Weakly Supervised Sea-floor Segmentation

Solomon Chibuzo Nwafor
Dept. of Engineering
University of Girona
Girona, Spain
u1999124@campus.udg.edu

Muhammad Faran Akram
Dept. of Engineering
University of Girona
Girona, Spain
u1999088@campus.udg.edu

Enis Hidri
Dept. of Engineering
University of Girona
Girona, Spain
u1999254@campus.udg.edu

***Abstract*—We address the problem of weakly supervised semantic segmentation (WSSS), where models are trained using only image-level annotations. Our approach improves the quality of pseudo-labels by enhancing both label refinement and loss function design. We begin by tuning dense Conditional Random Fields (dCRFs) to better align class activation maps with object boundaries, reducing noise and improving label consistency. Additionally, we evaluate and integrate advanced loss functions—namely Focal Loss and Lovász-Softmax Loss—to address class imbalance and directly optimize the intersection-over-union (IoU) metric. Applied to the dataset, our tuned model achieves up to a 1.4% improvement in mIoU compared to the untuned baseline, on metric that is not sensitive to class imbalance. These results highlight the importance of accurate pseudo-label refinement and segmentation-aware loss selection in weakly supervised settings.**

Keywords—Semantic segmentation, weak supervision, pseudo-labeling, loss functions.

I. Introduction

Semantic segmentation remains a core task in computer vision, with fully supervised methods achieving high performance when trained on pixel-level annotations. However, obtaining such detailed labels is time-consuming and expensive, making them impractical for many real-world applications. Weakly Supervised Semantic Segmentation (WSSS) addresses this limitation by leveraging cheaper, coarse annotations such as image-level labels.

Recent WSSS approaches commonly rely on Class Activation Maps (CAMs) to provide pseudo-labels. While effective for localization, CAMs often focus on the most discriminative object regions, leading to incomplete and noisy segmentation masks. To mitigate this, existing frameworks employ iterative self-training, pseudo-label refinement (e.g., with dense Conditional Random Fields), and multi-task learning.

In parallel, semantic segmentation has also gained relevance in non-visual sensing domains. For example, side-scan sonar (SSS) imagery is being widely used in underwater applications such as marine archaeology, structural inspection, and environmental monitoring [3]. Unlike optical sensors, SSS can operate in low-visibility and deep-sea conditions, enabling the acquisition of large-scale acoustic maps. However, analyzing these maps traditionally requires manual annotation of terrain and structural features, which is labor intensive and costly. Automating semantic segmentation for SSS data presents an important opportunity to improve the efficiency of marine surveys and enable real-time scene understanding for Autonomous Underwater Vehicles (AUVs).

In this work, we do not propose a new model, but instead fine-tune an existing WSSS framework to analyze the impact of label refinement and loss function design. Specifically, we tune the parameters of dense CRF and evaluate segmentation performance under different loss functions, including Cross-Entropy, Focal Loss, and Lovász-Softmax. Our analysis also reveals that commonly used evaluation metrics, such as mean Intersection-over-Union (mIoU), may obscure meaningful differences between models, especially under class imbalance. We report improvements that are better understood when interpreting both the metric and the qualitative segmentation behavior.

II. Methodology

This study builds on a recent weakly supervised semantic segmentation (WSSS) framework that leverages Class Activation Maps (CAMs) to generate pseudo-segmentation labels using only image-level annotations. Rather than proposing a new model, we fine-tune this existing framework with two specific goals: (1) improve the spatial coherence of pseudo-labels via dense Conditional Random Field (dCRF) tuning, and (2) analyze how different loss functions influence training dynamics and segmentation accuracy. In this section, we first summarize the baseline framework we adopted and then detail our tuning procedures and evaluation setup.

A. Baseline Framework

The base model we employ adopts an encoder-decoder segmentation architecture, such as DeepLabv3, trained in a multi-task setting to jointly predict classification and segmentation outputs. CAMs are extracted from the classification branch and thresholded to form initial pseudo-labels. These pseudo-labels are further refined using dCRF to improve boundary precision and reduce label noise. The refined masks are then used to supervise the decoder via a pixel-wise segmentation loss.

The process follows an iterative self-training strategy: pseudo-labels generated at one stage are used to update the segmentation branch, which in turn influences the next round of CAM generation. This framework implicitly aims to reduce the supervision gap by using image content and consistency constraints to progressively enhance pseudo-label quality.

In the original setup, standard cross-entropy loss is used for

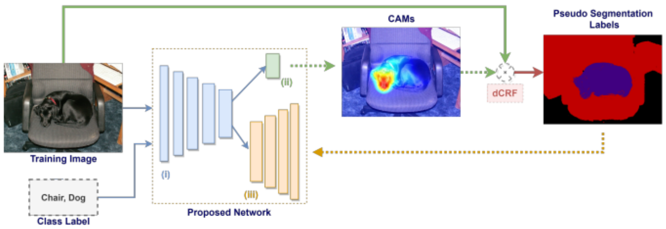

Fig. 1. Proposed Architecture: (i) Encoder network, (ii) Classification Branch, (iii) Decoder network

segmentation supervision. However, the model's reliance on CAMs and fixed thresholds often leads to incomplete object regions and poor background separation, especially in the presence of class imbalance. To address this, we re-evaluate the role of the segmentation loss function and refine the dCRF parameters, as detailed in the following subsections.

B. Dense CRF Tuning

Class Activation Maps (CAMs), while effective for highlighting object presence, tend to localize only the most discriminative parts of an object. As a result, the pseudo-labels derived from CAMs are often spatially fragmented and lack alignment with true object boundaries. To address this, the baseline framework incorporates a dense Conditional Random Field (dCRF) as a post-processing step to refine CAM outputs into more coherent segmentation masks.

In our work, we focus on tuning the dCRF [2] parameters to improve label quality without introducing new structural changes to the model. The dense CRF operates on the raw CAM-derived masks and uses low-level image features—such as pixel intensity and spatial location—to adjust label boundaries based on local consistency. Specifically, the energy function of the dCRF includes:

- A unary term derived from the soft CAM outputs.
- A pairwise bilateral term to encourage nearby pixels with similar color and position to share the same label.
- A spatial smoothness term to penalize isolated label changes.

We experimentally tune key hyperparameters of the dCRF, including:

- **Spatial standard deviation** (σ_α) and color standard deviation (σ_β), which control the influence radius of pairwise terms.
- **Weight coefficients** for bilateral and spatial kernels.
- **Number of inference iterations**, which affects convergence quality and computation time.

As shown in Figure 2, the application of dCRF leads to a substantial improvement in mask quality. The refined segmentation outputs exhibit smoother regions and more precise boundary delineation, demonstrating the effectiveness of dCRF tuning in enhancing weakly supervised label quality.

C. Loss Functions

In the baseline framework, pseudo-segmentation labels refined by dense CRF are used to supervise the segmentation branch through a standard pixel-wise cross-entropy (CE) loss. While effective in fully supervised settings, CE loss may underperform in weakly supervised scenarios, particularly in the presence of class imbalance and partial supervision from incomplete pseudo-labels.

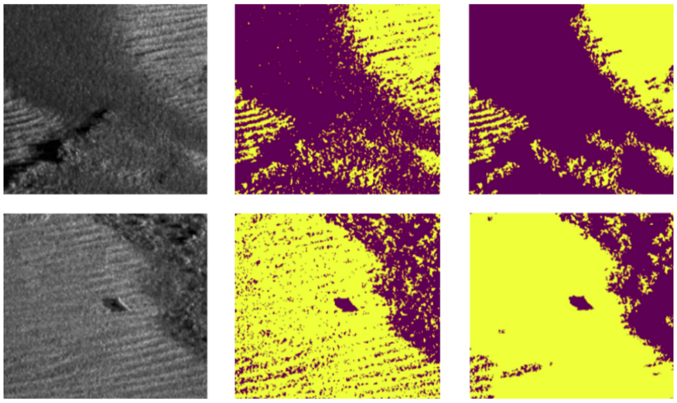

Fig. 2. Qualitative comparison of segmentation refinement. Each row shows (from left to right): the original input image, the initial pseudo-labels derived from CAMs, and the segmentation mask obtained after applying the tuned dense CRF. The final masks are significantly cleaner and better aligned with object boundaries.

To address these limitations, we explore the impact of alternative loss functions that are better suited for weak supervision. Specifically, we evaluate the following:

- **Focal Loss**: Designed to down-weight well-classified pixels and focus the training on hard examples, focal loss is useful for handling the background class imbalance commonly encountered in WSSS. The focal loss introduces a modulating factor to the standard cross-entropy, defined as:

$$\mathcal{L}_{\text{focal}} = -\sum_{c=1}^{C} \alpha_c (1 - p_c)^{\gamma} y_c \log(p_c) \tag{1}$$

  where:
  - p_c is the predicted probability for class c,
  - $y_c \in \{0, 1\}$ is the ground truth (pseudo-label),
  - $\gamma > 0$ is the focusing parameter that down-weights easy examples,
  - α_c is a weighting factor to balance class frequencies.
- **Lovász-Softmax Loss**: This loss directly optimizes the mean Intersection-over-Union (mIoU) metric by approximating the set-based Jaccard index with a convex surrogate, making it more aligned with evaluation criteria in segmentation tasks. The formulation is:

$$\mathcal{L}_{\text{lovasz}} = \frac{1}{C} \sum_{c=1}^{C} \text{LovaszHinge}(\mathbf{m}^{(c)}) \tag{2}$$

  where $\mathbf{m}^{(c)}$ is the vector of pixel-wise margin errors for class c.

Each loss function is integrated into the segmentation branch independently, while keeping all other training settings constant. We conducteded experiments to assess the qualitative and quantitative differences in the resulting segmentations.

III. Results

In this section, we present the experimental results obtained from training the baseline WSSS framework with different segmentation loss functions. Our analysis includes performance metrics, qualitative observations, and training dynamics to assess the impact of each loss. We place particular emphasis on the mean Intersection-over-Union (mIoU) metric and its reliability in reflecting actual segmentation quality.

A. *Cross-Entropy Loss and mIoU Trends*

We begin by analyzing the performance of the standard Binary Cross-Entropy (BCE) loss, which serves as the baseline. As shown in Fig. 3, the CE loss function converges well, gradually reducing the loss. It is evident that the model is neither under-fitting nor overfitting. The total number of training epochs is 100.

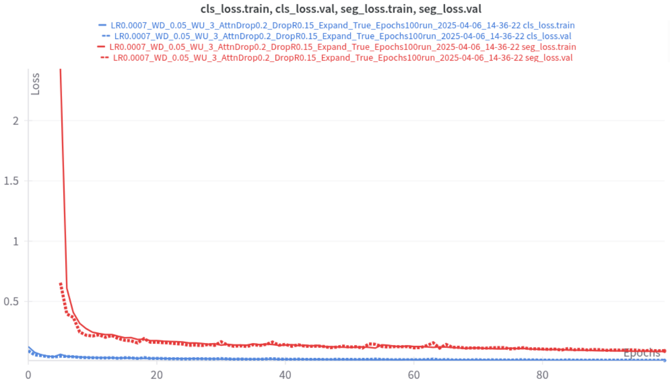

Fig. 3. Segmentation loss (CE) and classification loss (MLSM) curves over training epochs.

The mIOU curve can be seen in Fig. 4. Which shows maximum of 92.61% validation mIOU.

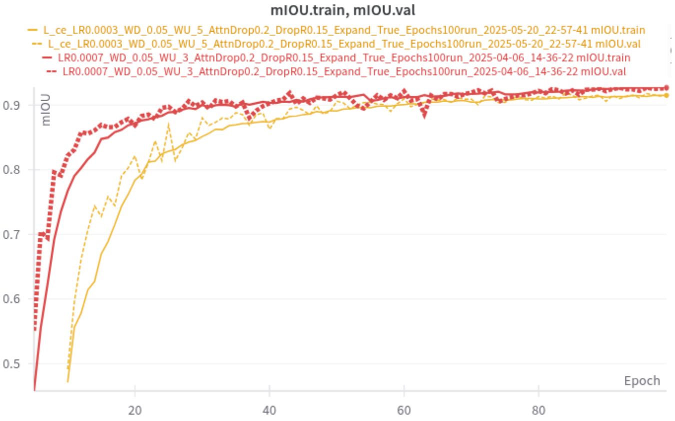

Fig. 4. mIOU curve for CE Loss.

The model was evaluated using the basic mIOU metric, which overlooks class imbalance in images. Since most images contain only a single class, the mIOU score can appear inflated, even when the model performs poorly on multi-class instances. Therefore, while computing mIOU, we excluded images that contained only a single class. The same model, using identical hyperparameters, was retrained to evaluate performance on this updated metric. The resulting curves, shown in Fig. 4, are represented by the yellow curve.

B. *Focal Loss Behavior and Comparison with BCE*

Next, we experimented Focal Loss to address the imbalance between background pixels. Fig. 5 presents the training curves, which show slower initial convergence compared to BCE but lead to more stable and meaningful segmentation results in later epochs.

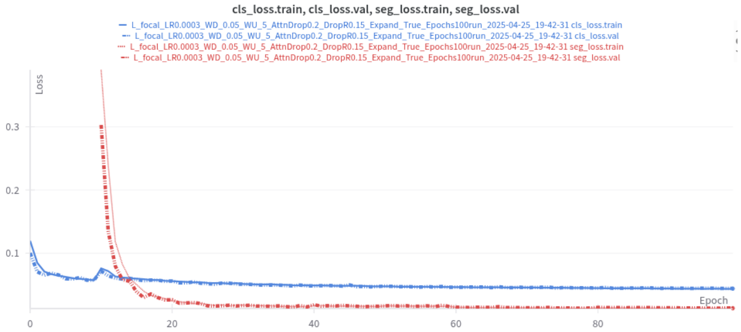

Fig. 5. Focal Loss curves

From Fig. 5, it is evident that the model converges gradually without signs of overfitting. However, as shown in Fig. 6, the final segmentation accuracy remains suboptimal. The likely reason is that Focal Loss, while reducing the contribution of easy background pixels, also suppresses gradients from moderately confident pixels. This effect may hinder the recovery of full object regions, especially in early training stages when pseudo-labels are incomplete.

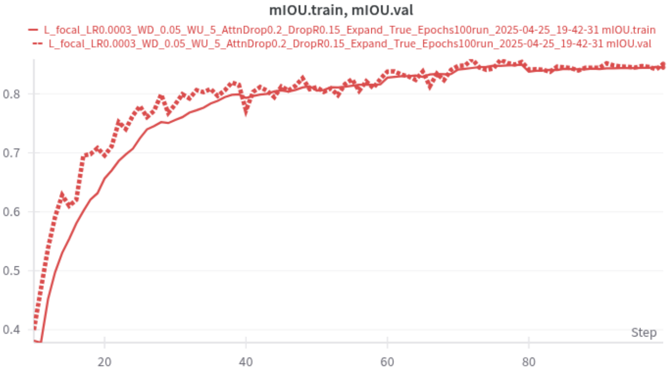

Fig. 6. mIOU curve for Focal Loss

The comparison Fig. 7 shows the resulting mIOU from CE (yellow) and focal (brown) losses.

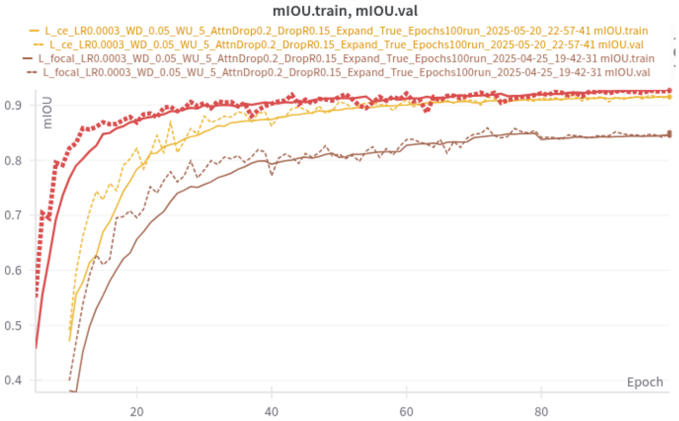

Fig. 7. mIOU for CE and Focal Losses

C. Lovász Loss Behavior and Comparison with BCE and Focal Losses

We further evaluate the Lovász-Softmax Loss, which directly optimizes the mean Intersection-over-Union (mIoU) metric. Unlike Cross-Entropy and Focal Loss, which operate on pixel-wise classification accuracy, Lovász Loss is a surrogate for the set-based Jaccard index and is thus better aligned with the segmentation evaluation objective.

Fig. 8 shows the segmentation and classification loss curves during training. The convergence is less smooth compared to BCE and Focal Loss, likely due to the ranking-based formulation of the Lovász surrogate. Nevertheless, the model remains stable and avoids overfitting.

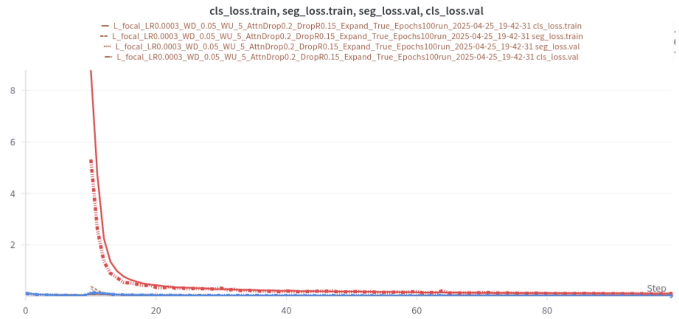

Fig. 8. Segmentation and classification loss curves using Lovász-Softmax Loss.

Fig. 9 illustrates the mIoU performance over training epochs.

Lovász Loss consistently outperforms both BCE and Focal Loss in terms of mIoU,can be seen in Fig. 10 It also shows better qualitative performance with more complete object regions and sharper boundaries as shown in Fig. 11

Table I summarizes the final mIoU scores across the three loss functions. Lovász loss achieves the best overall mIOU, demonstrating the importance of metric-aligned supervision in weakly supervised segmentation.

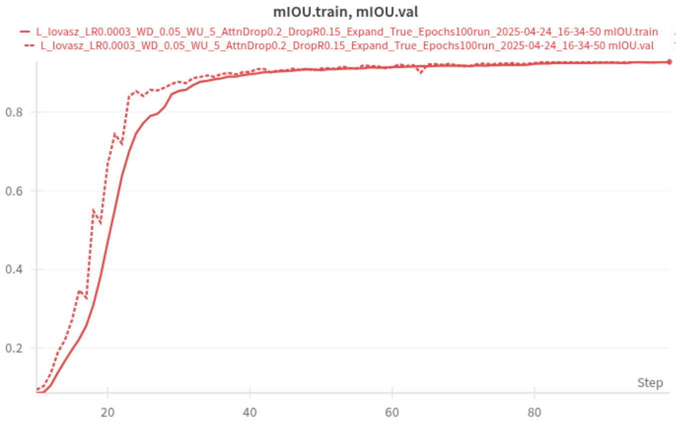

Fig. 9. mIoU curves for Lovász Loss.

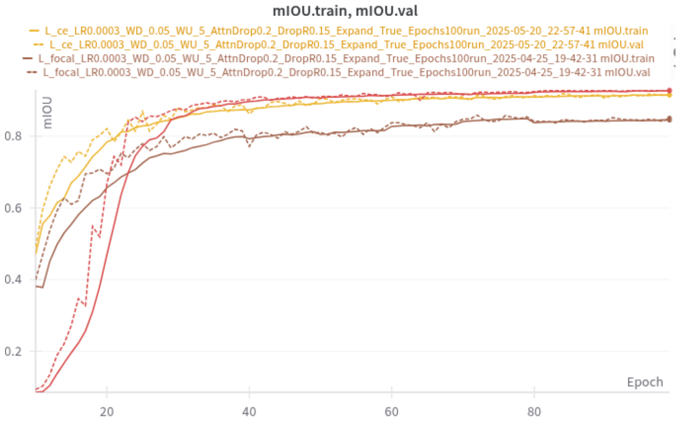

Fig. 10. mIOU curves for Lovasz vs Focal vs CE

IV. Conclusion

In this study, we investigated the impact of different loss functions on the performance of a weakly supervised semantic segmentation (WSSS) framework. While the base model used Class Activation Maps (CAMs) and dense Conditional Random Fields (dCRFs) for pseudo-label generation, our contributions focused on tuning the loss function to enhance segmentation performance under weak supervision.

We compared Cross-Entropy (CE), Focal Loss, and Lovász-Softmax Loss in terms of convergence behavior, segmentation quality, and metric alignment. CE served as a strong baseline, achieving high mIoU but overestimating performance due to class imbalance. Focal Loss handles class imbalance but showed reduced accuracy when evaluated with mIoU, possibly due to its emphasis on hard examples and underweighting

TABLE I
Comparison of Loss Functions on Final mIoU

Loss Function	**Final mIoU (%)**
Cross-Entropy	91.51
Focal Loss	85.06
Lovász-Softmax	92.94

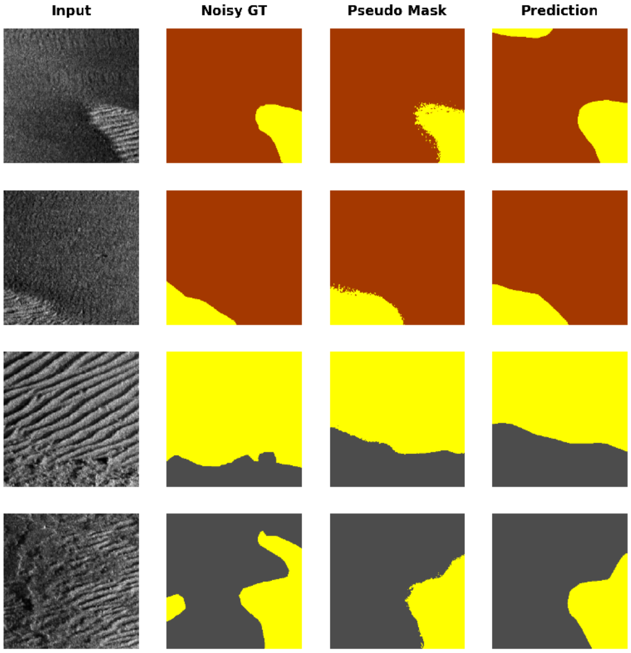

Fig. 11. Comparison between predicted segmentation mask and ground truth

moderate-confidence regions. Lovász-Softmax Loss, which is directly optimized for mIoU, consistently yielded the best quantitative and qualitative results.

Our findings highlight two key insights: (1) evaluation metrics like mIoU may not fully capture class imbalance effects unless carefully adapted; and (2) aligning the training objective with the evaluation metric, especially under noisy supervision, can significantly enhance performance. These insights can guide the design of more robust WSSS pipelines, especially in applications where annotation budgets are limited and segmentation quality remains critical.

References

[1] Cenk Bircanoglu and Nafiz Arica, “ISIM: Iterative Self-Improved Model for Weakly Supervised Segmentation,” *Springer Nature Computer Science*, 2021.

[2] Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum, “Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data,” Department of Computer Science, University of Massachusetts Amherst, MA, Technical Report, 2004.

[3] Hayat Rajani, Nuno Gracias, and Rafael Garcia, “A Convolutional Vision Transformer for Semantic Segmentation of Side-Scan Sonar Data,” *Computer Vision and Robotics Research Institute (ViCOROB), University of Girona*, 2022. [Online]. Available: https://arxiv.org/pdf/2202.09371