HAUNet3+: Enhancing UNet3+ with Hybrid Attention for SAMRefined Sinkhole Image Segmentation

Ayon Dey[1], Anav Katwal[1], Abdullah Naeem[1], Padam Jung Thapa[2], Kendall N. Niles[3], Steve Sloan[3], and Md Tamjidul Hoque[1,*]
[1]Department of Computer Science, Louisiana State University New Orleans, New Orleans, LA, USA
[2]University of Louisiana at Lafayette, Lafayette, Louisiana, USA
[3]U.S. Army Engineer Research and Development Center, Vicksburg, MS, USA
Emails: {adey, akatwal, aneem, thoque}@lsuneworleans.edu, tpadamjung@gmail.com, kendall.n.niles@erdc.dren.mil, steven.d.sloan@erdc.dren.mil
[*]Corresponding author: thoque@lsuneworleans.edu

Abstract

Semantic segmentation of sinkholes in levee systems from unmanned aerial vehicle (UAV) imagery presents a distinct and underexplored challenge: target structures are spatially sparse, morphologically irregular, and visually ambiguous relative to the surrounding terrain, with pronounced class imbalance. While UNet3+ offers representationally rich multi-scale feature aggregation through full-scale skip-connections, its indiscriminate fusion of features across all spatial scales risks incorporating redundant or task-irrelevant responses during decoding—a limitation that is particularly consequential for small, sparsely distributed targets. To address this, we propose HAUNet3+, a hybrid attention-enhanced variant of UNet3+ that introduces independent channel-spatial attention modules applied to each incoming skip pathway individually, prior to multi-scale concatenation. This per-pathway pre-aggregation strategy enables scale-specific noise suppression before feature fusion, allowing the network to selectively amplify the most task-relevant representations at each decoder stage. The architecture further employs deep supervision across all decoder outputs, using an optimized weighted loss that combines Binary Cross-Entropy and Dice objectives, tuned via systematic hyperparameter search to balance per-pixel accuracy against region-level overlap under class imbalance. Evaluated on the SAMRefined Sinkhole dataset under rigorous 10-fold cross-validation, HAUNet3+ achieves a mean Intersection over Union (mIoU) of 89.1%, pixel accuracy of 95.56%, balanced accuracy of 94.49%, AUC of 0.990, and Average Precision of 0.9750, outperforming all evaluated baselines, including UNet3+ with deep supervision, by 8.9 percentage points in mIoU. Ablation experiments confirm that pre-concatenation attention placement is superior to post-concatenation attention, and that the combination of channel and spatial attention is necessary to realize the full performance benefit. These results establish HAUNet3+ as an effective and principled architecture for automated sinkhole detection in complex, high-resolution UAV imagery of levee infrastructure.

Index Terms

Levee monitoring, semantic segmentation, sinkhole detection, UAV imagery, UNet3+, hybrid attention, deep supervision.

I. Introduction

Levees are among the most critical flood defense structures in modern civil infrastructure, safeguarding communities, agricultural land, and industrial assets from the destructive consequences of extreme hydrological events. Yet despite their importance, a significant portion of levee systems worldwide are aging, poorly documented, and infrequently inspected. One of the most insidious failure mechanisms in earthen levees is the formation of sinkholes, localized surface depressions that often serve as visible indicators of far more dangerous subsurface processes such as internal erosion or piping. Left undetected,

these defects can trigger rapid and catastrophic levee breaches with little warning [1], [2]. Traditional inspection workflows, which rely predominantly on manual field surveys and the judgment of experienced engineers, are not only time-consuming and costly but are also difficult to apply consistently across the vast networks of levees in flood-prone regions [3], [4].

To address these challenges, recent advances in remote sensing and deep learning have opened new avenues for automated sinkhole detection from aerial and satellite imagery. Progress in this domain, however, has been hampered by the scarcity of high-quality, finely annotated training datasets. Coarse or imprecise segmentation masks- a common artifact of large-scale manual labeling efforts- limit the ability of learned models to capture the precise boundaries and morphological characteristics of sinkhole features. To overcome this limitation, we construct the SAMRefined Sinkhole Dataset by systematically refining coarse sinkhole annotations using SAMRefiner [5], a prompt-based refinement pipeline built on the Segment Anything Model (SAM) [6]. The resulting dataset, which has been released publicly [7], provides high-fidelity semantic-level segmentation masks for levee sinkholes and serves as the primary benchmark for the detection and segmentation experiments presented in this work.

The increasing accessibility of unmanned aerial vehicle (UAV) platforms has opened new possibilities for infrastructure monitoring. UAVs can collect high-resolution optical imagery over large areas quickly and at relatively low cost, providing a level of spatial coverage and detail that was previously impractical [8]–[10]. However, the sheer volume of imagery generated by such surveys makes manual analysis a bottleneck. Translating raw aerial data into reliable, actionable defect maps demands automated analysis pipelines, a challenge for which deep learning methods are increasingly well-suited. Semantic segmentation, which assigns a class label to every pixel in an image, has proven particularly effective for detecting and delineating defects in civil infrastructure such as cracks in roads and bridges, spalling in concrete structures, and surface anomalies in earthen embankments [11]–[13].

The dominant paradigm for semantic segmentation is the encoder-decoder architecture, in which a contracting encoder extracts hierarchical feature representations and a symmetric decoder reconstructs dense spatial predictions, often assisted by skip-connections that preserve fine-grained spatial detail [14], [15]. Subsequent architectures such as UNet++ [16] and UNet3+ [17] refined this framework through progressively richer skip-connection designs, the latter connecting every decoder stage to all encoder and preceding decoder stages simultaneously to enable full-scale multi-scale feature aggregation. While this yields representationally rich feature maps, the unconstrained aggregation of features across all scales risks incorporating redundant or irrelevant responses into the decoding process, a concern that is amplified when target structures are small or visually ambiguous relative to the surrounding background.

Sinkhole detection from UAV imagery presents precisely this challenge. In typical levee survey imagery, sinkholes occupy a small fraction of the total scene area, often less than 1–3% of image pixels are irregularly shaped, and exhibit surface textures and color profiles that can be difficult to distinguish from natural terrain depressions, vegetation cover, or shadow artifacts [18], [19]. These properties make accurate segmentation acutely sensitive to the quality and selectivity of the feature representations used at each decoder stage, motivating architectures that can actively filter scale-specific noise rather than passively accumulate it.

Attention mechanisms offer a principled solution to this problem. By explicitly modeling which features are most informative - both across channels and across spatial locations - attention modules allow networks to concentrate their representational capacity on the most task-relevant responses. Channel-wise attention, as formalized by the Squeeze-and-Excitation Network (SENet) [20], and the combined channel-spatial formulation of the Convolutional Block Attention Module (CBAM) [21] have both demonstrated consistent improvements in classification, detection, and segmentation tasks, including when incorporated into encoder-decoder architectures [22], [23].

Motivated by these observations, this paper introduces **HAUNet3+**, a hybrid attention-enhanced variant of UNet3+ for binary semantic segmentation of levee sinkholes from UAV imagery. The principal contributions of this work are as follows:

- We propose a per-pathway pre-aggregation hybrid attention strategy in which independent channel-spatial attention modules are applied to each incoming skip connection prior to multi-scale fusion, enabling scale-specific noise suppression before feature aggregation.
- We show through ablation experiments that this pre-aggregation attention placement yields better segmentation performance than post-aggregation attention on a class-imbalanced levee sinkhole dataset.
- We provide a systematic ablation study isolating the effects of channel attention, spatial attention, deep supervision, and attention placement, offering reproducible insight into the design choices that influence small-object segmentation performance on this dataset.
- We propose a data pipeline that leverages SAMRefiner [5] to refine human-annotated segmentation masks of sinkhole images followed by augmentations to enhance model generalization on segmentation task.

Together, these design choices allow HAUNet3+ to better integrate fine-grained spatial localization with broader semantic context for automated sinkhole detection in complex levee environments.

II. Related Work

A. Encoder-Decoder Architectures for Segmentation

Encoder-decoder networks have established themselves as the dominant framework for semantic segmentation, particularly in scenarios where annotated data is scarce and precise spatial localization is required. U-Net [14] laid the groundwork for this family of models by introducing symmetric skip-connections between encoder and decoder stages, allowing high-resolution spatial features to bypass the bottleneck and directly inform the reconstruction process. This design proved remarkably effective in medical imaging contexts and has since been widely adapted for remote sensing and infrastructure inspection applications [24], [25].

UNet++ [16], [26] revisited the skip-connection design, arguing that the direct connections in the original U-Net create an unnecessarily large semantic gap between encoder and decoder at each scale. To address this, UNet++ introduced a series of nested, densely connected intermediate nodes along each skip pathway, enabling progressive feature fusion and reducing the representational gap through deep supervision at multiple decoder outputs. While effective, this added considerable architectural complexity and parameter count without fundamentally reconsidering which scales should contribute to each decoder stage.

UNet3+ [17] proposed a more holistic approach to multi-scale feature fusion. Rather than connecting each decoder stage only to its corresponding encoder stage, UNet3+ connects each decoder stage to all encoder stages and to all preceding decoder stages via full-scale skip-connections. This allows the network to simultaneously leverage both fine-grained spatial detail from shallow encoder layers and abstract semantic representations from deeper ones, with deep supervision applied across multiple decoder stages to reinforce hierarchical feature learning [27]. Despite the representational richness this offers, indiscriminately aggregating features across all scales can introduce redundant or irrelevant information into the decoding process. This limitation is particularly consequential when target structures are small and sparsely distributed, as the relevant signal constitutes only a minor component of the aggregated multi-scale feature map.

More recent efforts have sought to address the capacity limitations of purely convolutional encoders by incorporating transformer-based self-attention. TransUNet [28] augments a ResNet encoder with a ViT module to capture long-range dependencies, while Swin-UNet [29] replaces the entire encoder with a hierarchical Swin Transformer [30]. SegFormer [31] adopts a lightweight all-MLP decoder paired with a mix-transformer encoder, achieving competitive performance with a reduced parameter budget. While these architectures improve global context modeling, their gains are often most pronounced on dense, large-object segmentation benchmarks; for small, sparsely distributed targets in high-resolution UAV imagery, the inductive biases of convolutional encoders combined with targeted attention mechanisms frequently remain competitive or superior [32], [33].

More recent hybrid architectures have continued to push this direction. Swin–UNet++ [34] redesigns the decoder of Swin-UNet with nested dense skip connections inspired by UNet++, improving multi-scale feature reuse for fine-grained morphological segmentation. UNet3+STE [35] integrates a hybrid Swin Transformer and EfficientNet encoder within the UNet3+ full-scale skip-connection framework, demonstrating strong performance on remote sensing cloud segmentation in very-high-resolution satellite imagery with deep supervision. MSA-UNet3+ [36] augments UNet3+ with a Multi-Scale Dilated Bottleneck, a Contextual Attention Fusion Module, and a Supervised Prototypical Contrastive Loss to mitigate class imbalance in coronary vessel segmentation. Together, these architectures illustrate the active and diverse landscape of transformer-enhanced UNet3+ variants that HAUNet3+ is contextualized within.

B. Attention Mechanisms in Segmentation Networks

The integration of attention mechanisms into segmentation architectures has been an active research direction over the past several years. Attention gates, introduced in the context of U-Net by Oktay et al. [22], enable the network to suppress activations from irrelevant spatial regions during skip connection aggregation, improving sensitivity to small target structures with minimal computational overhead. Schlemper et al. [23] extended this concept and provided a more thorough theoretical analysis of how soft attention interacts with encoder feature maps during decoding.

Channel-wise attention, formalized by SENet [20], models inter-channel dependencies by using global average pooling to produce compact channel descriptors, which are then passed through a small fully connected network to generate recalibration weights. While SENet captures *what* features are important across channels, it does not explicitly address spatial selectivity, leaving open the question of *where* within the feature map attention should be directed [37].

CBAM [21] combined both perspectives into a single lightweight module, first applying channel attention using both average and max-pooled features through a shared MLP, then applying spatial attention via a convolutional layer operating on channel-pooled descriptors. The sequential application of these two complementary attention types has demonstrated consistent improvements across classification, detection, and segmentation benchmarks. Subsequent work has further extended dual-axis attention designs, including coordinate attention [38], triplet attention [39], and polarized self-attention [40], each offering different trade-offs between expressiveness and computational cost.

Several studies have explored the incorporation of CBAM into UNet-style architectures. Xu et al. [41] proposed Ref-UNet3+, a pruned variant of UNet3+ that reduces redundant full-scale connections and applies CBAM to the fused feature maps at each decoder stage *following* multi-scale concatenation. While this improves feature selectivity at the decoder level, applying attention after aggregation means that scale-specific noise and redundancy are not suppressed before they enter the fusion process. Tie et al. [42] introduced CBAM into the encoder modules of a UNet3+ network for agricultural image segmentation, demonstrating the utility of attention in the encoder pathway, though they did not specifically address the skip-connection fusion stage. HAUNet3+ is specifically designed to close this gap.

C. Sinkhole and Surface Defect Detection in Levee Systems

Automated detection of sinkholes and surface defects in levee systems remains comparatively underexplored relative to analogous segmentation tasks in medical imaging and urban scene understanding. Early remote sensing approaches relied predominantly on LiDAR-derived digital elevation models, employing morphological analysis to identify surface anomalies [43], [44]. While effective under favorable conditions, these methods exhibit pronounced sensitivity to data resolution and consistently struggle to detect shallow or early-stage depressions, where topographic signatures remain subtle and ambiguous. In the domain of sinkhole segmentation specifically, Alshawi et al. [45] proposed an enhanced UNet architecture incorporating depth-wise separable convolutional blocks with multiscale filters, enabling the network to simultaneously capture feature representations across multiple spatial scales and thereby improving detection of sinkholes with varying sizes and morphologies. UAV-based visual inspection

has since emerged as a scalable and operationally practical alternative to conventional ground-based and LiDAR-dependent assessment methodologies. Resop et al. [10] established the feasibility of UAV photogrammetry for levee condition monitoring, demonstrating that high-resolution aerial surveys could capture surface anomalies with sufficient fidelity for structural evaluation. Complementing this work, Hasan et al. [18] applied convolutional neural networks to UAV-derived orthomosaics for automated sinkhole identification, revealing class imbalance ratios exceeding 50:1 between background and target pixels - a severe distributional asymmetry that underscores the inherent difficulty of detecting sparse geotechnical features and directly motivates the attention-based architectural design proposed in this study. More recently, Alrabayah et al. [46] demonstrated an automated deep learning pipeline for sinkhole recognition in satellite imagery of the Eastern Dead Sea, using a U-Net trained first on high-resolution drone data and subsequently fine-tuned on Pleiades Neo satellite imagery, highlighting the generalizability challenges that arise when transferring sinkhole detection models across platforms and resolutions. The broader challenge of small object segmentation in aerial imagery has been systematically investigated across several infrastructure-related domains, including vehicle detection [47], building footprint extraction [48], and surface crack segmentation [49]. A consistent finding across these studies is that standard encoder-decoder architectures exhibit substantial performance degradation when target structures occupy less than 2% of the image area, due to progressive loss of spatial detail through successive downsampling stages. Addressing this limitation requires explicit architectural mechanisms that preserve fine-grained spatial representations and selectively enhance the saliency of small, structurally significant regions- capabilities that inform the design choices central to the proposed framework.

D. Positioning of HAUNet3+

HAUNet3+ is distinguished from prior UNet3+ based attention models by both the placement and granularity of its attention mechanism. Rather than applying a single attention block to the post-concatenation feature map at each decoder stage [41], HAUNet3+ applies independent hybrid channel-spatial attention modules to each incoming skip pathway individually, prior to aggregation. This allows the network to adaptively filter each scale's contribution before it enters the fusion process, enabling more discriminative and noise-robust multi-scale representations. This design choice is motivated directly by the class-imbalance and small-object properties of levee sinkhole imagery, where low signal-to-noise separation within aggregated feature maps can impair segmentation performance. Combined with deep supervision across all decoder stages, HAUNet3+ is specifically designed to address the challenges of detecting small, visually subtle defects in complex, high-resolution UAV imagery of levee systems.

III. Dataset Description

The **SAMRefined Sinkhole dataset** consists of 1,303 UAV-based remote sensing image–mask pairs collected from sinkhole-prone levee areas. Each image is paired with a binary segmentation mask, where sinkhole pixels represent the foreground class and all remaining pixels represent the background class. The dataset provides high-resolution images and highly pixel-accurate masks that were resized to 320×320 for training efficiency. Data pre-processing included random cropping and normalization. To enhance generalization, data augmentation was applied, including horizontal and vertical flips, random brightness/contrast adjustments, grid distortion, and elastic transformations. The original human-annotated masks were generated using polygon-based annotations and subsequently refined using SAMRefiner [5] to improve boundary alignment between the sinkhole regions and the corresponding visual structures in the images. Human-annotated segmentation masks generally use polygons with a low number of vertices to cover the objects of interest. Even though they are considered the ground truth, such masks are often imperfect due to the time consuming nature of annotating large datasets. Additionally, the cost of obtaining accurate labels adds to the already expensive task of making segmentation labels. Hence, segmentation dataset masks are inaccurate due to their high cost and human imperfection, especially when the object

of interest does not have a definitive boundary. This inaccuracy leads to performance degradation in segmentation models, and the impact is greater with boundary-localized errors [50].

Considering the limited annotation time, we implemented a data annotation pipeline that generates pixel-accurate segmentation masks from human-labeled masks annotated using polygons with low vertices. We use SAMRefiner [5], a mask refinement framework that leverages SAM (Segment Anything Model) [6] for mask refinement. The framework uses a multi-prompt excavation strategy that extracts diverse input prompts like points, boxes, and masks from a coarse mask and feeds these prompts to SAM along with the image to generate refined masks with better accuracy. The human-labeled Sinkhole dataset (images and masks) was fed to the SAMRefiner framework as input, and the obtained dataset was used for training the models. Figure 1 demonstrates the difference between human annotated masks and the mask output from SAMRefiner. It significantly confines the mask boundary to the actual object's boundary in the image reducing boundary-localized errors in the ground truth.

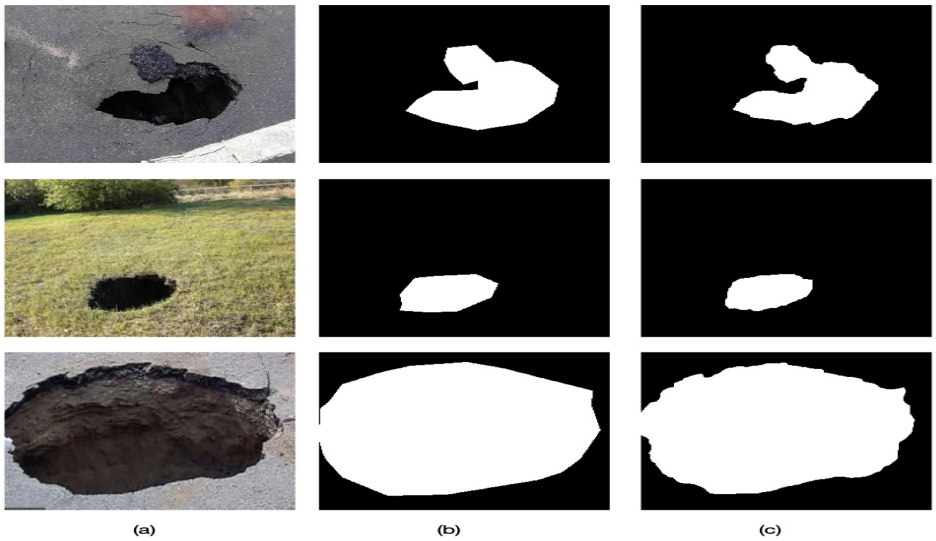

Figure 1: Examples from the SAMRefined Sinkhole dataset showing the difference between human-annotated masks and SAMRefiner-refined masks. (a) Original images, (b) human-annotated masks, and (c) SAMRefiner-refined masks.

IV. Methodology

A. Architecture

We propose HAUNet3+, a segmentation model $\mathcal{F}(x;\theta)$ that maps an input image $x \in \mathbb{R}^{H\times W\times 3}$ to a pixel-level probability map $\hat{y} \in [0,1]^{H\times W}$. The model builds on the UNet3+ backbone [17] and introduces a key modification: a hybrid channel-spatial attention module applied individually to every incoming feature pathway at each decoder stage, after scale alignment but before multi-scale concatenation. In addition, consistent with the deep-supervision philosophy of UNet3+, HAUNet3+ employs auxiliary supervision at all decoder stages and the bottleneck to encourage hierarchical feature learning throughout the network. The overall architecture is illustrated in Figure 2.

a) Encoder: The encoder follows the standard UNet3+ design [17], consisting of five stages that progressively increase feature abstraction while reducing spatial resolution. The first stage processes the input image directly; all subsequent stages apply spatial downsampling before feature extraction:

$$E_1 = \phi_1(x), \qquad E_i = \phi_i(\mathrm{MaxPool}_{2\times 2}(E_{i-1})), \quad i = 2,\dots,5, \tag{1}$$

where each block ϕ_i applies two consecutive $\mathrm{Conv}_{3\times 3} \rightarrow$ BatchNorm $\rightarrow$ ReLU operations [51]. Channel dimensions expand as $C_i \in \{64, 128, 256, 512, 1024\}$, and spatial resolution decreases by a factor of two

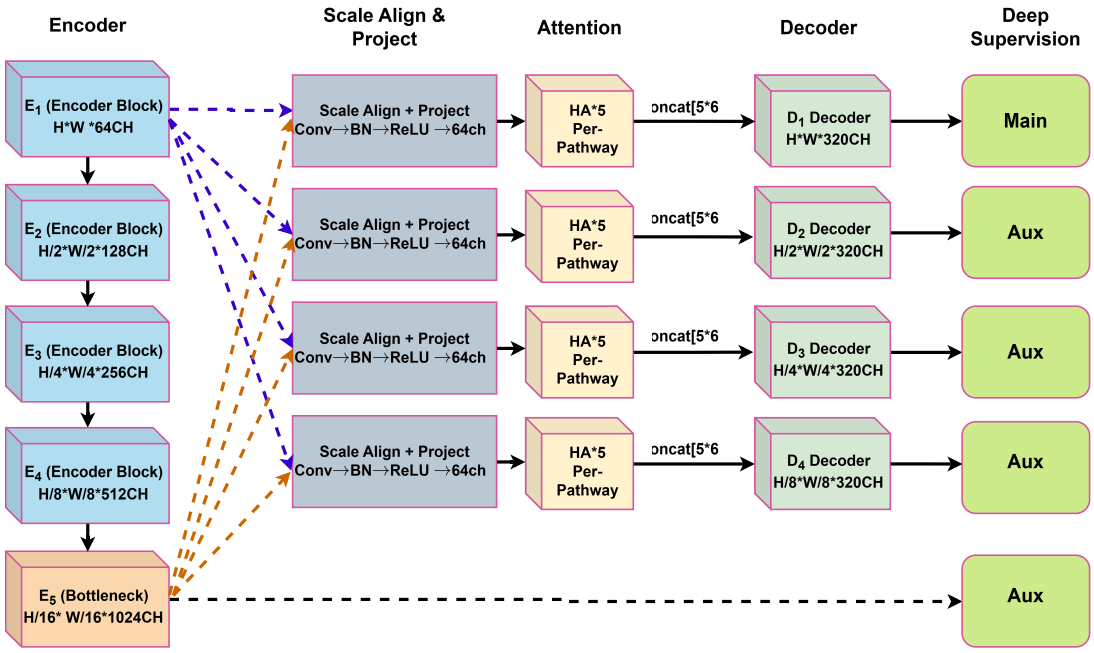

Figure 2: Full HAUNet3+ architecture. Encoder and bottleneck features are spatially aligned and projected to 64 channels, refined independently by five pre-concatenation Hybrid Attention modules at each decoder stage, and concatenated into a 320-channel decoder representation. Five prediction heads provide deep supervision during training; the primary inference output is $\hat{Y}_1$, while $\hat{Y}_2$–$\hat{Y}_5$ are auxiliary outputs upsampled to full input resolution.

at each transition, giving $H_i \times W_i = \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}$. An optional Dropout2d layer is inserted between the two convolutions in each block as a regularization measure [52]. All weights are initialized using Kaiming normal initialization [53], which is well suited to ReLU-activated networks and promotes stable gradient flow in early training.

b) Hybrid Attention Module: A central design choice in HAUNet3+ is the application of a hybrid attention operator $\mathcal{A}(\cdot)$ to each individual feature pathway before it contributes to a decoder stage. Rather than applying a single attention block to the concatenated feature map - as in [41] - we apply independent attention to every incoming skip-connection, regardless of whether it originates from an encoder stage or a deeper decoder stage. This allows the network to selectively suppress uninformative or noisy activations at each scale before they enter the fusion process. As illustrated in Figure 3, the module comprises two sequential sub-modules: channel attention followed by spatial attention, following the design philosophy of CBAM [21] but with a simplified single-branch channel formulation.

Channel Attention: Given a feature map $F \in \mathbb{R}^{C\times H\times W}$, channel attention identifies *which* channels carry the most relevant information. Global average pooling compresses the spatial dimensions into a compact descriptor:

$$z_c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{c,h,w}, \quad c = 1,\ldots,C, \tag{2}$$

producing $z \in \mathbb{R}^C$. This is passed through a two-layer FC bottleneck with reduction ratio $r = 16$, followed by a sigmoid to yield channel-wise recalibration weights:

$$s = \sigma(W_2\,\delta(W_1 z)), \quad s \in [0,1]^C, \tag{3}$$

where $W_1 \in \mathbb{R}^{(C/r)\times C}$, $W_2 \in \mathbb{R}^{C\times(C/r)}$, δ denotes ReLU, and σ denotes sigmoid. The recalibrated map is:

$$\mathcal{A}_c(F)_{c,h,w} = s_c \cdot F_{c,h,w}. \tag{4}$$

This follows the Squeeze-and-Excitation paradigm [20]. Unlike full CBAM [21], which additionally incorporates a max-pooled descriptor through a shared MLP, our single-branch formulation retains effective channel recalibration.

Spatial Attention: Following channel recalibration, spatial attention identifies *where* within the feature map the model should focus. Two descriptors are computed by pooling across the channel dimension of the channel-attended map $\mathcal{A}_c(F)$:

$$M_{\text{avg}}(h,w) = \frac{1}{C}\sum_{c=1}^{C} \mathcal{A}_c(F)_{c,h,w}, \qquad M_{\max}(h,w) = \max_c \mathcal{A}_c(F)_{c,h,w}. \tag{5}$$

These are concatenated and passed through a 7×7 convolution with same-padding, followed by a sigmoid, to produce a spatial attention map:

$$M_s = \sigma(\psi([M_{\text{avg}}\,;\, M_{\max}])) \in [0,1]^{H\times W}. \tag{6}$$

The spatially attended output is:

$$\mathcal{A}_s(F)_{c,h,w} = M_s(h,w)\cdot \mathcal{A}_c(F)_{c,h,w}. \tag{7}$$

The full hybrid attention operator is the sequential composition:

$$\mathcal{A}(F) = \mathcal{A}_s(\mathcal{A}_c(F))\,, \tag{8}$$

jointly modeling channel-wise importance and spatial saliency [21], [54].

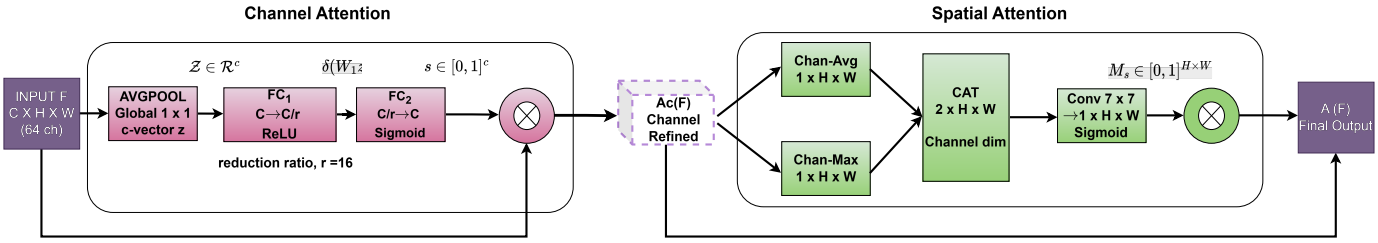

Figure 3: Internal structure of the Hybrid Attention module. Channel attention (left) uses global average pooling followed by a two-layer fully connected bottleneck to generate channel-wise recalibration weights. Spatial attention (right) computes channel-wise average and max descriptors, concatenates them, and applies a 7×7 convolution to produce a spatial attention map. Both sub-modules reweight the feature map through element-wise multiplication.

c) Multi-Scale Decoder with Pre-Concatenation Attention.: The decoder consists of four stages $\{D_k\}_{k=1}^{4}$, each operating at spatial resolution $H_k \times W_k = \frac{H}{2^{k-1}} \times \frac{W}{2^{k-1}}$, so that D_1 operates at full input resolution and D_4 at the coarsest decoder scale. Following UNet3+ [17], every decoder stage aggregates exactly five incoming feature pathways aligned to a common resolution. The coarsest stage D_4 receives all five encoder features $\{E_i\}_{i=1}^{5}$; intermediate and finer stages receive a mix of shallower encoder features and outputs of deeper decoder stages, always totaling five pathways. The detailed structure of a representative decoder [D3] is shown in Figure 4. In HAUNet3+, hybrid attention is applied to every incoming pathway *after* spatial alignment and channel projection but *before* concatenation.

We denote the n-th feature source feeding decoder stage D_k as $S_n^{(k)}$, and its 64-channel projected representation as:

$$\tilde{P}_n^{(k)} = \text{ReLU}\big(\text{BN}\big(\text{Conv}_{3\times3}\big(\mathcal{T}_n^{(k)}\big(S_n^{(k)}\big)\big)\big)\big) \in \mathbb{R}^{64\times H_k\times W_k}, \tag{9}$$

where $\mathcal{T}_n^{(k)}$ is the alignment operator: max-pooling by 2^{k-i} for encoder source E_i at a finer scale ($i<k$), bilinear upsampling by 2^{i-k} for encoder source E_i at a coarser scale ($i>k$), bilinear upsampling by 2^{j-k}

for deeper decoder source D_j ($j > k$), and identity when source and target share the same resolution. Hybrid attention is applied independently to each pathway:

$$\hat{P}_n^{(k)} = \mathcal{A}_n^{(k)}\left(\tilde{P}_n^{(k)}\right), \quad n = 1, \ldots, 5, \tag{10}$$

where $\mathcal{A}_n^{(k)}$ is a dedicated attention module for pathway n at stage k, giving five independent modules per decoder stage and 20 in total. The attended features are concatenated and fused:

$$D_k = \Phi_k\left(\left[\hat{P}_1^{(k)},\ \hat{P}_2^{(k)},\ \hat{P}_3^{(k)},\ \hat{P}_4^{(k)},\ \hat{P}_5^{(k)}\right]\right), \tag{11}$$

where concatenation produces a $5 \times 64 = 320$-channel tensor, and Φ_k is a $\text{Conv}_{3\times3} \rightarrow$ BatchNorm $\rightarrow$ ReLU block. As highlighted in Figure 4, the key distinction from [41] is that attention is applied to each of the five pathways *individually* before concatenation, rather than to the already-merged feature map. This per-pathway control is particularly valuable when features from distant scales carry very different levels of relevance to small target structures.

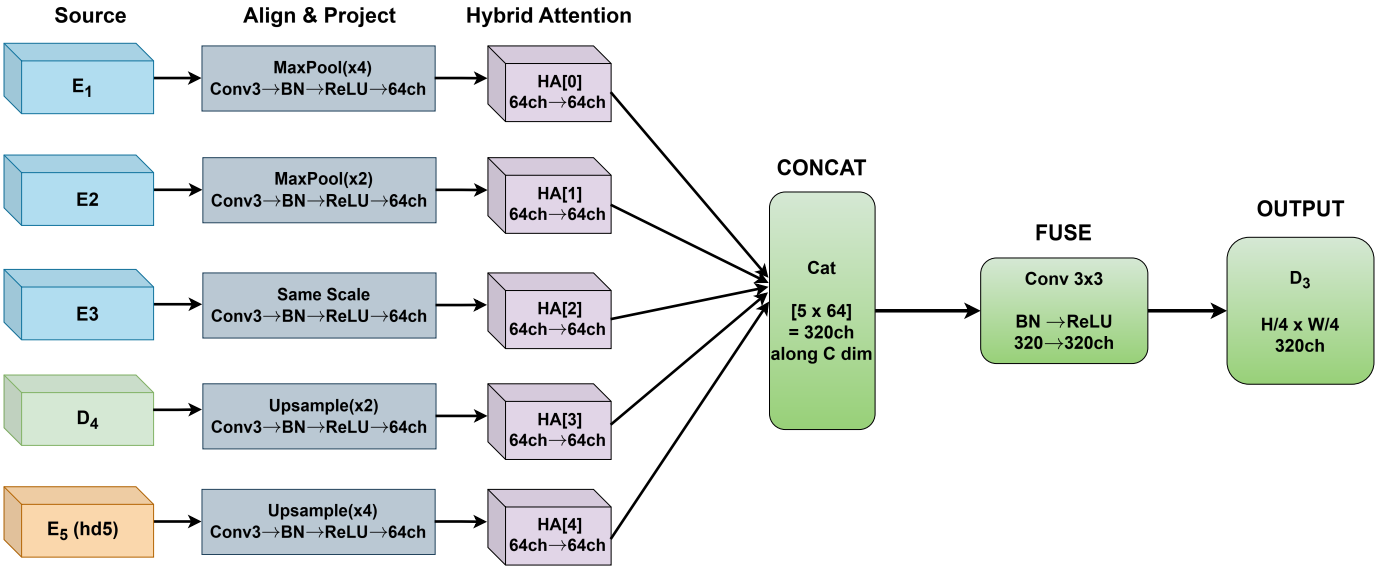

Figure 4: Detailed view of a single HAUNet3+ decoder stage D_3. Five input pathways from E_1 (max-pooled $4\times$), E_2 (max-pooled $2\times$), E_3 (same scale), D_4 (bilinear upsampled $2\times$), and E_5 (bilinear upsampled $4\times$) are each independently aligned, projected to 64 channels, and refined by a dedicated Hybrid Attention module before concatenation. The 320-channel concatenated tensor is fused by $\text{Conv}_{3\times3}$–BN–ReLU and passed to a deep supervision head.

d) Deep Supervision.: To encourage meaningful representations at every decoder level rather than relying solely on gradients from the final output stage [27], we attach auxiliary segmentation heads to all four decoder outputs and to the bottleneck stage E_5. Each head consists of a 3×3 convolutional layer followed by bilinear upsampling to full input resolution and a sigmoid activation:

$$\hat{y}_k = \sigma(\Psi_k(D_k)), \quad k = 1, \ldots, 4, \qquad \hat{y}_5 = \sigma(\Psi_5(E_5)), \tag{12}$$

where upsampling factors are $\times1$, $\times2$, $\times4$, $\times8$, $\times16$ for $\hat{y}_1$ through $\hat{y}_5$ respectively, consistent with the spatial scales $H_1 = H$, $H_2 = H/2$, $\ldots$, $H_5 = H/16$. The primary inference prediction is $\hat{y}_1$; the auxiliary outputs $\hat{y}_2$ through $\hat{y}_5$ contribute only during training through the loss described in Section IV-B.

e) Architecture Summary.: The complete forward pass of HAUNet3+ is:

$$E_1 = \phi_1(x), \quad E_i = \phi_i(\mathrm{MaxPool}(E_{i-1})), \quad i = 2, \ldots, 5, \tag{13}$$

$$\tilde{P}_n^{(k)} = \mathrm{Proj}\big(\mathcal{T}_n^{(k)}\big(S_n^{(k)}\big)\big), \quad \hat{P}_n^{(k)} = \mathcal{A}_n^{(k)}\Big(\tilde{P}_n^{(k)}\Big), \quad n = 1, \ldots, 5, \tag{14}$$

$$D_k = \Phi_k\Big(\Big[\hat{P}_1^{(k)}, \ldots, \hat{P}_5^{(k)}\Big]\Big), \quad k = 1, \ldots, 4, \tag{15}$$

$$\hat{y}_k = \sigma(\Psi_k(D_k)), \quad \hat{y}_5 = \sigma(\Psi_5(E_5)). \tag{16}$$

The architecture integrates a hierarchical convolutional encoder, scale-wise pre-concatenation attention, a full-scale multi-scale decoder, and deep supervision into a unified end-to-end trainable framework.

B. Loss Function

Choosing an appropriate loss function is critical for binary segmentation tasks in which the foreground class - levee sinkholes - may occupy only a small fraction of the image area. Training with a loss that is insensitive to class imbalance can cause the model to converge toward trivially predicting the dominant background class. To address this, we adopt a combined loss whose balance parameter α is treated as a hyperparameter and tuned via the search described in Section V-B.

1) Binary Cross-Entropy Loss: Binary Cross-Entropy (BCE) loss measures per-pixel classification accuracy by computing the log-likelihood of the predicted probability against the binary ground-truth label. For a single pixel with ground truth $y \in \{0, 1\}$ and predicted probability $\hat{y} \in (0, 1)$:

$$\mathcal{L}_{\mathrm{BCE}}(y, \hat{y}) = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right]. \tag{17}$$

Over an image with N pixels:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\right]. \tag{18}$$

BCE encourages well-calibrated pixel-wise probabilities. Its primary limitation under class imbalance is that the loss is dominated by the majority class: foreground boundaries receive disproportionately little gradient signal, leading to under-segmentation of small structures [55].

2) Dice Loss: Dice Loss directly optimizes the Dice Similarity Coefficient (DSC), which measures mask overlap as the *ratio* of intersection to total foreground area, making it inherently robust to class imbalance [55], [56]:

$$\mathrm{DSC}(Y, \hat{Y}) = \frac{2 \sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}. \tag{19}$$

The differentiable Dice loss is:

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2 \sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \varepsilon}, \tag{20}$$

where $\varepsilon = 10^{-6}$ prevents division by zero. While effective at driving high mask overlap, Dice loss can exhibit reduced sensitivity to precise boundary delineation when the foreground is extremely sparse, since boundary errors then constitute a disproportionately large fraction of the overlap ratio [57].

3) Combined Loss: To build on the complementary strengths of both losses - per-pixel accuracy from BCE and region-level overlap from Dice - we define:

$$\mathcal{L}_{\mathrm{combined}}(\hat{y}, y) = \alpha\, \mathcal{L}_{\mathrm{BCE}}(\hat{y}, y) + (1 - \alpha)\, \mathcal{L}_{\mathrm{Dice}}(\hat{y}, y), \tag{21}$$

where $\alpha \in (0, 1)$ is a tunable balance parameter. Values of $\alpha > 0.5$ place greater emphasis on per-pixel accuracy, while $\alpha < 0.5$ prioritizes regional overlap. The optimal value of α is determined through the hyperparameter search described in Section V-B, which identified $\alpha = 0.67$ as the best configuration on our dataset.

4) Deep Supervision Loss: Since HAUNet3+ produces five predictions $\{\hat{y}_k\}_{k=1}^{5}$ at training time, the combined loss is evaluated at each output and aggregated as a weighted, normalized sum. The weighting and normalization strategy are visualized in Figure 5. Finer decoder outputs receive higher weights, reflecting their greater spatial detail and closer proximity to the primary prediction:

$$\mathcal{L}_{\text{total}} = \frac{\sum_{k=1}^{5} w_k \cdot \mathcal{L}_{\text{combined}}(\hat{y}_k, y)}{\sum_{k=1}^{5} w_k}, \tag{22}$$

where $w_1 = 1.0,\ w_2 = 0.5,\ w_3 = 0.25,\ w_4 = 0.125,\ w_5 = 0.0625$, with $\hat{y}_1$ being the finest decoder output and $\hat{y}_5$ the bottleneck output. Normalizing by $\sum_k w_k$ maintains a consistent loss magnitude regardless of the number of supervision branches, following [58]. This formulation ensures that auxiliary branches provide meaningful gradient signal during early training while the primary output $\hat{y}_1$ dominates convergence [16], [26], [27].

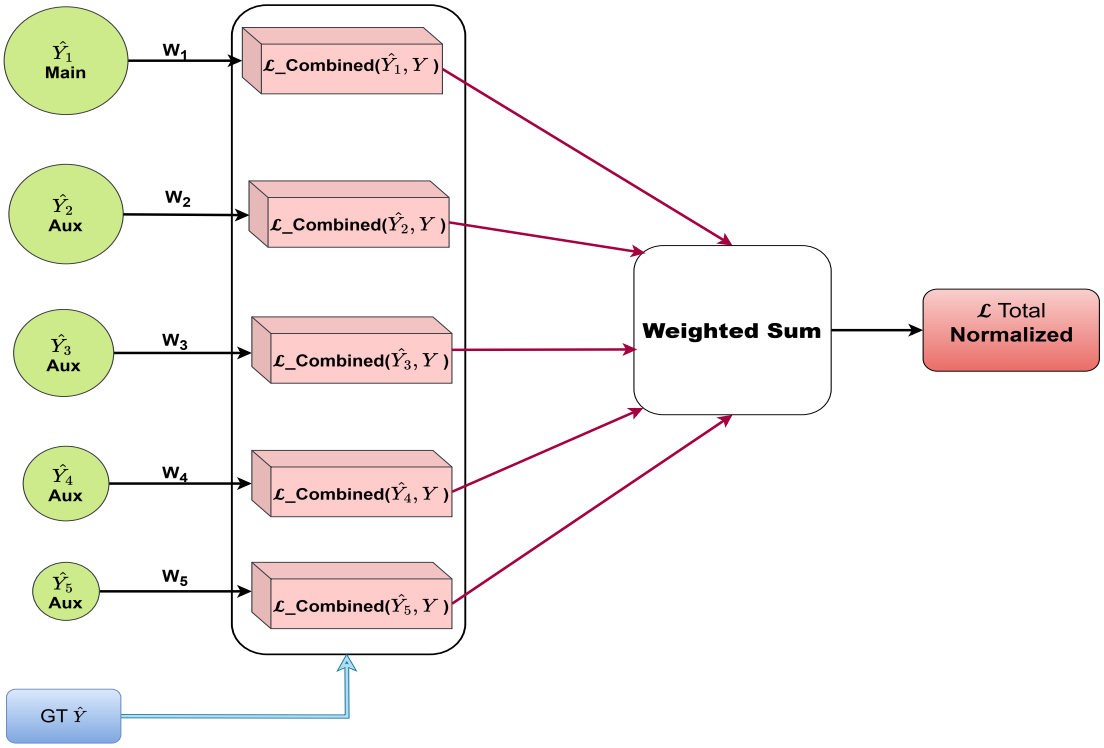

Figure 5: Deep supervision loss computation in HAUNet3+. Each of the five predictions ($\hat{y}_1$–$\hat{y}_5$) is independently evaluated against the ground truth using the combined BCE+Dice loss. The resulting losses are aggregated as a weighted, normalized sum. After hyperparameter search, the final model uses exponentially decreasing deep-supervision weights assigned to progressively coarser outputs ($w_1 = 1.0$, $w_2 = 0.5$, $w_3 = 0.25$, $w_4 = 0.125$, and $w_5 = 0.0625$). The total loss is optimized using AdamW with cosine annealing.

C. Optimization Strategy

The model is trained end-to-end using the AdamW optimizer [59] with an initial learning rate of 5×10^{-4} and weight decay of 10^{-4}. AdamW decouples the ℓ_2 weight decay penalty from the adaptive gradient update, providing more principled regularization than standard Adam [59]. The learning rate follows a cosine annealing schedule [60], which gradually reduces it from the initial value toward zero along a half-cosine curve, avoiding abrupt step-wise drops and promoting smoother convergence in later epochs.

Training was conducted for 500 epochs with a batch size of 16. Two complementary regularization strategies are employed to mitigate overfitting: ℓ_2 weight decay via AdamW, and early stopping based on

validation mIoU with a patience of 20 epochs. Standard data augmentation - random horizontal and vertical flipping, rotation, and brightness/contrast jittering - was applied during training to improve robustness to the viewpoint and illumination variations common in UAV-based levee imagery [61].

V. Experiments and Results

A. Experimental Setup

All experiments were conducted on an NVIDIA A100 GPU using PyTorch [62]. The input resolution was fixed at 320×320 pixels. Model performance was evaluated using three metrics: mean Intersection over Union (mIoU), pixel-wise accuracy (PA), and balanced accuracy (BA). Balanced accuracy is reported in addition to PA because the dataset exhibits a pronounced class imbalance, with background pixels substantially outnumbering foreground sinkhole pixels; in this regime, PA alone can be misleadingly high even for models that under-segment the foreground class.

B. Loss Hyperparameter Search

To identify the optimal combination of the loss-balancing parameter α and deep supervision (DS) weight scheme, we conducted a systematic grid search across multiple configurations. Recall that the combined loss is defined as:

$$\mathcal{L}_{\text{combined}} = \alpha\, \mathcal{L}_{\text{BCE}} + (1 - \alpha)\, \mathcal{L}_{\text{Dice}}, \tag{23}$$

so values of $\alpha < 0.5$ emphasize Dice loss, while values of $\alpha > 0.5$ prioritize BCE. For deep supervision, three weight-decay strategies were investigated: a linear scheme $[1.0, 0.8, 0.6, 0.4, 0.2]$, an exponential scheme $[1.0, 0.5, 0.25, 0.125, 0.0625]$, and a uniform scheme $[1.0, 1.0, 1.0, 1.0, 1.0]$. Each configuration was evaluated under 10-fold cross-validation on the training data. For each fold, the available data were partitioned into 90% training and 10% held-out testing. Within the training portion, a further 10% was reserved for validation, resulting in an effective split of approximately 81% training, 9% validation, and 10% testing per fold. Validation mIoU was used as the selection criterion. The coarse-grid results are presented in Figure 6 and the supplementary materials as Table S1. The combination $\alpha = 0.7$ with exponential DS weights achieved the highest validation mIoU of 88.8% in this initial sweep, identifying the $[0.65, 0.75]$ interval as the most promising region.

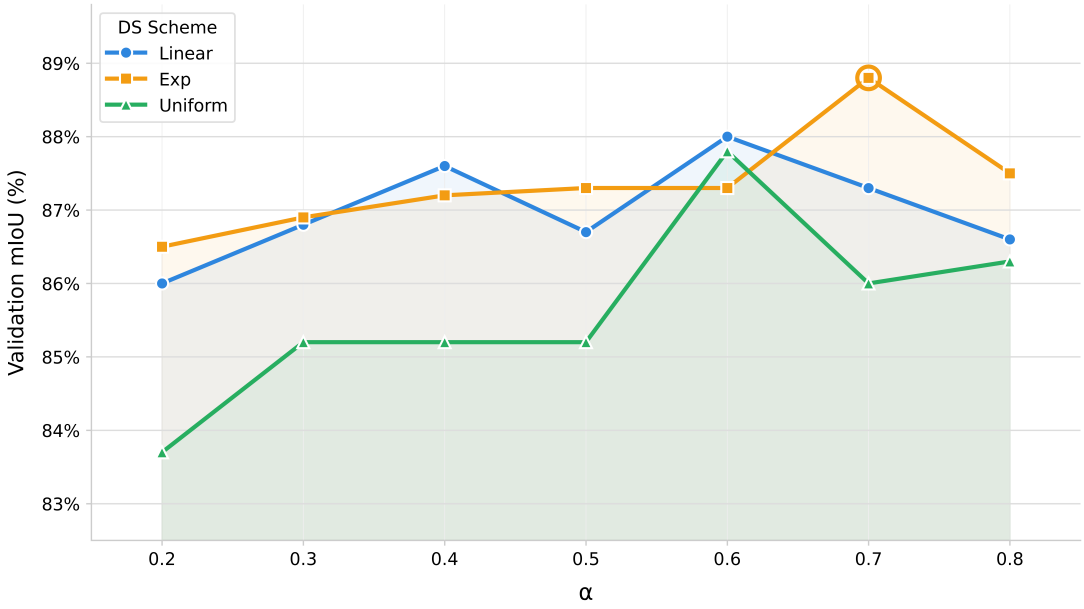

Figure 6: Coarse-grid search results for HAUNet3+: validation mIoU (%) across combinations of α and deep supervision weighting scheme.

Subsequently, a fine-grained search was performed over the interval $\alpha \in [0.65, 0.75]$ with a step size of 0.01, with fixed datasets of 80% training data, 10% validation data and 10% test data retaining all three DS weighting schemes. Results are presented in Figure 7 and in the supplementary materials as Table S2. The optimal configuration was $\alpha = 0.67$ with exponential DS weights (validation mIoU $= 88.95\%$), and this setting was used for all subsequent experiments.

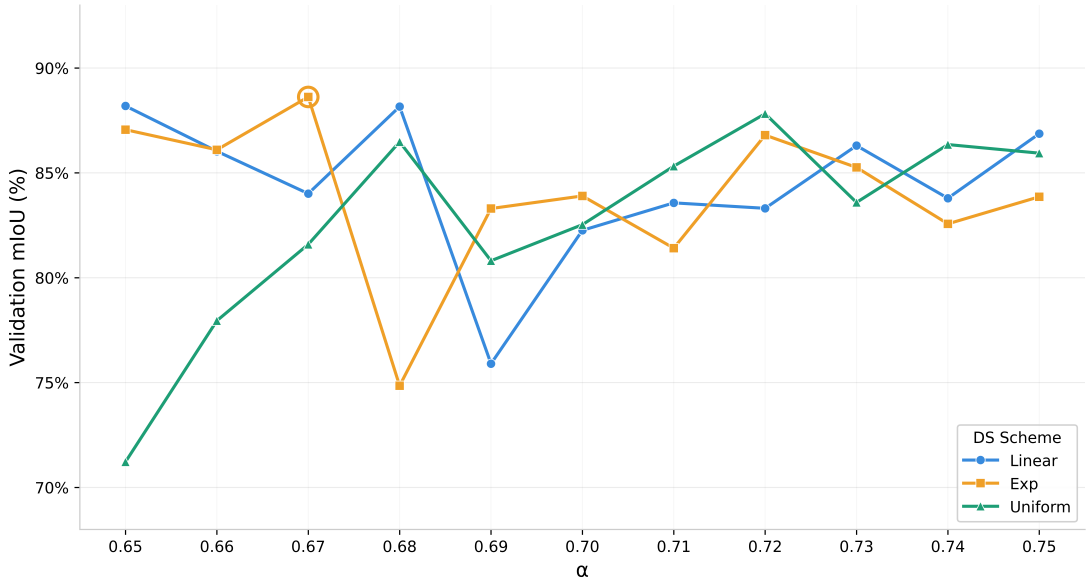

Figure 7: Fine-grained search results: validation mIoU (%) for $\alpha \in [0.65, 0.75]$ with step size 0.01.

C. Comparative Results

Using the optimal configuration ($\alpha = 0.67$, exponential DS weights), we compared HAUNet3+ against five baseline architectures under the same 10-fold cross-validation protocol: U-Net [14], Attention U-Net [22], UNet3+ [17], UNet3+ with deep supervision (DS) [17], and UNet3+ with DS and the classification-guided module (CGM) [17]. Mean mIoU, pixel-level accuracy (PA), and balanced accuracy (BA) averaged across folds are reported in Table I.

TABLE I: Comparison of segmentation performance on the SAMRefined Sinkhole dataset under 10-fold cross-validation ($\alpha = 0.67$, exponential DS weights).

Model	**mIoU (%)**	**PA (%)**	**BA (%)**
U-Net	78.0	94.53	93.04
Attention U-Net	65.8	90.78	88.64
UNet3+	78.7	94.85	93.70
UNet3+ with DS	80.2	94.96	93.53
UNet3+ with DS+CGM	70.8	88.40	83.67
HAUNet3+ (Ours)	**89.1**	**95.56**	**94.49**

HAUNet3+ achieves the highest mIoU (89.1%), PA (95.56%), and BA (94.49%) across all evaluated models. The gains over the closest UNet3+ based baseline (UNet3+ with DS, mIoU = 80.2%) amount to 8.9 percentage points in mIoU, demonstrating the effectiveness of per-pathway pre-concatenation hybrid attention in refining multi-scale feature representations.

D. ROC and Precision–Recall Analysis

Figure 8 presents the aggregated Receiver Operating Characteristic (ROC) curves obtained from 10-fold cross-validation for HAUNet3+ and all five baseline architectures. The ROC curve depicts the relationship

between the true positive rate (TPR) and false positive rate (FPR) across varying decision thresholds, providing a threshold-independent evaluation of discriminative performance [63].

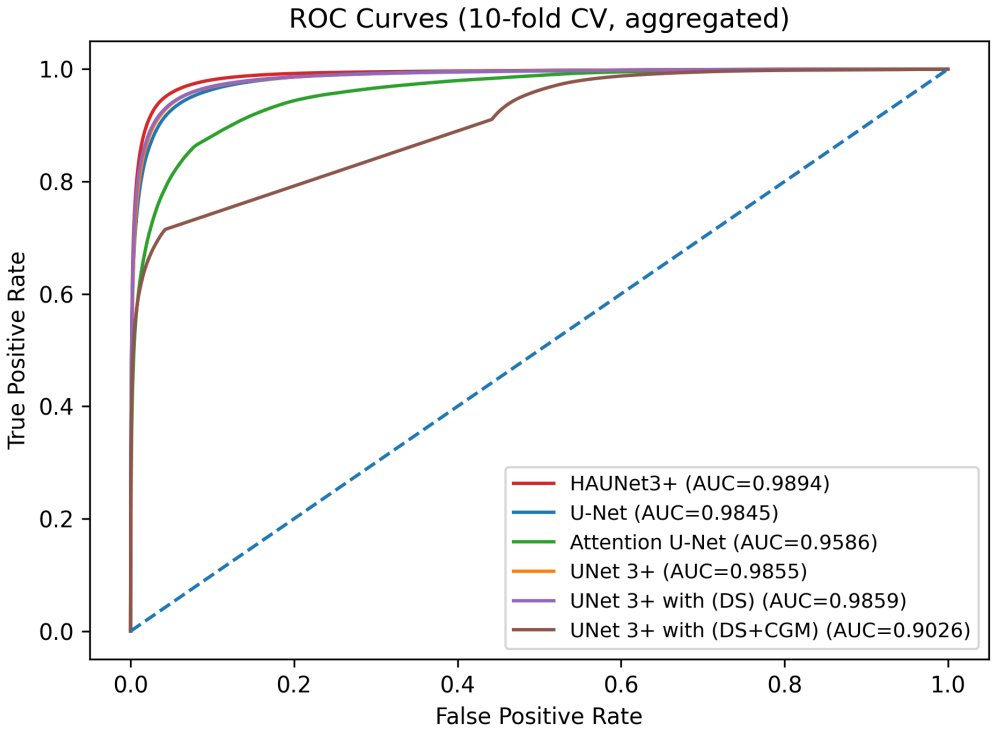

Figure 8: Aggregated ROC curves from 10-fold cross-validation. HAUNet3+ ($\alpha = 0.67$, exponential DS weights) achieves the highest AUC of 0.9894. The dashed diagonal represents a random classifier (AUC $= 0.5$).

Among all evaluated models, HAUNet3+ achieves the highest aggregated AUC of 0.9894, indicating that it has a 98.94% probability of assigning a higher confidence score to a randomly selected positive sample than to a randomly selected negative sample [64]. Its ROC curve consistently occupies the upper-left region of the plot, demonstrating that high sensitivity can be achieved while maintaining a very low false positive rate. The clear separation from competing models highlights the contribution of the hybrid attention mechanism and optimized loss weighting. Figure 9 presents the aggregated Precision–Recall (PR) curves under the same cross-validation protocol. The PR curve is particularly informative for imbalanced datasets, where negative samples dominate, because it directly characterizes the trade-off between detection precision and recall without being distorted by the large number of true negatives [65].

HAUNet3+ consistently maintains high precision across most recall levels, remaining close to 1.0 at low and moderate recall. As recall increases, precision degrades more gradually than for the baseline models, indicating that the model recovers a larger fraction of true positives while introducing comparatively few false alarms. A sharper decline in precision at near-complete recall is expected and observed across all models, as enforcing maximum recall entails accepting increasingly ambiguous predictions. HAUNet3+ achieves the highest aggregated Average Precision (AP) of 0.9750, outperforming U-Net (0.9651), Attention U-Net (0.9159), UNet3+ (0.9672), UNet3+ with DS (0.9688), and UNet3+ with DS+CGM (0.8205).

Both ROC and PR metrics were computed by aggregating prediction scores and ground-truth labels across all 10 validation folds rather than averaging fold-wise scalar metrics. This aggregation strategy is preferred because ROC and PR curves are nonlinear functions of the decision threshold, and averaging scalar summaries alone may obscure differences in classifier behavior at specific operating points [63], [65]. Pooling predictions from all folds captures the model's global ranking ability over the entire dataset,

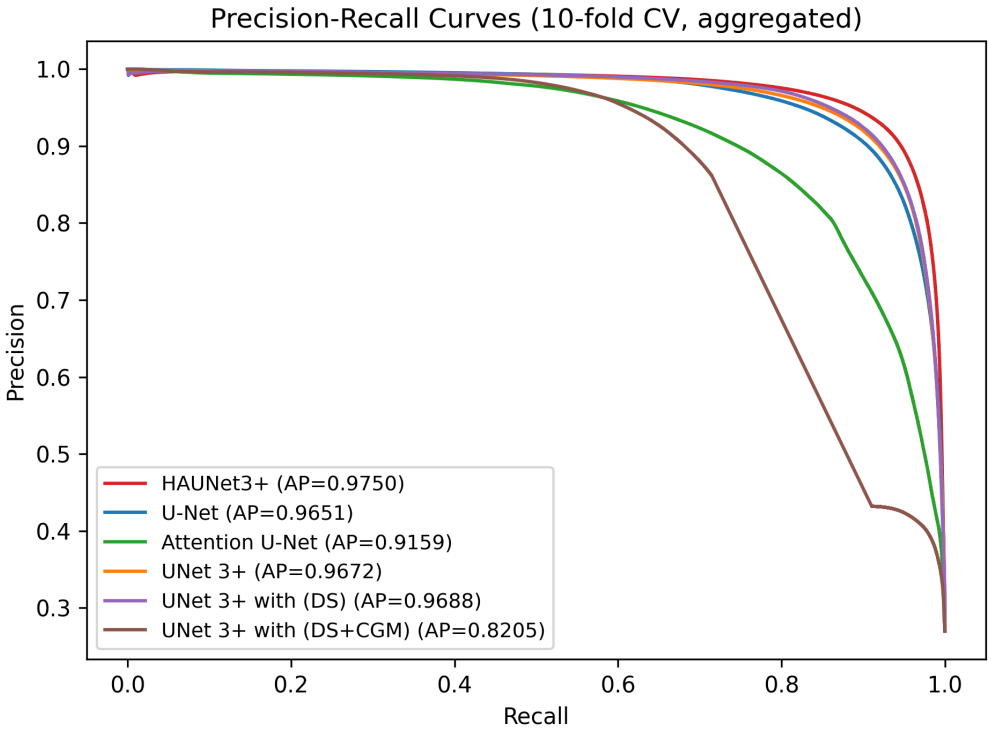

Figure 9: Aggregated Precision–Recall curves from 10-fold cross-validation. HAUNet3+ achieves the highest Average Precision (AP) of 0.9750.

producing smoother and more statistically robust curves. This is particularly important in class-imbalanced settings, where fold-specific class-ratio fluctuations can distort fold-averaged AP [66].

E. Statistical Significance Analysis

To rigorously evaluate the robustness and reliability of HAUNet3+, we computed 95% confidence intervals (CIs) for each metric using the Student t-distribution over the 10 outer cross-validation folds [67]. Results are presented in Table II. HAUNet3+ achieves a mean mIoU of 0.891 (95% CI: $[0.865, 0.916]$), PA of 0.956 (95% CI: $[0.953, 0.958]$), BA of 0.945 (95% CI: $[0.940, 0.949]$), AUC of 0.990 (95% CI: $[0.985, 0.994]$), and AP of 0.976 (95% CI: $[0.968, 0.984]$). The narrow confidence intervals across all metrics indicate stable, consistent performance across folds.

TABLE II: Performance of HAUNet3+ with 95% confidence intervals over 10-fold cross-validation.

Metric	**Mean**	**CI Low**	**CI High**
mIoU	0.891	0.865	0.916
Accuracy (PA)	0.956	0.953	0.958
Balanced Acc.	0.945	0.940	0.949
AUC	0.990	0.985	0.994
AP	0.976	0.968	0.984

To assess whether the observed improvements are statistically significant, paired t-tests were conducted between HAUNet3+ and each baseline over the 10 fold-level metric scores [68]. Results are presented in Table III. HAUNet3+ significantly outperforms all five baselines across all three metrics at the $p < 0.05$ significance level (denoted by p^* to avoid notational conflict with the loss-weighting parameter α). The improvements hold even against the strongest baselines (UNet3+ variants), confirming that the gains reflect consistent and meaningful architectural improvements rather than stochastic variation.

TABLE III: Paired t-test results comparing HAUNet3+ against baseline models across 10 cross-validation folds. Significance threshold: $p^* = 0.05$.

Model A	Model B	Metric	p-value	Significant ($p^* = 0.05$)
HAUNet3+	U-Net	mIoU	3.83×10^{-4}	Yes
HAUNet3+	U-Net	PA	1.88×10^{-4}	Yes
HAUNet3+	U-Net	Balanced Acc.	2.33×10^{-5}	Yes
HAUNet3+	Attention U-Net	mIoU	3.15×10^{-6}	Yes
HAUNet3+	Attention U-Net	PA	2.32×10^{-4}	Yes
HAUNet3+	Attention U-Net	Balanced Acc.	1.46×10^{-4}	Yes
HAUNet3+	UNet3+	mIoU	7.63×10^{-4}	Yes
HAUNet3+	UNet3+	PA	8.78×10^{-3}	Yes
HAUNet3+	UNet3+	Balanced Acc.	4.88×10^{-2}	Yes
HAUNet3+	UNet3+ (DS)	mIoU	3.31×10^{-4}	Yes
HAUNet3+	UNet3+ (DS)	PA	5.63×10^{-3}	Yes
HAUNet3+	UNet3+ (DS)	Balanced Acc.	4.20×10^{-3}	Yes
HAUNet3+	UNet3+ (DS+CGM)	mIoU	3.04×10^{-4}	Yes
HAUNet3+	UNet3+ (DS+CGM)	PA	2.50×10^{-4}	Yes
HAUNet3+	UNet3+ (DS+CGM)	Balanced Acc.	8.88×10^{-4}	Yes

F. Ablation Study

TABLE IV: Ablation study on the SAMRefined Sinkhole dataset under 10-fold cross-validation. Each row adds one component to the previous one. "Pre-concat" denotes applying hybrid attention to each pathway *before* multi-scale concatenation; "Post-concat" denotes applying a single attention block *after* concatenation (as in [69]). Best results are **bolded**.

Variant	Attention Type		Deep Sup.	Pre-concat	mIoU (%)	F1 (%)	PA (%)	BA (%)
	Channel	Spatial						
UNet3+ (baseline)	✗	✗	✗	—	78.7	91.29	94.85	93.70
UNet3+ + DS	✗	✗	✓	—	80.2	91.40	94.96	93.53
UNet3+ + DS + Channel Attn only	✓	✗	✓	✓	84.88	88.99	**95.71**	87.61
UNet3+ + DS + Spatial Attn only	✗	✓	✓	✓	79.76	84.33	94.54	85.99
UNet3+ + DS + Post-concat HA	✓	✓	✓	✗	85.23	90.26	95.63	87.47
HAUNet3+ (Pre-concat HA + DS)	✓	✓	✓	✓	**89.1**	**92.49**	95.56	**94.49**

Table IV reports the incremental contribution of each architectural component under 10-fold cross-validation. Deep supervision alone yields a modest but consistent gain over the vanilla UNet3+ baseline (+1.5 pp mIoU), confirming that auxiliary gradient signal at intermediate decoder stages prevents background-biased feature collapse when foreground pixels are sparse [70]. Introducing pre-concatenation channel attention delivers the largest single improvement (+4.68 pp), reflecting the primacy of channel-level feature recalibration in a task where sinkhole-discriminative spectral and textural cues must be selectively amplified before multi-scale fusion — without this prior filtering, spatial attention applied in isolation provides negligible benefit (79.76% mIoU), as it lacks a reliable basis for generating discriminative saliency maps from channel-noisy representations. The comparison between post-concatenation and pre-concatenation hybrid attention (85.23% vs. 89.1% mIoU) isolates the effect of attention placement: applying a single attention block to the already-merged 320-channel feature map forces the module to disentangle five spatially and semantically heterogeneous sources simultaneously, whereas per-pathway pre-concatenation attention assigns a dedicated module to each scale, enabling scale-specific noise suppression before irreversible feature interleaving occurs. The full HAUNet3+ configuration, combining both

attention types with pre-concatenation placement and deep supervision, achieves the best performance on three of four metrics (mIoU, F1, BA), confirming that the components are mutually reinforcing: channel recalibration enables effective spatial localization, and pre-concatenation placement ensures both operations act on clean, scale-specific representations.

Figure 10 presents qualitative segmentation results on selected test samples from the SAMRefined Sinkhole dataset, comparing HAUNet3+ against all five baseline models. Each row corresponds to a distinct test image; the columns show the original image, the ground-truth, and predictions from U-Net, Attention U-Net, UNet3+, UNet3+ with DS, UNet3+ with DS+CGM, and HAUNet3+ respectively.

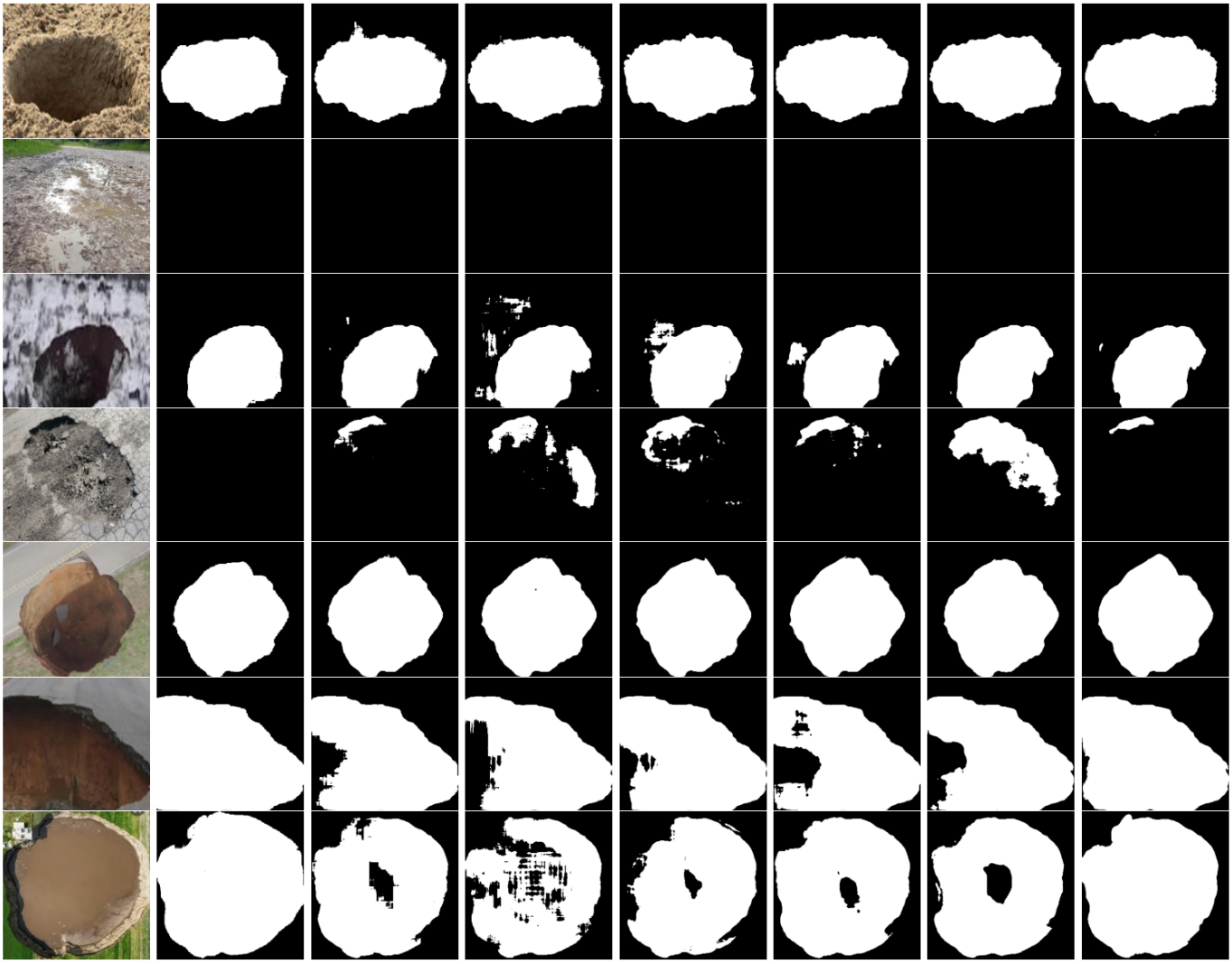

Figure 10: Qualitative segmentation results on selected test samples from the SAMRefined Sinkhole dataset. Each row corresponds to a different test image. Columns represent: (1) Original image, (2) Ground truth, (3) U-Net, (4) Attention U-Net, (5) UNet3+, (6) UNet3+ with DS, (7) UNet3+ with DS+CGM, (8) HAUNet3+ (Ours).

VI. Discussion and Conclusion

The experimental results validate the central hypothesis of this work: that applying independent hybrid channel-spatial attention to each incoming skip pathway *prior* to multi-scale feature concatenation yields more discriminative and noise-robust decoder representations than post-concatenation attention or unfiltered feature fusion. The 8.9 percentage point mIoU improvement over the strongest UNet3+-based

baseline, supported by statistically significant paired t-tests across all metrics and narrow 95% confidence intervals over ten cross-validation folds, confirms that the gains are consistent and attributable to principled architectural choices rather than stochastic variation. The ablation study further reveals that channel attention contributes more substantially than spatial attention in isolation — a finding that is interpretable given the nature of the task, where identifying *which* feature channels encode sinkhole-relevant spectral and morphological cues is a prerequisite for effective spatial localization of targets that occupy fewer than 1–3% of image pixels. The combined BCE+Dice loss with optimized deep supervision weighting addresses the severe class imbalance inherent in levee survey imagery, producing predictions that are well calibrated across both foreground and background classes, as evidenced by the small gap between pixel accuracy (95.56%) and balanced accuracy (94.49%). Despite these strengths, the study has notable limitations: all evaluations are conducted on a single dataset, and the framework does not currently provide uncertainty estimates for individual predictions — a capability that would be particularly valuable in safety-critical levee inspection workflows. Future work will address these gaps through multi-dataset validation across geographically diverse levee systems, integration of lightweight transformer modules for long-range contextual reasoning, and incorporation of calibrated uncertainty estimation to support field prioritization decisions. In summary, HAUNet3+ establishes a new performance benchmark for semantic segmentation of sinkholes in UAV-based levee imagery, and demonstrates that targeted architectural attention at the skip-connection level is an effective and principled strategy for small object segmentation under severe class imbalance.

AVAILABILITY

The code and datasets are publicly available at: https://github.com/ayonneub/HAUNet3Plus

ACKNOWLEDGMENTS

This research was supported by the U.S. Department of the Army, U.S. Army Corps of Engineers (USACE), under Contract No. W912HZ-23-2-0004. The views and conclusions expressed in this paper are those of the authors and do not necessarily reflect the official policies or positions of the U.S. Army Corps of Engineers or the U.S. Department of the Army.

REFERENCES

[1] U.S. Army Corps of Engineers, St. Louis District, "Pipe inspections: Why are they important?" https://www.mvs.usace.army.mil/Portals/54/docs/LeveeSafety/Documents/2024%20Pipe%20Inspections%20-%20Why%20They%20Are%20Important%20Final_pdf.pdf, 2024, accessed: 2025-11-19.

[2] R. Fell, P. MacGregor, D. Stapledon, and G. Bell, *Geotechnical Engineering of Embankment Dams*. A.A. Balkema, 2003.

[3] U.S. Army Corps of Engineers, "History of levees," https://levees.sec.usace.army.mil/levee-basics/history-of-levees/, n.d., accessed: 2025-11-11.

[4] M. E. Smith and K. L. Johnson, "Levee inspection practices and challenges in the united states," US Army Corps of Engineers, Engineer Research and Development Center, Tech. Rep. ERDC/CHL TR-20-8, 2020.

[5] Y. Lin, H. Li, W. Shao, Z. Yang, J. Zhao, X. He, P. Luo, and K. Zhang, "Samrefiner: Taming segment anything model for universal mask refinement," 2025. [Online]. Available: https://arxiv.org/abs/2502.06756

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[7] A. Dey, "Sinkhole dataset samrefiner," https://github.com/ayonneub/HAUNet3Plus/tree/main/Sinkhole_dataset_SAMRefiner, 2026, accessed: 2026-05-18.

[8] B. Brisco, A. Schmitt, and K. Murnaghan, "UAV-based flood and infrastructure monitoring," *Remote Sensing*, vol. 5, no. 6, pp. 2921–2940, 2013.

[9] C. Zhang and J. M. Kovacs, "The application of small unmanned aerial systems for precision agriculture and infrastructure inspection," *Precision Agriculture*, vol. 13, no. 6, pp. 693–712, 2019.

[10] J. P. Resop, L. Lehmann, and W. C. Hession, "Drone laser scanning for modeling riverscape topography and vegetation: Comparison with traditional aerial lidar," *Drones*, vol. 3, no. 2, p. 35, 2019.

[11] D. Feng, Z. Liu, and X. Wang, "Artificial intelligence-based methods for civil infrastructure inspection and monitoring," *Automation in Construction*, vol. 134, p. 104057, 2022.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[13] X. Bai, H. Shi, K. Zhang, X. Cao, and B. Kang, "A review of surface defect detection of industrial components based on deep learning," *Applied Sciences*, vol. 13, no. 2, p. 1039, 2023.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[17] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[18] M. Hasan, S. Serikawa, and Q. A. Memon, "Sinkhole detection using a UAV-based deep learning approach," *Sensors*, vol. 20, no. 21, p. 6304, 2020.

[19] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "A survey of small object detection based on deep learning," *International Journal of Computer Systems Science and Engineering*, vol. 44, no. 3, pp. 951–968, 2022.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[22] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[23] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[24] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[25] P. Iakubovskii, "Segmentation models," https://github.com/qubvel/segmentation_models, 2019.

[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip-connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[27] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 562–570, 2015.

[28] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[29] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision Workshops (ECCVW)*, 2022, pp. 205–218.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[31] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 12 077–12 090.

[32] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-resolution vision transformer for dense prediction," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 7281–7293.

[33] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[34] P. Liu, Y. Song, M. Chai, Z. Han, and Y. Zhang, "Swin–UNet++: A nested Swin Transformer architecture for location identification and morphology segmentation of dimples on 2.25Cr1Mo0.25V fractured surface," *Materials*, vol. 14, no. 24, p. 7504, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8703304/

[35] J. Choi, D. Seo, J. Jung, Y. Han, J. Oh, and C. Lee, "Cloud detection using a UNet3+ model with a hybrid Swin Transformer and EfficientNet (UNet3+STE) for very-high-resolution satellite imagery," *Remote Sensing*, vol. 16, no. 20, p. 3880, 2024. [Online]. Available: https://doi.org/10.3390/rs16203880

[36] R. M. Ahmed, A. Iltaf, M. Elmanna, G. Zhao, H. Li, Y. Du, B. Li, and S. Zhou, "MSA-UNet3+: Multi-scale attention UNet3+ with new supervised prototypical contrastive loss for coronary DSA image segmentation," *arXiv preprint arXiv:2504.05184*, 2025. [Online]. Available: https://arxiv.org/abs/2504.05184

[37] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5659–5667.

[38] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 713–13 722.

[39] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3138–3147.

[40] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," *Neurocomputing*, vol. 506, pp. 158–167, 2022.

[41] Y. Xu, S. Hou, X. Wang, D. Li, and L. Lu, "A medical image segmentation method based on improved UNet 3+ network," *Diagnostics*, vol. 13, no. 3, p. 576, 2023.

[42] J. Tie, W. Wu, L. Zheng, L. Wu, and T. Chen, "Improving walnut images segmentation using modified UNet3+ algorithm," *Agriculture*, vol. 14, no. 1, p. 149, 2024.

[43] F. Gutiérrez, A. H. Cooper, and K. S. Johnson, "Sinkholes and subsidence: Karst and cavernous rocks in engineering and environmental contexts," *Episodes*, vol. 31, no. 1, pp. 129–136, 2014.

[44] E. O. Lindsey, Y. Fialko, and Y. Bock, "Using aerial LiDAR to detect levee anomalies and improve inspection efficiency," *Natural Hazards and Earth System Sciences*, vol. 16, pp. 1–14, 2016.

[45] R. Alshawi, M. T. Hoque, and M. C. Flanagin, "A depth-wise separable u-net architecture with multiscale filters to detect sinkholes," *Remote Sensing*, vol. 15, no. 5, p. 1384, 2023.

[46] O. Alrabayah, D. Caus, R. A. Watson, H. Z. Schulten, T. Weigel, L. Rüpke, and D. Al-Halbouni, "Deep-learning-based automatic sinkhole recognition: Application to the eastern dead sea," *Remote Sensing*, vol. 16, no. 13, p. 2264, 2024.

[47] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2778–2788.

[48] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Building footprint extraction from high-resolution aerial imagery using a combined segmentation and regularization approach," *Remote Sensing*, vol. 14, no. 10, p. 2350, 2022.

[49] J. Fan, L. Bi, and B. Wu, "Automated crack detection and severity assessment for UAV-based bridge inspection," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 9, pp. 1110–1127, 2022.

[50] N. Heller, J. Dean, and N. Papanikolopoulos, "Imperfect segmentation labels: How much do they matter?" 2018. [Online]. Available: https://arxiv.org/abs/1806.04618

[51] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*. PMLR, 2015, pp. 448–456.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[54] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[55] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *International Workshop on Deep Learning in Medical Image Analysis*. Springer, 2017, pp. 240–248.

[56] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[57] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and Jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2019, pp. 92–100.

[58] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[60] ——, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[61] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[63] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[64] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[65] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 233–240.

[66] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[67] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

[68] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[69] Q. Xu, Z. Ma, N. He, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation," *Computers in Biology and Medicine*, vol. 154, p. 106626, Mar. 2023, arXiv preprint arXiv:2202.00972. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2023.106626

[70] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015, pp. 562–570.

Supplementary Material for HAUNet3+: Enhancing UNet3+ with Hybrid Attention for Samrefined Sinkhole Image Segmentation

Ayon Dey[1], Anav Katwal[1], Abdullah Naeem[1], Padam Jung Thapa[2], Kendall N. Niles[3], Steve Sloan[3], and Md Tamjidul Hoque[1,*]
[1]Department of Computer Science, Louisiana State University New Orleans, New Orleans, LA, USA
[2]University of Louisiana at Lafayette, Lafayette, Louisiana, USA
[3]U.S. Army Engineer Research and Development Center, Vicksburg, MS, USA
Emails: {adey, akatwal, aneem, thoque}@lsuneworleans.edu, tpadamjung@gmail.com, kendall.n.niles@erdc.dren.mil, steven.d.sloan@erdc.dren.mil
*Corresponding author: thoque@lsuneworleans.edu

SUPPLEMENTARY TABLES

TABLE S1: *Coarse-grid search results for HAUNet3+: validation mIoU (%) across combinations of α and deep supervision weighting scheme.*

α	DS Scheme	Validation mIoU (%)
0.2	Linear	86.0
0.2	Exp	86.5
0.2	Uniform	83.7
0.3	Linear	86.8
0.3	Exp	86.9
0.3	Uniform	85.2
0.4	Linear	87.6
0.4	Exp	87.2
0.4	Uniform	85.2
0.5	Linear	86.7
0.5	Exp	87.3
0.5	Uniform	85.2
0.6	Linear	88.0
0.6	Exp	87.3
0.6	Uniform	87.8
0.7	Linear	87.3
0.7	**Exp**	**88.8**
0.7	Uniform	86.0
0.8	Linear	86.6
0.8	Exp	87.5
0.8	Uniform	86.3

TABLE S2: *Fine-grained search results: validation mIoU (%) for* $\alpha \in [0.65, 0.75]$ *with step size* 0.01.

α	DS Scheme	Validation mIoU (%)
0.65	Linear	88.19
0.65	Exp	87.06
0.65	Uniform	71.22
0.66	Linear	86.03
0.66	Exp	86.10
0.66	Uniform	77.95
0.67	Linear	84.01
0.67	**Exp**	**88.95**
0.67	Uniform	81.59
0.68	Linear	88.16
0.68	Exp	74.85
0.68	Uniform	86.48
0.69	Linear	75.90
0.69	Exp	83.30
0.69	Uniform	80.81
0.70	Linear	82.26
0.70	Exp	83.90
0.70	Uniform	82.53
0.71	Linear	83.57
0.71	Exp	81.41
0.71	Uniform	85.32
0.72	Linear	83.31
0.72	Exp	86.80
0.72	Uniform	87.82
0.73	Linear	86.30
0.73	Exp	85.26
0.73	Uniform	83.59
0.74	Linear	83.79
0.74	Exp	82.57
0.74	Uniform	86.35
0.75	Linear	86.87
0.75	Exp	83.86
0.75	Uniform	85.94