Detection and Segmentation for Optical Remote Sensing: From Satellites to UAVs

Kangning Cui

Department of Mathematics
City University of Hong Kong

January 10, 2026

Outline

1. Introduction

2. Unsupervised Hyperspectral Image Clustering

3. Change Detection for ASGM Monitoring

4. UAV-Based Palm Localization and Spatial Analysis

5. Fetal Heart Tracking in Ultrasound Videos

6. Conclusion

Introduction

Fundamentals: Optical vs. Non-Optical Sensing

Optical Sensing (Our Focus)

- **Principle:** Detects reflected sunlight across spectral bands [1].
- **Provides:** Rich spectral and spatial detail.

Non-Optical Sensing

- **Principle:** Emits its own signal (SAR) or detects heat (Thermal) [2].
- **Provides:** Surface structure, moisture, and temperature.

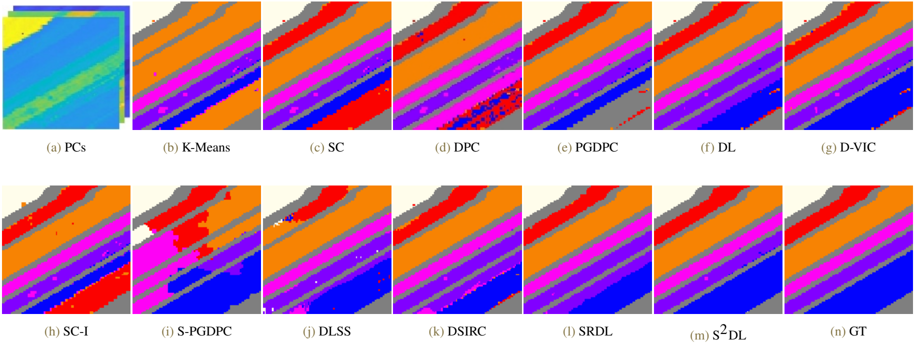
Rationale: Why This Thesis Focuses on Optical Sensing

This thesis employs optical sensing because its research questions require the rich spectral and spatial data needed to identify species, classify crops, and monitor ecological changes.

Main Goal

To design and implement computationally efficient, interpretable, and scalable frameworks for unsupervised clustering, change detection, and object localization in complex remote sensing data.

Contribution 1: S^{2}DL for HSI Clustering

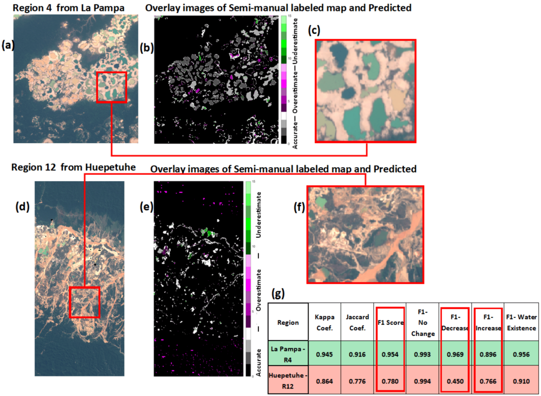

(a) PCs (b) K-Means (c) SC (d) DPC (e) PGDPC (f) DL (g) D-VIC

(h) SC-I (i) S-PGDPC (j) DLSS (k) DSIRC (l) SRDL (m) S^{2}DL (n) GT

Figure: Comparison of clustering results on the Salinas A dataset.

S^{2}DL successfully integrates spatial information to produce clean, accurate clusters that align with the ground truth for hyperspectral images (HSIs).

Contribution 2: Monitoring Small-Scale Gold Mining

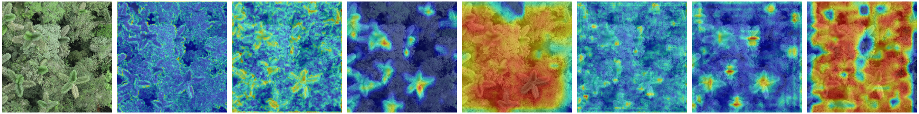

Region 4 from La Pampa
Overlay images of Semi-manual labeled map and Predicted
(a)
(b)
(c)
Accurate — Overestimate — Underestimate
Region 12 from Huepetuhe
Overlay images of Semi-manual labeled map and Predicted
(d)
(e)
(f)
Accurate — Overestimate — Underestimate
(g)

Region	Kappa Coef.	Jaccard Coef.	F1 Score	F1- No Change	F1- Decrease	F1- Increase	F1- Water Existence
La Pampa - R4	0.945	0.916	0.954	0.993	0.969	0.896	0.956
Huepetuhe - R12	0.864	0.776	0.780	0.994	0.450	0.766	0.910

Figure: Pond dynamics in two ASGM regions. Performance varies with mining typology.

E-ReCNN effectively detects land and water changes in the Peruvian Amazon using Sentinel-2 data, with results varying by mining typology.

Contribution 3: Palm Detection in UAV Imagery



(a) Input Image (b) Layer 1 (c) Layer 4 (d) Layer 7 (e) Layer 10 (f) Layer 16 (g) Layer 19 (h) Layer 22

Figure: Hierarchical Feature Learning in YOLOv10 through Grad-CAM Visualizations. Early layers capture low-level edges, intermediate layers integrate spatial context, while deep layers specialize in object-level features.

PRISM is a framework for detecting, mapping, and segmenting palm crowns in high-resolution UAV imagery. It further integrates ecological models to analyze the spatial patterns of these populations.

Unsupervised Hyperspectral Image Clustering

The Power and Problem of HSIs

The Power: Rich Spectral Information

- HSIs capture data across hundreds of spectral bands, enabling precise characterization of surface materials based on spectral signatures [3].
- Essential for applications like land cover classification, spectral unmixing, and environmental monitoring [4–6].

The Problem: The Need for Labels

- State-of-the-art deep learning methods are often **supervised**, requiring large amounts of expert-annotated training data.
- Acquiring ground truth for HSIs is expensive, time-consuming, and requires specialized expertise [7].
- This bottleneck drives the critical need for **unsupervised clustering** methods that can work without labels.

Key Challenges in Unsupervised HSI Clustering

Data Complexity

- **High Dimensionality:** The "curse of dimensionality" where 200+ bands lead to model overfitting and computational challenges [3].
- **Large Spatial Extent:** Scenes often exceed 10^6 pixels, making methods with quadratic complexity (like graph clustering) infeasible.

Data Quality

- **Sensor Noise:** Low SNR, especially in short-wave infrared bands, degrades data quality [8].
- **Spectral Variability:** Atmospheric effects and illumination changes cause the same material to have different spectral signatures (intra-class variability).

Core Issue

Treating each pixel independently ignores the fact that nearby pixels are often the same material. Any robust solution must leverage this spatial context.

Background: The Importance of Spatial Context

Key Insight

In HSI, neighboring pixels are highly correlated. Exploiting this spatial information is crucial for accurate clustering and noise reduction [3, 9].

Method 1: Spatially Regularized Graphs

- Edges are restricted to connect spatially nearby pixels [10].
- This encodes spatial coherence directly into the clustering method.

Method 2: Superpixel Segmentation

- Groups similar pixels into small, spatially closed regions [11].
- Reduces computational cost by working on superpixels.

Our Approach

S^{2}DL is novel in its integration of **both** superpixel segmentation **and** a spatially regularized graph within a diffusion learning framework.

Our Solution: The S^{2}DL Framework

A Three-Stage Approach: Superpixel Segmentation → Reduced Spatially Regularized Graph Construction → Diffusion-Based Clustering

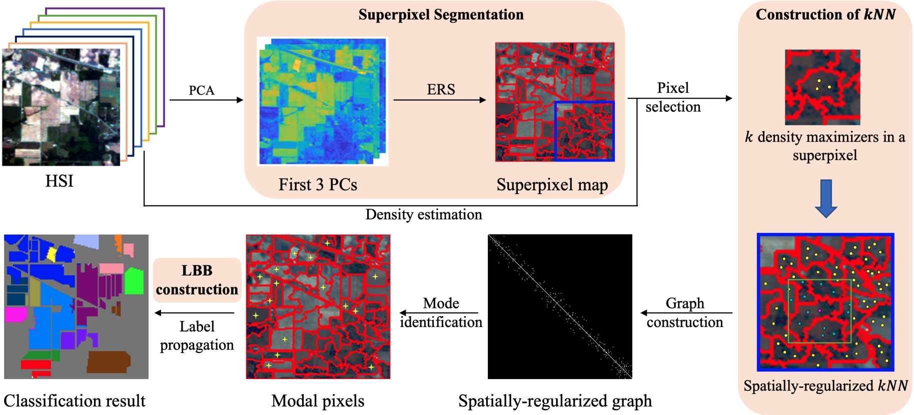


Figure: The S^{2}DL workflow [4].

Stage 1: Superpixel Segmentation & Representative Selection

Step 1.1: ERS Superpixel Segmentation

We first partition the HSI into N_s spatially coherent superpixels by optimizing an objective function on an image graph [11]. This balances segment compactness with size uniformity:

$$\max_{A} \mathcal{J}(A) = \underbrace{\mathcal{H}(A)}_{\text{Entropy Rate}} + \alpha \underbrace{\mathcal{B}(A)}_{\text{Balancing Term}}$$

where $\mathcal{H}(A)$ is the entropy rate (promoting compactness) and $\mathcal{B}(A)$ is a balancing term (promoting size uniformity).

Stage 1: Superpixel Segmentation & Representative Selection

Step 1.2: Representative Selection

Construct a reduced set $X_s \subset X$ by selecting the top k pixels from each superpixel S_j that maximize a Kernel Density Estimate (KDE) $\zeta(x)$:

$$\zeta(x) = \sum_{y \in k_n(x)} \exp\left(-\frac{||x-y||_2^2}{\sigma_0^2}\right)$$

$$X_s = \bigcup_{j=1}^{N_s} \underset{x \in S_j}{\operatorname{argmax}_k}(\zeta(x))$$

This reduces $|X|$ from N to $|X_s| = k \cdot N_s$.

Stage 2: Spatially Regularized Graph & Diffusion

Step 2.1: Spatially Regularized Graph Construction

Define a graph $G = (X_s, E_s)$ with an adjacency matrix $\mathbf{W} \in \{0,1\}^{|X_s|\times|X_s|}$. The spatial regularization is encoded directly into the connectivity:

$$\mathbf{W}_{ij} = \begin{cases} 1, & \text{if } x_j \in kNN(x_i) \text{ and } \mathrm{dist}_{\mathrm{spatial}}(i,j) \leq R \\ 0, & \text{otherwise} \end{cases}$$

This ensures edges only connect pixels that are close in both spectral and spatial domains.

Stage 2: Spatially Regularized Graph & Diffusion

Step 2.2: The Diffusion Distance

Given the row-normalized transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, the diffusion distance D_t is defined. It reveals the manifold geometry by averaging all paths of length t between nodes [12].

Definition:

$$D_t(x_i, x_j)^2 = \sum_{l=1}^{|X_s|} \frac{(p_t(i,l) - p_t(j,l))^2}{\pi_l}$$

where $p_t(i,l)$ is the probability of transitioning from i to l in t steps.

Computation (via Eigendecomposition):

$$D_t(x_i, x_j)^2 = \sum_{l=1}^{|X_s|} \lambda_l^{2t} \left[\psi_l(i) - \psi_l(j)\right]^2$$

where (λ_l, ψ_l) are the eigenpairs of $\mathbf{P}$.

Stage 3: Diffusion-Based Clustering

Step 3.1: Identify Cluster Modes

The set of K cluster modes, $\{x_{m_k}\}_{k=1}^{K}$, are the top K maximizers of the decision value $\Delta_t(x)$ over the representative set X_s.

$$\{x_{m_k}\}_{k=1}^{K} := \underset{x\in X_s}{\operatorname{argmax}_K} \left(\Delta_t(x)\right)$$

where the decision value is the product of local density $\zeta(x)$ and diffusion distance to the nearest higher-density point $d_t(x)$:

$$\Delta_t(x) = \zeta(x) \cdot d_t(x), \quad \text{with} \quad d_t(x) = \min_{y\in X_s:\zeta(y)>\zeta(x)} D_t(x,y)$$

Each mode is assigned a unique initial label, $\hat{C}(x_{m_k}) = k$.

Stage 3: Diffusion-Based Clustering

Step 3.2: Propagate Labels and Finalize

For non-modal points $x \in X_s$, labels are propagated iteratively in descending order of density $\zeta(x)$ according to the rule:

$$\hat{C}(x) := \hat{C}(x^*), \quad \text{where} \quad x^* = \mathop{\arg\min}_{\substack{y \in X_s \\ \zeta(y) \geq \zeta(x) \\ \hat{C}(y) > 0}} D_t(x, y)$$

The final class assignment $C(y)$ for any pixel y in a superpixel S_j is determined by a majority vote over the labeled representative pixels within that superpixel:

$$C(y) := \mathop{\arg\max}_{l \in \{1,\ldots,K\}} \left| \{x \in S_j \cap X_s \mid \hat{C}(x) = l\} \right|, \quad \forall y \in S_j$$

Numerical Results on Benchmark Datasets

Table: Performance Comparison on Four HSI Datasets.

Dataset		**Method**											
		K-Means	SC	DPC	PGDPC	DL	D-VIC	SC-I	S-PGDPC	DLSS	DSIRC	SRDL	S^{2}DL
Salinas A	OA	0.764	0.841	0.786	0.844	0.887	0.976	0.827	0.647	0.890	0.911	0.895	**0.996**
	AA	0.749	0.887	0.849	0.893	0.920	0.973	0.875	0.680	0.888	0.903	0.926	**0.996**
	κ	0.703	0.806	0.740	0.813	0.860	0.970	0.789	0.568	0.862	0.889	0.870	**0.995**
	Sum	2.216	2.534	2.375	2.550	2.667	2.919	2.491	1.895	2.640	2.703	2.691	**2.987**
	RT	0.05	1.59	2.66	1.63	1.93	4.89	6.43	0.10	5.27	26.39	14.99	1.78
Indian Pines	OA	0.386	0.382	0.391	0.428	0.404	0.471	0.496	0.477	0.467	0.620	0.640	**0.647**
	AA	0.398	0.368	0.376	0.399	0.401	0.376	0.304	0.530	0.462	0.549	0.553	**0.591**
	κ	0.315	0.313	0.304	0.351	0.313	0.383	0.394	0.431	0.400	0.573	0.596	**0.602**
	Sum	1.099	1.063	1.071	1.178	1.118	1.230	1.194	1.438	1.329	1.742	1.789	**1.840**
	RT	1.14	14.40	13.10	19.75	13.64	24.33	70.43	0.38	20.55	136.02	30.52	2.19
Salinas	OA	0.639	0.662	0.668	–	0.687	0.696	–	0.590	0.702	0.677	0.834	**0.889**
	AA	0.612	0.633	0.654	–	0.662	0.623	–	0.487	0.674	0.612	0.756	**0.776**
	κ	0.597	0.620	0.627	–	0.646	0.653	–	0.551	0.662	0.633	0.813	**0.876**
	Sum	1.848	1.915	1.949	–	1.995	1.972	–	1.628	2.038	1.922	2.403	**2.541**
	RT	4.81	414.44	432.98	–	450.82	496.37	–	1.20	504.88	3059.74	445.31	8.80
WHU	OA	0.625	0.743	**0.857**	–	0.857	0.779	–	0.364	0.837	0.829	0.771	0.822
	AA	0.487	0.507	0.540	–	0.540	0.468	–	0.515	0.523	0.415	0.480	**0.675**
	κ	0.545	0.674	**0.810**	–	0.810	0.710	–	0.276	0.784	0.764	0.698	0.766
	Sum	1.657	1.924	2.207	–	2.207	1.957	–	1.155	2.144	2.008	1.948	**2.263**
	RT	13.96	1896.55	1881.26	–	1851.27	1965.99	–	2.50	2059.11	9755.97	2881.06	15.46

Key Takeaway

S^{2}DL consistently achieves the best composite score while being quick.

Ablation Study: When Spatial Regularization Fails

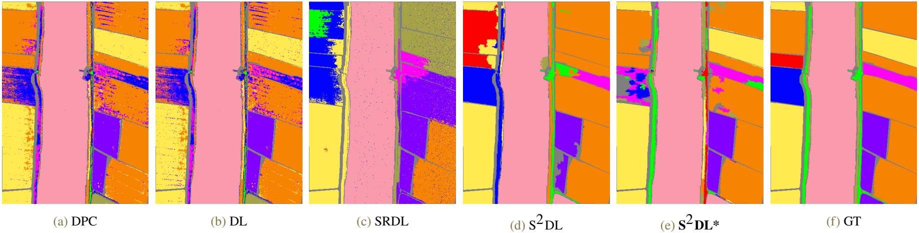
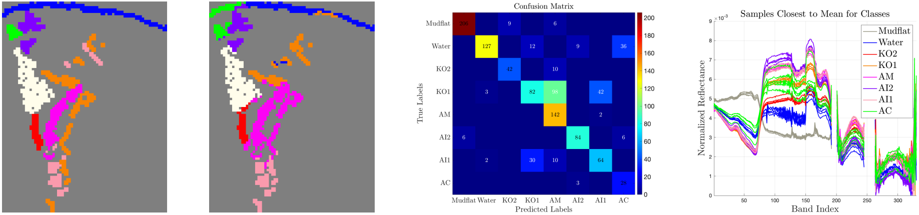

(a) DPC (b) DL (c) SRDL (d) S^2DL (e) **S^2DL*** (f) GT

Figure: Clustering on WHU-Hi. S^2DL with regularization (d) struggles with fragmented classes, while S^2DL* without regularization (e) succeeds.

- **The Trade-off:** Spatial regularization typically boosts performance by reducing noise, but may fail when a single class is spatially fragmented.
- **The Solution:** Our variant, **S^2DL***, removes spatial regularization to allow clustering based purely on spectral similarity across the entire image.
- **The Result:** S^2DL* successfully groups the fragmented regions, yielding a dramatic performance increase (OA: **+7.9%**, κ: **+10.4%**).

Real-World Application: Mangrove Mapping in Hong Kong



	Mudflat	Water	KO2	KO1	AM	AI2	AI1	AC
Mudflat	206		9		6			
Water		127		12		9		36
KO2			42		10			
KO1		3		82	98		42	
AM					142		2	
AI2	6					84		6
AI1		2		30	10		64	
AC						3		28

(a) Ground Truth (GT)

(b) S^{2}DL Result

(c) Confusion Matrix

(d) Mean Spectral Signatures

Figure: S^{2}DL effectively separates the six dominant mangrove species. The confusion matrix confirms high accuracy, though some spectral similarity persists between certain classes.

Summary of Results

Data from Gaofen-5 satellite HSI (330 bands, 30m resolution). S^{2}DL achieved the best performance compared to all competing methods, demonstrating its strong potential for unsupervised species mapping in complex ecological environments.

Unsupervised HSI Clustering Conclusion

Summary of Contribution

- Introduced **S^{2}DL**, a novel unsupervised framework integrating superpixel segmentation, spatial regularization, and diffusion geometry.
- Its core strategy—clustering a reduced graph of representative pixels—achieves both high accuracy and computational efficiency.
- Validated with state-of-the-art (SOTA) results on four benchmark datasets and a challenging mangrove mapping application.

Future Work

- Automated hyperparameter selection based on intrinsic data properties.
- Designing superpixel algorithms tailored for the high dimensionality of HSI data.
- Semi-supervised extensions where a few expert labels can guide the diffusion process to further boost accuracy.

Change Detection for ASGM Monitoring

The Challenge: Monitoring Artisanal Gold Mining (ASGM)

What is ASGM?

Artisanal and Small-Scale Gold Mining involves removing forest and disturbing alluvial sediments to extract gold. This practice leads to:

- Extensive deforestation (e.g., >120,000 ha in Madre de Dios, Peru by 2017 [13]).
- Creation of artificial ponds over time.
- Significant impacts on biogeochemistry and public health [14].

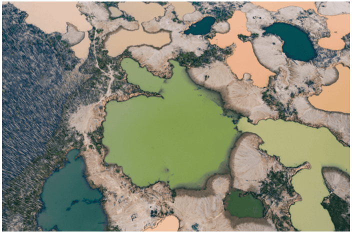


Figure: Mining ponds in La Pampa, Peru, with different status [15].

The Need for Automation

Automated change detection using remote sensing imagery is crucial for tracking these impacts, assessing policy, and guiding environmental management.

Our Approach: Dual Change Detection Strategies

This section explores automated methods for detecting land-cover changes from ASGM, focusing on the evolution of mining ponds.

Path 1: Semi-Supervised Learning

- **Goal:** Effective change detection with limited labeled data and computational resources.
- **Method:** Applies Support Vector Machine with Smoothed Total Variation (SVM-STV) [16].

Path 2: Supervised Learning

- **Goal:** Accurate detection of subtle, temporally-evolving features.
- **Method:** Introduces E-ReCNN, an extended Recurrent CNN architecture [15, 17].

Study Area & Dataset: Madre de Dios (MDD), Peru

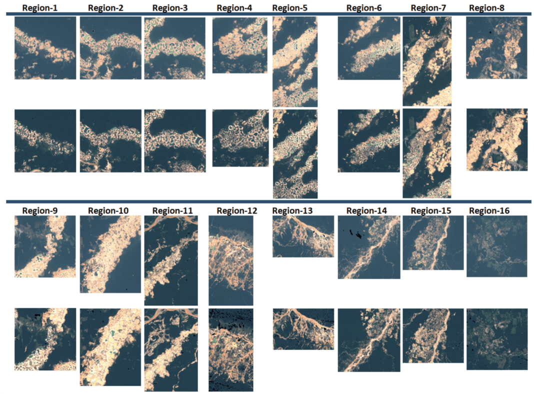


Figure: Bi-temporal Sentinel-2 imagery (2019 vs. 2021) showing ASGM impacts in La Pampa, MDD [15].

Region Focus:

- MDD, Peru: A global hotspot for ASGM activity.
- 16 sample regions (~70 km^2 each) selected in La Pampa.
- Captures varying mining intensity and policy enforcement.

Data Source:

- Sentinel-2 imagery via Google Earth Engine.
- Bi-temporal snapshots: Aug 2019 & Jul 2021.
- Preprocessing: Cloud masking, histogram matching.

Data Preparation & Change Classes

Image Datasets Created:

- RGB (3-channel): For visualization & efficient processing.
- Multispectral (6-channel): RGB + NIR + SWIR1 + SWIR2.
- Multispectral (10-channel): Additional bands.

Change Classes Derived from Pond States: (Active, Transition, Inactive)

1. Decrease (e.g., active → inactive)
2. Increase (e.g., inactive → active)
3. Water Existence/Absence
4. No Change

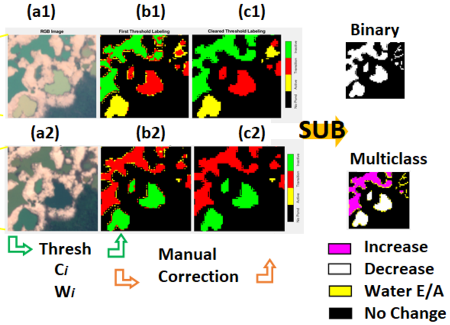


Figure: Semi-manual labeling process combining color index thresholding $Ci = \frac{(G-R)}{(G+R)}$ and MNDWI, followed by manual refinement [18].

Method 1: Semi-Supervised SVM-STV

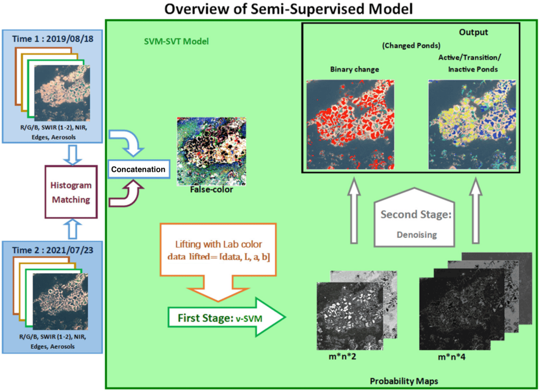


Figure: SVM-STV framework: Pixel-wise ν-SVM classification followed by STV spatial regularization [15].

Core Idea: Combines spectral classification with spatial refinement.

Stage 1: ν-SVM
Pixel-wise classification on concatenated bi-temporal spectral features.

Stage 2: STV
Smoothed Total Variation regularization applied to SVM probability maps.

Enhancement: Optional Lab color space "lifting" to boost feature discriminability.

Method 2: Supervised E-ReCNN

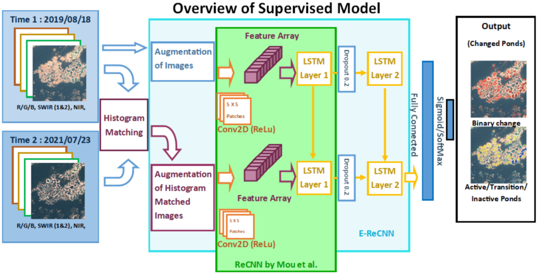


Figure: E-ReCNN architecture: CNN for spatial feature extraction, LSTM for temporal modeling [15].

Core Idea: Combines spatial and temporal features for joint analysis [17].

E-ReCNN Modifications:

- Additional LSTM layer with dropout to capture subtle temporal changes.
- Input layer separately processes the two temporal images.

Strength: Effective at detecting both large-scale and fine-grained transitions in features like ASGM ponds.

Experimental Results: SVM-STV

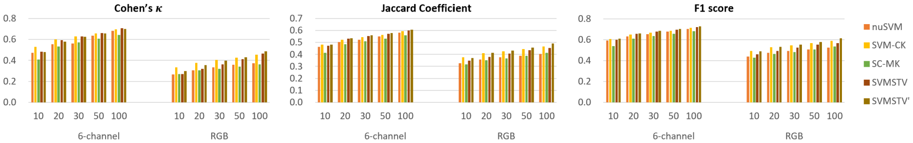


Figure: Average performance of SVM-based methods across MDD regions. Performance improves with more labels and more spectral channels.

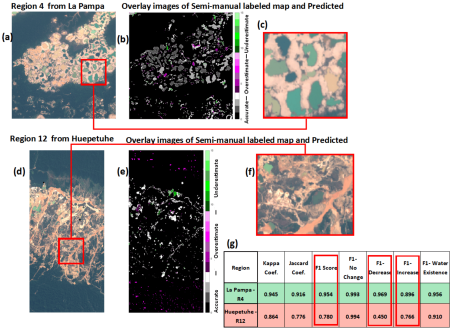
Key Findings (SVM-STV):

- Performance improves with more labeled samples per class.
- 6-channel imagery significantly reduces misclassifications compared to RGB.
- SVM-STV or SVM-STV' (with Lab lifting) generally achieve best results.
- Lifting effective for RGB, marginal for 6-channel.
- Computationally efficient (<1% labels, CPU training).

Experimental Results: E-ReCNN

Key Findings (E-ReCNN):

- Best performance achieved with 6-channel histogram-matched imagery (Kappa ≈ 0.92).
- Lab lifting showed limited benefit.
- High accuracy in MDD (F1 for No Change: 0.99, Water: 0.96).
- Performance varies with mining typology:
  - Higher F1 for distinct ponds (e.g., Region 4, La Pampa).
  - Lower F1 for diffuse sediment (e.g., Region 12, Huepetuhe).


Region 4 from La Pampa
Overlay images of Semi-manual labeled map and Predicted
(a)
(b)
(c)
Accurate — Overestimate — Underestimate
Region 12 from Huepetuhe
Overlay images of Semi-manual labeled map and Predicted
(d)
(e)
(f)
Accurate
Overestimate
Underestimate
(g)

Region	Kappa Coef.	Jaccard Coef.	F1 Score	F1- No Change	F1- Decrease	F1- Increase	F1- Water Existence
La Pampa - R4	0.945	0.916	0.954	0.993	0.969	0.896	0.956
Huepetuhe - R12	0.864	0.776	0.780	0.994	0.450	0.766	0.910

Figure: E-ReCNN performance comparison in Region 4 (distinct ponds) vs. Region 12 (diffuse sediment).

Generalizability: E-ReCNN on Global Mining Locations

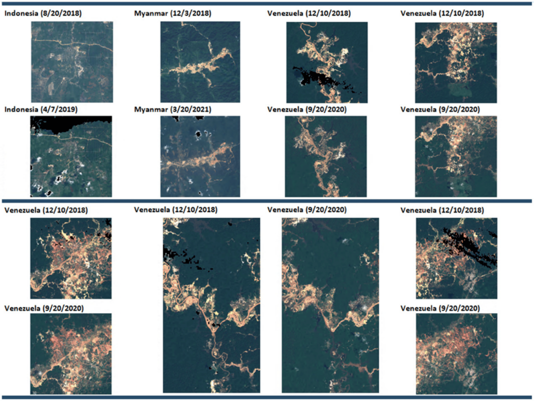


Figure: Test images from out-of-sample ASGM regions, 2018-2021.

Findings:

- E-ReCNN maintained competitive performance (Kappa, Jaccard > 0.9) for binary change.
- Model generalizes well for detecting ASGM-related water bodies globally.
- Detecting fine-grained turbidity changes (increase/decrease) is less effective without region-specific training.

Change Detection Conclusion

Summary of Contributions

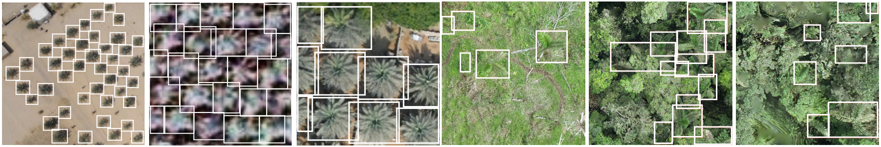
- **E-ReCNN** outperforms semi-supervised methods for detecting ASGM changes, especially with 6-band Sentinel-2 imagery and histogram matching.
- **SVM-STV** provides a practical, resource-efficient alternative when labeled data and multispectral inputs are limited. Effective with RGB + Lab lifting.
- A new, publicly available labeled dataset for ASGM detection was created.

Future Directions

- Explore active learning to reduce labeling effort for semi-supervised methods.
- Extend unsupervised, diffusion-based clustering to ASGM change detection.
- Assess E-ReCNN generalizability to other land-use changes (roads, agriculture).

UAV-Based Palm Localization and Spatial Analysis

The Importance of Palms in Tropical Forests



(a) Cases from existing studies

(b) Cases from our dataset

Figure: Comparative Samples of Manual Labels.

Ecological & Economic Significance of Palms:

- Vital to tropical forest ecology, biodiversity, and conservation planning [19].
- Support sustainable livelihoods and are key resources for tropical wildlife [20].
- Can serve as bioindicators of forest health and environmental impact.

Our Focus: Identifying and quantifying **naturally occurring palms** in complex forests, distinct from organized plantations.

Research Challenges & Our Contributions

Key Challenges

- **Image Variability:** Occlusion from overlapping canopies and inconsistent lighting degrade image quality.
- **Data Scarcity:** High-quality annotated datasets for tropical forests are rare and difficult to create.
- **Spatial Analysis Gap:** Most work focuses only on detection, not the large-scale spatial structure of populations.

Our Contributions

1. **New Dataset (PALMS):** Including 21 forest sites' orthomosaics in western Ecuador, with over 8,800 bounding boxes and 5,000 palm center annotations.
2. **PRISM Framework:** A flexible, interpretable pipeline for palm detection, segmentation, and counting.
3. **Spatial Modeling:** A Poisson-Gaussian model that simulates and provides insight into the ecological processes driving palm distribution.

The PRISM Framework

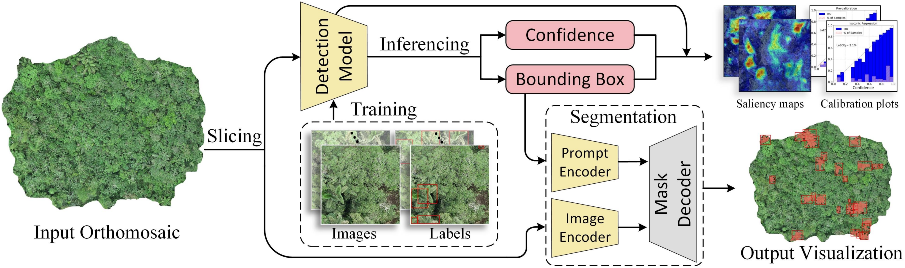


Figure: The PRISM Pipeline: From detection to segmentation and analysis [21].

Core Components:

- **Detection:** Fine-tuned models locate palms in orthomosaic patches.
- **Segmentation:** Detections are used as prompts for a zero-shot Segment Anything Model (SAM) to generate precise masks.
- **Mapping:** Outputs are georeferenced for landscape-scale analysis.
- **Interpretability:** Grad-CAM and calibration analysis enhance model reliability.

Detection and Segmentation Performance

Table: Detection model performance comparison.

Model	GFLOPS ↓	Params (M) ↓	FPS ↑	Precision ↑	Recall ↑	AP_{50} ↑	AP_{75} ↑	mAP ↑
DINO	1920.3	218.2	18.98 ± 0.95	0.7629 ± 0.0177	0.8494 ± 0.0071	0.8169 ± 0.0166	0.5455 ± 0.0150	0.5102 ± 0.0101
DDQ	1232.6	218.6	19.18 ± 0.96	0.7825 ± 0.0124	**0.8566 ± 0.0123**	0.8541 ± 0.0129	0.6354 ± 0.0137	0.5736 ± 0.0130
RT-DETR	222.5	65.5	151.49 ± 0.70	**0.8869 ± 0.0230**	0.7598 ± 0.0310	0.8416 ± 0.0181	0.6198 ± 0.0181	0.5769 ± 0.0145
YOLOv8	226.7	61.6	174.92 ± 0.86	0.8729 ± 0.0165	0.7997 ± 0.0203	0.8667 ± 0.0141	0.6777 ± 0.0137	0.6148 ± 0.0128
YOLOv9	**169.5**	53.2	114.96 ± 0.30	0.8763 ± 0.0176	0.7976 ± 0.0209	**0.8741 ± 0.0109**	0.6762 ± 0.0146	0.6162 ± 0.0122
YOLOv10	169.8	**31.6**	**177.04 ± 1.14**	0.8716 ± 0.0121	0.7968 ± 0.0089	0.8626 ± 0.0129	**0.6794 ± 0.0112**	**0.6173 ± 0.0090**
YOLO11	194.4	56.8	170.40 ± 0.95	0.8721 ± 0.0095	0.7896 ± 0.0127	0.8684 ± 0.0108	0.6677 ± 0.0180	0.6115 ± 0.0109

Key Findings:

- **YOLOv10 (Selected):** Best overall trade-off, achieving the highest mAP, AP_{75} and inference speed with the fewest parameters.
- **DDQ:** Highest recall, ideal when finding all possible instances is prioritized.
- **RT-DETR:** Highest precision, but misses more palms (lower recall).

Segmentation Performance: Comparing SAM Variants

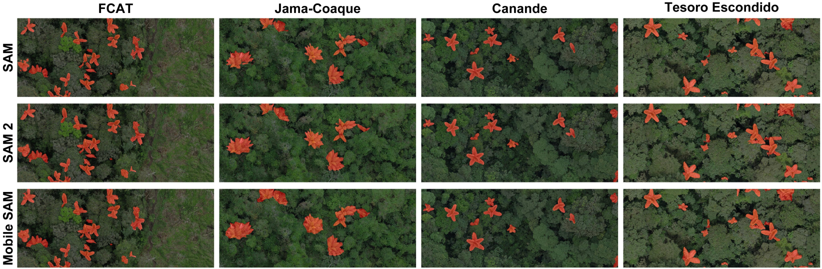
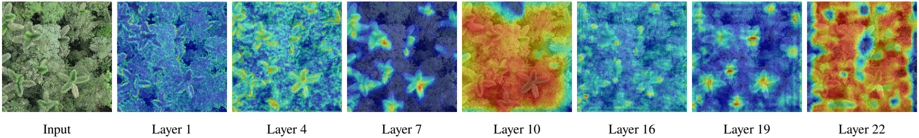


Figure: Visual comparison of SAM variants for zero-shot palm segmentation.

Key Findings

- We use the detector's bounding boxes as prompts for **zero-shot segmentation**.
- A comparison revealed distinct behaviors on our dataset:
  - **Original SAM:** Occasionally produces incomplete segments (under-segments).
  - **MobileSAM:** Tends to over-segment into non-palm areas.
  - **SAM 2 (Selected):** Provides the most balanced and accurate segmentation.

Visualizing What the Model "Sees" with Grad-CAM



Input | Layer 1 | Layer 4 | Layer 7 | Layer 10 | Layer 16 | Layer 19 | Layer 22

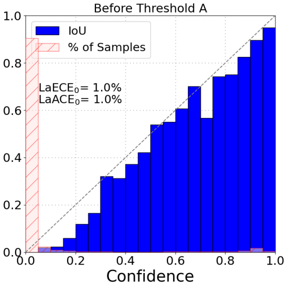
Figure: Hierarchical Feature Learning in YOLOv10 through Grad-CAM Visualizations.

Hierarchical Feature Learning

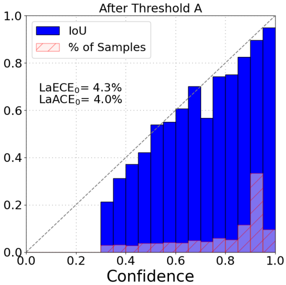
The analysis confirms the model learns a meaningful progression: early layers focus on low-level edges and textures; intermediate layers integrate spatial context; and deep layers exhibit focused activation over entire palm crowns.

Model Interpretability: A Step-by-Step Guide to Calibration

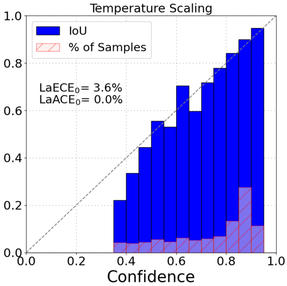
1. Initial Output



2. LRP Thresholding



3. Post-Hoc Calibration



The Problem

The uncalibrated model is unreliable; its confidence scores are poorly correlated with true accuracy (IoU).

The First Fix

LRP-based thresholding is first applied to prune the large number of unreliable, low-confidence predictions.

The Final Result

A post-hoc method (e.g., Temperature Scaling) is then applied to align confidence with accuracy.

Quantitative Analysis of Counting Performance

Table: Counting performance across four distinct ecological sites.

Site	Area (ha)	Counts	Pred2GT		GT2Pred	
			Ratio	Median (m)	Ratio	Median (m)
FCAT	21.62	471	0.9361	1.10	0.8854	0.77
Jama-Coaque	111.93	952	0.9348	1.50	0.8151	1.14
Canande	101.20	1,273	0.8956	0.82	0.7667	0.72
Tesoro Escondido	86.76	2,330	0.8981	1.09	0.9253	0.91

Key Metrics:

- **Pred2GT Ratio (Precision):** Proportion of predictions matched to a ground truth palm.
- **GT2Pred Ratio (Recall):** Proportion of ground truth palms matched by a prediction.

Summary of Findings:

- Precision is high across all sites, indicating the model generates few false positives.
- Recall is more variable, showing that detecting every true palm is harder and site-dependent.
- Sites like Tesoro Escondido show balanced performance, while Canandé reveals recall limitations (some palms are missed).

Background: Measuring Spatial Point Patterns

Goal

To quantify if a spatial point pattern is clustered, random, or regular by comparing it against a model of Complete Spatial Randomness (CSR).

Ripley's G & F Functions

These are cumulative distribution functions (CDFs) that measure nearest-neighbor distances at a given distance radius d:

- $G(d)$**:** CDF of distances from each point to its *nearest neighbor in the pattern*. It quantifies internal clustering.
- $F(d)$**:** CDF of distances from *random locations* to the nearest point in the pattern. It quantifies empty space.

Formal Definitions

Function	**Formulation**
$G(d)$	$\frac{1}{N_o}\sum_{i=1}^{N_o}\mathbb{1}(\hat{d}_i < d)$
$F(d)$	$\frac{1}{N_r}\sum_{j=1}^{N_r}\mathbb{1}(\tilde{d}_j < d)$
$J(d)$	$\frac{1-G(d)}{1-F(d)}$

Here, N_O is the number of observed points and $\hat{d}_i$ is the distance from point i to its nearest neighbor. N_r is the number of random points and $\tilde{d}_j$ is the distance from random point j to the nearest observed point.

Modeling Palm Distributions

Our Goal

To simulate palm spatial patterns that match observed distributions and to understand the ecological drivers of reproduction (e.g., long-range dispersal vs. local clustering).

The Core Mechanism: A Hybrid Generative Process

The model simulates palm propagation by combining two key ecological processes, controlled by two interpretable parameters [22]:

- **Global Dispersal (Poisson):** With probability $(1 - p)$, a new palm is placed randomly, representing animal-mediated or long-range dispersal.
- **Local Clustering (Gaussian):** With probability p, a new palm is placed near a parent, drawn from $\mathcal{N}(\mathbf{x}_{\text{parent}}, \sigma^2\mathbf{I})$, representing local seed drop.

Parameter Fitting

The optimal parameters (p^*, σ^*) are found by identifying the pair that generates simulated patterns whose spatial statistics (Ripley's G and F functions) most closely match those of the observed data.

The Poisson-Gaussian Algorithm: Implementation Details

Algorithm Pseudocode

```
Input: Candidate params $(\mathbf{p}, \boldsymbol{\sigma})$, Observed points $X$
Output: Optimal params $(p^*, \sigma^*)$
  Initialize: $d_{min} \leftarrow \infty$
  Pre-compute: Observed Ripley's stats $G_{obs}, F_{obs}$.
  for each $(p, \sigma)$ in grid do
      $d_{total} \leftarrow 0$
      for $i = 1$ to $N$ simulations do
          1. Generate simulated set $\hat{X}_i$ via the
        Poisson-Gaussian process.
          2. Compute simulated stats $G_{sim}, F_{sim}$.
          3. Calculate discrepancy $d_i$.
          4. Add $d_i$ to $d_{total}$.
      end for

      if $d_{total} < d_{min}$ then
          Update $d_{min}$, $p^* \leftarrow p$, $\sigma^* \leftarrow \sigma$.
      end if
  end for
  return $p^*, \sigma^*$
```

Discrepancy Metric

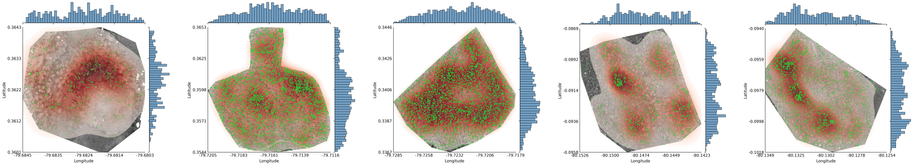
The discrepancy d_i for each simulation is the integrated absolute difference between observed and simulated Ripley's functions:

$$d_i = \int |\mathbf{g}_{obs} - \mathbf{g}_{sim}| + \int |\mathbf{f}_{obs} - \mathbf{f}_{sim}|$$

Optimization Process

A grid search is performed over the parameter space. The pair (p^*, σ^*) that minimizes the total discrepancy over all simulations is selected as the optimal fit.

Observed Spatial Patterns: Clustered or Random?



FCAT 1

FCAT 2

FCAT 3

Jama-Coaque 1

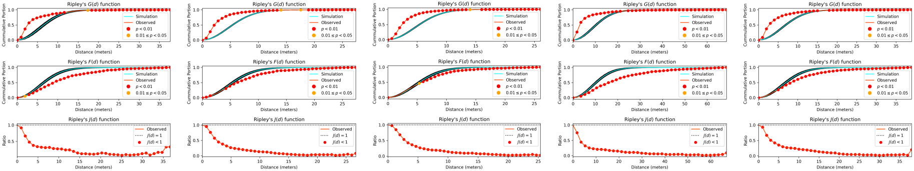
Jama-Coaque 2

Figure: Kernel Density Estimates (KDEs) of detected palm locations across five study sites, visually suggesting non-random clustering.

The Central Research Question

The spatial arrangement of palms appears clustered, but is this pattern statistically significant? We test the observed distributions against a null model of CSR.

Statistical Proof: Analysis with Ripley's Functions



FCAT 1 FCAT 2 FCAT 3 Jama-Coaque 1 Jama-Coaque 2

Figure: Ripley's function plots for each site. The observed pattern (red curve) is compared against the 95% confidence envelope of a CSR process (blue curve with shaded area).

Conclusion from the Analysis

The results confirm a **statistically significant** departure from randomness across all sites, with Ripley's functions revealing both dense internal clustering (*G*-function) and large empty spaces (*F*-function). This strong, non-random aggregation justifies our development of a more complex reproduction model.

Simulation Results: Replicating Observed Patterns

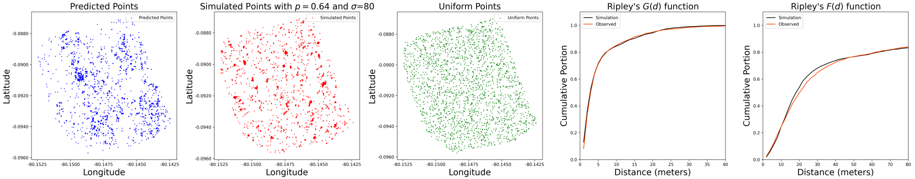


Figure: Visual and statistical comparison for the site Jama-Coaque 1. From left to right: PRISM prediction, model simulation, random distribution, G-, and F-function comparison.

Optimal Parameters

Fitted (p^*, σ^*) across sites.

Site	p^*	σ^*
FCAT 1	0.49	50
FCAT 2	0.52	70
FCAT 3	0.46	70
Jama-Coaque 1	0.64	80
Jama-Coaque 2	0.51	60

Key Findings

- The optimal parameters are highly consistent.
- This indicates a stable balance between local clustering (within a ∼2-4 meter radius) and random, long-range dispersal.
- The strong alignment between simulated and observed Ripley's functions (right panels) validates the model's fidelity.

Palm Detection and Distribution Conclusion

Summary of Contributions

- Created **PALMS**, a new, large-scale annotated dataset for palm detection in ecologically diverse tropical forests.
- Developed **PRISM**, an end-to-end framework for efficient palm detection, segmentation, and counting from UAV imagery.
- Introduced a simple, two-parameter **Poisson-Gaussian model** that successfully replicates the complex spatial dynamics of palm distribution, as validated by Ripley's functions.

Future Directions

- **Dataset Expansion:** A 1000 km^2 Amazonian region in Peru.
- **Model Enhancement:** Improve the localization quality of palm centers.
- **Deployment & Extension:** Real-time, on-device deployment (e.g., NVIDIA Jetson) and extension to species-level classification.

Fetal Heart Tracking in Ultrasound Videos

The Clinical Challenge: Congenital Heart Defects (CHDs)

A Global Health Concern

- CHDs affect up to **1.2% of all live births** globally and are a primary cause of neonatal mortality [23].
- Early diagnosis via Fetal Echocardiography (FE) is crucial for improving survival rates [24].

The Diagnostic Bottleneck

- Accurate interpretation of FE scans requires extensive expertise.
- A global scarcity of trained sonographers and cardiologists creates a significant barrier to early diagnosis, especially in low-resource settings [25].

Technical Challenges in Fetal Ultrasound Analysis

Data Heterogeneity and Quality

Real-world ultrasound data exhibits high variability due to different imaging machines, software, and scanning protocols, which can compromise model generalization.

Inherent Challenges of Fetal Cardiac Imaging

- The fetal heart is a **small, rapidly beating organ** with variable positioning and orientation.
- The heart occupies a widely varying portion of the frame (from 2.3% to 61% in our dataset).
- Videos are often screen-captured, including inconsistent graph user interfaces.

The Need for Temporal Awareness

Most prior work focuses on single frames, disregarding the rich information of the beating heart. Our work aims to address this by explicitly modeling temporal context.

Our Approach: A 3D CNN for Temporal Tracking

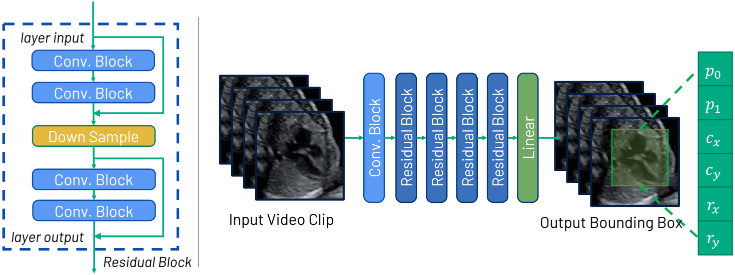


Figure: The modified 3D ResNet-18 architecture processes video clips to predict bounding boxes and presence indicators.

The Model:

- A modified 3D ResNet-18 architecture processes video clips (e.g., 64 frames).
- Preserves temporal resolution while downsampling spatial dimensions.

The Loss Function: A hybrid loss balances multiple objectives:

- Bounding box regression (MSE).
- Heart presence classification (Cross-Entropy).
- Temporal smoothness regularization (L^2-norm).

Dataset & Ground Truth Definition

Heart Tracking Set (Site 1)

- 738 scans from 401 healthy participants.
- Manual bounding box annotations reviewed by an expert fetal cardiologist.
- Includes multiple standard cardiac views (4CH, 3VV, etc.).

Anomaly Classification Set (Site 2)

- 3051 scans from 432 participants.
- Used to validate the downstream benefit of our tracking model for CHD detection.

Ground Truth Formulation:

For each frame, the ground truth g_i encodes the bounding box and heart presence:

$$g_i = (c_i^x, c_i^y, r_i^x, r_i^y, y_i^0, y_i^1)$$

- c: box center coordinates
- r: box half-dimensions
- y: one-hot vector for presence/absence

Results: Fetal Heart Tracking Performance

Table: Performance comparison for fetal heart tracking. Our 3D-based method with L2 regularization shows the best overall performance.

Method	Acc.↑	AP@50↑	AP@75↑	$MSE_{center}(\times 10^4)\downarrow$	$MSE_{radius}(\times 10^4)\downarrow$	mIoU↑
YOLO11 [26]	0.633	0.663	**0.543**	3.676±3.678	3.179±3.383	0.536±0.329
Conv2D	0.843	0.747	0.160	1.032±1.441	0.115±0.221	0.614±0.202
Conv2D+channel	0.876	0.824	0.223	0.826±1.322	0.099±0.173	0.643±0.184
Ours	0.911	0.838	0.319	0.616±1.022	0.103±0.179	0.678±0.173
Ours (w/ L2)	**0.918**	**0.866**	0.341	**0.525±0.801**	**0.093±0.169**	**0.693±0.161**

Key Takeaway

Explicitly modeling temporal information with a 3D architecture significantly outperforms frame-wise detection methods like YOLOv11. Adding a temporal smoothness regularization term (L2) further improves all performance metrics.

Qualitative Results: Visual Tracking Comparison

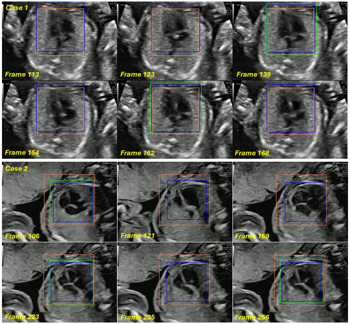


Figure: Tracking performance on two clinical cases. Bounding boxes: Ground Truth (blue), YOLOv11 baseline (green), and our proposed method (orange).

Baseline (YOLOv11)

- Struggles to maintain tracking consistency.
- Predictions can be unstable or missed.
- High spatial accuracy on frames it does detect.

Our Method

- Delivers significantly smoother and more consistent tracking across the cardiac cycle.
- Robust to variations in heart rate and motion.

Key Takeaway

The visual results highlight the critical importance of temporal modeling for reliable tracking in dynamic ultrasound videos.

Performance Analysis: Tracking and Downstream Impact

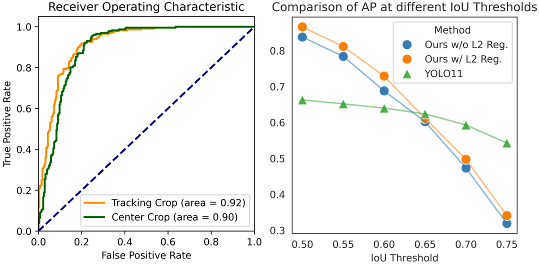


Figure: Left: AUC curves for CHD detection. Right: AP curves for the tracking task.

Key Findings

- **Downstream Impact:** Using our tracking model for preprocessing improves CHD detection, achieving a superior AUC of 0.92 over a standard center crop.
- **Tracking Performance:** Our 3D temporal model offers more consistent tracking (higher AP@50), while the frame-wise YOLOv11 shows better spatial localization when it succeeds (higher AP@75).

Fetal Heart Tracking Conclusion

Summary of Contribution

- Proposed a novel 3D deep learning method for fetal heart tracking in real-world ultrasound videos.
- The model demonstrated superior performance over strong baselines like YOLOv11 by combining spatial and temporal information.
- Showcased the method's utility as a pre-processing module, leading to improved accuracy in downstream CHD classification.

Future Directions

- Extend the framework to support detailed motion analysis by tracking multiple anatomical landmarks within the heart.
- Validate the method across a broader, more diverse patient population in collaboration with additional clinical institutions.

Conclusion

Thesis Conclusion: Summary of Contributions

A Common Theme

Across diverse domains, this thesis contributes data-efficient, interpretable, and spatially-aware machine learning frameworks to solve challenges with limited supervision and complex data structures.

Key Contributions

- **HSI Clustering:** Introduced **S^{2}DL**, an efficient unsupervised method combining superpixel segmentation and diffusion geometry.
- **Change Detection:** Applied supervised (**E-ReCNN**) and semi-supervised (**SVM-STV**) frameworks for detecting ASGM activities in Sentinel-2 imagery.
- **UAV Imagery Analysis:** Designed the **PRISM** pipeline for UAV-based palm detection and counting, complemented by a **Poisson-Gaussian model** for simulating spatial patterns.

Thank You

Questions & Discussion

Kangning Cui
City University of Hong Kong
ckn3.github.io

References I

[1] C. Gómez, J. C. White, and M. A. Wulder. “Optical remotely sensed time series data for land cover classification: A review”. In: *ISPRS J. Photogramm. Remote Sens.* 116 (2016), pp. 55–72.

[2] A. Beamish et al. “Recent trends and remaining challenges for optical remote sensing of Arctic tundra vegetation: A review and outlook”. In: *Remote Sens. Environ.* 246 (2020), p. 111872.

[3] H. Zhai et al. “Hyperspectral image clustering: current achievements and future lines”. In: *IEEE Geosci. Remote Sens. Mag.* 9.4 (2021), pp. 35–67.

[4] K. Cui et al. “Superpixel-based and spatially-regularized diffusion learning for unsupervised hyperspectral image clustering”. In: *IEEE Trans. Geosci. Remote Sens.* (2024).

[5] K. Cui and R. J. Plemmons. “Unsupervised classification of AVIRIS-NG hyperspectral images”. In: *Proc. WHISPERS*. IEEE. 2021, pp. 1–5.

[6] S. L. Polk et al. “Unsupervised detection of ash dieback disease (Hymenoscyphus fraxineus) using diffusion-based hyperspectral image clustering”. In: *Proc. IGARSS*. IEEE. 2022, pp. 2287–2290.

[7] S. Li et al. “Deep learning for hyperspectral image classification: an overview”. In: *IEEE Trans. Geosci. Remote Sens.* 57.9 (2019), pp. 6690–6709.

[8] B. Rasti et al. “Noise reduction in hyperspectral imagery: overview and application”. In: *Remote Sens.* 10.3 (2018), p. 482.

[9] K. Cui et al. “Unsupervised spatial-spectral hyperspectral image reconstruction and clustering with diffusion geometry”. In: *Proc. WHISPERS*. IEEE. 2022, pp. 1–5.

[10] J. M. Murphy and M. Maggioni. “Spectral–spatial diffusion geometry for hyperspectral image clustering”. In: *IEEE Geosci. Remote Sens. Lett.* 17.7 (2019), pp. 1243–1247.

[11] M.-Y. Liu et al. “Entropy rate superpixel segmentation”. In: *Proc. CVPR*. IEEE. 2011, pp. 2097–2104.

[12] R. R. Coifman and S. Lafon. “Diffusion maps”. In: *Appl. Comput. Harm. Anal.* 21.1 (2006), pp. 5–30.

[13] J. Caballero Espejo et al. “Deforestation and forest degradation due to gold mining in the Peruvian Amazon: A 34-year perspective”. In: *Remote Sens.* 10.12 (2018), p. 1903.

[14] J. R. Gerson et al. “Amazon forests capture high levels of atmospheric mercury pollution from artisanal gold mining”. In: *Nat Commun* 13.1 (2022), pp. 1–10.

[15] S. Camalan et al. “Change detection of amazonian alluvial gold mining using deep learning and sentinel-2 imagery”. In: *Remote Sens.* 14.7 (2022), p. 1746.

[16] X. Cai et al. “A three-stage approach for segmenting degraded color images: Smoothing, lifting and thresholding (SLaT)”. In: *J Sci Comput* 72.3 (2017), pp. 1313–1332.

References II

[17] L. Mou, L. Bruzzone, and X. X. Zhu. “Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery”. In: *IEEE Trans. Geosci. Remote Sens.* 57.2 (2018), pp. 924–935.

[18] K. Cui et al. “Semi-Supervised Change Detection Of Small Water Bodies Using Rgb And Multispectral Images In Peruvian Rainforests”. In: *Proc. WHISPERS*. IEEE. 2022, pp. 1–5.

[19] W. L. Eiserhardt et al. “Geographical ecology of the palms (Arecaceae): determinants of diversity and distributions across spatial scales”. In: *Ann Bot* 108.8 (2011), pp. 1391–1416.

[20] N. C. Pitman et al. “Distribution and abundance of tree species in swamp forests of Amazonian Ecuador”. In: *Ecography* 37.9 (2014), pp. 902–915.

[21] K. Cui et al. “Detection and Geographic Localization of Natural Objects in the Wild: A Case Study on Palms”. In: *Proc. IJCAI*. 2025.

[22] K. Cui et al. “Efficient Localization and Spatial Distribution Modeling of Canopy Palms Using UAV Imagery”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[23] W. Wu, J. He, and X. Shao. “Incidence and mortality trend of congenital heart disease at the global, regional, and national level, 1990–2017”. In: *Medicine* 99.23 (2020), e20593.

[24] B. Holland, J. Myers, and C. Woods Jr. “Prenatal diagnosis of critical congenital heart disease reduces risk of death from cardiovascular compromise prior to planned neonatal cardiac surgery: a meta-analysis”. In: *Ultrasound Obstet. Gynecol.* 45.6 (2015), pp. 631–638.

[25] L. Zühlke et al. “Congenital heart disease in low-and lower-middle–income countries: current status and new opportunities”. In: *Curr. Cardiol. Rep.* 21 (2019), pp. 1–13.

[26] N. Jegham et al. “Evaluating the Evolution of YOLO (You Only Look Once) Models: A Comprehensive Benchmark Study of YOLO11 and Its Predecessors”. In: *arXiv preprint arXiv:2411.00201* (2024).