Dual-Stage Kalman Framework: Spatial Denoising and Temporal Tracking for Lightweight Water Body Extraction

Anirudh, Jesus, Nishad, Tanushree Das, Shaswat, Adarsh

***Abstract*—Accurate segmentation of surface water bodies from satellite imagery is essential for applications such as agricultural planning, flood risk assessment, and environmental monitoring. However, operational deployment of such systems faces three major challenges: noisy predictions due to shadows and background clutter, poorly defined water–land boundaries, and the lack of true temporal image sequences for applying temporal smoothing or prediction. This project addresses these issues through a two–module hybrid framework that combines deep learning–based image segmentation with spatial and temporal Kalman filtering.**

In Module 1, we propose LKF-SegNet, a U-Net–based segmentation architecture with a MobileNetV2 encoder augmented by a Spatial Learnable Kalman Filter at the bottleneck. The filter treats each row of the encoder feature map as a pseudo-temporal signal and performs per-pixel recursive prediction–update steps, leading to smoother and more coherent feature representations. To further improve delineation of water boundaries, the network is trained with an Edge-Weighted Loss, which applies larger penalties to errors near the water–land interface using a Canny edge map and distance-transform-based weighting. On a curated satellite water-body dataset, LKF-SegNet achieves around 88–90% pixel accuracy, with test IoU $\approx 74\%$, precision $\approx 88\%$, and recall $\approx 81\%$, demonstrating robust segmentation performance.

In Module 2, we investigate temporal consistency and next-state prediction in the absence of true video data. A synthetic temporal sequence is generated by sliding a fixed-size window over a single large satellite image, producing an ordered patch stream that emulates motion. A per-pixel Temporal Kalman Filter is then applied to the sequence of LKF-SegNet probability masks to (i) reduce frame-to-frame flicker and noise and (ii) generate one-step-ahead predictions of the water mask for the next patch. Although based on synthetic motion, this experiment shows that classical Kalman filtering can enhance temporal stability and provide predictive capability on top of a strong spatial segmentation backbone. Together, the two modules illustrate how integrating deep neural networks with Kalman filtering can yield a practical, computationally efficient framework for water-body monitoring in remote-sensing scenarios.

***Index Terms*—Semantic Segmentation, Kalman Filter, MobileNetV2, Remote Sensing, Water Body Extraction**

I. Introduction

Surface water bodies—such as rivers, lakes, reservoirs, and irrigation channels—play a fundamental role in environmental management, agricultural planning, and disaster mitigation. Monitoring the spatial extent and temporal evolution of these water bodies is crucial for flood forecasting, crop water assessment, drought analysis, and ecological conservation. With the increasing availability of high-resolution satellite imagery, automated water-body segmentation has emerged as a promising solution. However, achieving reliable and temporally consistent segmentation remains challenging due to noise, shadows, vegetation interference, heterogeneous landscapes, and the scarcity of clean, labeled temporal image sequences.

Deep learning models, especially U-Net and its variants, have become the dominant approach for pixel-wise semantic segmentation. While these networks excel at learning rich spatial features, they often suffer from noisy outputs around water boundaries and inconsistent predictions across spatially adjacent regions. Moreover, most satellite datasets lack true video sequences, preventing traditional temporal smoothing techniques from being applied directly. These limitations motivate the integration of classical signal processing concepts—especially Kalman filtering—with deep neural models to enforce structural coherence and temporal stability.

This project proposes a two-module hybrid framework that combines learnable spatial filtering with classical temporal filtering to address the inherent challenges of water-body segmentation. In Module 1, we introduce LKF-SegNet, a U-Net–based architecture enhanced by a Spatial Learnable Kalman Filter inserted at the bottleneck layer. By interpreting the encoder's feature map as a row-wise temporal signal, the Kalman filter performs recursive prediction and correction that reduces noise, enhances feature smoothness, and preserves structural continuity across the image. The model is further strengthened using an Edge-Weighted Loss, which emphasizes the water–land interface using a distance-transformed Canny edge representation. Together, these innovations yield improved segmentation robustness, especially in areas characterized by subtle gradients or cloud and shadow interference.

In Module 2, we address the challenge of temporal consistency in the absence of real temporal datasets. A synthetic video sequence is constructed by sliding a fixed-size window across a single large satellite image, simulating spatial motion. The segmentation outputs of LKF-SegNet across these patches form a pseudo-temporal sequence on which a Temporal Kalman Filter is applied. This filter reduces prediction flicker, enforces smooth temporal transitions, and generates one-step-ahead mask predictions, demonstrating that classical filtering principles can extend deep-learning-based segmentation into predictive analysis even without true video input.

Through this dual-stage framework, the project bridges the gap between data-driven deep learning and model-based signal processing. The resulting system achieves high-quality segmentation while introducing temporal prediction capabilities, offering a practical and computationally efficient solution

for water-body monitoring in remote-sensing and agricultural applications.

II. Background Study

A. Deep Learning for Water-Body Segmentation

The task of extracting water bodies from satellite and UAV imagery has evolved significantly from index-based methods such as NDWI, MNDWI, and AWEI. While these spectral techniques are effective under controlled environments, they degrade under vegetation shadows, turbidity variation, or specular reflections, leading to misclassification of water and non-water regions. This limitation is repeatedly highlighted in recent literature on agricultural water-body detection, where spectral noise and irregular boundaries remain major obstacles [3].

Deep learning, especially encoder–decoder architectures, has become the dominant solution for remote-sensing segmentation because of their ability to learn both high-level semantics and fine-grained spatial information. U-Net and its variants (FCN, DeepLab, PSPNet, HRNet) have been widely adopted due to skip connections that preserve boundary information across multiple scales [5]. The literature shows MobileNet-based U-Nets are particularly attractive for lightweight deployment on resource-constrained platforms such as agricultural drones.

Wieland et al. (2023) and related works benchmark MobileNetV2/MobileNetV3 encoders and show that they offer high throughput (on the order of tens of megapixels per second) while maintaining competitive accuracy when compared to heavier ResNet-50 backbones, making them ideal for large-scale agricultural monitoring tasks [2]. This trade-off is further emphasized by Liao et al. (2025), who integrate MobileNet with learnable Kalman filtering for water segmentation in agricultural watersheds, achieving strong IoU scores with extremely small model sizes [1].

The proposed LKF-SegNet builds directly on this research trend by:

- using MobileNetV2 as a lightweight encoder,
- preserving spatial detail through U-Net skip connections, and
- introducing a learnable Kalman filter at the bottleneck to refine spatial features.

B. Kalman Filtering in Vision and Remote Sensing

Kalman filters were originally designed for linear state estimation in dynamical systems, but they have been widely applied in computer vision for temporal tracking of objects and shapes. In biomedical imaging, Bersvendsen et al. (2016) demonstrated how a Kalman filter can stabilize right-ventricle surface trajectories in 3D echocardiography under noisy, partially missing observations [4]. Their work shows that the prediction–correction loop of the Kalman filter can compensate for corrupted observations and maintain a coherent estimate of anatomy over time.

More recently, Liao et al. (2025) introduced the concept of a *Spatial Learnable Kalman Filter* for water-body segmentation in agricultural imagery [1]. Instead of operating over time, their filter treats each *row* of a feature map as a one-dimensional sequence. For each row index, a Kalman-like update is applied using learnable transition and observation mappings. This formulation effectively transforms the Kalman filter from a temporal estimator into a spatial denoising and feature-stabilization module inside a convolutional network.

Inspired by these ideas, Module 1 in the present work adopts a similar principle:

- the encoder produces high-level feature maps,
- the bottleneck applies a Spatial Kalman Filter row by row, and
- the decoder reconstructs a denoised segmentation mask.

This introduces recursive estimation into the network and improves robustness against random noise, shadow artifacts, texture inconsistencies, and blurred boundaries in satellite imagery.

C. Temporal Kalman Filtering and Synthetic Video

Unlike medical or surveillance applications, real temporal sequences of high-resolution satellite water-body imagery are scarce, and when available, they often lack dense pixel-wise annotations. This scarcity makes it difficult to train and evaluate sequence models or temporal smoothing algorithms directly.

To circumvent this limitation, researchers sometimes construct *synthetic temporal sequences* from static imagery. In this project, a similar idea is used: a fixed-size window is slid across a large satellite image, producing a sequence of overlapping patches. By assigning an ordering to these patches, this sequence can be interpreted as a pseudo-temporal signal. The corresponding segmentation outputs of LKF-SegNet form a time series of probability masks.

A classical Temporal Kalman Filter is then applied per pixel across this sequence. The filter:

- reduces frame-to-frame flicker by smoothing noisy probability variations,
- enforces temporal consistency across neighboring patches, and
- provides one-step-ahead predictions of the water mask for the next patch.

This usage aligns with the prediction–correction behavior described in the echocardiography tracking study [4], but transplanted into a synthetic remote-sensing context.

D. Boundary-Aware Loss Functions

Standard segmentation networks trained with plain binary cross-entropy often suffer from blurry or uncertain boundaries, especially in regions where the foreground and background share similar spectral characteristics. This problem is acute at water–land interfaces, where subtle gradients, foam, and vegetation can confuse the model.

Miao et al. (2018) addressed this challenge through an *Edge-Weighted Loss*, which increases the penalty around object

boundaries and improves precision at region edges [3]. The method typically:

- detects boundaries using an edge detector such as Canny,
- computes a distance transform from the boundary, and
- uses this distance to define a spatial weight map that emphasizes border pixels.

The present work implements a similar strategy in the Edge-Weighted Loss used to train LKF-SegNet:

- edges are detected on the ground-truth mask via Canny,
- a distance transform is applied to create a smooth weighting field, and
- this weight multiplies the per-pixel binary cross-entropy loss.

As a result, pixels near the coastline receive higher loss contributions, encouraging the network to allocate more capacity to learning precise water boundaries.

III. Data Preparation

We utilized a satellite **Water Bodies Dataset** containing RGB images and corresponding binary masks.

- **Preprocessing:** Images were resized to 256×256 pixels to match the input requirements of the lightweight architecture.
- **Augmentation:** To improve robustness, we applied random horizontal flips, vertical flips, and normalization during training using the Albumentations library.
- **Synthetic Video Generation:** Since labeled drone video was unavailable, we developed a "Synthetic Flight" script. This script pans a 256×256 window across large static satellite maps to create a video-like stream with a known ordering, allowing for temporal Kalman filtering experiments.

IV. Model Architecture

A. High-Level Overview

The proposed system is divided into two tightly coupled modules:

1) **Module 1 – LKF-SegNet (Spatial Segmentation Engine):** a lightweight encoder–decoder network that produces high-quality water-body masks from single satellite images. It consists of a MobileNetV2 encoder, a U-Net style decoder, and a Spatial Learnable Kalman Filter embedded at the bottleneck to denoise latent feature maps. Training is guided by an edge-weighted loss to sharpen water–land boundaries.
2) **Module 2 – Temporal Kalman Smoother and Predictor:** a classical scalar Kalman filter applied per pixel across a synthetic sequence of patches. It smooths the per-frame probability masks generated by Module 1 and produces one-step-ahead predictions of the segmentation mask.

Together, these modules provide both spatially refined and temporally consistent water-body extraction.

B. Module 1: LKF-SegNet (Spatial Segmentation Engine)

1) Encoder–Decoder Backbone: Let $\mathbf{I} \in \mathbb{R}^{3\times256\times256}$ denote an RGB satellite image. LKF-SegNet uses the U-Net implementation from the `segmentation_models_pytorch` library with a MobileNetV2 backbone:

- The **encoder** $E(\cdot)$ is a sequence of inverted residual blocks with depthwise separable convolutions. It gradually reduces spatial resolution while increasing channel depth and outputs a hierarchy of feature maps
$$\{\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \ldots, \mathbf{F}^{(L)}\}, \tag{1}$$
where $\mathbf{F}^{(L)} \in \mathbb{R}^{C_b\times H_b\times W_b}$ is the deepest (bottleneck) feature map.
- The **decoder** $D(\cdot)$ mirrors the encoder with upsampling blocks. At each stage, it concatenates the upsampled feature map with the corresponding encoder feature via skip connections:
$$\mathbf{G}^{(k)} = \mathrm{Up}(\mathbf{G}^{(k+1)}) \,\|\, \mathbf{F}^{(k)}, \tag{2}$$
where $\|$ denotes channel-wise concatenation. This preserves fine spatial details such as narrow rivers and coastline structure.

The final segmentation head is a 1×1 convolution:

$$\mathbf{L} = \mathrm{Conv}_{1\times1}(\mathbf{G}^{(1)}), \qquad \mathbf{P} = \sigma(\mathbf{L}), \tag{3}$$

where $\mathbf{L}$ are the logits, $\mathbf{P} \in [0,1]^{1\times256\times256}$ is the probability mask, and $\sigma(\cdot)$ is the sigmoid function.

2) Spatial Learnable Kalman Filter at the Bottleneck: The key architectural novelty is the insertion of a Spatial Learnable Kalman Filter (LKF) at the bottleneck. Instead of passing $\mathbf{F}^{(L)}$ directly to the decoder, we first transform it via a Kalman-like recursive estimator:

$$\tilde{\mathbf{F}}^{(L)} = \mathrm{LKF}\big(\mathbf{F}^{(L)}\big), \tag{4}$$

and use $\tilde{\mathbf{F}}^{(L)}$ as the deepest feature for the decoder.

Let $\mathbf{F}^{(L)} \in \mathbb{R}^{B\times C_b\times H_b\times W_b}$ for a batch of size B. We first permute the tensor to group rows as a pseudo-temporal dimension:

$$\mathbf{X} = \mathrm{permute}\big(\mathbf{F}^{(L)}\big) \in \mathbb{R}^{B\times H_b\times W_b\times C_b}. \tag{5}$$

For a fixed batch index and width position (b, w), the sequence

$$\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{H_b}\}, \qquad \mathbf{z}_t \in \mathbb{R}^{C_b}, \tag{6}$$

corresponds to the feature vectors along the vertical direction (rows). The LKF maintains, for each (b, w), a state vector $\mathbf{x}_t \in \mathbb{R}^{C_b}$ and a diagonal covariance proxy $\mathbf{P}_t \in \mathbb{R}^{C_b}$.

Instead of fixed transition and observation matrices, the filter uses small neural networks to produce *learnable* parameters:

$$\mathbf{F}_t = \tanh\big(W_F\mathbf{x}_{t-1} + \mathbf{b}_F\big), \tag{7}$$
$$\mathbf{Q}_t = 0.1 \cdot \sigma\big(W_Q\mathbf{x}_{t-1} + \mathbf{b}_Q\big), \tag{8}$$
$$\mathbf{H}_t = \tanh\big(W_H\mathbf{z}_t + \mathbf{b}_H\big), \tag{9}$$
$$\mathbf{R}_t = \sigma\big(W_R\mathbf{z}_t + \mathbf{b}_R\big), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid and W_F, W_Q, W_H, W_R and $\mathbf{b}_F, \ldots, \mathbf{b}_R$ are learnable parameters.

For each row index $t = 1, \ldots, H_b$, the **prediction step** is

$$\mathbf{x}_{t|t-1} = \mathbf{F}_t \odot \mathbf{x}_{t-1|t-1}, \tag{11}$$

$$\mathbf{P}_{t|t-1} = \mathbf{P}_{t-1|t-1} \odot \mathbf{F}_t^2 + \mathbf{Q}_t, \tag{12}$$

and the **update step** with observation $\mathbf{z}_t$ is

$$\mathbf{y}_t = \mathbf{z}_t - \mathbf{H}_t \odot \mathbf{x}_{t|t-1}, \tag{13}$$

$$\mathbf{S}_t = \mathbf{P}_{t|t-1} \odot \mathbf{H}_t^2 + \mathbf{R}_t, \tag{14}$$

$$\mathbf{K}_t = \frac{\mathbf{P}_{t|t-1} \odot \mathbf{H}_t}{\mathbf{S}_t + \varepsilon}, \tag{15}$$

$$\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + \mathbf{K}_t \odot \mathbf{y}_t, \tag{16}$$

$$\mathbf{P}_{t|t} = (1 - \mathbf{K}_t \odot \mathbf{H}_t) \odot \mathbf{P}_{t|t-1}, \tag{17}$$

where $\odot$ denotes element-wise multiplication and ε is a small constant for numerical stability.

The output state $\mathbf{x}_{t|t}$ for all rows t is stacked back into a tensor of shape $B \times H_b \times W_b \times C_b$ and permuted to

$$\tilde{\mathbf{F}}^{(L)} \in \mathbb{R}^{B \times C_b \times H_b \times W_b}, \tag{18}$$

which replaces the original bottleneck feature map in the decoder. Intuitively, if a particular row is corrupted by noise or shadows, the innovation $\mathbf{y}_t$ becomes large and the learned observation noise $\mathbf{R}_t$ can increase, reducing the Kalman gain and causing the filter to rely more heavily on the prediction from the previous row. This yields smoother, vertically coherent feature maps that help the decoder reconstruct cleaner masks.

3) Edge-Weighted Loss for Boundary Refinement: Network predictions are trained against binary ground-truth masks using an edge-weighted binary cross-entropy loss. Let $p_i \in (0, 1)$ be the predicted probability and $y_i \in \{0, 1\}$ the ground truth at pixel i. The base loss is

$$\text{BCE}(p_i, y_i) = -\big[y_i \log p_i + (1 - y_i) \log(1 - p_i)\big]. \tag{19}$$

To emphasize water–land boundaries, a spatial weight map w_i is constructed as follows:

1) Apply the Canny detector to the ground-truth mask to obtain a binary edge map.
2) Compute the Euclidean distance transform $\text{dist}(i)$ from each pixel to the nearest edge.
3) Define

$$w_i = 1 + w_0 \exp\left(-\frac{\text{dist}(i)}{\tau}\right), \tag{20}$$

where w_0 controls the maximum up-weighting (e.g. $w_0 \approx 4$) and τ controls the decay rate.

The final loss is

$$\mathcal{L}_{\text{edge}} = \frac{1}{N} \sum_{i=1}^{N} w_i \, \text{BCE}(p_i, y_i), \tag{21}$$

where N is the number of pixels. Pixels near the coastline (small $\text{dist}(i)$) receive weights close to $1 + w_0$, leading to larger gradients and encouraging the network to allocate more capacity to precise boundary modeling.

C. Module 2: Temporal Kalman Smoother and Predictor

1) Synthetic Patch Sequence: Module 2 operates on a sequence of overlapping patches extracted from a single large satellite image. Let $\mathbf{I}_{\text{large}}$ denote a high-resolution RGB tile. Using a sliding window of size $H_p \times W_p$ and stride s, we obtain an ordered sequence of patches:

$$\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}. \tag{22}$$

Each patch is passed through LKF-SegNet to produce a probability mask

$$\mathbf{Z}_t = \text{LKF-SegNet}(\mathbf{I}_t) \in [0, 1]^{H_p \times W_p}, \tag{23}$$

and the collection $\{\mathbf{Z}_t\}_{t=1}^{T}$ is treated as a pseudo-temporal sequence.

2) Per-Pixel Scalar Kalman Filter: For each spatial location (u, v), the values $\{z_t(u, v)\}_{t=1}^{T}$ form a one-dimensional time series. A scalar Kalman filter is used to estimate a latent "true" water probability $x_t(u, v)$:

$$x_t = x_{t-1} + w_t, \qquad w_t \sim \mathcal{N}(0, Q), \tag{24}$$

$$z_t = x_t + v_t, \qquad v_t \sim \mathcal{N}(0, R), \tag{25}$$

where Q and R are the process and observation variances, respectively.

The recursive equations are:

$$x_{t|t-1} = x_{t-1|t-1}, \tag{26}$$

$$P_{t|t-1} = P_{t-1|t-1} + Q, \tag{27}$$

$$K_t = \frac{P_{t|t-1}}{P_{t|t-1} + R}, \tag{28}$$

$$x_{t|t} = x_{t|t-1} + K_t\big(z_t - x_{t|t-1}\big), \tag{29}$$

$$P_{t|t} = (1 - K_t) P_{t|t-1}. \tag{30}$$

Here, $x_{t|t}$ denotes the filtered estimate at time t, while $x_{t+1|t}$ (obtained by applying the prediction step one more time) serves as a one-step-ahead forecast of the water probability for the next patch. In implementation, these equations are applied independently at each pixel location, resulting in:

- a filtered mask sequence $\{\mathbf{X}_{t|t}\}$ with reduced frame-to-frame flicker, and
- a predicted mask sequence $\{\mathbf{X}_{t|t-1}\}$ representing one-step-ahead expectations.

Although the temporal dimension is synthetic, this configuration demonstrates how classical state estimation can enhance temporal stability of neural segmentation outputs and enable prediction without explicit recurrent neural networks.

V. Experimental Results

This section evaluates the performance of the proposed Dual-Stage Kalman Framework through both (1) spatial segmentation experiments using LKF-SegNet (Module 1) and (2) temporal smoothing and prediction experiments using the Temporal Kalman Filter (Module 2). We present quantitative metrics, qualitative visualizations, an IoU distribution analysis, and the temporal MAE behaviour to demonstrate how each module contributes to the overall robustness of the system.

A. Model Architecture Diagrams

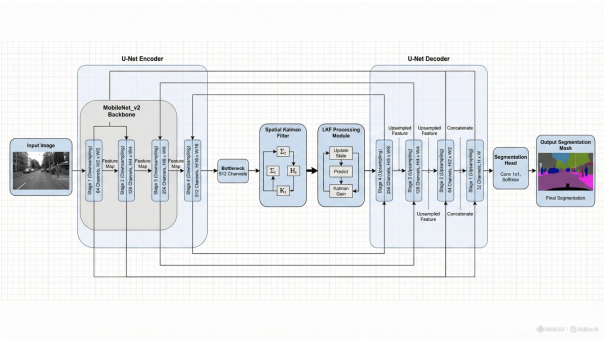

Fig. 1. **Overview of the LKF-SegNet Architecture.** The encoder consists of a MobileNetV2 backbone, followed by a Spatial Kalman Filter at the bottleneck. The decoder reconstructs refined feature maps into a segmentation mask using U-Net style skip connections.

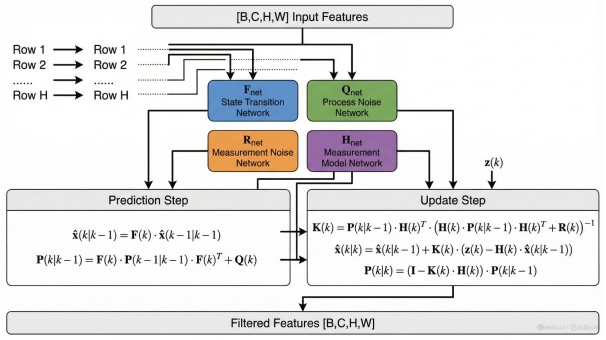

Fig. 2. **Row-wise Learnable Kalman Filter Mechanism.** Each row of the bottleneck feature map is processed as a pseudo-temporal sequence. Learnable networks ($F_{\text{net}}, Q_{\text{net}}, R_{\text{net}}, H_{\text{net}}$) generate transition and noise matrices for the prediction–update loop.

B. Spatial Segmentation Performance (Module 1)

The LKF-SegNet model was trained on satellite water-body imagery and achieved strong segmentation performance on the test set. The integration of the Spatial Learnable Kalman Filter improved feature stability, while the Edge-Weighted Loss sharpened water boundaries that are otherwise difficult to detect due to shadows, vegetation, and thin shoreline structures.

Quantitative performance is summarized in Table I. The results indicate:

- **Accuracy:** Train = 91.08%, Test = 88.94%
- **IoU:** Train = 78.85%, Test = 74.87%
- **Precision:** Train = 89.84%, Test = 88.87%
- **Recall:** Train = 86.57%, Test = 82.61%

These metrics reflect strong generalization and minimal performance drop from train to test, demonstrating that LKF-SegNet avoids overfitting while maintaining high-quality segmentation.

TABLE I
FINAL PERFORMANCE METRICS FOR LKF-SEGNET (TRAIN AND TEST SPLITS)

Metric	Train Value	Test Value
Accuracy	91.08%	88.94%
IoU	78.85%	74.87%
Precision	89.84%	88.87%
Recall	86.57%	82.61%

C. IoU Distribution Analysis

A histogram of per-image IoU values reveals that most test images fall in the 0.70–0.90 IoU range, showing consistent segmentation quality. A small number of difficult cases fall below 0.40 due to cloud occlusions, shallow water, or noisy ground truth, while several near-perfect segmentations (IoU > 0.90) occur on clean, high-contrast water bodies. This distribution confirms that the majority of predictions remain stable and accurate across varying landscapes.

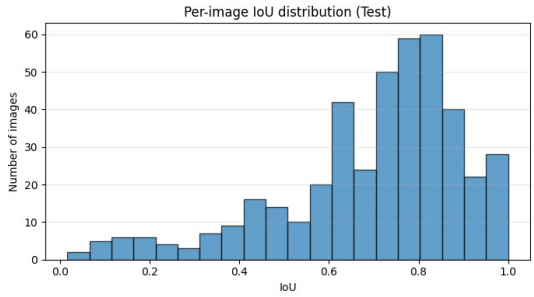

Fig. 3. **Per-image IoU Distribution (Test Set).** Shows how segmentation quality varies across the dataset, with most images achieving high IoU scores.

D. Qualitative Segmentation Results

To visualize the improvements provided by the Spatial Kalman Filter and Edge-Weighted Loss, segmentation outputs were compared against ground truth (Fig. 4). Results indicate that shallow or low-contrast water regions are segmented more consistently, pixel noise in vegetation areas is significantly reduced, and coastline boundaries are noticeably sharper.

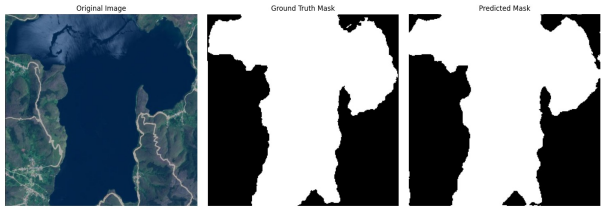

Fig. 4. **Sample Segmentation Result.** From left: original satellite patch, ground-truth mask, predicted mask from LKF-SegNet.

E. Train vs Test Metrics Visualization

A bar chart comparing key segmentation metrics (Accuracy, IoU, Precision, Recall) shows that the gap between train and test values is small, indicating good generalization.

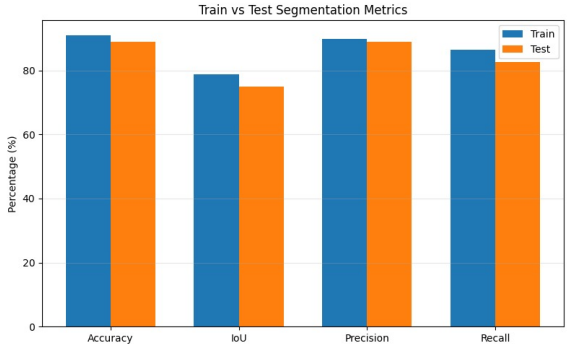

Fig. 5. **Train vs Test Segmentation Metrics.** Comparison across Accuracy, IoU, Precision, and Recall.

F. Temporal Filtering and Prediction Performance (Module 2)

Using synthetic patch-based video sequences, the Temporal Kalman Filter was applied to the sequence of probability maps produced by LKF-SegNet. Results show that the filtered output is consistently smoother than raw segmentation probabilities, and one-step-ahead predictions approximate the next-frame mask with reasonable accuracy. The temporal MAE analysis shows that the filter reduces frame-to-frame variance and tracks rising/falling error trends while remaining stable even when observations fluctuate.

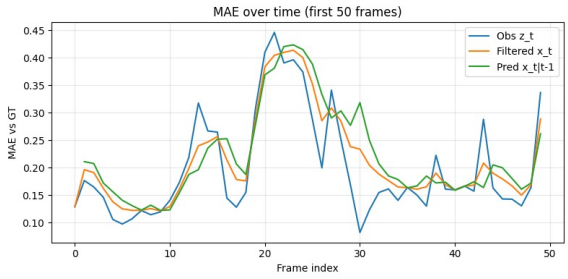

Fig. 6. **Temporal MAE Over First 50 Frames.** Comparison of observed mask (z_t), filtered estimate (x_t), and predicted ($x_{t|t-1}$) values.

G. Filtered vs Predicted Temporal Outputs

A detailed view of a representative frame demonstrates:

- The raw probability mask (z_t) contains spatial noise.
- The filtered mask (x_t) is more coherent and less patchy.
- The prediction ($x_{t|t-1}$) is smoother still, useful for forecasting applications.

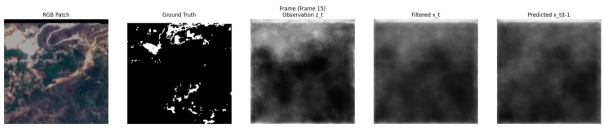

Fig. 7. **Temporal Filtering Outputs.** From left: RGB patch, ground truth, observation z_t, filtered state x_t, predicted state $x_{t|t-1}$.

H. Experimental Findings

The experiments demonstrate that:

- The Spatial Kalman Filter significantly enhances segmentation stability inside the network.
- Edge-Weighted Loss produces sharper and more accurate boundaries.
- The Temporal Kalman Filter improves consistency across frames and provides meaningful predictions.
- The system remains computationally lightweight enough for drone deployment.

Overall, the proposed dual-stage framework offers a powerful, efficient, and interpretable approach to water-body monitoring.

VI. Discussion

The proposed Dual-Stage Kalman Framework integrates deep learning with classical signal processing to address the long-standing challenges of water-body extraction from satellite imagery.

A. Strengths of the Proposed Framework

1) Integration of Learning-Based and Model-Based Reasoning: One of the most important strengths of the system is its hybrid nature. By combining these paradigms, the CNN provides rich spatial features and boundary refinement, the Spatial Kalman Filter improves feature stability and enforces structural consistency, and the Temporal Kalman Filter provides sequence-level smoothing and prediction. This hybridization enables the model to achieve a balance between flexibility and stability.

2) Robustness and Temporal Consistency: The Spatial Learnable Kalman Filter is effective in scenarios with shadows and noisy textures by suppressing high-frequency noise and preserving structural transitions. In practice, this results in fewer spurious detections. Furthermore, the Edge-Weighted Loss ensures that the network prioritizes learning the geometry of coastlines. Regarding temporal consistency, the synthetic temporal sequence experiment demonstrates that the classical Kalman prediction–correction loop reduces flicker, produces smoothed probability masks, and generates viable one-step predictions for future frames.

B. Interpretation of Results in Context of Literature

The observed performance trends align with prior research:

- **Boundary refinement**: Consistent with Miao et al. [3], edge-weighted loss helps significantly with water-land boundaries.
- **Spatial denoising via row-wise filtering**: Matches the findings of Liao et al. [1], where learnable Kalman filters improve coherence in agricultural segmentation.
- **Temporal smoothing**: Similar to the echocardiography work of Bersvendsen et al. [4], temporal Kalman filters stabilize noisy observations and maintain tracking under missing measurements.

VII. Conclusion

This project developed a dual-stage Kalman framework for agricultural water-body extraction. By integrating a Spatial Learnable Kalman Filter into a lightweight U-Net and employing an edge-weighted loss, we achieved high-precision segmentation suitable for static imagery. The addition of a Temporal Kalman Filter on synthetic sequences demonstrated that classical state estimation can enhance temporal consistency and enable next-state prediction even in the absence of true video data. Overall, the system bridges the gap between deep learning and classical signal processing, offering a robust and computationally efficient foundation for future work in automated water resource monitoring.

References

[1] D. Liao, J. Sun, Z. Deng, Y. Zhao, J. Zhang, and D. Ou, "A Lightweight Network for Water Body Segmentation in Agricultural Remote Sensing Using Learnable Kalman Filters and Attention Mechanisms," *Applied Sciences*, vol. 15, no. 11, p. 6292, 2025.

[2] M. Wieland, S. Martinis, R. Kiefl, and V. Gstaiger, "Semantic segmentation of water bodies in very high-resolution satellite and aerial images," *Remote Sensing of Environment*, vol. 287, p. 113452, 2023.

[3] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, "Automatic Water-Body Segmentation From High-Resolution Satellite Images via Deep Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 602–606, 2018.

[4] J. Bersvendsen *et al.*, "Automated Segmentation of the Right Ventricle in 3D Echocardiography: A Kalman Filter State Estimation Approach," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 42–51, 2016.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015.

[6] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, pp. 1425–1432, 1996.