ADAPTIVE DICE LOSS FOR EXTREMELY IMBALANCED SEGMENTATION IN WETLAND DELINEATION

Sipeng Chen, Xu Zheng, Zeda Yin, Qiang Chen, Yuepeng Li, Jason Liu, Dongsheng Luo
Florida International University, Miami, United States
{schen131,xzhen019,zyin005,qchen,yuepli,liux,dluo}@fiu.edu

ABSTRACT

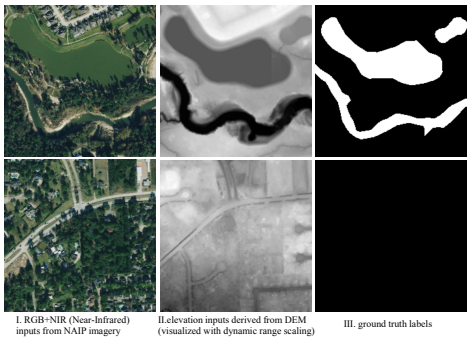
Wetlands play an essential role in mitigating climate change through their remarkable capacity for carbon sequestration. As such, their global degradation underscores an urgent need for precise mapping and monitoring. Deep learning has emerged as a promising solution for automated wetland delineation, enabling large-scale ecosystem monitoring. However, sparse spatial distribution of wetlands poses a significant challenge for segmentation methods, since many satellite imagery regions contain little to no wetland presence. Traditional loss functions such as Dice Loss fail to provide meaningful gradients in these wetland-sparse scenarios. To address this limitation, we introduce a novel formulation of Flipped Dice Loss that transforms the original pixel-wise relationships to enable gradient propagation in wetland-sparse regions. Building upon this method, we develop an Adaptive Dice Loss framework that can dynamically adjust the balance between standard Dice Loss and Flipped Dice Loss using a shifted sigmoid function. Experiments on our newly created Houston Wetland Dataset demonstrate that our method significantly improves wetland detection accuracy compared to state-of-the-art approaches. To facilitate future research in climate-oriented machine learning, we plan to release our multi-modal Houston Wetland Dataset.

1 INTRODUCTION

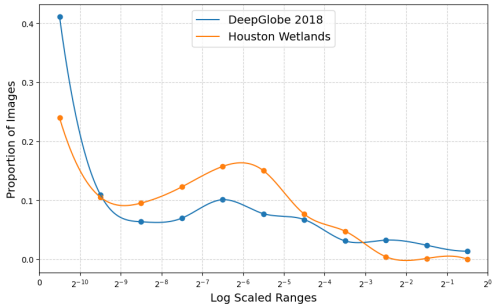
Wetlands are crucial ecosystems for biodiversity, water management, and carbon storage, covering about 6% of the Earth's surface and holding roughly 12% of the global carbon (Erwin, 2009). However, they have suffered severe degradation, with an estimated 64–71% loss since 1900 AD, and up to 87% since 1700 AD (Davidson, 2014). This rapid decline underscores the urgent need for accurate delineation of wetland locations and compositions to support effective conservation and restoration efforts, as these losses have been largely driven by direct human activities, such as conversion to agriculture (Murray et al., 2022).

In recent years, deep learning methods have been applied for capturing geographic information and delineating wetlands (Lin et al., 2023; Pham et al., 2022). Common segmentation models, such as U-Net (Ronneberger et al., 2015) and Swin-Unet (Cao et al., 2022), are often adapted to address the unique characteristics of wetland imagery. These models frequently incorporate multi-source data—such as satellite imagery (including RGB and Near-Infrared channels) and elevation data—to enhance segmentation accuracy. Compared to other semantic segmentation tasks, water distribution tends to exhibit higher spatial continuity (Ramos et al., 2022; Rakhimov et al., 2020). As shown in Fig. 1, many areas contain no water at all, resulting in all-background labels. Even in regions with water presence, labels are typically positive-sparse, with a very small proportion of positive pixels.

Various loss functions have been proposed to address this class imbalance, with Weighted Cross-Entropy and Dice Loss (Milletari et al., 2016) emerging as predominant approaches. Dice Loss computes the spatial overlap between predicted and ground truth foreground regions, functioning as a region-based metric analogous to IoU (Intersection over Union). This property makes it particularly suitable for handling imbalanced segmentation tasks (Liu et al., 2024a). However, Dice Loss exhibits a critical limitation in scenarios with all-background labels, where it generates zero gradients and fails to penalize false positive predictions effectively. While recent modifications to Dice Loss have explored various improvements—such as enhanced true positive emphasis and prediction



(a) Examples of images used for training



(b) Distribution of foreground pixel proportions

Figure 1: (a) Examples of images used for training and dataset distribution statistics; (2) Distributions of two datasets: Houston Wetlands and DeepGlobe 2018.

smoothing—these adaptations have not fundamentally resolved the gradient degeneration problem that occurs in all-background conditions.

To solve this problem, we introduce a novel formulation, called Flipped Dice Loss, that modifies the standard Dice Loss computation by treating background pixels as foreground and vice versa, thus enabling meaningful gradients in wetland-sparse regions. On top of that, we introduce an adaptive framework that automatically adjusts the balance between the conventional Dice Loss and Flipped Dice Loss methods using a sigmoid-based weighting mechanism. This dynamic adjustment enables the loss function to respond appropriately across varying densities of wetland presence, We validate our method on a newly constructed Houston Wetland Dataset, which combines high-resolution satellite imagery with elevation data to capture the complex characteristics of urban and suburban wetland distributions. Our contributions are summarized as follows:

- We provide theoretical analysis of Dice Loss gradient behavior in positive-sparse scenarios and propose Flipped Dice Loss as a solution to gradient degeneration.
- We develop an adaptive weighting mechanism that dynamically balances between standard Dice Loss and Flipped Dice Loss based on the density of wetland presence.
- We demonstrate that our adaptive framework consistently improves segmentation performance when applied to various existing Dice Loss variants.
- We construct and annotate the Houston Wetland Delineation Dataset, a multi-modal dataset combining satellite imagery and elevation data, to facilitate research in wetland segmentation under realistic conditions.

2 METHOD

2.1 NOTATIONS

Consider an image segmentation task where each image consists of $N = H \times W$ pixels, where H and W are the image height and width, respectively. The model predicts a vector of probabilities $\mathbf{p} \in \mathbb{R}^N$. The ground truth labels for all pixels are denoted as $\mathbf{r} \in \{0, 1\}^N$, where $r = 1$ represents foreground and $r = 0$ represents background. The proportion of positive pixels is defined as $\rho = \frac{1}{N}\sum_{i=1}^{N} r_i$. A small positive constant ϵ is added in loss function calculations to ensure numerical stability.

2.2 GRADIENT DEGENERATION OF DICE LOSS

The Dice Coefficient (Cohen, 1960) is a widely used metric for image segmentation tasks to quantify the overlap between predicted regions and ground truth (Fan et al., 2024). Dice Loss is derived from the Dice Coefficient to directly optimize region overlap (Milletari et al., 2016). For binary

classification, its formulation is as follows:

$$\mathcal{L}_{\text{Dice}} = 1 - DiceCoefficient = 1 - \frac{2\mathbf{p}^\top \mathbf{r} + \epsilon}{\mathbf{p}^\top \mathbf{1} + \mathbf{r}^\top \mathbf{1} + \epsilon}. \tag{1}$$

Specifically, when all ground truth labels are zeros, the gradient of $\mathcal{L}_{\text{Dice}}$ with respect to the predicted probabilities $\mathbf{p}$ is as follows:

$$\left.\frac{\partial \mathcal{L}_{\text{Dice}}}{\partial \mathbf{p}}\right|_{\mathbf{r}=\mathbf{0}} = \frac{\epsilon\,\mathbf{1}}{\left(\mathbf{p}^\top \mathbf{1} + \epsilon\right)^2}, \tag{2}$$

This expression shows that the gradient is a constant vector close to $\mathbf{0}$ when $\mathbf{p} \neq \mathbf{0}$ and exhibits significant gradient variance when $\mathbf{p}$ transitions from non-zero to zero.

Flipped Dice Loss treats background as foreground and vice versa by negating both labels and predicted probabilities. The vector form of Flipped Dice Loss for binary classification is as follows:

$$\mathcal{L}_{\text{Flipped Dice}} = 1 - \frac{2(\mathbf{1} - \mathbf{p})^\top (\mathbf{1} - \mathbf{r}) + \epsilon}{2N - \mathbf{p}^\top \mathbf{1} - \mathbf{r}^\top \mathbf{1} + \epsilon}. \tag{3}$$

Considering the gradients of $\mathcal{L}_{\text{Flipped Dice}}$ with respect to $\mathbf{p}$ when the label is all-background cases, we have:

$$\left.\frac{\partial \mathcal{L}_{\text{Flipped Dice}}}{\partial \mathbf{p}}\right|_{\mathbf{r}=\mathbf{0}} = \frac{2N\mathbf{1} + \epsilon\mathbf{1}}{\left(2N - \mathbf{p}^\top \mathbf{1} + \epsilon\right)^2}. \tag{4}$$

We propose that $\mathcal{L}_{\text{Flipped Dice}}$ provides three gradient benefits over standard $\mathcal{L}_{\text{Dice}}$: (1) While the small constant ϵ often drives the gradient of $\mathcal{L}_{\text{Dice}}$ toward zero, $\mathcal{L}_{\text{Flipped Dice}}$ maintains a more substantial gradient. (2) As p_i increases, the gradient of $\mathcal{L}_{\text{Dice}}$ diminishes, impeding the correction of false positives. In contrast, $\mathcal{L}_{\text{Flipped Dice}}$ yields a higher gradient for larger p_i, encouraging more vigorous error corrections. (3) Even when $\mathbf{r} \neq \mathbf{0}$, small ρ in $\mathcal{L}_{\text{Dice}}$ induces high gradient variance, destabilizing training. By comparison, $\mathcal{L}_{\text{Flipped Dice}}$ mitigates this issue, offering more stable gradient updates.

2.3 Adaptive Weighting Function

While $\mathcal{L}_{\text{Flipped Dice}}$ effectively addresses gradient issues in positive-sparse scenarios, its approach of treating background as foreground becomes suboptimal when foreground pixels are more prevalent. A straightforward solution is to combine both loss terms with a constant weight:

$$\mathcal{L}_{\text{Combined}} = \lambda \cdot \mathcal{L}_{\text{Dice}} + (1 - \lambda) \cdot \mathcal{L}_{\text{Flipped Dice}}, \tag{5}$$

where $\lambda \in [0, 1]$ is a weighting parameter. While this combination mitigates the gradient issues discussed in Sec. 2.2, using a fixed λ fails to adapt to varying distributions of foreground pixels across different images.

We further propose an adaptive weighting function using a shifted sigmoid $g(x) = \frac{1}{1+e^{-k(x-a)}}$. The parameter k controls the steepness of the function, while a controls its horizontal shift. We set $a = 0.002$ and $k = 5000$ as general design choices to ensure smooth adaptation across different segmentation tasks.

For any variant of Dice loss represented as $\mathcal{L}_{\text{Original}}(\hat{\mathbf{p}}, \mathbf{r})$, the corresponding adaptive loss function can be expressed as:

$$\mathcal{L}_{\text{Adaptive}} = g(\rho) \cdot \mathcal{L}_{\text{Original}}(\mathbf{p}, \mathbf{r}) + (1 - g(\rho)) \cdot \mathcal{L}_{\text{Flipped}}(1 - \mathbf{p}, 1 - \mathbf{r}). \tag{6}$$

This formulation applies to the entire family of Dice-based losses. As a result, $\mathcal{L}_{\text{Adaptive}}$ dynamically adjusts the loss function based on the foreground pixel proportion in each image, preventing gradient degradation while stabilizing training in positive-sparse scenarios.

3 Experiments

3.1 Datasets

We introduce a new wetland delineation dataset focused on the Houston metropolitan area. The dataset combines DEM (Digital Elevation Model) and NAIP (National Agriculture Imagery Pro-

Method	IoU↑	F1↑	Precision↑	Recall↑	Accuracy↑
Standard Loss Functions					
$\mathcal{L}_{\text{BCE}}$	0.7809	0.8770	0.8777	0.8763	0.9967
$\mathcal{L}_{\text{Focal}}$	0.7890	0.8821	0.8754	0.8889	0.9968
$\mathcal{L}_{\text{IoU}}$	0.7174	0.8354	0.7744	0.9070	0.9952
Dice-Based Methods					
$\mathcal{L}_{\text{Dice}}$	0.7701	0.8701	0.8474	**0.8941**	0.9964
$\mathcal{L}_{\text{Adaptive Dice}}$ (ours)	**0.7890**	**0.8821**	**0.8754**	0.8889	**0.9968**
Tversky-Based Methods					
$\mathcal{L}_{\text{Tversky}}$	0.7803	0.8766	0.8809	0.8723	0.9967
$\mathcal{L}_{\text{Adaptive Tversky}}$ (ours)	**0.8006**	**0.8892**	**0.8845**	**0.8941**	**0.9970**
$\mathcal{L}_{\text{Focal Tversky}}$	0.7875	0.8811	0.8932	0.8694	0.9968
$\mathcal{L}_{\text{Adaptive Focal Tversky}}$ (ours)	**0.8067**	**0.8930**	**0.9035**	**0.8828**	**0.9972**
Advanced Loss Combinations					
$\mathcal{L}_{\text{Weighted BCE}} + \mathcal{L}_{\text{Dice}}$	0.7718	0.8712	**0.8521**	0.8912	0.9965
$\mathcal{L}_{\text{Weighted BCE}} + \mathcal{L}_{\text{Adaptive Dice}}$ (ours)	**0.7870**	**0.8808**	0.8413	**0.9241**	**0.9966**
Other Dice Variants					
$\mathcal{L}_{\text{Generalized Dice}}$	0.0217	0.0425	0.0257	0.1230	0.9255
$\mathcal{L}_{\text{Adaptive Generalized Dice}}$ (ours)	**0.8009**	**0.8895**	**0.8924**	**0.8865**	**0.9970**
$\mathcal{L}_{\text{Log-Cosh Dice}}$	0.7893	0.8822	0.8807	**0.8837**	0.9968
$\mathcal{L}_{\text{Adaptive Log-Cosh Dice}}$ (ours)	**0.7954**	**0.8860**	**0.8939**	0.8784	**0.9970**

Table 1: Comparison of different loss functions on the test set. Best results in each category are shown in **bold**. Our adaptive methods consistently improve upon their baseline counterparts across different metrics.

gram) imagery, providing rich multi-modal information for wetland detection. Each sample contains 5 channels: a single-channel DEM and 4-channel NAIP data (RGB and Near-Infrared). Starting from the original $32,810 \times 13,070$ resolution imagery, we preprocess the data into 560×560 patches for efficient processing. Ground truth labels are manually annotated through careful cross-referencing with Google Maps. To ensure data quality, we discard patches where over 70% of the area lies outside the Houston metropolitan region. This dataset will be made publicly available to facilitate future research in wetland delineation.

3.2 Main Results

We conducte extensive experiments evaluating our adaptive loss framework against established segmentation losses. Table 1 demonstrates performance metrics across different loss functions, with IoU serving as our primary evaluation criterion.

The experimental results reveal systematic improvements of our adaptive approach across various baseline methods. The adaptive formulation enhances performance for both standard losses and their combinations, showing particular effectiveness when integrated with Tversky-based variants. In the Tversky family, both standard and focal versions demonstrate marked improvements in precision and recall when adapted with our method. The most notable improvement appears in the Generalized Dice variant, where our adaptive approach transforms an underperforming baseline into a competitive performer, indicating robust handling of severe class imbalance scenarios. These quantitative results demonstrate that our adaptive framework successfully addresses gradient degeneration issues while maintaining high segmentation accuracy across different loss formulations.

4 Conclusion

We propose an adaptive loss framework to improve segmentation in positive-sparse scenarios, addressing the gradient degeneration issue of Dice Loss through Flipped Dice Loss and an adaptive weighting function. Our method consistently enhances performance across various loss functions, as demonstrated in wetland delineation experiments. Additionally, we introduce a high-resolution

wetland dataset to support future research in environmental monitoring. By improving wetland segmentation, our approach contributes to large-scale ecosystem analysis and conservation efforts.

ACKNOWLEDGMENTS

This project was partially supported by NSF grant IIS-2331908. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

REFERENCES

Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pp. 683–687. IEEE, 2019.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4413–4421, 2018.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Nick C Davidson. How much wetland has the world lost? long-term and recent trends in global wetland area. *Marine and Freshwater Research*, 65(10):934–941, 2014.

Kevin L Erwin. Wetlands and global climate change: the role of wetland restoration in a changing world. *Wetlands Ecology and management*, 17(1):71–84, 2009.

Xin Fan, Xiaolin Wang, Jiaxin Gao, Jia Wang, Zhongxuan Luo, and Risheng Liu. Bi-level learning of task-specific decoders for joint registration and one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11726–11735, June 2024.

K Ruwani M Fernando and Chris P Tsokos. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2021.

Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pp. 1–7. IEEE, 2020.

Minhyeok Lee, Suhwan Cho, Dogyoon Lee, Chaewon Park, Jungho Lee, and Sangyoun Lee. Guided slot attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3807–3816, June 2024.

Xufeng Lin, Youwei Cheng, Gong Chen, Wenjing Chen, Rong Chen, Demin Gao, Yinlong Zhang, and Yongbo Wu. Semantic segmentation of china's coastal wetlands based on sentinel-2 and segformer. *Remote Sensing*, 15(15):3714, 2023.

Bingyuan Liu, Jose Dolz, Adrian Galdran, Riadh Kobbi, and Ismail Ben Ayed. Do we really need dice? the hidden region-size biases of segmentation losses. *Medical Image Analysis*, 91:103015, 2024a.

Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3491–3500, June 2024b.

Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11642–11651, June 2024.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016.

Nicholas J Murray, Thomas A Worthington, Pete Bunting, Stephanie Duce, Valerie Hagger, Catherine E Lovelock, Richard Lucas, Megan I Saunders, Marcus Sheaves, Mark Spalding, et al. High-resolution mapping of losses and gains of earth's tidal wetlands. *Science*, 376(6594):744–749, 2022.

Hanh Nguyen Pham, Kinh Bac Dang, Thanh Vinh Nguyen, Ngoc Cuong Tran, Xuan Quy Ngo, Duc Anh Nguyen, Thi Thanh Hai Phan, Thu Thuy Nguyen, Wenshan Guo, and Huu Hao Ngo. A new deep learning approach based on bilateral semantic segmentation models for sustainable estuarine wetland ecosystem management. *Science of The Total Environment*, 838:155826, 2022.

Shavkat Rakhimov, Aybek Seytov, Nasiba Rakhimova, and Bahrom Xonimqulov. Mathematical models of optimal distribution of water in main channels. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–4. IEEE, 2020.

Helena M Ramos, Maria Cristina Morani, Armando Carravetta, Oreste Fecarrotta, Kemi Adeyeye, P Amparo López-Jiménez, and Modesto Pérez-Sánchez. New challenges towards smart systems' efficiency by digital twin in water distribution networks. *Water*, 14(8):1304, 2022.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2980–2988, 2017.

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pp. 379–387. Springer, 2017.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pp. 240–248. Springer, 2017.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Guotai Wang, Wenqi Li, Sebastien Ourselin, and Tom Vercauteren. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–560. Springer, 2018.

Zixue Xiang, Wei Peng, Xu Liu, and Wen Yao. Self-adaptive loss balanced physics-informed neural networks. *Neurocomputing*, 496:11–34, 2022.

Lingxi Xie, Jianzhong He, Yutong Bai, Yizhe Zhang, Yanfeng Wang, and Alan Yuille. Hausdorff distance loss for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10407–10416, 2020.

Yi-Fan Zhang, Weiqiang Ren, Zhang Zhang, Zhen Jia, Liang Wang, and Tieniu Tan. Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing*, 506:146–157, 2022.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12993–13000, 2020.

A Related Work

A.1 Dice Loss Family

Tversky Loss (Salehi et al., 2017). The Tversky index (Tversky, 1977) generalizes the Dice coefficient and F_β scores, enabling an improved trade-off between precision and recall for segmenting highly unbalanced data. By adjusting the hyperparameters α and β, we can control the balance between false positives and false negatives:

$$\mathcal{L}_{\text{Tversky}} = 1 - \frac{\mathbf{p}^\top \mathbf{r} + \epsilon}{\mathbf{p}^\top \mathbf{r} + \alpha \mathbf{p}^\top (\mathbf{1} - \mathbf{r}) + \beta (\mathbf{1} - \mathbf{p})^\top \mathbf{r} + \epsilon}. \tag{7}$$

Focal Tversky Loss (Abraham & Khan, 2019). Similar to Focal Loss (Ross & Dollár, 2017), Focal Tversky Loss introduces a parameter γ to focus learning on hard examples and handle class imbalance more effectively. The parameter γ adjusts the contribution of easy versus hard examples:

$$\mathcal{L}_{\text{Focal Tversky}} = \left(\mathcal{L}_{\text{Focal Tversky}}\right)^\gamma. \tag{8}$$

Generalized Dice Loss (Sudre et al., 2017). $\mathcal{L}_{\text{GDL}}$ introduces class weights inversely proportional to the label frequencies. This weighting balances the influence of each class, mitigating the impact of class imbalance:

$$\mathcal{L}_{\text{GDL}} = 1 - \frac{2\left(w_1 \mathbf{p}^\top \mathbf{r} + w_0 \left((1 - \mathbf{p})^\top (1 - \mathbf{r})\right)\right)}{w_1 \left(\mathbf{p}^\top \mathbf{1} + \mathbf{r}^\top \mathbf{1}\right) + w_0 \left((1 - \mathbf{p})^\top \mathbf{1} + (1 - \mathbf{r})^\top \mathbf{1}\right)}, \tag{9}$$

with class weights defined as:

$$w_1 = \frac{1 + \epsilon}{\left(\mathbf{r}^\top \mathbf{1}\right)^2 + \epsilon}, \quad w_0 = \frac{1 + \epsilon}{\left((\mathbf{1} - \mathbf{r})^\top \mathbf{1}\right)^2 + \epsilon}, \tag{10}$$

Note that for very sparse labels, w_1 excessively emphasizes the positive class, leading to difficulties in updating the model's weights effectively. This issue is demonstrated in Tab. 1.

Log-Cosh Dice Loss (Jadon, 2020). By applying the logarithm of the hyperbolic cosine to $\mathcal{L}_{\text{Dice}}$, this formulation combines the advantages of logarithmic and hyperbolic cosine functions to provide a smooth and robust loss function:

$$\mathcal{L}_{\text{Log-Cosh Dice}} = \ln\left(\cosh\left(\mathcal{L}_{\text{Dice}}\right)\right). \tag{11}$$

However, none of these Dice Loss variations effectively resolve the vanishing gradient issue encountered when all labels are zero, which limits their ability to handle background-dominated images.

A.2 Other Loss Functions for Sparse Label Semantic Segmentation

Loss functions for semantic segmentation can be broadly classified into four categories (Ma et al., 2021; Jadon, 2020): Distribution-based Losses, Region-based Losses, Boundary-based Losses, and Compounded Losses that combine these types.

Distribution-based losses primarily include Cross-Entropy (CE) and its variations, such as Weighted Cross-Entropy and Focal Loss (Ross & Dollár, 2017). These methods address class imbalances by emphasizing different class distributions or focusing on hard-to-classify examples, making them effective for imbalanced datasets.

Region-based losses include the Dice Loss family introduced in Section A.1, as well as IoU Loss and its variants such as Efficient IoU Loss (Zhang et al., 2022), Generalized IoU Loss (Rezatofighi et al., 2019; Lee et al., 2024), and Lovász-Softmax Loss (Berman et al., 2018). Distance-IoU Loss (Zheng et al., 2020), which converges faster during training than IoU and GIoU losses, also falls into this category. These loss functions directly optimize for region overlap, making them particularly effective for addressing class imbalance in segmentation tasks.

Boundary-based losses, such as Hausdorff Distance Loss (Xie et al., 2020) and Shape-aware Loss (Wang et al., 2018), focus on capturing accurate boundary details. These losses are especially beneficial for fine-grained segmentation where boundary precision is critical.

Compounded losses combine multiple types of loss functions to leverage their strengths. For instance, in medical image segmentation, BCEDiceLoss (Liu et al., 2024b; Ma et al., 2024) is widely used to address class imbalance and improve region overlap. However, such combinations introduce a new challenge: balancing the contributions of different losses. Adaptive weighting strategies have been proposed to tackle this issue, such as dynamically adjusting weights based on class frequency or the predicted probability of the ground-truth class (Xiang et al., 2022; Fernando & Tsokos, 2021).

B The Gradient Degeneration Problems for the Dice Loss Family

In the main paper, we discussed the gradient issues of $\mathcal{L}_{\text{Dice}}$ but omitted its variants due to space constraints. Here, we provide additional proofs of gradient degeneration in the dice loss family. We define $\overline{\mathbf{p}} = \mathbf{1} - \mathbf{p}$ and $\overline{\mathbf{r}} = \mathbf{1} - \mathbf{r}$ as the complements of the predicted probabilities and ground truth labels, respectively.

Tversky Loss

The Tversky Loss (Salehi et al., 2017) is defined as:

$$\mathcal{L}_{\text{Tversky}} = 1 - \frac{\mathbf{p}^\top \mathbf{r} + \epsilon}{\mathbf{p}^\top \mathbf{r} + \alpha \mathbf{p}^\top \overline{\mathbf{r}} + \beta \overline{\mathbf{p}}^\top \mathbf{r} + \epsilon}. \tag{12}$$

The gradient of the Tversky Loss with respect to $\mathbf{p}$ is:

$$\frac{\partial \mathcal{L}_{\text{Tversky}}}{\partial \mathbf{p}} = \frac{-\mathbf{r}}{\mathbf{p}^\top \mathbf{r} + \alpha\, \mathbf{p}^\top \overline{\mathbf{r}} + \beta\, \overline{\mathbf{p}}^\top \mathbf{r} + \epsilon} + \frac{(\mathbf{p}^\top \mathbf{r} + \epsilon)\,(\alpha \overline{\mathbf{r}} - \beta\, \mathbf{r})}{\left(\mathbf{p}^\top \mathbf{r} + \alpha\, \mathbf{p}^\top \overline{\mathbf{r}} + \beta\, \overline{\mathbf{p}}^\top \mathbf{r} + \epsilon\right)^2}.$$

When $\mathbf{r} = \mathbf{0}$:

$$\left.\frac{\partial \mathcal{L}_{\text{Tversky}}}{\partial \mathbf{p}}\right|_{\mathbf{r}=\mathbf{0}} = \frac{\epsilon\, \alpha\, \mathbf{1}}{\left(\alpha\, \mathbf{p}^\top \mathbf{1} + \epsilon\right)^2}. \tag{13}$$

Similar to the gradient of $\mathcal{L}_{\text{Dice}}$ with respect to the predicted probabilities $\mathbf{p}$ in Eqn. 2, the gradient of $\mathcal{L}_{\text{Tversky}}$ approaches zero when $\mathbf{p} \neq \mathbf{0}$. However, when $\mathbf{p} = \mathbf{0}$, the gradient becomes a very large constant, causing $\mathcal{L}_{\text{Tversky}}$ to directly shift from 1 to 0, which could lead to very large gradient variance.

Focal Tversky Loss

The Focal Tversky Loss (Abraham & Khan, 2019) is defined as:

$$\mathcal{L}_{\text{Focal Tversky}} = \left(1 - \mathcal{L}_{\text{Tversky}}\right)^{\gamma}. \tag{14}$$

The gradient of the Focal Tversky Loss with respect to $\mathbf{p}$ is:

$$\frac{\partial \mathcal{L}_{\text{Focal Tversky}}}{\partial \mathbf{p}} = -\gamma (1 - \mathcal{L}_{\text{Tversky}})^{\gamma - 1} \frac{\partial \mathcal{L}_{\text{Tversky}}}{\partial \mathbf{p}}. \tag{15}$$

When $\mathbf{r} = \mathbf{0}$:

$$\left.\frac{\partial \mathcal{L}_{\text{Focal Tversky}}}{\partial \mathbf{p}}\right|_{\mathbf{r}=\mathbf{0}} = \gamma \left(1 - \mathcal{L}_{\text{Tversky}}\right)^{\gamma - 1} \frac{\epsilon\, \alpha\, \mathbf{1}}{\left(\alpha\, \mathbf{p}^\top \mathbf{1} + \epsilon\right)^2}. \tag{16}$$

For all-background labels, $\mathcal{L}_{\text{Tversky}}$ becomes 0 when $\mathbf{p}$ is zero; otherwise, $\mathcal{L}_{\text{Tversky}}$ is 1. Hence, the gradient approaches $\mathbf{0}$ when $\mathbf{p} \neq \mathbf{0}$, and it approaches a large constant when $\mathbf{p} = \mathbf{0}$. This behavior is consistent with the gradient of $\mathcal{L}_{\text{Dice}}$ under similar conditions.

GENERALIZED DICE LOSS

The Generalized Dice Loss (GDL) (Sudre et al., 2017) is defined as:

$$\mathcal{L}_{\text{GDL}} = 1 - \frac{2\left(w_1\,\mathbf{p}^\top\mathbf{r} + w_0\,\overline{\mathbf{p}}^\top\overline{\mathbf{r}}\right)}{w_1\left(\mathbf{p}^\top\mathbf{1} + \mathbf{r}^\top\mathbf{1}\right) + w_0\left(\overline{\mathbf{p}}^\top\mathbf{1} + \overline{\mathbf{r}}^\top\mathbf{1}\right)}, \tag{17}$$

where w_1 and w_0 are class weights defined as:

$$w_1 = \frac{1+\epsilon}{\left(\mathbf{r}^\top\mathbf{1}\right)^2 + \epsilon}, \quad w_0 = \frac{1+\epsilon}{\left(\left(\overline{\mathbf{r}}^\top\mathbf{1}\right)^2 + \epsilon\right.}. \tag{18}$$

The gradient of the GDL with respect to $\mathbf{p}$ is:

$$\frac{\partial\mathcal{L}_{\text{GDL}}}{\partial\mathbf{p}} = \frac{-2\left(w_1\mathbf{r} - w_0\overline{\mathbf{r}}\right)}{w_1\left(\mathbf{p}^\top\mathbf{1} + \mathbf{r}^\top\mathbf{1}\right) + w_0\left(\overline{\mathbf{p}}^\top\mathbf{1} + \overline{\mathbf{r}}^\top\mathbf{1}\right)} + \frac{2\left(w_1\mathbf{p}^\top\mathbf{r} + w_0\overline{\mathbf{p}}^\top\overline{\mathbf{r}}\right)\left(w_1\mathbf{1} - w_0\mathbf{1}\right)}{\left[w_1\left(\mathbf{p}^\top\mathbf{1} + \mathbf{r}^\top\mathbf{1}\right) + w_0\left(\overline{\mathbf{p}}^\top\mathbf{1} + \overline{\mathbf{r}}^\top\mathbf{1}\right)\right]^2}.$$

When $\mathbf{r} = \mathbf{0}$:

$$\left.\frac{\partial\mathcal{L}_{\text{GDL}}}{\partial\mathbf{p}}\right|_{\mathbf{r}=\mathbf{0}} = \frac{2w_0\mathbf{1}}{w_1\mathbf{p}^\top\mathbf{1} + w_0\left(\overline{\mathbf{p}}^\top\mathbf{1} + N\right)} + \frac{2w_0\left(\overline{\mathbf{p}}^\top\mathbf{1}\right)\left(w_1 - w_0\right)\mathbf{1}}{\left[w_1\mathbf{p}^\top\mathbf{1} + w_0\left(\overline{\mathbf{p}}^\top\mathbf{1} + N\right)\right]^2},$$

Consider $\mathbf{p} = \mathbf{0}$:

$$\begin{aligned}\left.\frac{\partial\mathcal{L}_{\text{GDL}}}{\partial\mathbf{p}}\right|_{\mathbf{r}=\mathbf{0},\mathbf{p}=\mathbf{0}} &= \frac{\mathbf{1}}{N} + \frac{(w_1 - w_0)\mathbf{1}}{2Nw_0} \\ &\approx \frac{\mathbf{1}}{N} + \frac{N\mathbf{1}}{2}\left(\frac{1}{\epsilon} - \frac{1}{N^2}\right) \\ &\approx \frac{N\mathbf{1}}{2\epsilon},\end{aligned} \tag{19}$$

indicating a large constant gradient for perfect true negative predictions.

When $\mathbf{p} \neq \mathbf{0}$:

$$\begin{aligned}\left.\frac{\partial\mathcal{L}_{\text{GDL}}}{\partial\mathbf{p}}\right|_{\mathbf{r}=\mathbf{0},\mathbf{p}\neq\mathbf{0}} &\approx \frac{2w_0\mathbf{1}}{w_1\mathbf{p}^\top\mathbf{1}} + \frac{2w_1w_0(\mathbf{p}^\top\mathbf{1})\mathbf{1}}{(w_1\mathbf{p}^\top\mathbf{1})^2} \\ &= \frac{2w_0\mathbf{1}}{w_1\mathbf{p}^\top\mathbf{1}} \\ &\approx \frac{4\epsilon\mathbf{1}}{N^2\mathbf{p}^\top\mathbf{1}},\end{aligned} \tag{20}$$

showing that the gradient approaches $\mathbf{0}$ because ϵ is a small constant. This indicates there is no effective gradient for false positive cases.

LOG-COSH DICE LOSS

The Log-Cosh Dice Loss (Jadon, 2020) is defined as:

$$\mathcal{L}_{\text{Log-Cosh}} = \log\left(\cosh\left(\mathcal{L}_{\text{Dice}}\right)\right). \tag{21}$$

The gradient of $\mathcal{L}_{\text{Log-Cosh}}$ with respect to $\mathbf{p}$ is:

$$\frac{\partial\mathcal{L}_{\text{Log-Cosh}}}{\partial\mathbf{p}} = \tanh\left(\mathcal{L}_{\text{Dice}}\right)\cdot\frac{\partial\mathcal{L}_{\text{Dice}}}{\partial\mathbf{p}}, \tag{22}$$

where the gradient of $\mathcal{L}_{\text{Dice}}$ is:

$$\frac{\partial \mathcal{L}_{\text{Dice}}}{\partial \mathbf{p}} = -2\frac{\left(\mathbf{r}\left(\mathbf{p}^\top \mathbf{1} + \mathbf{r}^\top \mathbf{1} + \epsilon\right) - \left(2\,\mathbf{p}^\top \mathbf{r} + \epsilon\right)\mathbf{1}\right)}{\left(\mathbf{p}^\top \mathbf{1} + \mathbf{r}^\top \mathbf{1} + \epsilon\right)^2}, \tag{23}$$

therefore, the gradient becomes:

$$\frac{\partial \mathcal{L}_{\text{Log-Cosh}}}{\partial \mathbf{p}} = -2\,\tanh\left(\mathcal{L}_{\text{Dice}}\right)\left(\frac{\mathbf{r}}{\mathbf{p}^\top \mathbf{1} + \mathbf{r}^\top \mathbf{1} + \epsilon} - \frac{\left(2\,\mathbf{p}^\top \mathbf{r} + \epsilon\right)\mathbf{1}}{\left(\mathbf{p}^\top \mathbf{1} + \mathbf{r}^\top \mathbf{1} + \epsilon\right)^2}\right).$$

When $\mathbf{r} = \mathbf{0}$:

$$\left.\frac{\partial \mathcal{L}_{\text{Log-Cosh}}}{\partial \mathbf{p}}\right|_{\mathbf{r}=\mathbf{0}} = 2\,\tanh\left(\mathcal{L}_{\text{Dice}}\right)\frac{\epsilon\,\mathbf{1}}{\left(\mathbf{p}^\top \mathbf{1} + \epsilon\right)^2}. \tag{24}$$

Note that when $\mathbf{p} = \mathbf{0}$, the gradient of $\mathcal{L}_{\text{Log-Cosh}}$ with respect to $\mathbf{p}$ becomes $\mathbf{0}$, indicating that $\mathcal{L}_{\text{Log-Cosh}}$ does not suffer from the issue of large constant gradients for perfect true negative predictions. However, when $\mathbf{p} \neq \mathbf{0}$, $\mathcal{L}_{\text{Log-Cosh}}$ still decreases as the number of false positives increases. This behavior implies that as false positives grow, the model becomes less effective in correcting errors.

Adaptive Loss

The proposed $\mathcal{L}_{\text{Adaptive}}$ is defined as:

$$\mathcal{L}_{\text{Adaptive}} = g(\rho) \cdot \mathcal{L}_{\text{Original}}(\mathbf{p}, \mathbf{r}) + (1 - g(\rho)) \cdot \mathcal{L}_{\text{Flipped}}(\overline{\mathbf{p}}, \overline{\mathbf{r}}). \tag{25}$$

When $\mathbf{r} = \mathbf{0}$, the positive pixel ratio ρ becomes 0. Since $g(0)$ is set to 0, we have:

$$\mathcal{L}_{\text{Adaptive}}|_{\mathbf{r}=\mathbf{0}} = \mathcal{L}_{\text{Flipped}}. \tag{26}$$

This allows the loss to handle all-background cases effectively, as $\mathcal{L}_{\text{Flipped}}$ is designed to perform well in such scenarios (see Section 2.2).

C Details of the Houston Wetland Delineation Dataset

The Houston Wetland Delineation Dataset was constructed to address the challenges of real-world positive-sparse conditions in segmentation tasks. Below, we provide additional details about its data sources, preprocessing steps, and annotation process, which were omitted in the main paper due to space limitations.

Data Sources

The dataset integrates two complementary data modalities:

Digital Elevation Model (DEM): Sourced from the United States Geological Survey (USGS[1]), the DEM data is LiDAR-based, offering precise elevation measurements. LiDAR (Light Detection and Ranging) is a remote sensing technology that uses laser pulses to measure distances, providing high-resolution and accurate elevation data. The DEM consists of single-channel images with pixel values ranging from 13.69 to 95.20 meters, capturing the topographical variation of the Houston region.

National Agriculture Imagery Program (NAIP): The NAIP imagery, also provided by the USGS, contains four-channel data (RGB + Near-Infrared). The pixel intensity values range from 0 to 255, enabling detailed analysis of vegetation, water bodies, and urban structures. This multi-spectral information is critical for wetland delineation tasks.

[1] https://www.usgs.gov

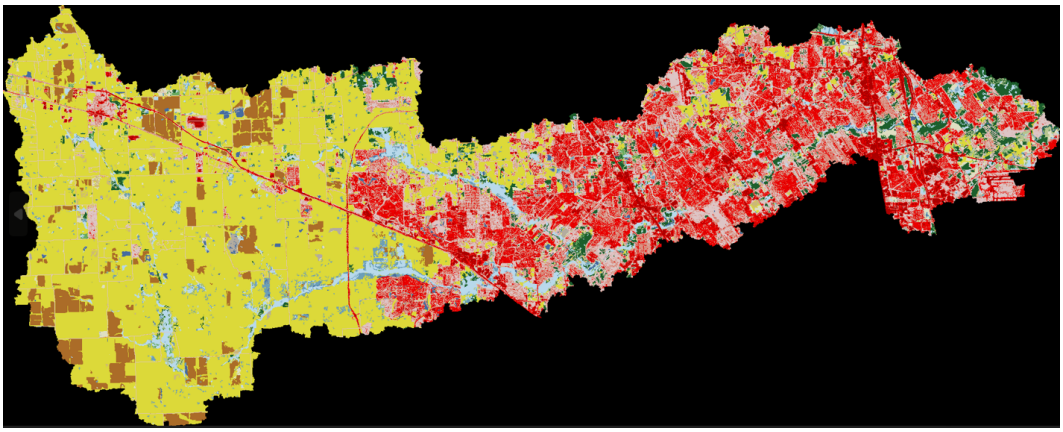


Figure 2: NAIP image showing an example of the irregular shape of the Houston area. Patches with over 70% background (background is shown as the black area in this figure) are excluded from the dataset.

Figure 3: Examples of cropped images used for training. (a) Images discarded due to over 70% non-Houston area, (b) images retained for training.

Annotation and Ground Truth Generation

Ground truth annotations were created using QGIS[2], guided by high-resolution satellite imagery from Google Maps. The annotation team comprised Ph.D. students with expertise in environmental science and remote sensing. Each patch was manually labeled, with particular care taken to ensure accuracy in delineating wetland boundaries.

The ground truth data highlights the sparse distribution of positive regions, with valid annotations covering only 1.464% of the total area. This extreme sparsity underscores the difficulty of the task and the necessity of adaptive loss functions tailored for such conditions.

Preprocessing and Filtering

To enhance the relevance of training data, the dataset was divided into 560×560 patches. However, the irregular shape of the Houston area resulted in many patches containing substantial amounts of irrelevant background. To mitigate this, patches where more than 70% of the area fell outside the

[2] https://qgis.org/

Houston boundary were excluded, as shown in Fig. 3. This preprocessing step reduces noise and ensures that the model is trained on meaningful regions, improving overall robustness.

Implementation Details

Our experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 24GB of VRAM, utilizing PyTorch as the deep learning framework. The system operates on Ubuntu 24.04 LTS, with CUDA 12.0 and cuDNN 8.7.0. We use U-Net (Ronneberger et al., 2015) as the primary model for segmentation tasks. Stochastic Gradient Descent (SGD) with a momentum of 0.9 serves as the optimizer, along with a Cosine Annealing Learning Rate scheduler for a gradual reduction of the learning rate during training. The batch size is set to 16. Data augmentation is performed using random rotations. The training process spans 150 epochs, with each dataset split into training, validation, and testing sets in a 7:1:2 ratio. A fixed random seed of 42 is used to ensure reproducibility of results. Hyperparameters are optimized using the Optuna framework [3] to ensure effective tuning and performance enhancement. The model is saved based on the best IoU performance on the validation set, while final evaluation metrics are computed over the entire test set.

C.1 Ablation Study on Adaptive Weighting Strategies

We conduct a systematic analysis of different weighting strategies in our loss function formulation to understand the contribution of each component. As shown in Tab. 1, we investigate several variants: the standard $\mathcal{L}_{\text{Dice}}$ ($\lambda = 1$), $\mathcal{L}_{\text{Flipped Dice}}$ ($\lambda = 0$), and two combined versions with different weighting schemes ($\lambda = 0.5$ and $\lambda = \rho$).

The results reveal a clear performance progression. While $\mathcal{L}_{\text{Flipped Dice}}$ alone performs poorly (IoU: 0.6130, F1: 0.7412) due to its overemphasis on background regions, a naive combination with equal weights ($\lambda = 0.5$) shows limited improvement (IoU: 0.7037, F1: 0.8235). Using the proportion of positive pixels as the weight ($\lambda = \rho$) yields better results (IoU: 0.7209, F1: 0.8348), outperforming both individual losses. However, our proposed $\mathcal{L}_{\text{Adaptive Dice}}$ with the adaptive weighting function $g(\rho)$ achieves the best performance (IoU: 0.7563, F1: 0.8583), demonstrating that a more sophisticated weighting mechanism is crucial for handling the varying distributions of positive pixels in wetland delineation tasks.

Loss Function	IoU↑	F1↑
$\mathcal{L}_{\text{Dice}}$ ($\lambda = 1$)	0.7144	0.8118
$\mathcal{L}_{\text{Flipped Dice}}$ ($\lambda = 0$)	0.6130	0.7412
$\mathcal{L}_{\text{Combined}}$ ($\lambda = 0.5$)	0.7037	0.8235
$\mathcal{L}_{\text{Combined}}$ ($\lambda = \rho$)	0.7209	0.8348
$\mathcal{L}_{\text{Adaptive Dice}}$ ($g(\rho)$)	**0.7563**	**0.8583**

Table 2: Ablation study of different weighting strategies in our loss function on the Houston Wetland Delineation dataset.

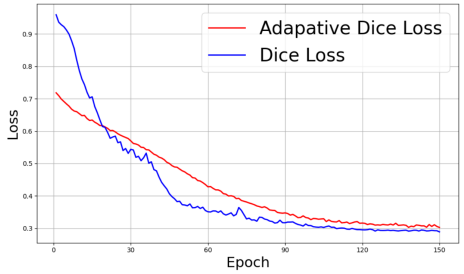


Figure 4: Comparison of $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{Adaptive Dice}}$ on the Houston Wetland Delineation dataset. $\mathcal{L}_{\text{Adaptive Dice}}$ exhibits smoother behavior, enhancing stability during training.

[3] https://optuna.org

C.2 Parameter Sensitivity Experiment

To assess the sensitivity of our model to the parameters a and k, we conducted a series of experiments varying these parameters while keeping all other variables constant. The IoU and F1 scores for different combinations of a and k are presented separately in Tabs. 3–4.

Effect of Parameter a. From Tabs. 3–4, we observe that increasing the value of a generally leads to higher and more stable IoU and F1 scores across different k values. Specifically:

- At $a = 0.0001$, the IoU and F1 scores fluctuate significantly with varying k, indicating sensitivity to parameter k at low a values.
- As a increases to 0.001, the performance improves, especially for higher k values ($k \geq 3000$), where IoU and F1 scores reach approximately 0.7570 and 0.8587, respectively.
- For $a \geq 0.002$, the IoU and F1 scores become consistently high (around 0.755 and 0.858) across all k values tested, demonstrating that higher values of a mitigate the sensitivity to k.

Effect of Parameter k. The impact of parameter k on performance varies depending on the value of a:

- At lower a values (e.g., $a = 0.0001$ and $a = 0.0005$), the model's performance is more sensitive to changes in k, with IoU and F1 scores showing noticeable fluctuations.
- When a is increased to 0.001, the performance becomes more stable for higher k values, but some variability remains at lower k.
- At $a \geq 0.002$, the influence of k on performance is minimal, indicating that a higher a value effectively reduces the model's sensitivity to k.

Optimal Parameter Settings. The parameter sensitivity analysis indicates that a is a crucial parameter influencing the model's performance and stability. Higher values of a lead to improved and consistent results, effectively mitigating the sensitivity to k. However, considering that $a \times k$ will be exponentiated, selecting excessively large values of k may cause numerical instability or computational challenges. Therefore, we recommend selecting $a \geq 0.002$ for optimal performance, while choosing k within a reasonable range to balance performance and computational stability.

$a \backslash k$	**500**	**1000**	**3000**	**5000**	**10000**
0.0001	0.6213	0.7031	0.6219	0.6879	0.6095
0.0005	0.7201	0.6159	0.7552	0.6615	0.7559
0.001	0.7040	0.6811	0.7570	0.7570	0.7569
0.002	0.7554	0.7558	0.7556	0.7559	0.7558
0.003	0.7552	0.7558	0.7558	0.7557	0.7557

Table 3: IoU scores for different k and a values in the parameter sensitivity experiment.

$a \backslash k$	**500**	**1000**	**3000**	**5000**	**10000**
0.0001	0.7629	0.8223	0.7613	0.8114	0.7517
0.0005	0.8344	0.7565	0.8575	0.7917	0.8579
0.001	0.8230	0.8070	0.8587	0.8587	0.8586
0.002	0.8576	0.8580	0.8578	0.8580	0.8580
0.003	0.8575	0.8579	0.8578	0.8578	0.8578

Table 4: F1 scores for different k and a values in the parameter sensitivity experiment.

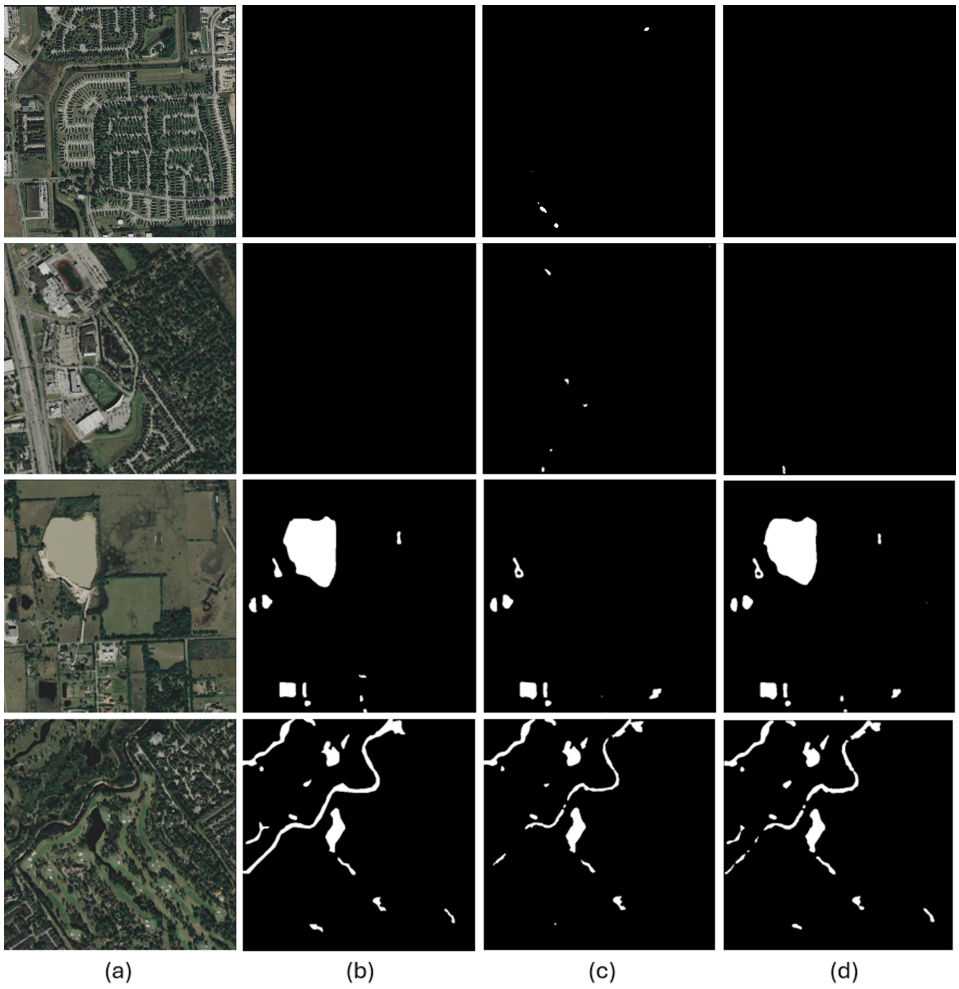


Figure 5: Visualization of segmentation results on the Houston Wetland Delineation dataset. (a) Remote sensing inputs in RGB, (b) Ground truth labels, (c) Segmentation results with Dice Loss, (d) Segmentation results with Adaptive Dice Loss. Adaptive Dice Loss demonstrates better handling of sparse water regions and improved delineation of boundaries.

C.3 Loss Analysis

We analyze the training dynamics of $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{Adaptive Dice}}$ on the Houston Wetland Delineation dataset. As shown in Fig. 4, $\mathcal{L}_{\text{Adaptive Dice}}$ demonstrates notably smoother behavior compared to standard $\mathcal{L}_{\text{Dice}}$. The standard Dice Loss exhibits significant fluctuations, particularly in positive-sparse scenarios where gradient instability occurs. In contrast, our Adaptive Dice Loss maintains more stable gradients through its dynamic weighting mechanism, leading to more consistent optimization during training.

This improved stability directly translates to better segmentation performance, as demonstrated in our case studies below. The smooth training behavior of $\mathcal{L}_{\text{Adaptive Dice}}$ enables the model to better handle both false positive and false negative cases, which are common challenges in wetland delineation tasks.

C.4 Case Studies

We present four representative cases from the Houston Wetland Delineation dataset to demonstrate the advantages of our Adaptive Dice Loss over standard Dice Loss, as shown in Fig. 5. The first two cases highlight how standard Dice Loss can lead to false positives in all-background regions. In these

instances, Dice Loss fails to provide effective gradients, resulting in spurious wetland predictions. Our Adaptive Dice Loss successfully suppresses these false detections by leveraging the flipped formulation when the positive pixel proportion is low.

The latter two cases illustrate the opposite scenario where Dice Loss misses actual wetland regions, producing false negatives. This typically occurs in positive-sparse scenarios where the gradient variance of Dice Loss becomes unstable. Our Adaptive Dice Loss maintains more stable gradients through its dynamic weighting mechanism, enabling better detection of these sparse wetland regions while preserving precise boundary delineation.