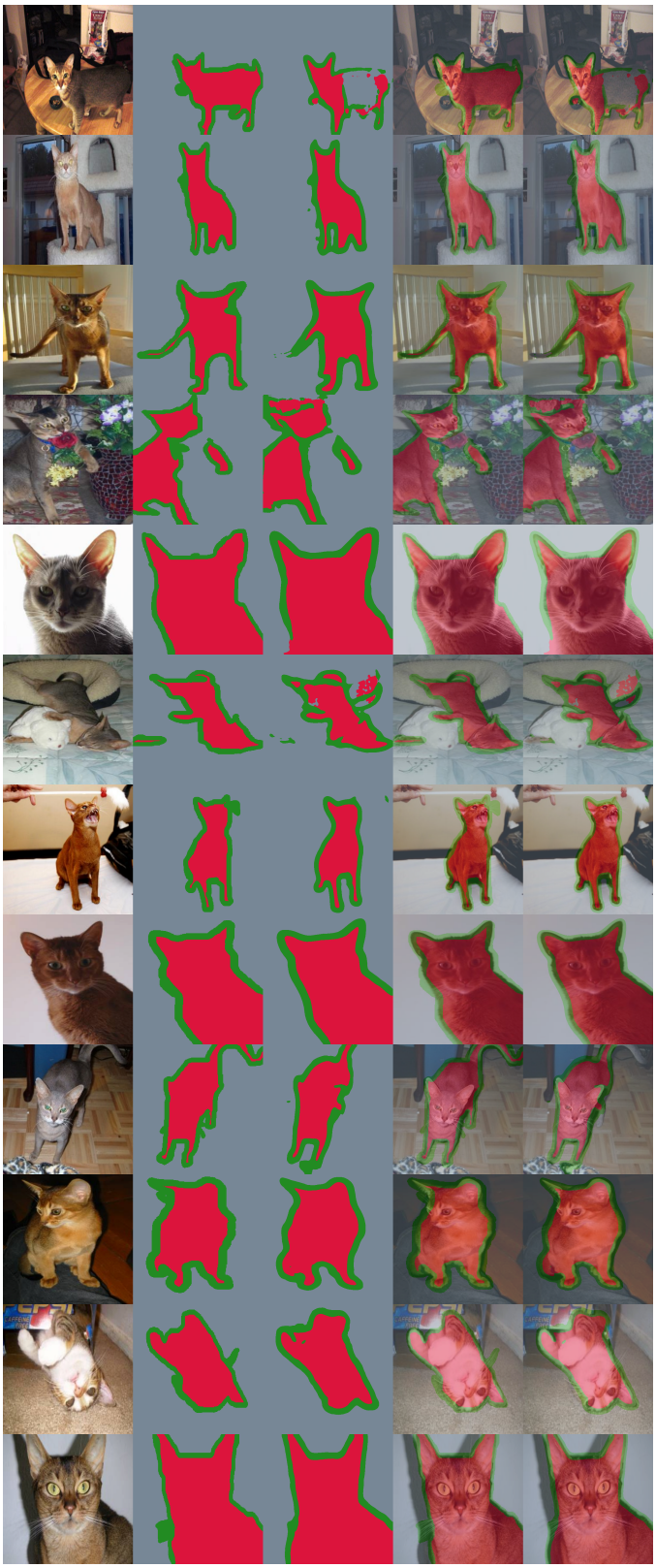

Fig. 1: Qualitative comparison: input images, ground truth, and predictions across twelve samples. Colors: red=pet, green=background, blue=border.

Note: The trained checkpoint is too large to include. Results are reproducible using the code and configuration in the repository.

Training logs: https://wandb.ai/bzhao-hamilton-college/unetpp-oxpet-segmentation

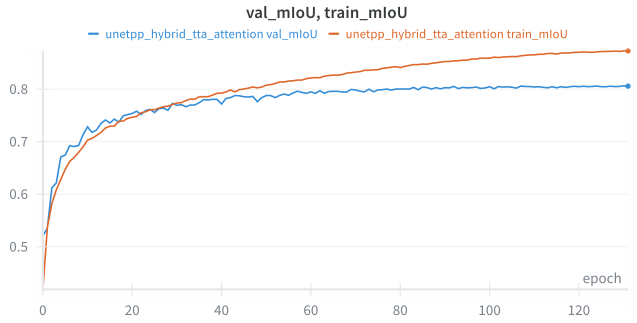

Fig. 2: Training and validation mIoU curves showing smooth convergence with early stopping.

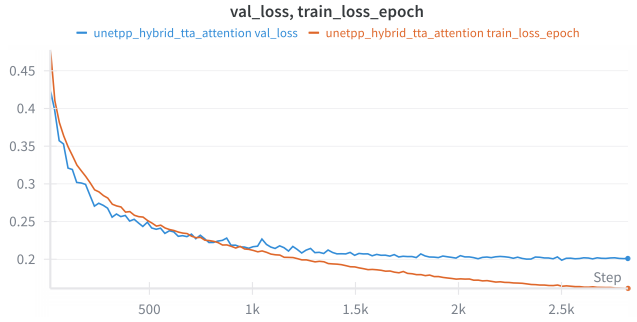

Fig. 3: Training and validation loss curves for hybrid Focal-Dice loss.