Leveraging Kolmogorov-Arnold Networks (KANs) for Mixed-Domain Satellite Imagery Segmentation

Abdul Hadi Mazbah, Dr. Safiza Suhana Binti Kamal Baharin, Md. Shadman Zoha
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka Durian Tunggal, 76100, Melaka, Malaysia

***Abstract*—Semantic segmentation of satellite imagery requires models that capture global context while preserving sharp object boundaries. Convolutional Neural Networks (CNNs) excel at local feature extraction, but often struggle with long-range dependencies. Transformers provide global context but may blur edges and rely on opaque classifier heads. This study aims to develop an interpretable hybrid segmentation model that improves boundary accuracy and generalization across mixed-domain satellite imagery. This study presents SwinKANet, a hybrid segmentation model that combines a transformer encoder with boundary-aware decoding and an interpretable prediction head. SwinKANet employs a Swin Transformer (SwinV2-Tiny) encoder to extract multi-scale features, while a Convolutional Block Attention Module (CBAM) at the bottleneck refines channel and spatial responses. Skip connections equipped with SharpBlock units enhance edge detail, and an FPN-like lateral fusion module aligns and merges decoder features. The conventional multilayer perceptron head is replaced with a Kolmogorov–Arnold Network (KAN) head, enabling flexible function approximation and class-wise interpretability. We evaluate SwinKANet on a mixed-domain LoveDA dataset (urban + rural) for diverse spatial learning and on the urban-only ISPRS Vaihingen dataset for city-scale benchmarking. SwinKANet achieves 0.5269 mIoU on LoveDA and 0.7645 mIoU on Vaihingen, delivering sharper boundaries and more consistent class regions than CNN, Mamba, and transformer baselines. The KAN head further enhances explainability by revealing feature contributions for each class, supporting interpretable remote sensing applications.**

Keywords*—*Satellite imagery; Kolmogorov-Arnold Network; semantic segmentation; attention; mixed-domain

I. Introduction

Remote sensing has become a central technology in observing and understanding the Earth's surface. Today, applications like urban planning, climate monitoring, disaster management, and agriculture use satellite imagery for multiple tasks. Semantic segmentation is one of the most important components in these tasks [1–3]. Through it, we can understand what every pixel represents, whether it is a tree, building, or another object. By doing this, researchers and agencies can reliably distinguish between true land cover classes, which is a fundamental requirement for most remote sensing tasks.

Deep learning has been one of the major ways for semantic segmentation. Convolutional Neural Networks (CNNs) has been very successful for semantic segmentation, but it lacks global context, where the model cannot capture relationships between close and far neighboring pixels, which breaks the stream in the output maps [4]. Transformer [5] can capture global context where it knows the close and far neighboring pixels, so it can recognize and make maps with better accuracy and precision [6]. Moreover, most available satellites use selective and isolated domains, like some may only have urban, and some only have rural. For example, the LoveDA [7] dataset has rural and urban separately, and ISPRS Vaihingen and Potsdam both have only urban. However, the real-world shows otherwise. The real-world is a mix of urban and rural areas where urban and rural places are located side by side and in some cases, in mixed formation [8]. It is important to let the model know about this real situation.

Moreover, current models, Multilayer Perceptrons (MLPs), are used in the final output head, which is now in a leading position for good accuracy. However, it is a black box, and we do not know why it decides that one pixel should be building but not a road [9]. This raises trust issues with the model. Without dedicated XAI, it is impossible to know the reason behind the decisions [10]. Researchers are working on this interpretability and searching for an alternative to MLP. One of the most recent methods introduced was Kolmogorov-Arnold Networks (KANs) [11]. It is based on the Kolmogorov-Arnold Theorem [12] that says any multivariate continuous function can be represented as a finite composition of continuous univariate functions and addition. MLPs work on fixed activation functions on neurons, where KANs use learnable nonlinear functions on weights, and every weight parameter is replaced by a univariate function parameterized as a spline. This makes KANs an interpretable mechanism within the model that can tell us reasons before the decisions.

In this research, we propose a U-Net-inspired hybrid architecture named SwinKANet to address these gaps and limitations. The model is evaluated on the LoveDA dataset for mixed-domain segmentation and in the ISPRS Vaihingen dataset for high-resolution urban scene segmentation.

The main contributions are:

- Using a Swin transformer as encoder for global-local context, Convolutional Block Attention Module (CBAM) [13] and SharpBlock [14] for feature refinement, a special design for decoder, including Feature Pyramid Network (FPN), and finally, KANs at the output head.
- A mixed-domain dataset (LoveDA) and an urban-based dataset (ISPRS Vaihingen) approach.
- Use interpretability of KAN to understand model decisions.

The remainder of this study is structured as follows: Section II reviews previous works in a similar domain. Section III outlines the methodology of the proposed SwinKANet architecture. Section IV presents the results of various experiments conducted to validate the proposed model, Section V presents the discussion of the study, and Section VI provides the conclusion.

II. Related Work

A. Semantic Segmentation of Satellite Imagery

Semantic segmentation in satellite imagery refers to the process of assigning a class label to every pixel, producing detailed maps of features such as roads, buildings, vegetation, and water bodies. This capability transforms raw satellite data into actionable knowledge, enabling applications in urban planning, environmental monitoring, disaster management, agriculture, and infrastructure development [15–18]. Its societal impact has been repeatedly highlighted. Wieland et al. [19] demonstrated its value in flood detection for emergency response, Wu et al. [20] emphasized its role in mapping road systems and land parcels, Singh and Nongmeikapam [21] connected it to sustainable development, Wurm et al. [22] illustrated its utility in identifying informal settlements, and Bagwari et al. [23] described it as a critical enabler of high-resolution geospatial intelligence.

Methodologically, progress has been shaped by an iterative sequence of challenges and solutions. Early Convolutional Neural Networks (CNNs), particularly Fully Connected Networks (FCNs) and U-Net, enabled pixel-level classification of high-resolution images, but were constrained by limited contextual reasoning and poor adaptability to diverse domains [1]. Subsequent research explored semi-supervised and weakly supervised learning to alleviate dependence on extensive annotations, showing that deep models can generalize even with minimal labels [24,25]. To address the complexity of very high-resolution imagery, attention-based mechanisms such as SSAtNet preserved fine spatial structures [26], while multi-scale frameworks like A²-FPN and adaptive feature selection modules refined semantic consistency across scales [27,28]. These innovations significantly advanced segmentation accuracy in complex land-cover scenarios.

The introduction of transformer-based architectures marked another turning point. Models such as ST-UNet [29] integrated Swin Transformers with convolutional layers to capture global and local features, simultaneously, achieving state-of-the-art performance. However, the rising complexity of such models also raised interpretability concerns, prompting calls for explainable AI [30]. In parallel, researchers began adapting foundation models such as the Segment Anything Model (SAM) for remote sensing [31], with further refinement like RSPrompter automating prompt generation to improve adaptability [32]. Alongside, works on cross-domain generalization [33,34], optimization for difficult imaging conditions [35,36], and interpretable thematic mapping [37] have broadened the methodological landscape, each addressing specific weaknesses left by earlier approaches.

These advances, while impressive, also point to the limitations of relying on single paradigms. This naturally leads to the exploration of hybrid architectures, where the complementary strengths of CNNs, transformers, and attention mechanisms are combined to address the remaining challenges in remote sensing segmentation. Recent studies have also explored more advanced hybrid and transformer-based segmentation models for remote sensing, showing that combining global context modelling with boundary-aware refinement can further improve performance in high-resolution imagery [38–41].

B. Hybrid Architectures

Hybrid models blend multiple types of mechanisms to reconcile the strengths of local feature extraction with global semantic reasoning. A recurring theme in this evolution has been the iterative resolution of persistent problems such as pixel-level ambiguity, small object detection, spectral variability, and computational inefficiency.

One early step in this direction was HMANet by Niu et al. [42], which addressed the restricted receptive fields of CNNs. By combining spatial, channel, and category-based attention through modules such as Class Augmented Attention (CAA) and Region Shuffle Attention (RSA), HMANet improved feature correlations and reduced redundancy. On ISPRS Vaihingen and Potsdam, it reported a higher mean intersection over union (mIoU) and sharper semantic boundaries. Building on this, Li et al. [43] introduced HCANet, which emphasized hierarchical fusion using a Cross-level Contextual Representation Module (CCRM) and a Hybrid Representation Enhancement Module (HREM). Together with DUpsampling, this approach enhanced structural preservation and improved accuracy on the same benchmarks.

Addressing the challenge of multi-scale detection, particularly for irregular or small disturbance regions, Lv et al. [44] proposed HAssNet. Its spatial and channel attention modules captured global positions and emphasized task-relevant features, leading to a 6.7% improvement in mIoU over state-of-the-art baselines. Meanwhile, Zhong et al. [45] introduced FAENet, which incorporated discrete wavelet transformations to separate frequency bands. Through Inner-Component and Cross-Component Channel Attention (ICCA and CCCA), the network highlighted discriminative spectral responses before reprojecting them into the spatial domain, achieving superior boundary delineation on ISPRS Potsdam and LoveDA.

Parallel developments explored CNN-Transformer hybrids. Zhang et al. [46] combined Swin Transformer encoders with CNN-based decoders, including depth-wise separable atrous spatial pyramid pooling (SASPP), squeeze-and-excitation (SE) blocks, and skip connections, while an auxiliary boundary branch improved edge accuracy. This model ranked among the top-performing methods on Vaihingen and Potsdam.

Other researchers extended hybrid attention into weakly supervised and multimodal contexts. Chen et al. [47] proposed SASM-Net, which fused multi-scale feature extraction with semantic attention and structured model guidance, improving mAP under limited labels while maintaining fine contours. Li et al. [48] introduced HAENet with a Similarity-Hybrid Attention Module (SHAM) that fused Euclidean and hyperbolic similarities, outperforming conventional attention models on

ISPRS Potsdam and DeepGlobe. Their EDENet, leveraging Edge Distribution Attention (EDA) and a Hybrid Attention Module (HAM), further sharpened boundary precision and reduced misclassification in regions of high intra-class variance.

Subsequent work focused on multi-scale fusion and efficiency. Srivastava et al. [49] designed MC-SegNext with Sequential Atrous Spatial Pyramid Pooling (S-ASPP), Convolutional Block Attention Module (CBAM), and multi-head self-attention. This preserved multi-scale contributions and improved mIoU by 4-5% across Potsdam, Vaihingen, and LoveDA datasets. Du and Liang [50] followed with a cascade fusion model employing multi-core pooling, cross-stage integration, and attention modules, reporting 86.68% mIoU on PASCAL VOC2021 and 61.55% on SUIM. Li et al. [51] added DASSN_RSI, which integrated Deep Layer Channel Attention and Shallow Layer Spatial Attention with DUpsampling and Weight-Adaptive Focal Loss, excelling on the Gaofen Image Dataset (GID).

Toward operational scalability, Islam et al. [52] introduced Attentive U-Net, which incorporated attention layers into U-Net and achieved an IoU of 0.7959. Huynh-The et al. [53] advanced this with HBSeNet, a bilateral dual-path architecture that combined spatial and contextual flows. On ISPRS Potsdam, HBSeNet reported 92.04% global accuracy, 83.57% mIoU, and 90.23% mean boundary F1, surpassing baselines such as DeepLabV3+ and ST-UNet. Similarly, Xaio et al. [54] proposed EMRT, fusing CNNs with deformable self-attention to dynamically adjust receptive fields. Results included 50.89% mIoU on LoveDA, 73.62% on Potsdam, and 69.79% on Vaihingen, with enhanced boundary sharpness. Dimitrovski et al. [55] extended this trajectory into ensemble strategies, combining Multi-Axis ViT, ConvFormer, and EfficientNet-based-U-Nets, fused through a geometric mean. This ensemble achieved strong generalization across heterogeneous datasets, including ISPRS, UAVid, and LandCover.ai.

While these advances deliver state-of-the-art performance, they still depend heavily on MLP heads. This reliance leaves open questions regarding interpretability and alternative methods like Kolmogorov-Arnold Networks (KANs).

C. Kolmogorov-Arnold Networks (KANs)

MLPs, while effective in fusing features, act as opaque black boxes, offering little transparency into the reasoning behind segmentation or classification. This constraint has spurred interest in alternatives that combine expressive modelling power with intrinsic interpretability. One of the most notable advances in this regard is the Kolmogorov-Arnold Network (KAN), introduced by Li et al. [11]. Unlike traditional MLPs that use fixed activation functions on weighted edges, KANs replace these scalars with learnable spline-parameterized univariate components, making the network both highly expressive and inherently explainable.

The potential of KANs has been quickly recognized in remote sensing, where model transparency is particularly important due to the high-stakes nature of applications such as land cover mapping, infrastructure monitoring, and anomaly detection. Ma et al. [56] addressed the difficulty of fully exploiting high-dimensional encoder outputs and preserving fine-grained boundaries in semantic segmentation. They introduced DeepKANSeg, which integrates KANs at two critical stages: a refinement module at the encoder output and a global-local reconstruction module in the decoder. By transforming encoder features into univariate B-spline functions and employing GLKAN linear layers for pixel-level reconstruction, the model allowed each decision path to be directly visualized. On ISPRS Vaihingen, it achieved 84.05% mIoU, while on Potsdam it reached 85.44% mIoU, outperforming state-of-the-art CNN and Transformer models. Importantly, its clear boundary predictions demonstrated that accuracy gains could be coupled with transparent reasoning.

While DeepKANSeg focused on refining encoder-decoder interpretability, Li et al. [57] explored how KANs could enhance attention mechanisms for dense urban scenes. Their model MKLANet, embedded multi-scale KAN-based linear attention (MKLA) blocks into the decoder. These blocks captured both linear and nonlinear dependencies while exposing each univariate mapping used in attention calculations, effectively demystifying the global-local aggregation process. A deformable convolution module further enriched the encoder features. Tested on four datasets, MKLANet achieved 83.68% mIoU, 92.98% OA, and 91.08% mean F1 (mF1) on ISPRS Vaihingen; 69.78% mIoU and 96.51% OA on UAVid; 51.53% mIoU, 86.42% OA, and 67.19% mF1 on LoveDA; and 97.14% mIoU, 92.64% OA, and 93.80% mF1 on ISPRS Potsdam. Beyond strong results, the explicit univariate mappings provided unique insight into how ambiguous or small-scale targets were segmented.

The interpretability of KANs was also extended to remote sensing scene classification by An et al. [58]. They observed that CNN-Transformer hybrids with standard MLP heads obscure classification logic, particularly when modelling complex spatial patterns. To address this, they proposed the Swin Kansformer, which replaced MLP layers with KAN modules and introduced asymmetric convolution groups for enhanced local feature extraction. The KAN module decomposed multivariate inputs into univariate functions, revealing each component's contribution to classification. On the AID and NWPU-RESISC45 datasets, the model achieved 97.78% and 94.90% accuracy, outperforming ViT+PA baselines. This demonstrates that KANs could embed transparency not only in segmentation pipelines but also in classification tasks, making decisions paths more accessible for practitioners.

Li et al. [59] further broadened the scope by applying KANs to anomaly detection in ionospheric power spectrum imagery. These data exhibit irregular patterns that confound conventional CNN or Transformer models due to their high parameter counts and limited generalization. The proposed Kans-UNet replaced conventional convolutional weights with learnable B-spline functions through a KAN-Conv module embedded in the encoder. Combined with Feature Pyramid Attention and CBAM attention mechanisms, this design improved feature refinement across scales while enabling direct interpretation of transformations. On a custom dataset derived from satellite generated power spectrum images, Kans-UNet achieved an mIoU improvement of about 10% over PSPNet and 7.8% over PAN, while also reducing training time and complexity. Its

interpretability proved particularly valuable for tracing pixel-level predictions in small, patch-shaped anomaly regions.

Despite these advances, existing methods often improve accuracy at the cost of interpretability, and few works combine global context modelling, boundary refinement, and transparent prediction in a unified framework.

III. Methodology

The proposed SwinKANet architecture integrates a hierarchical encoder, attention-driven bottleneck, multi-stage decoder with transposed convolutions, a feature pyramid fusion mechanism, and a Kolmogorov-Arnold Network (KAN) decision head. Each component is designed to address specific limitations of traditional segmentation models, such as loss of spatial detail, lack of global context, and absence of interpretability, while ensuring computational feasibility. The workflow proceeds sequentially from input image processing in the encoder to interpretable pixel-wise predictions generated by the KAN head.

A. SwinKANet Architecture

As illustrated in Fig. 1, the SwinKANet architecture is a hybrid model consisting of Swin Transformer V2 [60] as encoder, CBAM at bottleneck, Sharpblock between skip connections, custom decoder with KAN at the output head.

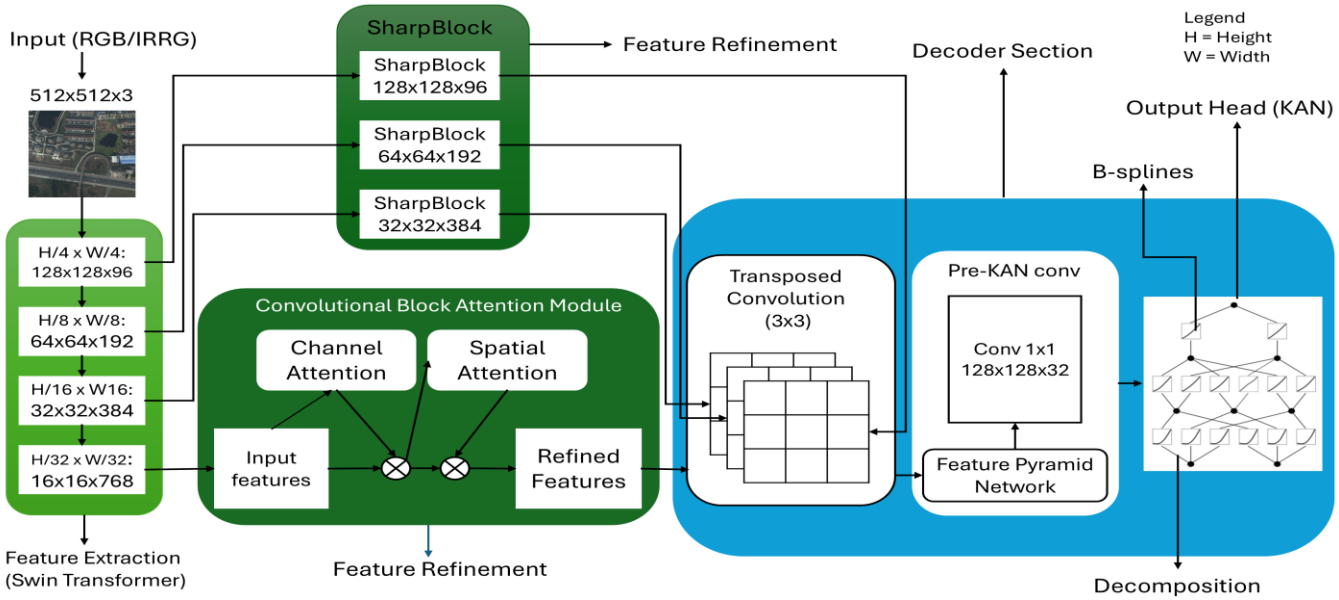

Fig. 1. SwinKANet architecture.

The encoder is built upon the SwinV2 Transformer backbone, which partitions an input satellite image $X \epsilon R^{B\times 3\times 512\times 512}$ into non-overlapping patches of size 16x16. Each patch is linearly projected and passed through successive Swin Transformer blocks, producing hierarchical feature representations at multiple spatial resolutions. The outputs are multi-resolution feature maps:

$$F_s \epsilon R^{B\times C_s\times H_s\times W_s},\ s\epsilon\{1,2,3,4,5\} \quad (1)$$

where, B is the batch size, C_s are channels, and H_s, W_s are spatial sizes. Specifically:

- F_1: 128×128,
- F_2:64×64,
- F_3:32×32,
- F_4,F_5:16×16.

As shown in Eq. (1), the encoder progressively reduces spatial size while increasing channel richness. In the architecture, F_1, F_2 *and* F_3 are used for skip connections, while F_5 forms the bottleneck.

Although F_5 captures semantics, it may contain noisy features. To refine it, a Convolutional Block Attention Module (CBAM) is applied in two stages.

- Channel Attention: Enhances discriminative channels by computing channel weights.

$$F_c=\sigma(W_2\delta(W_1 GAP(F_5)))\odot F_5 \quad (2)$$

where, GAP is global average pooling, δ is Rectified Linear Unit (ReLU), and σ is the sigmoid function. This operation in Eq. (2) ensures informative channels dominate.

- Spatial Attention: Highlights critical regions within the feature map.

$$F_b= \sigma(Conv_{7\times 7}(F_c))\odot F_c). \quad (3)$$

Furthermore, as shown in Eq. (3), the 7×7 convolution learns spatial dependency patterns. Together, both attention filter F_5 into bottleneck representation F_b, enriched with focus.

Then, bottleneck features must be upsampled to match the original input resolution. A three-stage decoder uses transposed convolutions for progressive up sampling:

$$U_d=\delta(BN(Conv(\delta(BN(Conv^T(F_{d-1})))))) \quad (4)$$

where, $Conv^T$ denotes transposed convolution, BN batch normalization, and δ ReLU. It goes through following:

- Stage 1: 768->384 channels, resolution 16->32.
- Stage 2: 384->192 channels, resolution 32->64.
- Stage 3: 192->96 channels, resolution 64->128.

Eq. (4) illustrates this progressive recovery. Each stage doubles resolution and reduces channel depth.

During decoding, spatial detail often blurs [14]. To mitigate this, SharpBlock is placed in between skip connections to refine features by applying a sharpening kernel, as shown in Eq. (5):

$$S_d^{'}=\delta(BN(Conv(S_d)+\alpha(K_{sharp}*S_d))) \quad (5)$$

where, $K_{sharp}=\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$, and α is a learnable scaling parameter. The sharpened skip $S_d^{'}$ is added to the decoder output.

$$D_d=U_d+S_d^{'} \quad (6)$$

Moreover, as shown in Eq. (6), this fusion enhances edges and boundaries. The rationale is that skip features from earlier encoder stages carry spatial fidelity, while SharpBlock sharpens them before merging. Next, decoder outputs $\{D_1,D_2,D_3\}$ capture features at increasing scales. To unify them, we use an FPN-like fusion:

$$M_s=Upsample(Conv_{1\times1}(D_s),(128,128)) \quad (7)$$

where, each map is resized to a common resolution, as shown in Eq. (7). The fused feature is:

$$F_{FPN}=\sum_{s=1}^{3} M_s \quad (8)$$

Eq. (8) ensures balanced integration of fine *(D3)* and coarse *(D1)* features. This multi-scale aggregation improves recognition of objects of varying size, which is essential for mixed-domain remote sensing imagery. KANs are computationally heavy. Directly applying them on F_{FPN} would not be feasible as the computation will be very heavy and would need very high-performance GPU. Therefore, we reduce dimensions with 1×1 convolutions and dropout, as defined in Eq. (9):

$$F_{pre}=\delta(Conv_{1\times1}(F_{FPN})) \quad (9)$$

yielding,

$$F_{pre}\epsilon R^{B\times32\times128\times128} \quad (10)$$

As shown in Eq. (10), compression reduces 96 channels to 32, while dropout regularizes. This step is crucial for tractable KAN execution without sacrificing too much representation power.

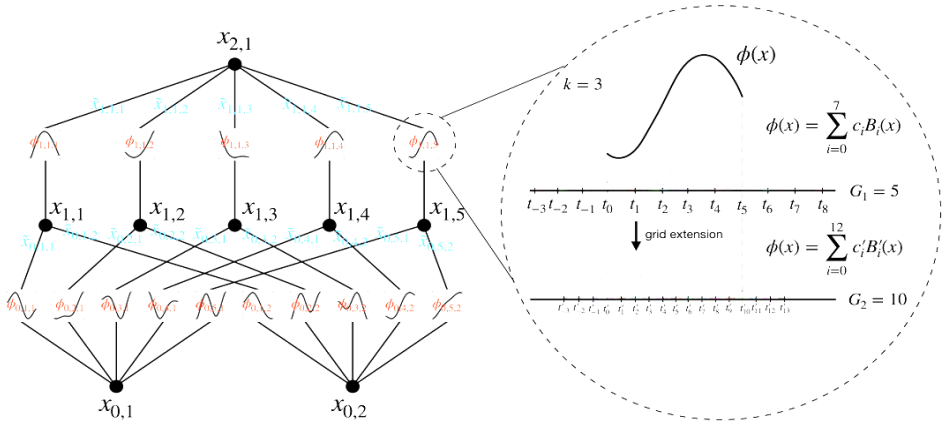

Fig. 2. Kolmogorov-Arnold network [11].

Fig. 2 represents KAN, where we send data from the pre-KAN conv after reducing the size. Upon receiving, the compressed tensor is flattened [see Eq. (11)]:

$$X=Reshape(F_{pre})\epsilon R^{B\times d_{in}\times N} \quad (11)$$

With $d_{in}=32$, $N=128\times128$. Each KAN layer computes:

$$Y=\sum_{i=1}^{d_{in}}\sum_{j=1}^{d_{out}}(w_{ij}^{spline}B_{ij}(X_i)+w_{ij}^{silu}\sigma(X_i)) \quad (12)$$

where, B_{ij} are spline basis functions, w_{ij}^{spline} and w_{ij}^{silu} are learnable weights. Eq. (12) decomposes mappings into interpretable univariate components. Finally, the final segmentation output is reconstructed:

$$Y\hat{}=Softmax(Reshape^{-1}(Y)),Y\hat{}\epsilon R^{B\times C\times512\times512} \quad (13)$$

with C = 7 classes. Eq. (13) produces per-pixel class probabilities.

B. Loss Function and Training Strategy

Accurate segmentation of mixed-domain satellite imagery requires addressing two main issues:

- Class Imbalance, where underrepresented classes are overshadowed by the dominant one.
- Boundary uncertainty, where class transitions are ambiguous.

To mitigate these challenges, we adopt a hybrid loss that combines Focal Loss and Dice Loss. Moreover, we used median frequency balancing strategy to make sure both dominant and rare classes get equal attention for LoveDA, as it is a heavy class

imbalance dataset, however, very relatable with real-world. Focal loss modulates the standard cross-entropy by down-weighting easy examples and emphasizing hard misclassified pixels. The formulation is:

$$L_{focal}=-\frac{1}{N}\sum_{i=1}^{N}\alpha yi\left(1-pi,_{yi}\right)^{\gamma}\log(pi,_{yi}) \quad (14)$$

where, N is the number of valid pixels, $p_{i,yi}$ is the predicted probability of the true class y_i, α_{yi} is the class-balancing weight, γ is the focusing parameter. In Eq. (14), the term $(1-p_{i,yi})^{\gamma}$ ensures that well-classified examples contribute less to the loss, focusing the model on challenging pixels. While Focal loss focuses on pixel-wise difficulty, Dice loss measures overlap between prediction and ground truth, directly optimizing segmentation quality. It is defined as:

$$L_{dice}=1-\frac{2\sum_{c=1}^{C}\sum_{i=1}^{N}p_{i,c}g_{i,c}+\epsilon}{\sum_{c=1}^{C}\sum_{i=1}^{N}p_{i,c}+\sum_{c=1}^{C}\sum_{i=1}^{N}g_{i,c}+\epsilon} \quad (15)$$

where, C is the number of classes, $p_{i,c}$ is the predicted probability of class c, $g_{i,c}$ is the one-hot encoded ground truth for class c, ϵ is a smoothing constant. Eq. (15) ensures better handling of class imbalance by maximizing the overlap for each class.

The final hybrid loss combines the two as follows:

$$L_{hybrid}=\lambda L_{focal}+L_{dice} \quad (16)$$

With $\lambda = 0.7$. As shown in Eq. (16), this weighting ensures Focal loss primarily drives learning on difficult pixels, while Dice loss regularizes global structure and class overlap. Together, the two complement each other by aligning pixel-level precision with region-level accuracy. Moreover, as shown in Eq. (17), we compute class weights using a median frequency balancing strategy. Let f_c denote the frequency of class c. Then the weight w_c for class c is given by:

$$w_c=\frac{\text{median}\{f_j|f_j>0,\ j=1,\ldots,C\}}{f_c+\epsilon} \quad (17)$$

where, C is the number of valid classes, $f_c=\frac{n_c}{\sum_{j=1}^{C}n_j}$ is the normalized frequency of class c, with n_c being the total pixel count for that class, ϵ is a small constant to prevent division by zero. Eq. (17) ensures that rare classes receive higher weights, while common classes receive lower weights. After computing, weights are clipped to a maximum cap at 10.0 to prevent extreme imbalance and are normalized to maintain stable optimization. Basically, it counts pixels for each class across all segmentation masks then compute normalized class frequency and calculate the median of non-zero frequencies. Then it applies Eq. (17) to compute weights and normalize weights to ensure stability in training. This weighting scheme is integrated directly into the Focal and Dice loss functions during training to ensure that underrepresented classes contribute more strongly to the optimization process.

Next, the study adopted AdamW optimizer, which decouples weight decay from adaptive moment estimation, improving regularization. The update rule is:

$$\theta_{t+1}=\theta_t-\eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t+\epsilon}}+\lambda\theta_t\right) \quad (18)$$

where, $\hat{m}_t$ and $\hat{v}_t$ are bias corrected estimates of first and second moments, η is learning rate, and λ is weight decay. Eq. (18) enforces stability while preventing overfitting. To balance exploration and convergence, we use Cosine Annealing with Warm Restarts (CAWR):

$$\eta_t=\eta_{min}+\frac{1}{2}(\eta_{max}-\eta_{min})\left(1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \quad (19)$$

where, η_t is learning rate at iteration t, η_{min} and η_{max} are bounds, T_{cur} is the number of iterations since the last restart, T_i is the length of the current restart cycle. As shown in Eq. (19), it allows the learning rate to periodically restart, avoiding local minima and enhancing generalization across diverse domains. Onwards, training is monitored using validation mIoU. If no improvement is observed for certain epochs, training halts early. This prevents overfitting and reduces unnecessary computation.

C. Implementation Details

The proposed SwinKANet model was implemented in PyTorch v2. All experiments were conducted on Google Colab using NVIDIA L4 GPU which has 22.5 GB VRAM. The environment configuration included Python 3.12, torchvision, numpy, scikit-learn, and wandb for experiment logging and monitoring. Table I shows implementation details that were used in this research.

TABLE I. IMPLEMENTATION DETAILS

Parameter	Value/Description
Batch Size	8
Learning Rate	3×10^{-4}
Weight Decay	0.5
Label Smoothing	0.1
Schedular	Cosine Annealing with Warm Restarts
Optimizer	AdamW
Input size	LoveDA (512×512×3), ISPRS Vaihingen (256×256×3)

TABLE II. MODEL COMPLEXITY

SwinKANet			
ISPRS Vaihingen (256x256)		**LoveDA (512x512)**	
Parameters (M)	37.67	Parameters (M)	37.67
Model Size (MB)	143.71	Model Size (MB)	143.71
MACs (Billion)	11.62	MACs (Billion)	46.43
Latency (ms)	41.12	Latency (ms)	83.64
Throughput (Images/sec)	24.32	Throughput(Images/sec)	11.96
GPU Memory Allocated (MB)	321.74	GPU Memory Allocated (MB)	162.87

As summarized in Table II, SwinKANet contains 37.67 million parameters and size is 143.7 MB for both settings. For ISPRS Vaihingen, input of 256x256, SwinKANet performs 11.62 billion multiply accumulate operations (MACs) and achieves an average inference latency of 41.12 milliseconds per image, yielding a throughput of 24.32 images per second. For

the LoveDA dataset with 512x512 input resolution, the computational cost increases to 46.43 billion MACs, with an average latency of 83.64 milliseconds per image and a throughput of 11.96 images per second. The GPU memory consumption remains relatively low, indicating that SwinKANet is efficient and suitable for deployment on moderately resourced systems.

IV. Experimental Results and Analysis

To validate the effectiveness of our proposed SwinKANet model, our training and validation followed a supervised learning paradigm with the LoveDA dataset (urban-rural mixed) and ISPRS Vaihingen dataset (urban).

A. Dataset Overview

For LoveDA dataset, the input size was 512x512x3, and we mixed both urban and rural images together and made one single train/validation/test set and we resized the images to 512x512 from original 1024x1024 because of heavy computation of KAN. LoveDA dataset contains around 5987 high resolution satellite images (0.3m Ground Sample Distance (GSD)) taken from three cities of China which are Nanjing, ChangZhou, and Wuhan. Training set had 2522 images, Validation set had 1669 images, and Test set had 1796 images. We predicted 7 classes for LoveDA which are: Background, Building, Road, Water, Barren, Forest, Agriculture.

For ISPRS Vaihingen, we split original 33 high resolutions remote sensing images of 2494x2064 pixels to 256x256 pixels patches. The GSD for it is 0.09 m. The near infrared, red, and green channels are provided in it. We predicted 5 classes, which are: Impervious Surface, Building, Low Vegetation, Tree, Car. Training set had 1645 samples, Validation set had 280 samples, and Test set had 329 samples.

B. Evaluation Metrics

To comprehensively assess the performance of SwinKANet on the semantic segmentation of remote sensing imagery, we employ four widely recognized evaluation metrics: Intersection over Union (IoU), mean Intersection over Union (mIoU), overall Accuracy (OA), and F1-score. These metrics are selected because they directly capture both pixel-level correctness and class specific segmentation quality, which are critical in urban-rural mixed-domain applications. The IoU metric evaluates the overlap between the predicted segmentation and ground truth for a given class. It is defined as shown in Eq. (20):

$$\mathrm{IoU_c}=\frac{\mathrm{TP_c}}{\mathrm{TP_c+FP_c+FN_c}} \tag{20}$$

where, TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively. IoU, thus measures the ratio of correctly classified pixels to the Union of predicted and ground-truth pixels for that class. Higher IoU values indicate stronger per class segmentation quality. To provide an aggregated measure of performance across all semantic classes, we report the (mIoU), as shown in Eq. (21):

$$\mathrm{mIoU}=\frac{1}{\mathrm{N}}\sum_{\mathrm{C=1}}^{\mathrm{N}}\mathrm{IoU_c} \tag{21}$$

where, N is the number of semantic classes. mIoU ensures fair evaluation by averaging performance equally across frequent and infrequent classes, making it a primary benchmark for segmentation in imbalanced remote sensing datasets. Overall Accuracy quantifies the proportion of correctly classified pixels across the entire image and is computed, as shown in Eq. (22):

$$\mathrm{OA}=\frac{\sum_{\mathrm{c=1}}^{\mathrm{N}}\mathrm{TP_c}}{\sum_{\mathrm{c=1}}^{\mathrm{N}}(\mathrm{TP_c+FP_c+FN_c})} \tag{22}$$

Unlike, IoU and mIoU, OA aggregates all classes into a single measure, which is useful for gauging general model reliability but can be biased toward majority classes in imbalanced datasets. The F1-score complements IoU by balancing precision and recall for each class. It is defined as the harmonic mean, as shown in Eq. (23):

$$\mathrm{F1_c}=\frac{2\cdot\mathrm{TP_c}}{2\cdot\mathrm{TP_c+FP_c+FN_c}} \tag{23}$$

where, precision is $\frac{TP_c}{TP_c+FP_c}$ and recall is $\frac{TP_c}{TP_c+FN_c}$. $F1$ is particularly informative in segmentation tasks where both false positives which is miscalculation, and false negatives which has missed detections, significantly impact performance.

C. Quantitative and Qualitative Analysis for LoveDA Dataset

To evaluate the effectiveness of the proposed SwinKANet model, we conducted a comprehensive comparison with several baseline models, including U-Net [61], ABCNet [62], U-NetFormer [63], DC-Swin [64], Rs^{3}Mamba [65], and MANet [66] on the LoveDA dataset. The quantitative results are summarized in Table III, which reports the metrics.

TABLE III. Experimental Comparison of Different Models for LoveDA. Best Results are Highlighted in Bold

Model	Background (IoU)	Building (IoU)	Road (IoU)	Water (IoU)	Barren (IoU)	Forest (IoU)	Agriculture (IoU)	OA	F1	mIoU
U-Net	0.4594	0.5825	0.5322	0.5613	0.1961	0.4261	0.4937	0.6389	0.6237	0.4645
ABCNet	0.4991	0.5530	0.5438	0.6720	0.2871	0.3961	0.5188	0.6770	0.6547	0.4957
UNetFormer	0.4909	0.5057	0.5415	0.6648	0.3581	0.4025	0.5143	0.6714	0.6589	0.4968
DC-Swin	0.3396	0.3704	0.4379	0.4767	0.2348	0.3033	0.4332	0.5440	0.5361	0.3708
Rs^{3}Mamba	0.4746	0.4769	0.5582	0.6426	0.3683	0.3909	0.5258	0.6634	0.6540	0.4910
MANet	0.5091	0.5905	0.5692	0.6641	0.2803	0.3839	0.5390	0.6885	0.6620	0.5051
SwinKANet	0.4789	0.6161	0.5711	0.6695	0.3505	0.4464	0.5557	0.6887	0.6843	0.5269

The results demonstrate that SwinKANet consistently outperforms the baselines across most evaluation criteria. In terms of overall metrics, SwinKANet achieves the highest mIoU 0.5269, surpassing the closest competitor MANet by more than two percentage points, while also yielding the best F1-score 0.6843 and OA 0.6887. This indicates that SwinKANet delivers superior segmentation consistency across diverse landcover categories. From a class-wise perspective, SwinKANet provides the strongest performance for Building (0.6161), Road (0.5711), Forest (0.4464), and Agriculture (0.5557), outperforming both transformer-based models, such as DC-Swin and hybrid models like U-NetFormer. The ability of SwinKANet to capture fine-grained structural features is particularly evident in the Forest and Agriculture classes, where it records significant gains over MANet. Although ABCNet and U-NetFormer achieve marginally higher IoU in the Water and Barren categories, respectively, SwinKANet remains highly competitive while maintaining superiority in aggregate measures. In contrast, the transformer-only DC-Swin shows the weakest performance with mIoU of 0.3708, highlighting that attention-based mechanisms alone are insufficient to ensure generalization in urban-rural mixed environments. To further substantiate the quantitative superiority of SwinKANet, we present a qualitative comparison against baseline models in Fig. 4, showcasing representative samples from the LoveDA dataset across both urban and rural domains. The figure contrasts the segmentation prediction of the same models shown in quantitative comparison in Table III, alongside the corresponding ground truth. The first images were taken from rural areas and the last one in the bottom was taken from urban areas. The results clearly show that SwinKANet generates segmentation maps that are visually closer to the ground truth, particularly in complex regions where fine boundaries and heterogeneous textures coexist. For example, in urban environments with dense building structures, SwinKANet preserves roof edges and inter-building separations with minimal over-segmentation, unlike U-Net and Rs^{3}Mamba, which tend to blur adjacent classes. Similarly, in rural landscapes characterized by agricultural and barren land, SwinKANet demonstrates sharper delineation between vegetation patches and non-vegetated regions, outperforming U-NetFormer and ABCNet, which exhibit class bleeding and misclassification. Moreover, DC-Swin suffers from fragmented predictions and inconsistent labelling, underscoring the limitations of attention-only designs in capturing multi-scale spatial dependencies. In contrast, SwinKANet successfully integrates Swin Transformer with Kolmogorov-Arnold Networks (KANs), thereby achieving both global-local coherence and detail preservation. This dual capability is especially evident in road and water boundaries, where SwinKANet maintains continuity and smoothness, leading to more faithful structural reconstructions.

Moreover, adding contrast in skip features does give us an edge in better boundaries, along with the special attention added by CBAM in the bottleneck of SwinKANet. In Fig. 3, the confusion matrices on the LoveDA dataset further substantiate the performance gap. While baseline models exhibit noticeable inter-class confusion, particularly between roads, buildings and forest, agricultural, SwinKANet achieves clearer separation with stronger diagonal dominance. This reflects its ability to preserve class boundaries and reduce ambiguities, validating the advantage of the SwinKANet architecture. This improvement is practically important for real-world remote sensing applications, where mixed urban-rural environments are common and accurate land-cover boundaries are required for reliable spatial analysis. More consistent segmentation of vegetation, roads, and built structures can support tasks such as environmental monitoring, land management, and automated map generation, where small classification errors may propagate into incorrect geographic assessments.

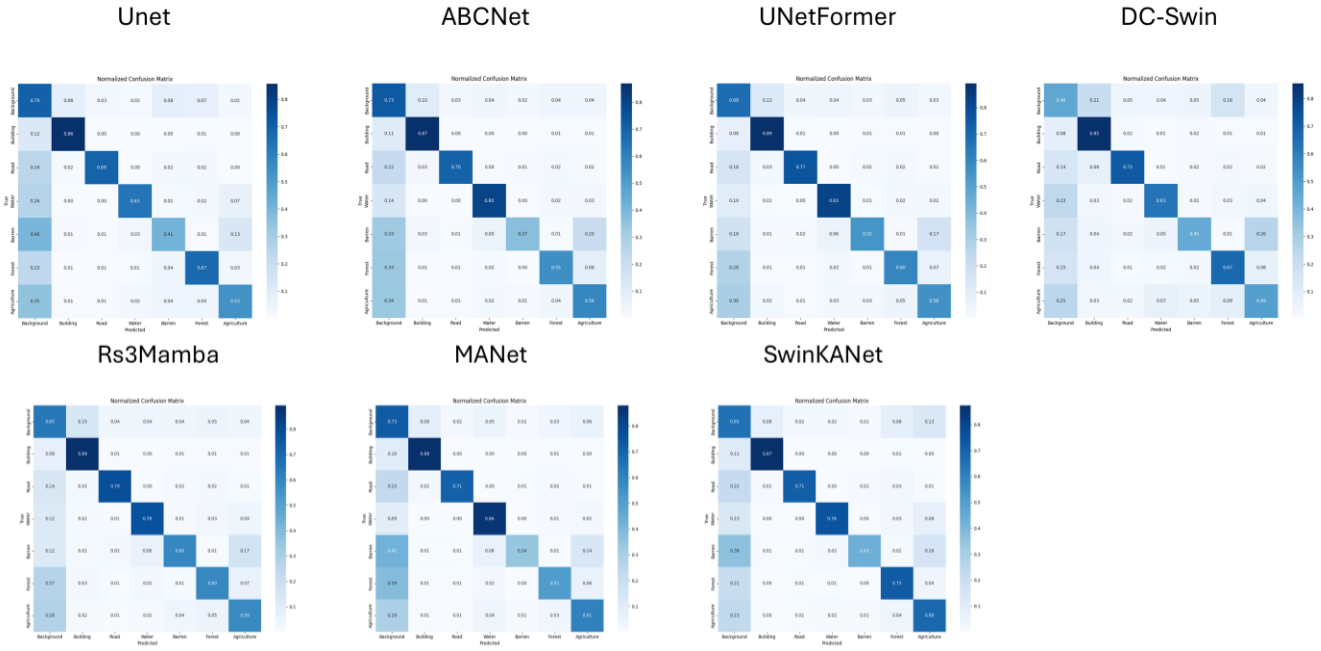

Fig. 3. Confusion matrix for different models for LoveDA.

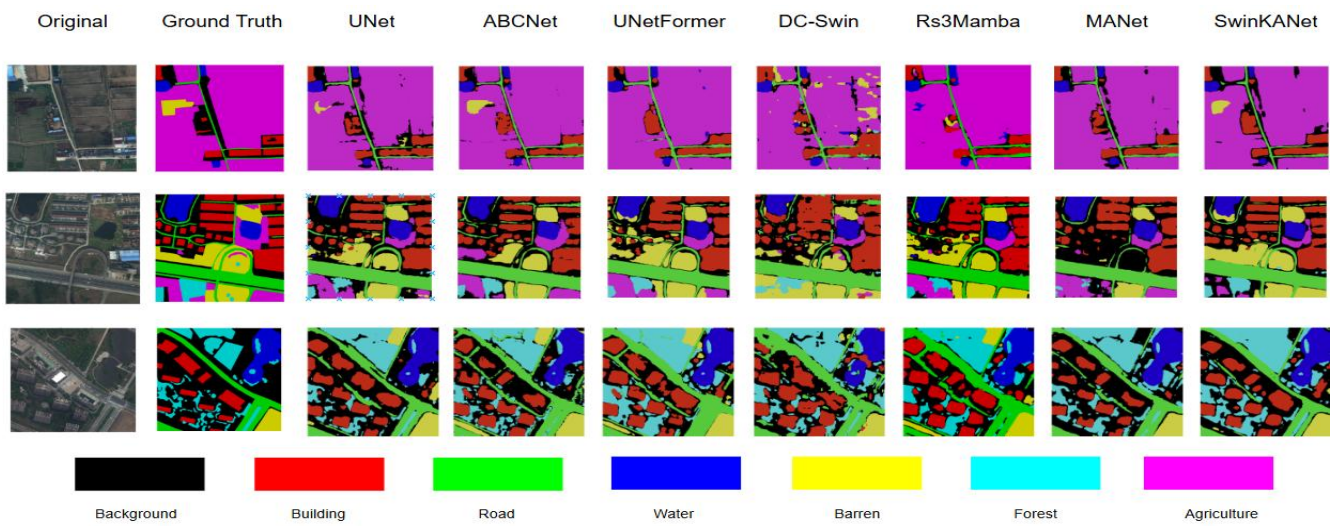

Fig. 4. Comparison between different models for LoveDA.

D. *Quantitative and Qualitative Analysis for Vaihingen dataset*

Onwards, the experimental results on the ISPRS Vaihingen dataset provide both qualitative and quantitative insights into the comparative performance of SwinKANet against the same established baseline models we used in the LoveDA dataset comparison. Quantitatively, as shown in Table IV, SwinKANet consistently outperforms competing models across nearly all metrics, achieving the highest mIoU of 0.7645, along with superior OA of 0.8788 and F1-score of 0.8646. These improvements are particularly evident in the segmentation of critical classes such as Impervious Surface with 0.8242 IoU and Building with 0.8780 IoU, where SwinKANet demonstrates clear robustness over transformer, Mamba and CNN-based alternatives. Although some models like UNetFormer and Rs3Mamba perform competitively in individual classes, their overall performance lags, highlighting the advantage of a hybrid with hierarchical contextual reasoning from Swin Transformer, CBAM attention, sharpblock contrast enhancer and expressive functional decomposition of KANs. Qualitative comparisons reinforce these findings, as shown in Fig. 6, SwinKANet generates segmentation maps that more closely align with ground truth, preserving fine boundaries and reducing misclassifications in challenging classes such as cars and low vegetation. Competing models frequently exhibit boundary fragments like in U-Net and DC-Swin or under-segmentation in MANet, which limits their capacity for reliable urban scene interpretation. Furthermore, the confusion matrix analysis shown in Fig. 5 demonstrates that the SwinKANet is superior in discrimination across all five classes, with notably higher diagonal dominance, reflecting fewer misclassifications compared to baselines. These results underscore SwinKANet's capability to deliver both high segmentation accuracy and consistent class level reliability, validating its design choice of integrating KANs in the output head along with other methods used in the model. Such improvements are useful for high-resolution urban mapping and several other applications.

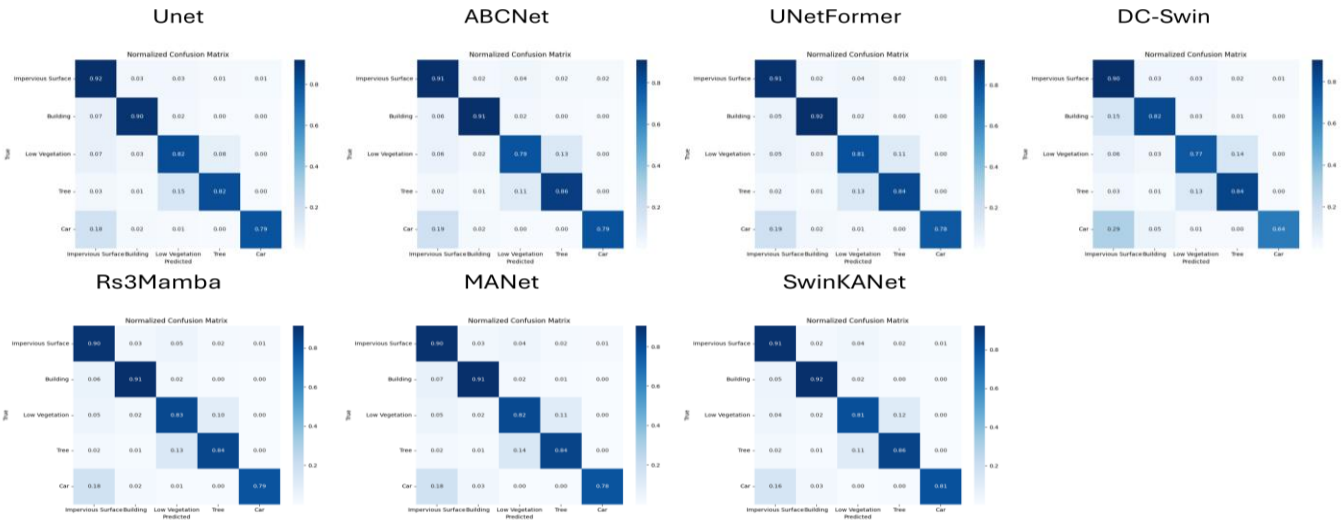

Fig. 5. Confusion matrix for different models for ISPRS Vaihingen.

TABLE IV. EXPERIMENTAL COMPARISON OF DIFFERENT MODELS FOR ISPRS VAIHINGEN. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Model	Impervious Surface (IoU)	Building (IoU)	Low Vegetation (IoU)	Tree (IoU)	Car (IoU)	OA	F1	mIoU
U-Net	0.7926	0.8490	0.6935	0.7275	0.6816	0.8680	0.8549	0.7488
ABCNet	0.8008	0.8652	0.6775	0.7221	0.6435	0.8678	0.8493	0.7419
UNetFormer	0.8146	0.8706	0.6855	0.7195	0.6923	0.8729	0.8594	0.7565
DC-Swin	0.7342	0.7610	0.6507	0.6933	0.5352	0.8290	0.8031	0.6749
Rs^3 Mamba	0.7997	0.8613	0.6946	0.7287	0.6778	0.8710	0.8570	0.7524
MANet	0.7924	0.8518	0.6887	0.7211	0.6693	0.8663	0.8519	0.7447
SwinKANet	0.8242	0.8780	0.6976	0.7302	0.6924	0.8788	0.8646	0.7645

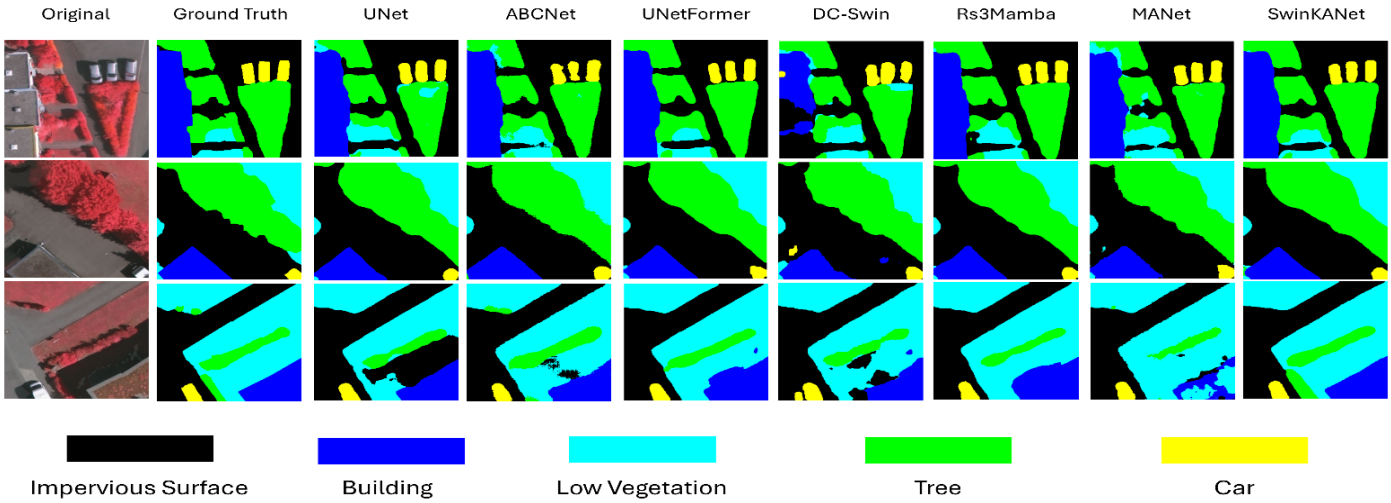

Fig. 6. Comparison between different models for ISPRS Vaihingen.

E. Interpretability of Kolmogorov-Arnold Networks (KANs)

A key motivation for integrating Kolmogorov-Arnold Networks (KANs) into SwinKANet, as mentioned before is their intrinsic interpretability. KANs use learnable spline-based univariate functions, allowing each class decision to be traced back to specific feature channels. This makes the model not only more expressive but also more transparent.

As shown in Fig. 7, the top three contributing channels for each class in the first KAN layer. For instance, the 'Road' class heavily depends on Channel 2 where weight is 0.497, whose spline captures consistent structural patterns, while 'Building' relies on Channels 29 and 25, highlighting sharp transitions around edges and boundaries. Such spline shapes reveal how KAN decomposes features into explicit, interpretable mappings offering insights into which channels drive predictions for each class that something that is not possible in standard MLP heads.

Moreover, in Fig. 8, we demonstrate how KAN refines logits before the final decision. The 'Background' class confidence increases substantially where mean shift from 2.4790 to 4.0510, while competing classes like 'Water' and 'Agriculture' are strongly suppressed, Δ=*-3.5269 and* Δ=*-2.6913*. This selective amplification and suppression improve discrimination in ambiguous regions and reduces class confusion, particularly in mixed-domain settings. Onwards, we observe similar behavior for ISPRS Vaihingen dataset. For example, as shown in Fig. 9, the 'impervious surface' class is primarily influenced by channels 8, 16, 27, with distinct weight polarities reflecting its structural rigidity. Similarly, 'Buildings' exhibit strong reliance on channel 27 with the highest weight magnitude of 0.886, underscoring KAN's ability to isolate dominant structural cues. For natural classes such as 'Low vegetation' and 'Tree', the spline functions highlight smoother but discriminative channel variations, while the 'car' class shows high weight localized responses with channel 26, weight = 0.411, critical for distinguishing small objects in dense urban scenes. This explicit mapping of univariate spline transformations clarifies which channels drive class decisions, transforming opaque activations into interpretable signals.

Moreover, as shown in Fig. 10, before refinement, logits were inflated and poorly separated like for 'building', pre-KAN mean is 6.8567. After KAN, they became re-centered and class specific like *Impervious Surfaces = 2.2016 (Δ=-4.551), Buildings= -0.6470* (Δ=*-7.5037*), *and Cars = -2.7013 (Δ=-9.5579)*. This suppression of spurious activations and reinforcement of discriminative evidence results in sharper, cleaner predictions. The refinement maps confirm this shift, showing diffuse pre-KAN logits transformed into sharper boundaries and clearer separations, especially for small objects like 'Cars' and complex vegetation.

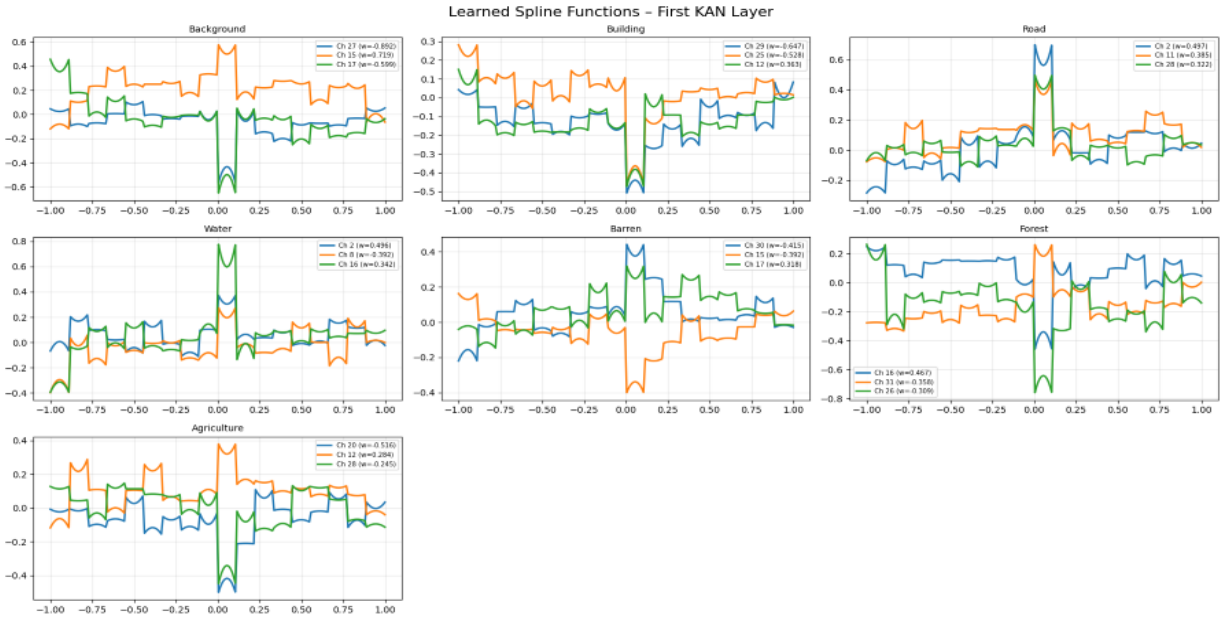

Fig. 7. Learned spline functions in the first KAN layer (LoveDA).

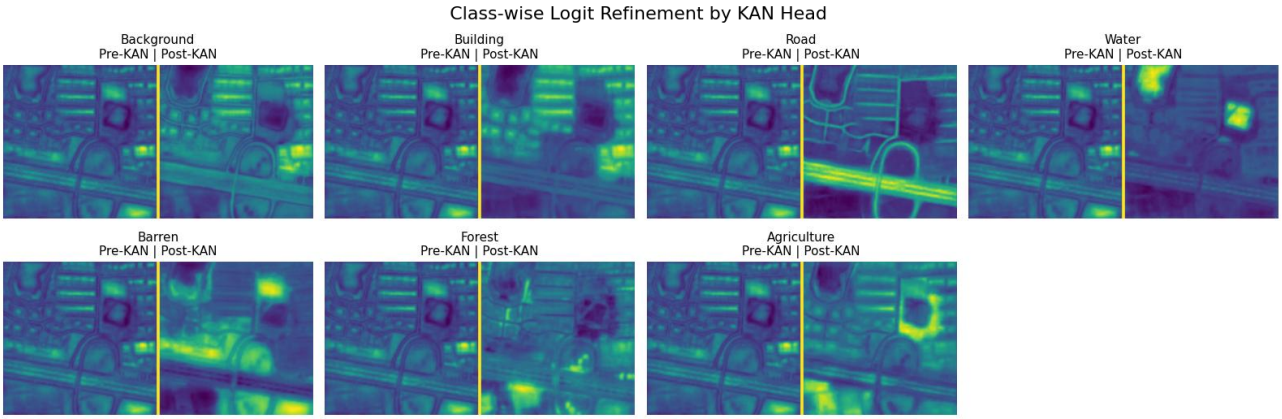

Fig. 8. Class-wise logit refinement by the KAN head (LoveDA).

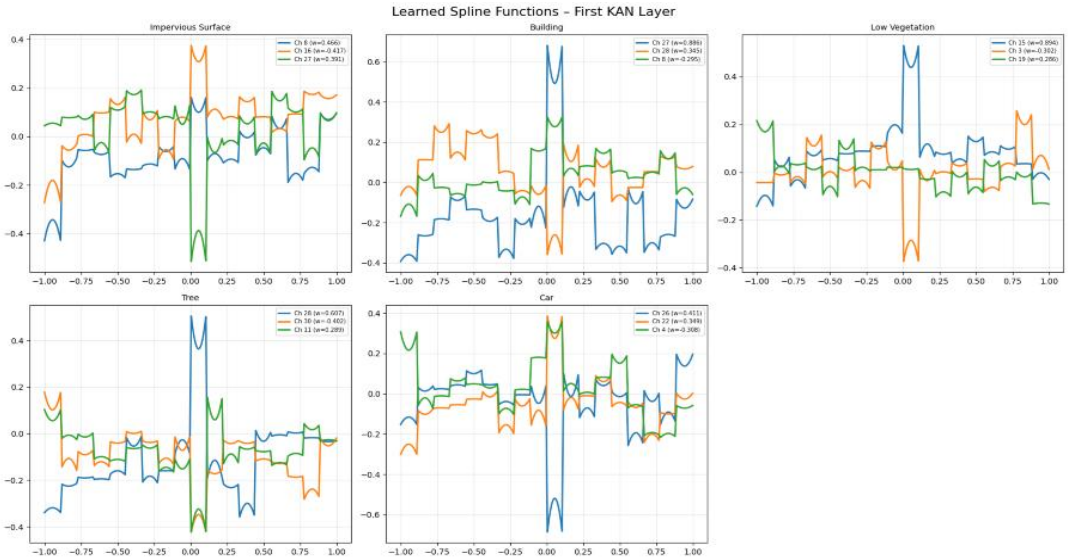

Fig. 9. Learned spline functions in the first KAN layer (ISPRS Vaihingen).

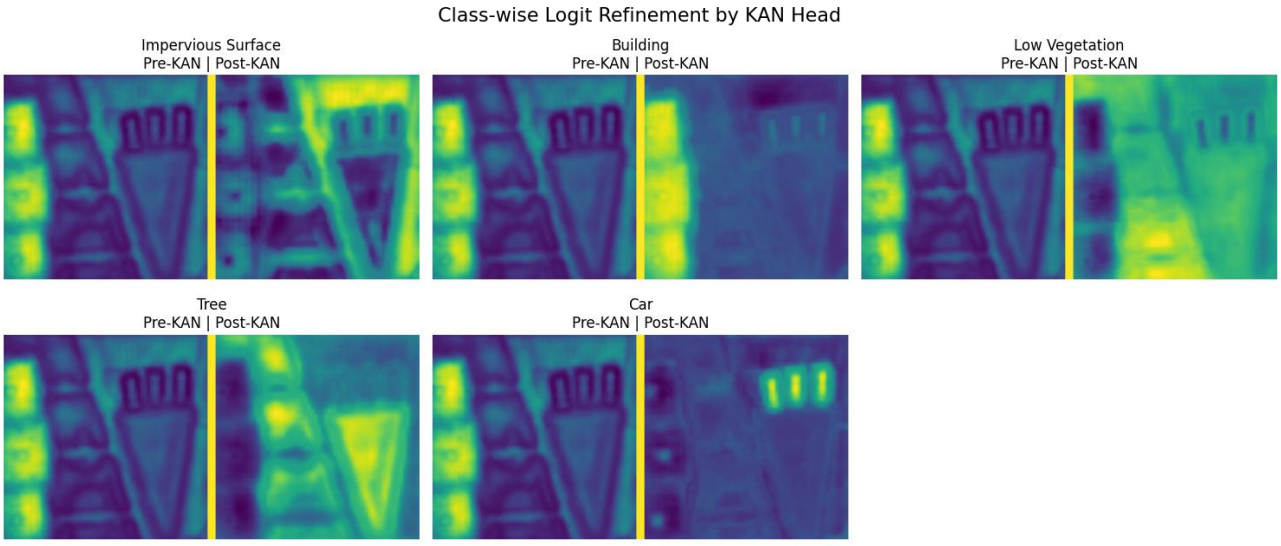

Fig. 10. Class-wise logit refinement by the KAN head (ISPRS Vaihingen).

V. Discussion

The experimental results demonstrate that the proposed SwinKANet architecture provides a balanced improvement in both segmentation accuracy and interpretability across mixed-domain and urban remote sensing datasets. The integration of Swin Transformer features with a KAN-based output head enables the model to preserve spatial boundaries while maintaining transparent decision behavior. This combination is particularly beneficial for real-world remote sensing applications, where reliable land-cover mapping and consistent object detection are required. The results also suggest that hybrid architectures incorporating interpretable components can improve the trustworthiness of deep learning models without sacrificing performance. These findings support the idea that hybrid architectures combining transformer-based feature extraction with interpretable prediction layers can provide both high performance and more transparent decision behavior, which is important for reliable deployment in real-world remote sensing systems.

VI. Conclusion

This work introduced SwinKANet, a novel hybrid architecture that combines the hierarchical contextual reasoning of SwinKANet, CBAM for spatial and channel attention, SharpBlock in the skip connection, and the interpretability of Kolmogorov-Arnold Network (KANs) in the output head. Extensive experiments on two benchmark datasets demonstrated its effectiveness. On the LoveDA dataset, SwinKANet achieved a mIoU of 52.69%, while on the ISPRS Vaihingen dataset, it attained an mIoU of 76.45%, consistently outperforming all baseline models compared. These results highlight the strengths of the proposed framework in addressing mixed-domain segmentation and advancing explainable AI in remote sensing.

A. Limitations

Despite its success, several limitations remain. In addition, the experiments are limited to the LoveDA and ISPRS Vaihingen datasets, which may not fully represent all geographic conditions or sensor variations found in real-world remote sensing scenarios. Classes with fewer samples, such as Cars or Barren, remain difficult to segment accurately, leading to confusion in predictions. Additionally, while KAN enhances interpretability and accuracy, it is computationally expensive and not yet optimized enough. A fully KAN-based network remains infeasible with current hardware.

B. Future Work

Future research should explore optimized implementation of KAN to reduce computational cost and enable end-to-end deployment. Moreover, extending explainability from the input stage to the final decision layer would strengthen predictions, making semantic segmentation more reliable in real-world high-stakes applications.

Acknowledgment

This research was supported by the Fundamental Research Grant Scheme from the Ministry of Higher Education Malaysia, grant FRGS/1/2023/WAB07/UTEM/02/1 and UTeM Kesidang scholarship.

References

[1] L. Huang, B. Jiang, S. Lv, Y. Liu, and Y. Fu, "Deep-learning-based semantic segmentation of remote sensing images: a survey," IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 17, pp. 8370–8396, 2024.

[2] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-net for semantic segmentation of fine-resolution remotely sensed images," IEEE Geosci. Remote Sensing Lett., vol. 19, pp. 1–5, 2022.

[3] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for earth observation," IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–11, 2022.

[4] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: fast semantic segmentation network." arXiv, 2019. [Online]. Available: https://arxiv.org/abs/1902.04502

[5] 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. [Online]. Available: https://ieeexplore.ieee.org/document/9710580/

[6] Y. Li, Z. Zhou, G. Qi, G. Hu, Z. Zhu, and X. Huang, "Remote sensing micro-object detection under global and local attention mechanism," Remote Sensing, vol. 16, no. 4, p. 644, Feb. 2024.

[7] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: a remote sensing land-cover dataset for domain adaptive semantic segmentation." arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2110.08733

[8] D. I. Alexander Wandl, V. Nadin, W. Zonneveld, and R. Rooij, "Beyond urban–rural classifications: characterising and mapping territories-in-between across europe," Landscape and Urban Planning, vol. 130, pp. 50–63, Oct. 2014.

[9] X. Hu and Q. Weng, "Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks," Remote Sensing of Environment, vol. 113, no. 10, pp. 2089–2102, Oct. 2009.

[10] D. Gaspar, P. Silva, and C. Silva, "Explainable ai for intrusion detection systems: lime and shap applicability on multi-layer perceptron," IEEE Access, vol. 12, pp. 30164–30175, 2024.

[11] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "KAN: kolmogorov-arnold networks." arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2404.19756

[12] J. Schmidt-Hieber, "The kolmogorov-arnold representation theorem revisited." arXiv, 2020. [Online]. Available: https://arxiv.org/abs/2007.15884

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module." arXiv, 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[14] M. R. Goni, A. Olalekan Ige, and N. I. Raihana Ruhaiyem, "TL-attsharpnet: automated lung image segmentation using transfer learning with depthwise convolution and attention," in 2023 IEEE 2nd National Biomedical Engineering Conference (NBEC), Melaka, Malaysia: IEEE, Sept. 2023, pp. 133–137. [Online]. Available: https://ieeexplore.ieee.org/document/10352617/

[15] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: a meta-analysis and review," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 152, pp. 166–177, June 2019.

[16] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 145, pp. 60–77, Nov. 2018.

[17] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery." arXiv, 2016. [Online]. Available: https://arxiv.org/abs/1606.02585

[18] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: a comprehensive review and list of resources," IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, Dec. 2017.

[19] 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, Mar. 2023.

[20] M. Wu, C. Zhang, J. Liu, L. Zhou, and X. Li, "Towards accurate high resolution satellite image semantic segmentation," IEEE Access, vol. 7, pp. 55609–55619, 2019.

[21] N. J. Singh and K. Nongmeikapam, "Semantic segmentation of satellite images using deep-unet," Arab J Sci Eng, vol. 48, no. 2, pp. 1193–1205, Feb. 2023.

[22] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 150, pp. 59–69, Apr. 2019.

[23] N. Bagwari, S. Kumar, and V. S. Verma, "A comprehensive review on segmentation techniques for satellite images," Arch Computat Methods Eng, vol. 30, no. 7, pp. 4325–4358, Sept. 2023.

[24] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," IEEE Geosci. Remote Sensing Lett., vol. 19, pp. 1–5, 2022.

[25] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," Remote Sensing, vol. 12, no. 2, p. 207, Jan. 2020.

[26] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–13, 2022.

[27] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2 -fpn for semantic segmentation of fine-resolution remotely sensed images," International Journal of Remote Sensing, vol. 43, no. 3, pp. 1131–1155, Feb. 2022.

[28] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," IEEE Geosci. Remote Sensing Lett., vol. 19, pp. 1–5, 2022.

[29] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–15, 2022.

[30] Q. Chong, M. Ni, J. Huang, G. Wei, Z. Li, and J. Xu, "Rethinking high-resolution remote sensing image segmentation not limited to technology: a review of segmentation methods and outlook on technical interpretability," International Journal of Remote Sensing, vol. 45, no. 11, pp. 3689–3716, June 2024.

[31] L. P. Osco, Q. Wu, E. L. de Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. M. Junior, "The segment anything model (sam) for remote sensing applications: from zero to one shot." arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2306.16623

[32] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "RSPrompter: learning to prompt for remote sensing instance segmentation based on visual foundation model." arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2306.16269

[33] [33] Q. Xu, X. Yuan, and C. Ouyang, "Class-aware domain adaptation for semantic segmentation of remote sensing images," IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–17, 2022.

[34] Y. Chen, C. Wei, D. Wang, C. Ji, and B. Li, "Semi-supervised contrastive learning for few-shot segmentation of remote sensing images," Remote Sensing, vol. 14, no. 17, p. 4254, Aug. 2022.

[35] S. Pare, H. Mittal, M. Sajid, J. C. Bansal, A. Saxena, T. Jan, W. Pedrycz, and M. Prasad, "Remote sensing imagery segmentation: a hybrid approach," Remote Sensing, vol. 13, no. 22, p. 4604, Nov. 2021.

[36] E. Basaeed, H. Bhaskar, and M. Al-Mualla, "Supervised remote sensing image segmentation using boosted convolutional neural networks," Knowledge-Based Systems, vol. 99, pp. 19–27, May 2016.

[37] B. Cardone, F. Di Martino, and V. Miraglia, "A novel fuzzy-based remote sensing image segmentation method," Sensors, vol. 23, no. 24, p. 9641, Dec. 2023.

[38] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. khan, "Transformers in remote sensing: a survey." arXiv, Sept. 02, 2022. [Online]. Available: http://arxiv.org/abs/2209.01206

[39] J. Song, A.-X. Zhu, and Y. Zhu, "Transformer-based semantic segmentation for extraction of building footprints from very-high-resolution images," Sensors, vol. 23, no. 11, p. 5166, May 2023.

[40] A. H. Mazbah, S. S. K. Baharin, M. R. Goni, and Md. S. Zoha, "SwinKANet: global-local feature extraction via attentions and kolmogorov arnold network for satellite imagery," in 2025 2nd International Conference on Electronic and Computer Engineering (ECE), Johor Bahru, Malaysia: IEEE, Aug. 2025, pp. 64–68. [Online]. Available: https://ieeexplore.ieee.org/document/11276263/

[41] X. Zhou, L. Zhou, S. Gong, H. Zhang, S. Zhong, Y. Xia, and Y. Huang, "Hybrid cnn and transformer network for semantic segmentation of uav remote sensing images," IEEE J. Miniat. Air Space Syst., vol. 5, no. 1, pp. 33–41, Mar. 2024.

[42] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images." arXiv, 2020. [Online]. Available: https://arxiv.org/abs/2001.02870

[43] X. Li, F. Xu, R. Xia, X. Lyu, H. Gao, and Y. Tong, "Hybridizing cross-level contextual and attentive representations for remote sensing imagery semantic segmentation," Remote Sensing, vol. 13, no. 15, p. 2986, July 2021.

[44] N. Lv, Z. Zhang, C. Li, J. Deng, T. Su, C. Chen, and Y. Zhou, "A hybrid-attention semantic segmentation network for remote sensing interpretation in land-use surveillance," Int. J. Mach. Learn. & Cyber., vol. 14, no. 2, pp. 395–406, Feb. 2023.

[45] J. Zhong, T. Zeng, Z. Xu, C. Wu, S. Qian, N. Xu, Z. Chen, X. Lyu, and X. Li, "A frequency attention-enhanced network for semantic segmentation of high-resolution remote sensing images," Remote Sensing, vol. 17, no. 3, p. 402, Jan. 2025.

[46] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–20, 2022.

[47] M. Chen, K. Xu, E. Chen, Y. Zhang, Y. Xie, Y. Hu, and Z. Pan, "Semantic attention and structured model for weakly supervised instance segmentation in optical and sar remote sensing imagery," Remote Sensing, vol. 15, no. 21, p. 5201, Nov. 2023.

[48] X. Li, F. Xu, F. Liu, R. Xia, Y. Tong, L. Li, Z. Xu, and X. Lyu, "Hybridizing euclidean and hyperbolic similarities for attentively refining representations in semantic segmentation of remote sensing images," IEEE Geosci. Remote Sensing Lett., vol. 19, pp. 1–5, 2022.

[49] N. Srivastava, A. Rai, S. K. Prasad Kushwaha, and K. Jain, "Advancing multi-class semantic segmentation of high-resolution satellite imagery through enhanced aspp and attention mechanisms," in IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece: IEEE, July 2024, pp. 8423–8427. [Online]. Available: https://ieeexplore.ieee.org/document/10640382/

[50] Z. Du and Y. Liang, "Research on image semantic segmentation based on hybrid cascade feature fusion and detailed attention mechanism," IEEE Access, vol. 12, pp. 62365–62377, 2024.

[51] X. Li, F. Xu, X. Lyu, H. Gao, Y. Tong, S. Cai, S. Li, and D. Liu, "Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images," International Journal of Remote Sensing, vol. 42, no. 9, pp. 3583–3610, May 2021.

[52] N. Islam, Md. F. Hossain, and Md. A. Hossain, "Semantic segmentation in satellite imagery: an attentive u - net approach," in 2024 2nd International Conference on Information and Communication Technology (ICICT), Dhaka, Bangladesh: IEEE, Oct. 2024, pp. 259–263. [Online]. Available: https://ieeexplore.ieee.org/document/10839725/

[53] T. Huynh-The, S. N. Truong, and G.-V. Nguyen, "HBSeNet: a hybrid bilateral network for accurate semantic segmentation of remote sensing images," IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 17, pp. 14179–14193, 2024.

[54] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," IEEE Trans. Geosci. Remote Sensing, vol. 61, pp. 1–16, 2023.

[55] I. Dimitrovski, V. Spasev, S. Loshkovska, and I. Kitanovski, "U-net ensemble for enhanced semantic segmentation in remote sensing imagery," Remote Sensing, vol. 16, no. 12, p. 2077, June 2024.

[56] X. Ma, Z. Wang, Y. Hu, X. Zhang, and M.-O. Pun, "Kolmogorov-arnold network for remote sensing image semantic segmentation." arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2501.07390

[57] Y. Li, S. Liu, J. Wu, W. Sun, Q. Wen, Y. Wu, X. Qin, and Y. Qiao, "Multi-scale kolmogorov-arnold network (kan)-based linear attention network: multi-scale feature fusion with kan and deformable convolution for urban scene image semantic segmentation," Remote Sensing, vol. 17, no. 5, p. 802, Feb. 2025.

[58] S. An, L. Zhang, X. Li, G. Zhang, P. Li, K. Zhao, H. Ma, and Z. Lian, "Global–local feature fusion of swin kansformer novel network for complex scene classification in remote sensing images," Remote Sensing, vol. 17, no. 7, p. 1137, Mar. 2025.

[59] X. Li, Z. Li, J. Huang, Y. Han, K. Zhu, B. Hao, J. Song, and Y. Huo, "Kans-unet model and its application in image patch-shaped detection," IEEE Access, vol. 13, pp. 80508–80519, 2025.

[60] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: scaling up capacity and resolution," 2021, [Online]. Available: https://arxiv.org/abs/2111.09883

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation." arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[62] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 181, pp. 84–98, Nov. 2021.

[63] 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, Aug. 2022.

[64] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," IEEE Geosci. Remote Sensing Lett., vol. 19, pp. 1–5, 2022.

[65] X. Ma, X. Zhang, and M.-O. Pun, "RS3 mamba: visual state space model for remote sensing image semantic segmentation," IEEE Geosci. Remote Sensing Lett., vol. 21, pp. 1–5, 2024.

[66] X. Ma, X. Zhang, M.-O. Pun, and B. Huang, "MANet: fine-tuning segment anything model for multimodal remote sensing semantic segmentation." arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2410.11160