DA-SEGFORMER: DAMAGE-AWARE SEMANTIC SEGMENTATION FOR FINE-GRAINED DISASTER ASSESSMENT

Kevin Zhu, William Tang, Raphael Hay Tene, Zesheng Liu, Nhut Le, and Maryam Rahnemoonfar
Bina Labs, Lehigh University
Bethlehem, PA, USA

***Abstract*—Rapid and accurate damage assessment following natural disasters is critical for effective emergency response. However, identifying fine-grained damage levels (e.g., distinguishing minor from major roof damage) in UAV imagery remains challenging due to the degradation of texture cues during resizing and extreme class imbalance. In this work, we propose DA-SegFormer, a damage-aware adaptation of the SegFormer architecture optimized for high-resolution disaster imagery. Our method introduces a Class-Aware Sampling strategy to guarantee exposure to rare damage features, and it integrates Online Hard Example Mining (OHEM) with Dice Loss to dynamically focus on underrepresented classes. In addition, we employ a resolution-preserving inference protocol that maintains native texture details. We evaluate our method on the RescueNet dataset, achieving a Mean IoU of 74.67% and outperforming the baseline SegFormer by 4.37%. Notably, our improvements yield double-digit gains in critical damage classes: Minor Damage (+13.9%) and Major Damage (+12.5%).**

***Index Terms*—Semantic Segmentation, RescueNet, SegFormer, OHEM, Disaster Response, Remote Sensing.**

I. Introduction

Natural disasters such as hurricanes, floods, and earthquakes cause significant human and economic losses worldwide. While the frequency of these events continues to rise, historical data underscores the severity of the threat; for instance, in 2025 alone, the United States experienced 23 natural disasters costing approximately 115 billion dollars [1]. A critical step in minimizing these losses is rapid and accurate damage assessment, which enables rescue teams to allocate resources efficiently and prioritize areas requiring immediate attention. However, traditional manual assessment methods involving field supervision and damage reports are time-consuming and often impossible in heavily affected areas [2].

Semantic segmentation, defined as the task of assigning a class label to every pixel in an image, has emerged as a powerful tool for automated damage assessment from aerial imagery. Unlike image-level classification, semantic segmentation provides precise spatial localization of damage, enabling responders to identify exactly which structures are affected and to what degree. UAV (Unmanned Aerial Vehicle) imagery is particularly valuable for this task, as it provides higher resolution than satellite imagery and can capture detailed damage patterns such as missing shingles, debris distribution, and structural collapse [2, 6].

The RescueNet dataset [2] provides a comprehensive benchmark for semantic segmentation in post-disaster scenarios. Collected after Hurricane Michael using DJI Mavic Pro quadcopters at 200 feet above ground level, the dataset contains approximately 2000 high-resolution images (3000×4000 pixels) with pixel-level annotations for 10 classes. Buildings are annotated at four damage levels: No Damage, Minor Damage, Major Damage, and Total Destruction. Following [2], these levels are defined as: No Damage (building unharmed), Minor Damage (parts damaged but coverable with blue tarp), Major Damage (significant structural damage requiring extensive repairs), and Total Destruction (complete failure of two or more major structural components).

Prior work on RescueNet [2] identified two fundamental challenges. First, differentiating between damage levels is extremely difficult from a top-down aerial view, as the visual differences between Minor and Major damage are often subtle texture variations. Second, the dataset exhibits severe class imbalance: background pixels constitute 52.51% of the data, while critical damage classes like Major Damage (1.68%) and Total Destruction (1.44%) are significantly underrepresented.

To address these challenges, we propose DA-SegFormer, a damage-aware adaptation of the SegFormer architecture [4] specifically designed for fine-grained disaster damage assessment. Our contributions are threefold: 1) We integrate Online Hard Example Mining (OHEM) with Dice Loss to dynamically focus on difficult pixels. 2) We introduce a Class-Aware Sampling strategy combined with a native-resolution inference protocol to preserve texture details. 3) We demonstrate significant improvements on RescueNet, with double-digit gains on Minor and Major Damage classes.

II. Related Work

A. Semantic Segmentation for Disaster Assessment

Deep learning has transformed automated damage assessment from aerial and satellite imagery. Rahnemoonfar et al. [2] introduced the RescueNet dataset and evaluated CNN-based architectures including PSPNet [11], DeepLabv3+ [12], and ENet [13] for comprehensive scene segmentation. Their experiments demonstrated that pyramid pooling approaches capture global context more effectively than encoder-decoder architectures for disaster imagery, achieving 79.43% mIoU

with PSPNet. However, they also identified persistent challenges in distinguishing between intermediate damage levels (Minor vs. Major), where all models showed significantly lower performance.

Subsequent work explored attention mechanisms for this domain. Chowdhury and Rahnemoonfar [5] applied self-attention methods to UAV imagery for damage assessment, demonstrating improved feature extraction for complex disaster scenes. Safavi et al. [6] conducted a comparative study between real-time and non-real-time segmentation models on the FloodNet dataset [3], showing that while lightweight models like UNet-MobileNetV3 achieve reasonable accuracy (59.3% mIoU), non-real-time models like PSPNet (79.7% mIoU) significantly outperform them.

Recent benchmarking on FloodNet [7] demonstrated that transformer-based architectures, specifically SegFormer, outperform CNN-based methods for post-disaster aerial imagery. This finding motivates our adoption of SegFormer as the backbone architecture, while addressing its limitations through targeted algorithmic improvements.

B. Handling Class Imbalance

Class imbalance is pervasive in remote sensing segmentation, where background and common classes dominate the pixel distribution. Standard Cross-Entropy loss causes models to converge toward predicting majority classes, ignoring rare but critical categories [10].

Online Hard Example Mining (OHEM), originally proposed for object detection [8], addresses this by dynamically selecting high-loss examples during training. Rather than treating all pixels equally, OHEM focuses gradient updates on the most difficult samples, typically class boundaries and minority class pixels. Dice Loss [9], derived from the Dice coefficient, provides complementary benefits by optimizing for region overlap on a per-class basis, giving equal weight to rare and common classes. Recent studies [10] have shown that combining region-based losses with hard example mining improves minority class recall in segmentation tasks.

III. Methodology

A. SegFormer Architecture Details

We adopt SegFormer [4] as our backbone architecture. Unlike standard Vision Transformers (ViT) that generate single-resolution feature maps, SegFormer utilizes a hierarchical *Mix Transformer* (MiT) encoder. This hierarchy is crucial for damage assessment as it captures both high-resolution coarse features (essential for locating small debris) and low-resolution fine features (essential for semantic context).

The MiT encoder generates multi-scale features through four stages, producing feature maps at resolutions $\{1/4, 1/8, 1/16, 1/32\}$ of the input image $H \times W$. A key innovation in MiT is the *Overlapped Patch Merging* process. Standard ViT uses non-overlapping patch embeddings, which often results in a loss of local continuity around patch boundaries. SegFormer employs a convolution-based patch merging with an overlap, formalized as a convolution with

TABLE I
RescueNet Class Distribution. Damage classes constitute less than 8% of total pixels combined.

Class	Freq (%)	Class	Freq (%)
Background	52.51	Major Damage	1.68
Tree	22.05	Road-Blocked	1.59
Water	8.13	Total Destruction	1.44
Minor Damage	2.63	Pool	0.06

kernel size $K = 7$, stride $S = 4$, and padding $P = 3$. This preserves the local continuity of building edges and road networks, which are critical in the RescueNet dataset.

Furthermore, the encoder replaces fixed Positional Encodings (PE) with *Mix-FFN*, a convolutional feed-forward network. This allows the model to handle variable input resolutions during inference without performance degradation, supporting our resolution-preserving inference strategy.

The lightweight *All-MLP* decoder aggregates these multi-scale features. Features F_i from each stage i are first passed through an MLP layer to unify the channel dimension. Then, they are upsampled to the $1/4$ resolution and concatenated:

$$F_{\text{concat}} = \text{Concat}(\forall i : \text{Upsample}(\text{MLP}(F_i))) \tag{1}$$

Finally, a prediction layer projects the concatenated features to the semantic segmentation mask $\in \mathbb{R}^{H/4 \times W/4 \times N_{\text{cls}}}$. We utilize the SegFormer-B4 variant, which features a deeper Stage 3 (27 transformer blocks) compared to lighter variants, providing the necessary capacity to model complex post-disaster scenes.

B. Online Hard Example Mining (OHEM)

To address the severe class imbalance in RescueNet (Table I), we integrate Online Hard Example Mining into our training pipeline. OHEM dynamically identifies and prioritizes the most difficult pixels during each forward pass.

During each forward pass, we compute the pixel-wise Cross-Entropy loss for all N pixels in the batch. We then rank pixels by loss value and select the top k hardest pixels (we use $k = 100,000$) to form the hard example set $\mathcal{K}$. The OHEM loss is computed only over this subset:

$$\mathcal{L}_{\text{OHEM}} = \frac{1}{k} \sum_{p \in \mathcal{K}} -\log(P(y_p|x_p)) \tag{2}$$

This approach forces the optimizer to focus on decision boundaries where the model is uncertain, such as the transitions between damage levels (e.g., Minor to Major) and edges between buildings and background. Pixels that are easily classified (e.g., clear background regions) contribute minimally to gradient updates.

C. Dice Loss

While OHEM addresses pixel-level difficulty, Dice Loss handles class imbalance at the region level. The Dice coefficient measures overlap between predicted and ground truth regions:

$$\text{Dice}_c = \frac{2 \sum_i p_{i,c} \cdot g_{i,c}}{\sum_i p_{i,c} + \sum_i g_{i,c}} \tag{3}$$

where $p_{i,c}$ is the predicted probability for class c at pixel i, and $g_{i,c}$ is the ground truth. Dice Loss is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{C}\sum_{c=1}^{C} \text{Dice}_c \tag{4}$$

Importantly, Dice Loss computes overlap per class then averages across classes, giving equal weight to Minor Damage (2.63% of pixels) and Background (52.51%). This prevents the model from ignoring rare classes to minimize overall loss.

Our total loss function combines both components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{OHEM}} \tag{5}$$

D. Class-Aware Sampling Strategy

Standard random cropping is inefficient for disaster imagery where damage pixels constitute less than 8% of the data; the vast majority of random crops contain only background or vegetation. To counter this, we implement a Class-Aware Sampling strategy.

Instead of uniform sampling, we enforce a bias where 50% of training crops are centered on pixels belonging to underrepresented damage classes (Minor Damage, Major Damage, and Total Destruction). This guarantees that the model receives a consistent learning signal for critical features, ensuring that texture cues associated with damage are not overwhelmed by the dominant background classes.

E. Resolution-Preserving Inference

Prior work [7] demonstrated that fine-grained damage assessment depends on high-frequency texture cues (missing shingles, debris patterns) that are destroyed by aggressive downsampling. Standard practice resizes images to fixed dimensions (e.g., 512×512) during inference, creating a distribution shift when training uses higher resolution crops.

We address this through a resolution-preserving inference protocol. During inference, rather than resizing, we apply a sliding window approach with a patch size of 1024×1024 and a stride $S = 768$, resulting in a 25% overlap between adjacent patches. Predictions in overlapping regions are averaged uniformly to suppress boundary artifacts:

$$\mathbf{Y}_{\text{pred}} = \frac{1}{N_{ov}}\sum_{n=1}^{N_{ov}} f_\theta(x_n) \tag{6}$$

where N_{ov} is the number of overlapping predictions for a given pixel. This ensures the model never encounters resolution-degraded inputs, maintaining consistency between the training and inference distributions.

IV. Experiment Details

A. RescueNet Dataset

We evaluate on the RescueNet dataset [2], which contains 1,973 high-resolution UAV images (3000×4000 pixels) captured after Hurricane Michael. Images are annotated with 10 semantic classes: Background, Water, Building-No-Damage, Building-Minor-Damage, Building-Major-Damage, Building-Total-Destruction, Road-Clear, Road-Blocked, Tree, Pool, and Vehicle. Following the standard split, we use 1,591 images for training, 199 for validation, and 183 for testing.

For preprocessing, we apply the Class-Aware Sampling described in Section III-D to generate 1024×1024 patches during training. Data augmentation includes random horizontal and vertical flips, rotation (±15°), and photometric distortion (brightness, contrast, saturation adjustments) to improve generalization.

B. Evaluation Metrics

To quantitatively evaluate the performance of our proposed method, we utilize the Intersection over Union (IoU) and the Mean Intersection over Union (mIoU). Given the severe class imbalance in the RescueNet dataset, pixel accuracy is an insufficient metric, as a model predicting only "Background" would still achieve high accuracy.

The IoU for a specific class c is defined as the ratio of true positives to the sum of true positives, false positives, and false negatives:

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{7}$$

where TP_c, FP_c, and FN_c represent the number of true positive, false positive, and false negative pixels for class c, respectively.

The Mean IoU (mIoU) is calculated by averaging the IoU across all C classes (here $C = 11$):

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \text{IoU}_c \tag{8}$$

We report pixel-level IoU scores for all classes to ensure transparency regarding the model's performance on underrepresented damage categories.

C. Training Details

We implement our method in PyTorch using the MMSegmentation framework. The SegFormer-B4 encoder is initialized with weights pretrained on ADE20K [15]. Training was conducted on NVIDIA A100 GPUs provided by Jetstream2 [16] for 300 epochs with batch size 2.

We use the AdamW optimizer [14] with an initial learning rate of 6×10^{-5} and a Cosine Annealing scheduler. Weight decay is set to 0.01. For OHEM, we select the top $k = 100,000$ hardest pixels per batch. The Dice Loss and OHEM Loss are combined with equal weighting.

For comparison, we train a baseline SegFormer-B4 with standard Cross-Entropy loss and input resizing to 512×512, representing the conventional approach.

V. Results

We compare our improved SegFormer against the baseline SegFormer-B4 trained with standard Cross-Entropy loss and 512×512 resizing. All models use identical encoder initialization and training epochs.

TABLE II
PER-CLASS IOU COMPARISON. DA-SEGFORMER OUTPERFORMS THE BASELINE, WITH CRITICAL GAINS IN MINOR (+13.9%) AND MAJOR (+12.5%) DAMAGE.

Method	Bkgd	Water	No Dmg	Minor	Major	Destr.	Road	Blkd	Tree	Pool	Veh.	mIoU
SegFormer	89.50	78.30	69.80	58.10	59.60	59.00	77.80	55.40	85.20	76.70	63.20	70.30
DA-SegFormer	**91.20**	**80.10**	**70.30**	**72.00**	**72.10**	**60.70**	**83.50**	**58.20**	**87.30**	**88.20**	**67.10**	**74.67**
Gain	**+1.70**	**+1.80**	**+0.50**	**+13.90**	**+12.50**	**+1.70**	**+5.70**	**+2.80**	**+2.10**	**+11.50**	**+3.90**	**+4.37**

Fig. 1. Qualitative comparison on RescueNet test images. Columns: (a) Original Image, (b) Ground Truth, (c) SegFormer Baseline, (d) DA-SegFormer, (e) Overlay. Our method produces sharper building boundaries and more accurate damage level classification.

A. Quantitative Results

Table II presents per-class IoU and mean IoU for both methods. DA-SegFormer achieves 74.67% mIoU, a 4.37% improvement over the baseline SegFormer (70.30%). The most significant improvements occur in the damage severity classes critical for emergency response: Minor Damage improves by 13.90% (58.10% to 72.00%) and Major Damage by 12.50% (59.60% to 72.10%). These classes require distinguishing subtle texture differences visible only at high resolution; this is precisely what our resolution-preserving inference and hard example mining address. The Pool class also shows substantial improvement (+11.50%), as pools are small objects whose features are degraded by downsampling. Prior work [2] noted that distinguishing Minor and Major damage levels was the most difficult task; our OHEM-based approach specifically targets these difficult decision boundaries.

B. Qualitative Results

Figure 1 presents qualitative comparisons on test images showing the original UAV image, ground truth, baseline SegFormer prediction, DA-SegFormer prediction, and our prediction overlaid on the original. The baseline SegFormer frequently confuses Minor and Major damage levels, producing inconsistent predictions within single buildings. In contrast, DA-SegFormer produces more coherent, building-level predictions with sharper boundaries. The baseline also generates fragmented predictions on small structures due to texture degradation from downsampling, while our resolution-preserving inference maintains the sharp features necessary for accurate segmentation.

VI. CONCLUSION

We presented DA-SegFormer, a damage-aware adaptation of the SegFormer architecture for fine-grained disaster damage assessment. By integrating Online Hard Example Mining with Dice Loss, we address the severe class imbalance in disaster imagery datasets where damage pixels constitute less than 8% of the data. Our training-aligned inference strategy preserves native resolution texture features critical for distinguishing subtle damage levels.

Evaluated on the RescueNet dataset, DA-SegFormer achieves 74.67% mIoU, outperforming baseline SegFormer by 4.37%. Notably, we achieve double-digit improvements on the critical damage severity classes: Minor Damage (+13.9%) and Major Damage (+12.5%). These are precisely the classes that prior work [2] identified as most challenging, where visual differences are subtle and class imbalance is severe.

Our results confirm that for fine-grained disaster analysis, algorithmic handling of class imbalance and preservation of texture resolution are as important as the choice of network architecture. Future work will explore extending these techniques to multi-temporal damage assessment, comparing pre- and post-disaster imagery for change detection.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under grants #2423211 and #2401942, the Consortium for Enhancing Resilience and Catastrophe Modeling, and NSF ACCESS resources. This work used Jetstream2 at Indiana University through allocation CIS251039 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Climate Central, "2025 in Review: U.S. Billion-Dollar Disasters," *Climate Central*, Jan. 8, 2026. [Online]. Available: https://www.climatecentral.org/climate-matters/2025-in-review

[2] M. Rahnemoonfar, T. Chowdhury, R. Murphy, and O. Fernandes, "Comprehensive semantic segmentation on high resolution UAV imagery for natural disaster damage assessment," in *Proc. IEEE Int. Conf. Big Data*, Atlanta, GA, USA, 2020, pp. 3726–3735.

[3] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, M. Varshney, M. Yari, and R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post-flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NeurIPS*, 2021.

[5] T. Chowdhury and M. Rahnemoonfar, "Attention based semantic segmentation on UAV dataset for natural disaster damage assessment," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Brussels, Belgium, 2021, pp. 2325–2328.

[6] F. Safavi, T. Chowdhury, and M. Rahnemoonfar, "Comparative study between real-time and non-real-time segmentation models on flooding events," in *Proc. IEEE Int. Conf. Big Data*, Orlando, FL, USA, 2021, pp. 4199–4207.

[7] M. Rahnemoonfar *et al.*, "Real-time semantic segmentation of aerial imagery for emergency response," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4–20, 2022.

[8] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE CVPR*, 2016, pp. 761–769.

[9] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571.

[10] Z. Xu *et al.*, "A comparative study of loss functions for road segmentation in remote sensing imagery," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 116, 2023.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE CVPR*, 2017, pp. 2881–2890.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.

[13] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[15] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE CVPR*, 2017, pp. 633–641.

[16] D. Hancock *et al.*, "Jetstream2: Accelerating cloud computing via Jetstream," in *Proc. PEARC*, 2021.