Bridging Classical and Modern Computer Vision: PerceptiveNet for Tree Crown Semantic Segmentation

Georgios Voulgaris
University of Oxford
georgios.voulgaris@biology.ox.ac.uk

Abstract

The accurate semantic segmentation of tree crowns within remotely sensed data is crucial for scientific endeavours such as forest management, biodiversity studies, and carbon sequestration quantification. However, precise segmentation remains challenging due to complexities in the forest canopy, including shadows, intricate backgrounds, scale variations, and subtle spectral differences among tree species. Compared to the traditional methods, Deep Learning models improve accuracy by extracting informative and discriminative features, but often fall short in capturing the aforementioned complexities.

To address these challenges, we propose PerceptiveNet, a novel model incorporating a Logarithmic Gabor-parameterised convolutional layer with trainable filter parameters, alongside a backbone that extracts salient features while capturing extensive context and spatial information through a wider receptive field. We investigate the impact of Log-Gabor, Gabor, and standard convolutional layers on semantic segmentation performance through extensive experimentation. Additionally, we conduct an ablation study to assess the contributions of individual layers and their combinations to overall model performance, and we evaluate PerceptiveNet as a backbone within a novel hybrid CNN-Transformer model. Our results outperform state-of-the-art models, demonstrating significant performance improvements on a tree crown dataset while generalising across domains, including two benchmark aerial scene semantic segmentation datasets with varying complexities.

This work has been accepted for publication at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025 EarthVision Workshop.
Vision and Learning, AI for Science

1. Introduction

Tree crowns are crucial indicators of tree health, directly impacting photosynthesis, transpiration, and nutrient absorption (Bakalo et al. [1]; Li et al. [2]). Accurate semantic segmentation of tree crowns (Figure 1) enables quantitative analysis of forest characteristics, supporting applications in biodiversity assessment, carbon sequestration measurement, and ecological monitoring.

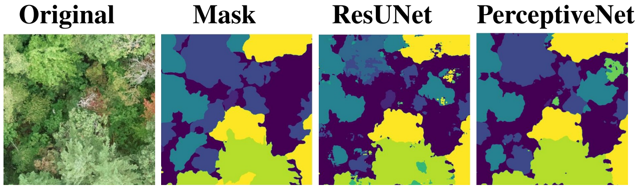

Figure 1. Tree Crown Semantic Segmentation, depicting (from left to right): Original Image; Corresponding Mask; ResUNet; and proposed model's semantic segmentation. Each colour represents a different tree species. The images portray densely packed trees with complex boundaries due to partial overlap. Moreover, tree crown similarities, complex occlusion, combined with light variations and shadows further augment the scene's complexity.

Dense forests present unique challenges for semantic segmentation: (1) variable lighting conditions and shadows that obscure crown boundaries (Bargas et al. [3]); (2) complex backgrounds with overlapping canopies (Hu et al. [4]); (3) scale variations from different Unmanned Aerial Vehicle (UAV) heights (Ocer et al. [5]); and (4) subtle spectral differences between species (Cao and Zhang [6]). Figure 2 demonstrates these challenges, showing how lighting conditions affect crown appearance of different species.

Current aerial imagery segmentation largely relies on U-Net variants (Ronneberger et al. [7]), ranging from basic implementations (Ye et al. [8]; Zhang et al. [9]; Cao and Zhang [6]; Wanger et al. [10]; Schiefer et al. [11]) to transformer-enhanced architectures (Scheibenreif et al. [12]; Vinod et al. [13]; Alshammari and Shahin [14]). While specialised modules like dilated convolutions and pyramid pooling (He et al. [15]; Zhao et al. [16]) improve performance, these approaches fail to address the unique spatial and spectral characteristics of aerial forest scenes. One thing that all the aforementioned works have in common is that they do not consider the challenges that tree crown semantic segmentation in dense forests presents.

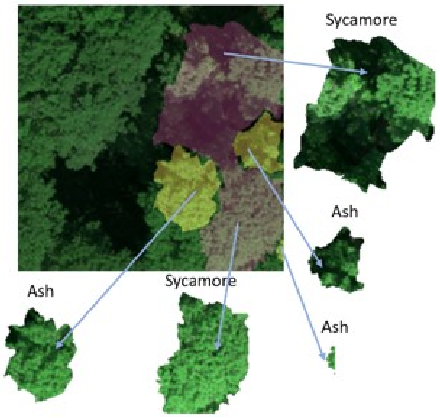

Figure 2. Dense forest canopy, demonstrating the impact of shadows, light variations, overlapping tree crowns, and weak distinctive features among tree species on the tree crown segmentation.

These challenges translate into specific issues for computer vision algorithms and principles. Additionally, they often ignore the Deep Learning models' characteristics, feature extraction capabilities, and inherent biases.

Despite advancements in Deep Learning for image processing, tree crown semantic segmentation in dense forest aerial imagery remains challenging. While Convolutional Neural Networks (CNNs) are inherently biased towards learning hierarchies of localised features (Fukushima [17]), studies have shown they tend to over-rely on texture rather than shape features when trained on standard (Geirhos et al. [18]) or aerial (Voulgaris et al. [19]) datasets. Relying solely on texture features is particularly problematic in aerial imagery, where texture features vary significantly with lighting, shadows and atmospheric conditions. This exposes a critical gap in current approaches, which often overlook the need for robust shape-based features in aerial scene analysis. Adding fixed Gabor filters helps networks avoid relying on texture, leading to more structured representations (Evans et al. [20]). However, this approach restricts the filters to non-adaptive Gabor functions, preventing the network from learning and adapting.

Contributions. This work's contributions are as follows: 1) A novel convolutional layer parameterised by trainable Log-Gabor functions, evaluated both qualitatively and quantitatively against Gabor-based and standard convolutional layers; 2) A new backbone architecture that extracts more salient features, as demonstrated through ablation studies and Class Activation Maps; 3) A novel hybrid CNN-Transformer model leveraging our proposed backbone; and 4) Comprehensive evaluation across multiple aerial datasets and benchmarking against leading models, with extensive qualitative and quantitative analyses demonstrating superior performance in tree crown segmentation and strong generalisation to diverse aerial scene segmentation tasks.

Theoretical Insights. Gabor filters combine sinusoidal waves with Gaussian envelopes (Gabor [21]; Daugman [22]), enabling simultaneous spatial and frequency domain analysis. This property makes them particularly suitable for texture analysis and feature extraction in complex imagery: *Real:* $g(x, y, \omega, \theta, \psi, \sigma, \gamma) =$

$$\exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos(\omega x' + \psi) \quad (1)$$

Imaginary: $g(x, y, \omega, \theta, \psi, \sigma, \gamma) =$

$$\exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2})\sin(\omega x' + \psi) \quad (2)$$

$$x' = x\cos\theta + y\sin\theta; \quad y' = -x\sin\theta + y\cos\theta$$

While Gabor filters offer differentiable shapes (Nava et al. [23] and Boukerroui et al. [24]), they suffer from several limitations:

- *Non-Zero DC Component:* The cosine component's non-zero integration limits sensitivity to high-frequency components [25], [26].
- *Non-Orthogonality:* Overlapping frequency bands across scales reduce discriminative power [27].
- *Non-Uniform Fourier Domain Coverage:* Uneven frequency distribution leads to inadequate high-frequency coverage [24], [28].
- *Lack of True Quadrature:* Inconsistent phase differences affect precision in phase-sensitive tasks [28].

Log-Gabor filters address these limitations through logarithmic frequency scaling (Fischer et al. [28]):

$$G_{\text{log-gabor}}(f) = \exp\left(-\frac{(\log(f/f_0))^2}{2(\log(\sigma/f_0))^2}\right) \quad (3)$$

This design provides zero DC components, improved orthogonality across scales, uniform Fourier domain coverage, and superior spatial localisation.

Convolutional filters in the first layer of CNNs often resemble Gabor filters, capturing basic patterns like edges and textures (Luan et al. [29]; Alekseev and Bobe [30]; Evans et al. [20]). Both types of filters perform spatial and frequency domain analysis, enhancing the architecture's ability to capture spatial localisation, orientation, and spatial frequency selectivity. This is particularly beneficial for semantic segmentation of tree crowns in dense forests from aerial images, as it improves the network's ability to localise tree crowns, detect various orientations, and differentiate between textures, leading to more accurate and robust segmentation results. Moreover, Log-Gabor filters, with their zero DC components, improved orthogonality, uniform Fourier domain coverage, and superior spatial localisation, further enhance these capabilities, resulting in even more precise and reliable segmentation of tree crowns. In the next sections, we perform a comparative analysis, both quantitatively and qualitatively, to evaluate how each implementation capitalises on the theoretical background and impacts semantic segmentation performance.

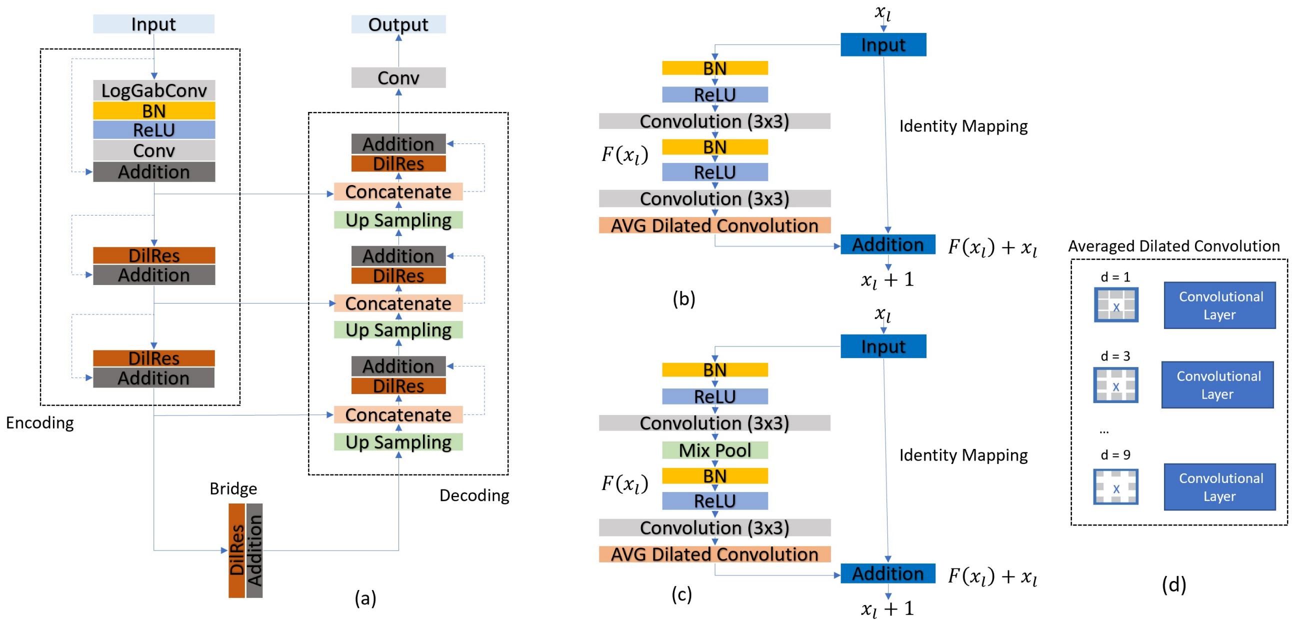

Figure 3. Building blocks of the proposed Architecture: (a) PerceptiveNet architecture, (b) Decoder proposed dilated residual unit (DilRes), (c) Encoder/Bridge proposed dilated residual unit (DilRes), comprised of a mixture of average and maximum pooling layer (Mix Pool) and an averaged Dilated convolutional layer, (d) Dilated convolutional layer. Noticeably, the Mix Pool layer is not present on the decoder.

2. Methods

This section, introduces the proposed model PerceptiveNet and the hybrid CNN-Transformer PerceptiveNeTr, followed by the test datasets. As PerceptiveNet is based on a ResUNet, we first describe it, and then discuss the changes we have made. We end with a brief description of the models we test in which we isolate each of the changes we have made.

2.1. Models

ResUNet. This architecture (Zhang et al. [31]) extends the U-Net model by incorporating residual units (He et al. [32]). It is comprised of an encoder and a decoder. The encoder feature maps low-level fine-grained information, whilst the decoder feature maps high-level, coarse-grained semantic information. Skip-connections between low- and high-level feature maps enhance semantic extraction within the encoder and decoder framework.

PerceptiveNet. Dilated convolution was first introduced by Chen et al. [33] [34] as a way of increasing the receptive field for the task of semantic segmentation. According to Wei et al. [35], convolutional kernel receptive fields are enlarged when employing varying dilation rates, which results in transferring the surrounding discriminative information to the discriminative scene regions. In this work, we propose a novel convolutional layer parameterised by trainable Log-Gabor functions and explore how it performs when combined with averaged dilated convolutions and a mixture of maximum and average pooling layers to impact semantic segmentation performance. Specifically, for the Log-Gabor-parameterised convolutional layer, we used:

$$g(x, y) = g_r(r) \cdot g_\theta(\theta) \cdot \cos(2\pi f_0 r + \psi) \cdot \frac{1}{2\pi\sigma^2} \tag{4}$$

where the radial and angular components are defined as:

$$g_r(r) = \exp\left(-\frac{(\log(r/f_0))^2}{2(\log(\sigma/f_0))^2}\right) \tag{5}$$

$$g_\theta(\theta) = \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma^2}\right) \tag{6}$$

and the variables are defined as:

$$r = \sqrt{x'^2 + y'^2 + \delta}$$

$$x' = x\cos\theta + y\sin\theta; \qquad y' = -x\sin\theta + y\cos\theta$$

Here, (x, y) represents the spatial position, f_0 is the centre frequency, θ is the orientation, θ_0 is the reference orientation, σ is the bandwidth parameter, ψ is the phase offset, and δ is a small constant to prevent division by zero.

Log-Gabor layer weights were initialised by setting the bandwidth parameter σ, centre frequency f_0, and frequency f for each filter. The phase offset ψ is set by uniform distribution Unif.$(0, \pi)$. Notably, the Log-Gabor function parameters (θ, σ, ψ, f_0, θ_0) are learnable and updated during backpropagation as part of the model's optimisation.

This implementation offers several advantages over the standard Gabor filter. The logarithmic nature of the radial component allows for a more even coverage of spatial frequencies and can be designed with arbitrary bandwidth.

Additionally, Log-Gabor functions have no DC component, which can be beneficial in certain image processing tasks. The key differences from a standard Gabor filter include:

- The use of a logarithmic term in the radial component, which provides better frequency coverage.
- A normalisation factor of $\frac{1}{2\pi\sigma^2}$, which ensures proper scaling of the filter.

These modifications allow the Log-Gabor filter to capture a wider range of spatial frequencies and orientations, potentially improving its performance in tasks such as edge detection, texture analysis, and feature extraction for semantic segmentation.

In addition, due to the information complexity in aerial images, a method that combines maximum and average pooling was applied:

$$f_{\text{mix}}(x) = \alpha_l \cdot f_{\max}(x) + (1 - \alpha_l) \cdot f_{\text{avg}}(x) \tag{7}$$

where scalar mixing portion $\alpha_l \in [0, 1]$ indicates the max and average combination per layer l. For the purpose of this work, we chose a scalar mixing portion $\alpha_l = 0.8$.

The use of pooling layers reduces the spatial resolution of the feature map, while it increases the receptive field of feature points. Thus, each feature point in the feature map is influenced by a larger portion of the input image. By increasing the receptive field, the network can capture larger contextual information, including spatial relationships and dependencies between objects, and this benefits the semantic segmentation tasks where objects may appear at different scales and positions within the image.

Similar to [35], an averaged dilated convolutional layer was added on the last convolutional layer of the residual block. Specifically, convolutional blocks with multiple dilated rates (i.e. d = 1, 3, 6, 9) were appended to the final convolutional layer, thus localising scene-related regions observed by different receptive fields. Using high dilation rates (i.e. d = 9) can cause inaccuracies by mistakenly highlighting scene-irrelevant regions. To avoid such scenarios, we used equation 8, where the average over the localisation maps H_i (i.e i = 3, 6, 9) generated by different dilated convolutional blocks was summed to the localisation map H_0 of the convolutional block with dilation d = 1.

$$H = H_o + \frac{1}{n_d} \sum_{n=1}^{n_d} H_i \tag{8}$$

Figure 3 illustrates the architecture under consideration. We replaced the initial ResUNet convolutional layer with a convolutional layer parameterised by trainable Log-Gabor functions. Additionally, we propose a residual block for the encoding and bridge parts of the network, where we replaced stride 2 convolutional layers with stride 1. To maintain effective downsampling, we added a mixture of Maximum and Average pooling layers to halve the feature map size, along with an averaged dilated convolutional layer for enhanced feature extraction. For decoding, we employed a residual block with an averaged dilated convolutional layer.

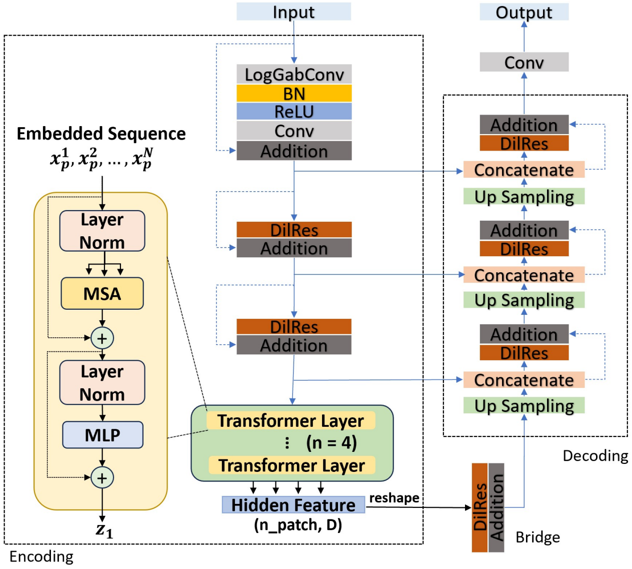

Figure 4. PerceptiveNeTr: Hybrid CNN (PerceptiveNet) - Transformer architecture, leveraging long-range dependencies and global context.

LGMPResUNet. To enhance feature extraction, 1) we replaced the first convolutional with a Logarithmic Gabor-parameterised convolutional layer; 2) the encoding path is comprised of three residual blocks. Instead of using the first convolutional layer stride 2 to downsample the feature map size, we used stride 1 and added a mixture of maximum and average pooling layers to reduce the feature map by half.

DilResUNet. This model employs dilated convolutional blocks to enhance the representation capacity of the convolution layers when compared to the ResUNet. Thus, added dilated convolutional blocks with multiple dilated rates (i.e. d = 1, 3, 6, 9). This enables the extraction of fine-grained detailed and coarse-grained semantic information.

PerceptiveNeTr. We propose an architecture that combines elements from CNNs and transformer models, enabling the capture of long-range dependencies and global context [36]. It is structured into three main components: Encoding, Decoding, and a Bridge (Figure 4). The Encoding section begins the proposed Log-Gabor parameterised convolutional layer, followed by the proposed backbone described above. An embedded sequence processing component is included, representing the Patch Embedding, and features Layer Normalisation, Multi-Head Self-Attention (MSA), and Multi-Layer Perceptron (MLP) layers. Finally, the encoding section consists of a stack of four Transformer layers (n=4), which correspond to the Transformer Encoder described in the model. The Bridge and Decoding section remain as described previously.

ViTResUNet. This model is the same as the hybrid CNN-Transformer PerceptiveNeTr described above, but uses a ResUNet as the backbone instead.

Model Training. Due to novelty, all models were trained from scratch for 130 epochs (based on when train/validation results stopped improving), using Adam Optimiser and Cross Entropy loss and a batch size of 16. The data was split as 80% train/validation (of which, 80% train; 20% validation) and 20% for testing. The only data augmentations applied during training were geometric transformations, i.e rotation with 90% probability and horizontal and vertical flip with 50% and 10% probabilities respectively.

Performance Metrics. To measure model performance two scores are used, Pixel accuracy (Acc) and mean Intersection of a Union (mIoU) score. As pixel accuracy (calculated as the proportion of correctly predicted pixels) might not accurately reflect a model's performance when e.g. water or ground scenes dominate an image, we additionally use mIoU as it takes into account the area covered by each of the k classes via:

$$\text{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{9}$$

2.2. Data

Alongside TreeCrown, two additional datasets were chosen to evaluate model generalisation across diverse aerial scenes of varying complexity. While differing from dense forests, both exhibit similar challenges, such as complex spatial boundaries and occlusions caused by canopy and shadows.

TreeCrown. The dataset by Cloutier et al. [37] covers a temperate-mixed forest in the Laurentides region of Québec, Canada, during 2021. It includes 1,360 (768×768-pixels) UAV images acquired monthly from May to August. The dataset contains 23,000 segmented tree crowns with per-pixel species annotations for 14 tree species classes.

Landcover.AI. The dataset by Boguszewski et al. [38] depicts RGB aerial images and annotated buildings, forests, water bodies and roads in Poland. The dataset is comprised of 33 orthophotos with 25 cm per pixel resolution (9000×9500-pixels) and 8 orthophotos with 50 cm per pixel resolution (4200×4700-pixels) of a total area of 216.27 km^2. We sub-sampled the orthophotos to produce 21,924 (256×256-pixels) images and their corresponding masks.

UAVid. This is a UAV semantic segmentation dataset by Lyu et al. [39], that focuses on urban scenes with two spatial resolutions (3840×2160 and 4096×2160-pixels) and eight class labels (Building, Road, Static car, Tree, Low vegetation, Human, Moving car, and Background). Each image was padded and cropped into eight 256×256-pixel patches.

3. Results and Discussion

3.1. Comparative Analysis: Standard, Gabor and Log-Gabor Convolutional Layers

In this section, we evaluate the impact of using a standard initial convolutional layer, a Gabor-based, and a Log-Gabor parameterised initial convolutional layer in tree crown semantic segmentation. The theoretical advantages of Log-Gabor filters—including improved frequency coverage, zero DC component, reduced artefacts, enhanced edge detection, and adaptability to shape variability—suggest that the Log-Gabor parameterised convolutional layer is likely to extract more salient features compared to both the Gabor-based and standard convolutional layers. By conducting qualitative and quantitative analyses, we investigate how these theoretical insights are reflected in the experimental results, thereby illustrating the effectiveness of our proposed Log-Gabor parameterised convolutional layer in accurately segmenting tree crowns in dense forests.

Quantitative Analysis. We evaluate the impact of using a standard, Gabor-based, and trainable Log-Gabor function-parameterised initial convolutional layer within our proposed backbone on semantic segmentation performance. Table 1 shows that in the complex TreeCrown dataset, the parameterised Log-Gabor outperforms both the convolution and the Gabor one by 4.6% and 3.7% mIoU respectively. Experiments on Landcover.AI and UAVid datasets confirm this trend, with Log-Gabor outperforming conventional and Gabor implementations by 3.5% and 2.8% mIoU on Landcover.AI, and 4.5% and 1.5% mIoU on UAVid, respectively.

Table 1. Convolutional vs Gabor vs Log-Gabor Parameterised Initial Convolutional Layer Semantic Segmentation Performances.

Dataset	TreeCrown		Landcover.AI		UAVid	
Architecture	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**
PerceptiveNet$^{\text{Conv}}$	82.9	43.5	91.3	78.1	85.3	64.1
PerceptiveNet$^{\text{Gab}}$	82.2	44.4	92.1	78.8	86.2	67.1
PerceptiveNet$^{\text{LogGab}}$	**84.4**	**48.1**	**92.5**	**81.6**	**87.1**	**68.6**

Qualitative Analysis. We perform a visual inspection of the segmentation masks predicted when using the two initial Log-Gabor and Gabor parameterised convolutional layers under review. As is evident in Figure 6, incorporating a Log-Gabor convolutional layer allows the model to capture finer spatial details, resulting in more accurate segmentation across different tree crown sizes. Moreover, the Log-Gabor convolutional layer is very efficient at differentiating closely packed trees, effectively managing complex crown shapes and smaller, irregular crowns. Particularly in the presence of dead trees, Log-Gabor outperforms Gabor or normal convolutional layers by capturing more high-frequency features, enhancing its overall segmentation performance. This demonstrates the importance of zero DC components, improved orthogonality across scales, uniform

Fourier domain coverage, and spatial localisation. These properties enhance the model's ability to perform semantic segmentation tasks in complex, challenging dense forests.

3.2. Ablation Study

As the proposed architecture comprises of two components, a texture-biased part for extracting salient features and a wider receptive field, we evaluated how each of these components contributes to segmentation performance.

Table 2. Ablation Study - Semantic Segmentation Performance.

Dataset	TreeCrown		Landcover.AI		UAVid	
Architecture	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**
ResUNet	79.3	37.6	91.3	72.6	82.6	59.7
DilResUNet	80.6	40.5	89.4	76.8	82.9	63.1
LGMPResUNet	82.2	44.4	90.6	78.4	85.8	66.4
PerceptiveNet	**84.4**	**48.1**	**92.5**	**81.6**	**87.1**	**68.6**

Table 2 results indicate that DilResUNet architecture, with averaged dilated convolutional residual blocks, improves mIoU by 2.9% (TreeCrown), 4.2% (Landcover.AI), and 3.4% (UAVid) over ResUNet. LGMPResUNet, integrating an initial Log-Gabor parameterised convolutional layer and residual blocks comprised of a mixture of average and maximum pooling layers, further boosts mIoU by 6.8%, 5.8%, and 6.7%, respectively. The proposed PerceptiveNet, combining an initial Log-Gabor parameterised convolutional layer, residual blocks consisting of a mixture of average and maximum pooling layers, and averaged dilated convolutional layer, achieves the best mIoU scores of 10.5% (TreeCrown), 9.0% (Landcover.AI), and 8.9% (UAVid) compared to ResUNet. This study demonstrates that the proposed layers complement each other to further enhance the semantic segmentation performance.

3.3. Comparative Feature Extraction Analysis

In this section, we employ Class Activation Mapping (CAM, Zhou et al. [40]) to gain insight into the model's decision process by overlaying a heatmap on the original image, indicating the discriminative region used by the model when predicting that an image belongs to a particular class. Figure 5 illustrates the CAMs for images containing forests and agricultural land, comparing the feature extraction capabilities of a standard CNN with those of the proposed model. In the ResUNet column, the standard model's activation patterns are diffuse, with focus areas spread across both forested regions and agricultural lands. This scattered attention suggests inefficient feature capture, missing critical details and leading to less robust representations.

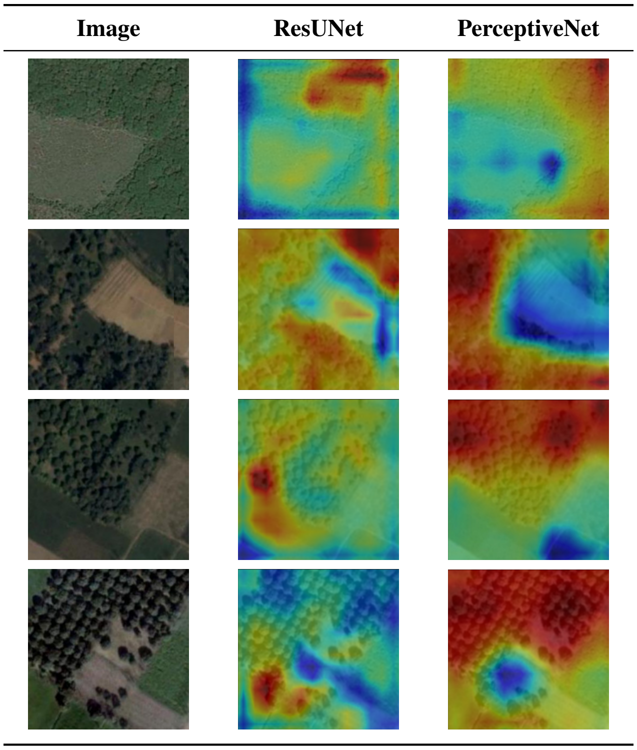

Figure 5. Class Activation Maps, ResUNet vs PerceptiveNet Encoder feature extraction capabilities. PerceptiveNet: More focused and detailed activation pattern.

In contrast, the PerceptiveNet model demonstrates a more focused and detailed activation pattern, particularly in areas with dense tree coverage. This is expected due to the model's enhanced ability to extract salient features combined with a wider receptive field. Thus, the activation maps display more defined structures and higher contrast, especially in forested regions. This targeted focus indicates that our model develops more robust and specific representations of forest structures. The enhanced ability to concentrate on relevant features, particularly trees, suggests better out-of-distribution generalisation, crucial for tasks like dense forest tree crown segmentation. Comparing these activation maps, we infer that the proposed model's architecture allows for more efficient and targeted feature extraction, potentially improving performance.

3.4. Quantitative Analysis

We evaluate the semantic segmentation performance of our proposed architectures through two analyses. First, we compare the PerceptiveNet against State-Of-The-Art (SOTA) CNN models including U-Net [7], MACUNet [41], ResUNet [31], and DeepLabV3+ [42]. Second, we investigate how semantic segmentation performance benefits from the combination of capturing long-range dependencies and global context through our hybrid CNN-Transformer model (ViTResUNet), comparing it against SOTA Transformer architectures UNETR_2D [43] and SwinUNet [44], as well as examine the generalisation of the proposed backbone (PerceptiveNeTr) within such a hybrid framework. All evaluations are conducted across three complex aerial scene SOTA datasets: TreeCrown, Landcover.AI, and UAVid.

3.4.1. Convolutional Models

For the TreeCrown dataset in Table 3, ResUNet achieved an mIoU of 37.6%. U-Net improved this by 5.6%. MACUNet further enhanced the results to 45.9%, an 8.3% improvement, while DeeplabV3+ showed a similar improvement with an mIoU of 45.7%, 8.1% higher than ResUNet. The proposed PerceptiveNet outperformed all models with an mIoU of 48.1%, a 10.5% improvement over ResUNet.

Table 3. Semantic Segmentation Performance in Various Datasets.

Dataset	TreeCrown		Landcover.AI		UAVid	
Architecture	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**
ResUNet	79.3	37.6	91.3	72.6	82.6	59.7
UNet	81.9	43.2	91.1	73.1	85.7	64.9
MACUNet	83.7	45.9	91.4	74.3	85.7	66.1
DeepLabV3+	81.9	45.7	91.9	78.6	86.7	65.8
PerceptiveNet	**84.4**	**48.1**	**92.5**	**81.6**	**87.1**	**68.6**

In the Landcover.AI dataset, ResUNet recorded an mIoU of 72.6%. U-Net achieved 73.1% mIoU, 0.5% higher. MACUNet achieved an mIoU of 74.3%, which is 1.7% higher. DeepLabV3+, with a mIoU score of 78.6% improved by 6%. PerceptiveNet model achieved the best performance, with an mIoU of 81.6%, 9% higher than ResUNet.

For the UAVid dataset, ResUNet attained an mIoU of 59.7%. The U-Net improved to 64.9% mIoU, 5.2% higher. MACUNet achieved 66.1%, 6.4% higher. DeepLabV3+ reached 65.8%, a 6.1% improvement. The PerceptiveNet model led with 68.6% mIoU, 8.9% higher than ResUNet.

3.4.2. Transformer Models

The summarised results in Table 4 indicate 2 key findings:

Table 4. Transformer vs Hybrid CNN-Transformer Semantic Segmentation Performance in Various Datasets.

Dataset	TreeCrown		Landcover.AI		UAVid	
Architecture	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**	**Acc(%)**	**mIoU(%)**
UNETR_2D	75.2	27.1	81.2	52.4	80.4	55.9
SwinUNet	76.9	36.5	89.1	67.3	82.3	58.3
ViTResUNet	79.9	37.8	89.1	68.2	83.9	60.1
PerceptiveNeTr	**81.8**	**42.0**	**91.2**	**75.3**	**84.8**	**64.3**

Proposed Hybrid CNN-Transformer: The ViTResUNet model outperforms the pure Transformer models UNETR_2D and SwinUNet by 10.7% & 1.3% mIoU (TreeCrown), 15.8% & 0.9% mIoU (Landcover.AI), and 4.2% & 0.8% mIoU (UAVid). This demonstrates that hybridising CNN and Transformer elements benefits from capturing long-range dependencies and global context, leading to enhanced performance in semantic segmentation tasks.

PerceptiveNet Backbone Impact: The proposed Log-Gabor parameterised convolutional layer and backbone significantly enhance the performance of the hybrid model. PerceptiveNeTr achieves higher mIoU scores compared to ViTResUNet, improving even further the performance by 4.2% mIoU for TreeCrown, 7.1% mIoU for Landcover.AI, and 4.2% mIoU for UAVid datasets. This demonstrates the generalisability of our PerceptiveNet model, indicating the benefits of implementing the proposed backbone even within a hybrid CNN-Transformer framework.

Overall, the proposed architectures demonstrated superior performance, with PerceptiveNet achieving the highest mIoU scores amongst all models and datasets and PerceptiveNeTr outperforming all Transformer architectures. The integration of Log-Gabor parameterised convolutional layer, residual blocks with mixed pooling layers and averaged dilated convolutions enhanced semantic segmentation in both standalone and hybrid implementations, proving effective at capturing complex features across various models.

3.5. Qualitative Analysis

To assess the performance of our proposed segmentation model, we conduct a visual comparison of tree crown segmentation results across the original ResUNet and proposed PerceptiveNet (LogGab PerceptiveNet). Furthermore, to asses the effect of the proposed initial Log-Gabor parameterised convolutional layer, we provide a visual comparison of the proposed model with an initial parameterised Gabor convolutional layer (Gabor PerceptiveNet). The segmentation results, presented in Figure 6, highlight each model's effectiveness in capturing tree crowns with varying characteristics, including living and decaying trees. While accuracy in segmenting tree crowns is the main focus, the red-circled dead trees aid analysis as reference points.

ResUNet demonstrates a reasonable ability to segment large tree crowns, yet it encounters challenges when it comes to finer boundary delineations, particularly with smaller and irregularly shaped trees. The model often over-segments larger crowns while under-segmenting smaller ones, leading to a reduction in segmentation accuracy. Dead trees, which exhibit sparse or absent foliage, are especially problematic for ResUNet, with frequent misclassifications or merging with nearby crowns observed in the results.

The Gabor PerceptiveNet, shows improved capability in capturing texture and boundary details. This is apparent in the clearer segmentation of tree crowns compared to ResUNet. While some inaccuracies persist—most notably with smaller crowns and more complex crown structures—the model provides improvements, particularly in separating adjacent tree crowns. Despite errors, the Gabor filters help the model differentiate between trees more effectively, including those exhibiting signs of decay or death.

PerceptiveNet exhibits the highest accuracy in segmenting tree crowns, particularly when it comes to delineating boundaries. The use of convolutional layer parameterised by trainable Log-Gabor functions enables the model to cap-

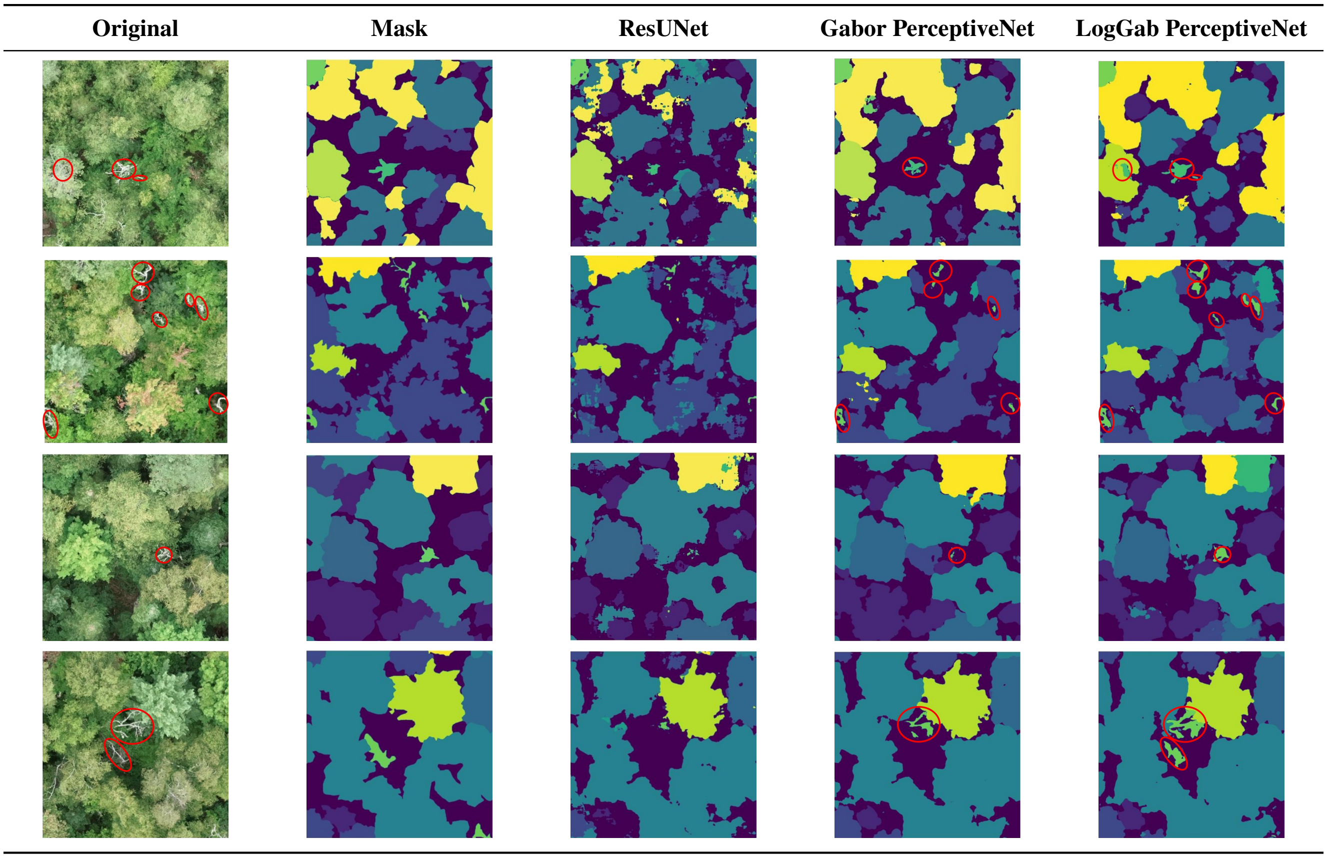

Figure 6. Aerial Tree Crown Semantic Segmentation of a Dense Forest (TreeCrown), Comprised of Visually Similar Tree Species. The left column shows the original image; the right columns show the labelled masks and the segmentation results. Red circles indicate dead trees.

ture finer spatial details, translating into more precise segmentation across various tree crown sizes. PerceptiveNet excels at distinguishing closely situated trees and minimising over-segmentation, effectively handling complex crown shapes and smaller, irregular crowns. Notably, while dead trees pose a challenge for all models, PerceptiveNet outperforms them by capturing more high-frequency features, contributing to its superior segmentation performance.

Overall, PerceptiveNet delivers the most proficient segmentation of tree crowns, including those of varying sizes and health statuses. The Log-Gabor initial convolutional layer enhances feature extraction, allowing for clearer boundary detection and more accurate segmentation. These visualisations demonstrate the importance of the Log-Gabor properties: zero DC components, improved orthogonality across scales, uniform Fourier domain coverage, and superior spatial localisation in semantic segmentation.

4. Conclusion

While tree crown semantic segmentation is essential for ecology, forestry, agriculture, and biodiversity studies, it is significantly impacted by factors such as shadows, light variations, overlapping tree crowns, and weak distinctive features among tree species. These challenges are further exacerbated in dense forests during the green leaf season. To address these issues, we proposed PerceptiveNet, a model that extracts salient features while capturing contextual and spatial information through a wider receptive field. Our model significantly outperforms SOTA models on the TreeCrown and two benchmark aerial scene datasets. Qualitative analysis shows enhanced feature extraction, enabling clearer boundary detection and more accurate segmentation.

Additionally, quantitative and qualitative analyses demonstrate that the proposed convolutional layer, parameterised by trainable Log-Gabor functions, outperforms both traditional Gabor-based and standard convolutional layers by effectively leveraging the strengths of Log-Gabor filters for semantic segmentation. Moreover, an investigation into the proposed individual layers and their combinations reveals that while each contributes to performance gains, their synergistic integration leads to even greater improvements.

Finally, the proposed hybrid CNN-Transformer model, PerceptiveNeTr, illustrates the advantages of capturing long-range dependencies and global context. Although PerceptiveNeTr's performance metrics are lower compared to our pure CNN PerceptiveNet, it establishes a foundation for future research aimed at integrating advanced transformer models, such as the Hierarchical Vision Transformer, along with pre-trained ones to improve performance. Importantly, this comparison highlights the robust generalisation of our PerceptiveNet backbone across various applications.

Acknowledgements: This work was funded in part by UKRI Frontiers grant EP/X024520/1.

References

[1] M. Bokalo, K. J. Stadt, P. G. Comeau, and S. J. Titus, "The validation of the mixedwood growth model (mgm) for use in forest management decision making," *Forests*, vol. 4, no. 1, pp. 1–27, 2013. 1

[2] Y. Li, W. Wang, W. Zeng, J. Wang, and J. Meng, "Development of crown ratio and height to crown base models for masson pine in southern china," *Forests*, vol. 11, no. 11, p. 1216, 2020. 1

[3] J. R. G. Braga, V. Peripato, R. Dalagnol, M. P. Ferreira, Y. Tarabalka, L. E. OC Aragão, H. F. de Campos Velho, E. H. Shiguemori, and F. H. Wagner, "Tree crown delineation algorithm based on a convolutional neural network," *Remote Sensing*, vol. 12, no. 8, p. 1288, 2020. 1

[4] G. Hu, T. Wang, M. Wan, W. Bao, and W. Zeng, "Uav remote sensing monitoring of pine forest diseases based on improved mask r-cnn," *International Journal of Remote Sensing*, vol. 43, no. 4, pp. 1274–1305, 2022. 1

[5] N. E. Ocer, G. Kaplan, F. Erdem, D. Kucuk Matci, and U. Avdan, "Tree extraction from multi-scale uav images using mask r-cnn with fpn," *Remote sensing letters*, vol. 11, no. 9, pp. 847–856, 2020. 1

[6] K. Cao and X. Zhang, "An improved res-unet model for tree species classification using airborne high-resolution images," *Remote Sensing*, vol. 12, no. 7, p. 1128, 2020. 1

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015. 1, 6

[8] Z. Ye, J. Wei, Y. Lin, Q. Guo, J. Zhang, H. Zhang, H. Deng, and K. Yang, "Extraction of olive crown based on uav visible images and the u2-net deep learning model," *Remote Sensing*, vol. 14, no. 6, p. 1523, 2022. 1

[9] J. Zhang, T. Xie, C. Yang, H. Song, Z. Jiang, G. Zhou, D. Zhang, H. Feng, and J. Xie, "Segmenting purple rapeseed leaves in the field from uav rgb imagery using deep learning as an auxiliary means for nitrogen stress detection," *Remote Sensing*, vol. 12, no. 9, p. 1403, 2020. 1

[10] F. H. Wagner, A. Sanchez, Y. Tarabalka, R. G. Lotte, M. P. Ferreira, M. P. Aidar, E. Gloor, O. L. Phillips, and L. E. Aragao, "Using the u-net convolutional network to map forest types and disturbance in the atlantic rainforest with very high resolution images," *Remote Sensing in Ecology and Conservation*, vol. 5, no. 4, pp. 360–375, 2019. 1

[11] F. Schiefer, T. Kattenborn, A. Frick, J. Frey, P. Schall, B. Koch, and S. Schmidtlein, "Mapping forest tree species in high resolution uav-based rgb-imagery by means of convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 170, pp. 205–215, 2020. 1

[12] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1422–1431, 2022. 1

[13] P. Vinod, M. Behera, A. J. Prakash, R. Hebbar, and S. Srivastav, "A novel multitask transformer deep learning architecture for joint classification and segmentation of horticulture plantations using very high-resolution satellite imagery," *Computers and Electronics in Agriculture*, vol. 227, p. 109540, 2024. 1

[14] H. H. Alshammari and O. R. Shahin, "An efficient deep learning mechanism for the recognition of olive trees in jouf region," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 9249530, 2022. 1

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015. 1

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017. 1

[17] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, pp. 267–285, Springer, 1982. 2

[18] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018. 2

[19] G. Voulgaris, A. Philippides, J. Dolley, J. Reffin, F. Marshall, and N. Quadrianto, "Seasonal domain shift in the global south: Dataset and deep features analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2115–2123, 2023. 2

[20] B. D. Evans, G. Malhotra, and J. S. Bowers, "Biological convolutions improve dnn robustness to noise and generalisation," *Neural Networks*, vol. 148, pp. 96–110, 2022. 2

[21] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946. 2

[22] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985. 2

[23] R. Nava, B. Escalante-Ramírez, and G. Cristóbal, "Texture image retrieval based on log-gabor features," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17*, pp. 414–421, Springer, 2012. 2

[24] D. Boukerroui, J. A. Noble, and M. Brady, "On the choice of band-pass quadrature filters," *Journal of Mathematical Imaging and Vision*, vol. 21, pp. 53–80, 2004. 2

[25] F. Heitger, L. Rosenthaler, R. Von Der Heydt, E. Peterhans, and O. Kübler, "Simulation of neural contour mechanisms: from simple to end-stopped cells," *Vision research*, vol. 32, no. 5, pp. 963–981, 1992. 2

[26] C. Ronse, "On idempotence and related requirements in edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 5, pp. 484–491, 1993. 2

[27] G. H. Granlund and H. Knutsson, *Signal processing for computer vision*. Springer Science & Business Media, 2013. 2

[28] S. Fischer, F. Šroubek, L. Perrinet, R. Redondo, and G. Cristóbal, "Self-invertible 2d log-gabor wavelets," *International Journal of Computer Vision*, vol. 75, pp. 231–246, 2007. 2

[29] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018. 2

[30] A. Alekseev and A. Bobe, "Gabornet: Gabor filters with learnable parameters in deep convolutional neural network," in *International Conference on Engineering and Telecommunication*, pp. 1–4, 2019. 2

[31] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018. 3, 6

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 3

[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014. 3

[34] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 3

[35] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7268–7277, 2018. 3, 4

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021. 4

[37] M. Cloutier, M. Germain, and E. Laliberté, "Influence of temperate forest autumn leaf phenology on segmentation of tree species from uav imagery using deep learning," *Remote Sensing of Environment*, vol. 311, p. 114283, 2024. 5

[38] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1102–1110, June 2021. 5

[39] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, 2020. 5

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016. 6

[41] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021. 6

[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018. 6

[43] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022. 6

[44] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, pp. 205–218, Springer, 2022. 6