ICLR
International Conference On Learning Representations

SuoiAI: Building a Dataset for Aquatic Invertebrates in Vietnam

Tue Vo[1], Lakshay Sharma[2], Tuan Dinh[1], Khuong Dinh[3], Trang Nguyen[4], Trung Phan[5], Minh Do[1], Duong Vu[6]

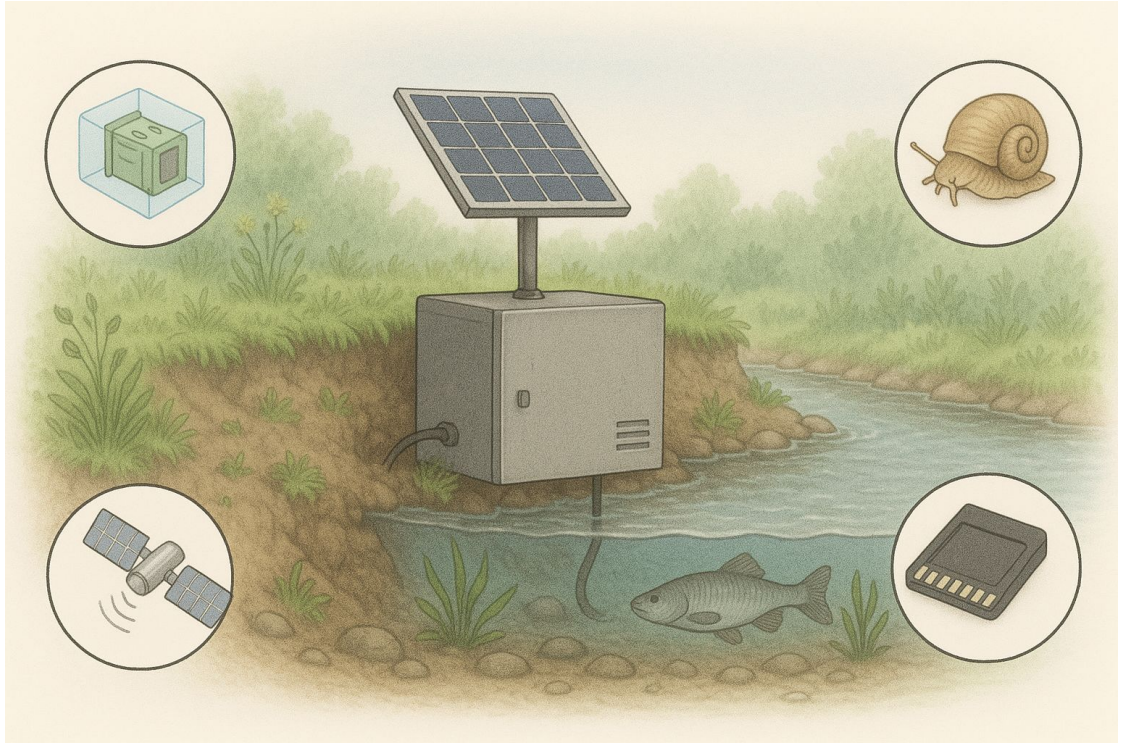
1: Nuoc Solutions, 2: Microsoft 3: Oslo University, 4: Bowdoin University, 5: Fulbright University Vietnam, 6: Westerdijk Fungal Biodiversity Institute

Vietnam's Biodiversity Challenge

- Vietnam's aquatic ecosystems are underrepresented in global biodiversity databases.
- Only 2,000 of 100,000+ freshwater invertebrate species have been identified in Vietnam.
- 800 aquatic species lack systematic documentation.
- This limits ability to assess aquatic health and understand climate change's impact.
- Building Vietnam's first aquatic invertebrate database will address this gap.

SuoiAI: An End-to-End Pipeline

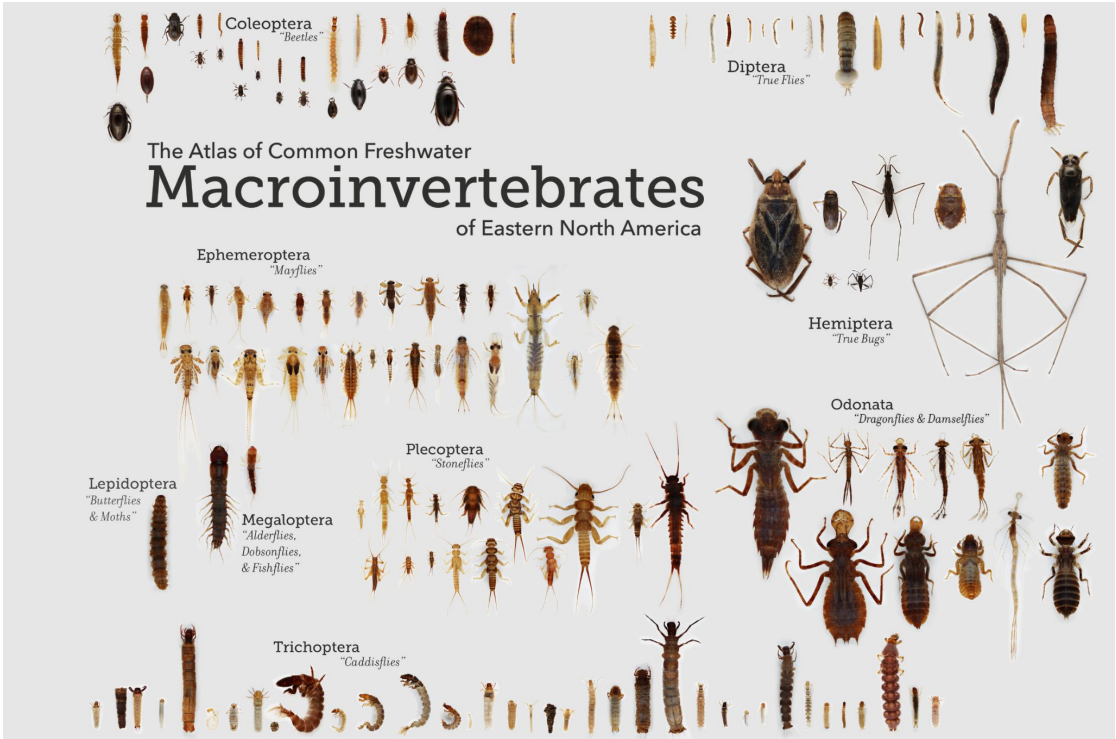
- SuoiAI is proposed to build a robust dataset for aquatic invertebrates and use machine learning for species classification.
- Underwater cameras will be deployed to collect data, and a combination of manual and automated annotation strategies will be used.
- Supervised, semi-supervised, and training-free machine learning models will be used for object detection and classification.
- The system will be deployed to monitor biodiversity, identify species, track populations, and assess water quality and ecosystem health.



Illustrative rendering of SuoiAI autonomous station

Related Work

- Existing datasets focus on American or European species.
- iNaturalist, Benthic Macroinvertebrate Database, Freshwater Biological Traits Database.
- Automatic camera projects (SPARROW, Heuschele et al. 2019, Albini et al. 2024)
- SuoiAI can collaborate with these projects to develop solar-powered cameras.



Macroinvertebrate database

Data Acquisition

- Underwater cameras will be deployed in key aquatic ecosystems (rivers, lakes, and coastal areas).
- 1080p resolution cameras will be used.
- 1-5 specimens per image will be captured.
- Initial field test sites include Cat Tien and Cuc Phuong national parks.
- System can expand to 135 sites nationwide.
- Automated systems will generate ~3 million data points per site annually.

Cat Tien national park, a biodiversity reserve in Vietnam

Data Annotation Strategy

- Manual labeling of a few hundred high-quality images for genus and species identification.
- Annotation tools: LabelImg and VIA.
- Minimize manual annotation by using teacher-student model.
- Few-shot and zero-shot learning explored to utilize pre-trained models.
- Clustering with vision embeddings to group similar images for annotation.

Leopard shrimp (Caridina rubropunctata), an indigenous species from Vietnam, only found in limited range in North Vietnam

Modeling Techniques

- Baseline object detection: YOLO, Faster R-CNN
- Training-free and semi-supervised methods: Segment Anything Model (SAM), teacher-student paradigm
- Fine-grained image classification: genus and species level classification

Baseline Object Detection

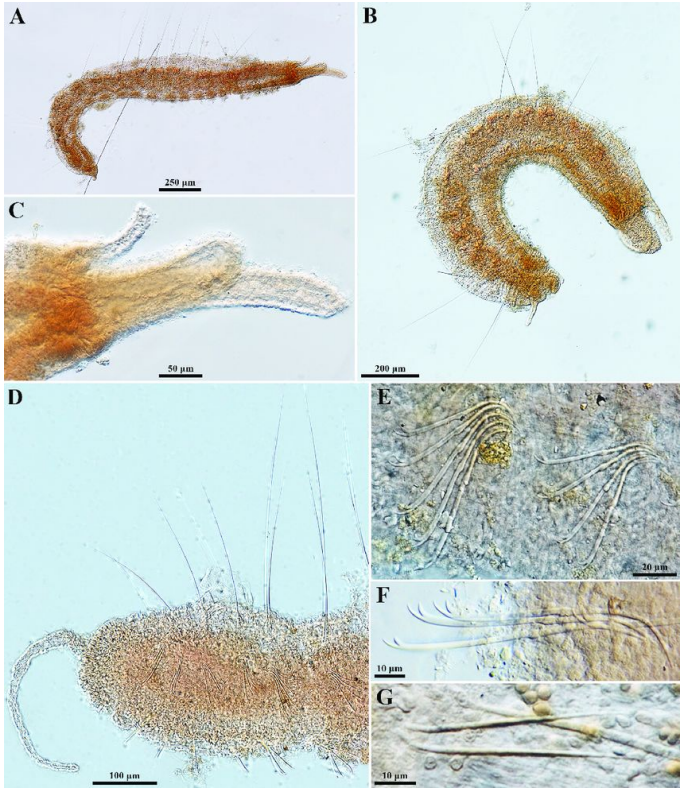
- Older models are small enough for on-device deployment (YOLO, Faster R-CNN)
- Newer models are more accurate, but only operable in the cloud (SwinTransformer, DETR (DEtection TRansformer))

Training-Free & Semi-Supervised

- Segment Anything Model (SAM) for prompt-based segmentation
- SAM is scalable and training-free
- May have limitations in accuracy for novel species
- Open vocabulary object detection using language-vision models
- Semi-supervised learning employs teacher-student paradigms
- Refines model predictions iteratively on unlabeled data

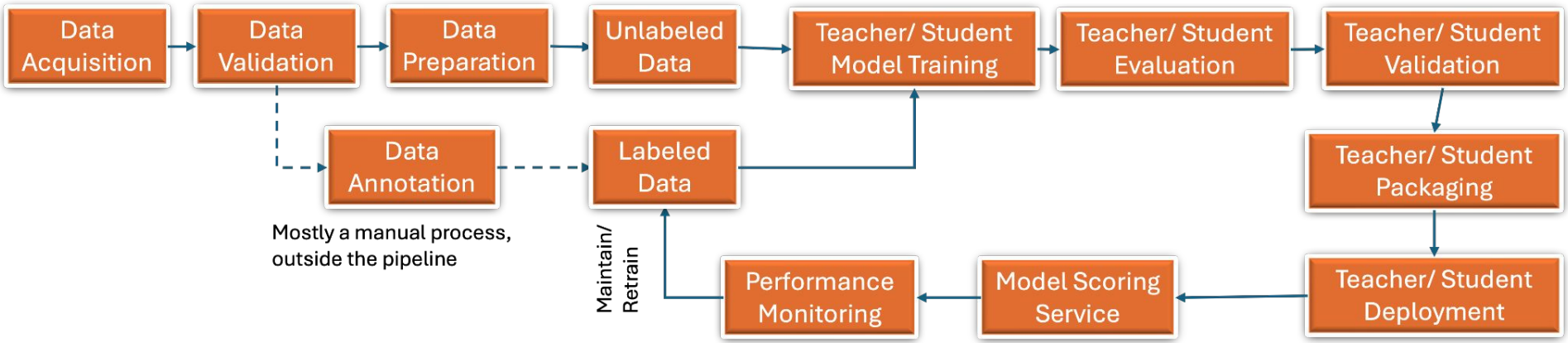
Fine-Grained Image Classification

- Fine-grained image classification techniques will be used to address high intra-class similarity.
- Classification at genus and species level.
- 20-50 genus-level classes, 100-200 species-level classes.
- ~1000 labeled samples per class.
- Techniques will handle long-tail distributions and improve performance on rare classes.
- Unknown species and genus can be classified using deep hierarchical Bayesian learning.



Aquatic invertebrates captured in the wild (Trieminentia sp. from Vietnam. (A,B): entire body, lateral view, head to the left; (C): posterior appendages, lateral view; (D): anterior part of the body, ventrolateral view; (E): ventral chaetae of segments V-VI; (F): ventral chaetae of segment XX; (G): dorsal needle chaetae.)

Pipeline for data collection and annotation using teacher/student model



Practical Considerations

- Deployment challenges addressed through a dual-pronged approach
- Lightweight models for on-device processing in field conditions
- Cloud-based analysis for large-scale data processing
- Image super-resolution techniques to enhance low-quality images
- System must be robust in diverse aquatic conditions

SuoiAI Applications

- In-situ biodiversity monitoring
- Species identification and classification
- Population dynamics tracking
- Water quality and ecosystem health assessment
- Discovering and documenting new species
- Quantifying aquatic invertebrate biomass
- Foundation Model for aquatic invertebrates
- Possible for scaling to other global hotspots

Mangrove forest, a key biodiversity region of Vietnam

Conclusion

- SuoiAI offers an integrated approach to building a dataset and pipeline for aquatic invertebrates in Vietnam.
- The project will enhance biodiversity monitoring and conservation efforts in the region and beyond.
- SuoiAI can be deployed in other tropical and equatorial regions.
- The data collected and analyzed will uncover hidden biodiversity and ecological functions on a global scale.
- Contact: tue@nuoc.solutions