Grenze International Journal of Engineering and Technology, June Issue

IndiRoad-Net: An AI-based Architecture for Automatic Road Network Extraction from Indian Satellite Imagery

Isra[1] and S. Pravinth Raja[2]
[1]PG Scholar, School of Computer Science & Engineering, Presidency University, Bengaluru, Karnataka, India
Email: isramomina@gmail.com
[2]Professor & HOD, School of Computer Science and Engineering, Presidency University, Bengaluru, Karnataka, India
Email: pravinth.raja@presidencyuniversity.in

***Abstract*— Segmentation of roads from satellite imagery is particularly challenging in dense Indian urban areas due to canopy occlusion, shadows, and severe class imbalance (road pixels below 5% per tile, rising to 50:1 background-to-road ratio in peri-urban scenes). Existing models trained on Western benchmarks such as DeepGlobe and Massachusetts Roads do not generalise well to Indian imagery conditions including monsoon cloud cover and dense deciduous canopies lining arterial roads. To address this, we propose IndiRoad-Net: a five-stage deep learning pipeline comprising an EfficientNet-B3 encoder, Attention U-Net++ decoder, Partial Convolution-based occlusion handling, topology correction via Zhang-Suen skeletonization and Douglas-Peucker simplification, and GeoJSON export for GIS integration. Trained on 850 DeepGlobe tiles and evaluated on 150 held-out tiles at threshold $\tau = 0.3$, the model achieves Recall of 0.823 and AUC-ROC of 0.950, prioritising road detection completeness for routing applications. Zero-shot transfer to Sentinel-2 imagery over Bengaluru correctly identifies major arterial roads without Indian fine-tuning.**

***Index Terms*—road extraction, Indian satellite imagery, deep learning, EfficientNet-B3, Attention U-Net++, occlusion handling, partial convolution, topology correction, GeoJSON, DeepGlobe, Sentinel-2, Bengaluru.**

I. INTRODUCTION

India's road network spans over 6.4 million kilometers, ranking third largest in the world, yet a significant portion remains unmapped or outdated in public datasets. OpenStreetMap contributions consistently lag behind actual construction — roads such as the Bengaluru Peripheral Ring Road and Chennai Outer Ring Road were either absent or incomplete at the time of writing. Automated road extraction from satellite imagery is the natural solution, but applying it to Indian scenes introduces challenges that standard benchmarks do not capture.

A. Problem Statement

The central problem is: how can road networks be reliably extracted from Indian satellite imagery, given three structural conditions that cause standard pipelines to fail? First, dense canopies of banyan and rain trees fragment visible road surfaces, and monsoon cloud cover persists four to five months annually — a model with no mechanism for missing pixels produces broken, unusable road maps. Second, road pixels constitute only ~4.5% of DeepGlobe tiles (21:1 imbalance), rising to ~50:1 in peri-urban Indian imagery; without an explicit correction, models learn to predict background.

Grenze ID: 01.GIJET.12.2.42_26

Third, available benchmarks (Massachusetts Roads, DeepGlobe) were collected under clear weather with broad, well-defined roads — properties that do not hold in Indian urban imagery, making models trained on them poor starting points.
When D-LinkNet, winner of the 2018 DeepGlobe challenge, is applied directly to Sentinel-2 imagery over Bengaluru, it produces fragmented road maps in canopy-heavy corridors — not due to architectural flaws, but because its training data never presented these conditions.

B. Proposed Solution

IndiRoad-Net is a five-stage pipeline that maps each challenge to a specific module: (1) a threshold-based occlusion detector produces a validity mask that suppresses shadow, canopy, and haze pixels via Partial Convolution before encoding; (2) a weighted loss ($\omega = 10.0$) counteracts the 50:1 class imbalance; (3) a topology correction module skeletonizes predictions and reconnects endpoints within 25 pixels, restoring occlusion-induced breaks. A lightweight material classification head assigns surface labels (asphalt, concrete, unpaved) using RGB pseudo-labels with no additional annotation. The final output is a GeoJSON road network ready for QGIS or ArcGIS.
The key contributions of this work are:

- Explicit problem formulation mapping three Indian-imagery challenges — occlusion, class imbalance, and domain mismatch — to targeted architectural solutions.
- Partial Convolution-based occlusion handling to preserve road continuity under dense canopy and cloud cover.
- Attention U-Net++ decoder with dual heads for road segmentation and surface material classification.
- Post-processing topology correction via Zhang-Suen skeletonization and linear gap-bridging.
- Quantitative evaluation on 150 DeepGlobe tiles (Recall 0.823, AUC-ROC 0.950) and qualitative zero-shot transfer to Sentinel-2 Bengaluru imagery, with full GeoJSON pipeline output.

II. Related Work

A. From Hand-Crafted Features to Learned Representations

Until the invention of convolutional networks, techniques such as morphology-based filters, active contours, and Markov random fields were applied to extract roads. These methods incorporated domain-specific information but were brittle under varying lighting and occlusions. Mnih and Hinton demonstrated that a fully-convolutional network trained on aerial imagery surpasses hand-crafted feature extractors on the Massachusetts Roads dataset [18].

B. Encoder-Decoder Architectures and Skip Connections

Long et al. transformed image classification networks into fully convolutional models producing spatial probability maps [1].
U-Net added symmetric skip connections between encoder and decoder stages [2], crucial for preserving thin road edge information lost during downsampling. D-LinkNet, winner of the 2018 DeepGlobe challenge, combines a ResNet-50 encoder with dilated bottleneck convolutions, achieving a Dice score of 0.628 on the full challenge set — the primary baseline for IndiRoad-Net [11].

C. Attention Mechanisms and Multi-Scale Decoding

Attention U-Net modifies skip connections with learned gating signals, reducing weight on areas unlikely to contain the target structure [3]. U-Net++ replaces each skip connection with a dense sub-network capturing features from multiple decoder depths simultaneously [4]. IndiRoad-Net adopts the Attention U-Net++ decoder for its gating and multi-scale fusion capabilities, while keeping the EfficientNet-B3 encoder's parameter count manageable compared to ResNet-50 alternatives.

D. Occlusion and Domain Shift in Indian Imagery

Most published road extraction work uses DeepGlobe or Massachusetts Roads, neither representative of Indian urban occlusion conditions where monsoon cloud cover persists four to five months and dense canopy lines major arterial roads year-round.
Partial convolution [5], originally proposed for image inpainting, addresses how to convolve feature maps when some pixels are masked. We borrow this mechanism to suppress occlusion-masked pixels before they propagate through the EfficientNet encoder.

E. Topology Correction and Vectorization

Binary segmentation masks are not directly usable in routing or GIS workflows. Batra et al. [8] showed that predicting road orientation jointly with segmentation reduces topological errors. Our topology correction stage uses Zhang-Suen skeletonization [9] followed by linear gap-bridging for endpoint pairs within 25 pixels. Vectorization uses Douglas-Peucker simplification [10] at two-pixel tolerance, producing GeoJSON LineStrings annotated with material class and mean confidence.

III. DATASET

A. DeepGlobe Road Extraction Dataset

The DeepGlobe Road Extraction Dataset [7] includes 6,226 high-resolution RGB image-mask pairs at 0.5 m/pixel from Thailand, Indonesia, and India. Each 1024×1024 tile has a binary road mask. For these experiments, 1,000 pairs were randomly selected and split 850/150 for training and validation. All images were resized to 512×512 pixels. Road pixels represent 4.48% of the training set on average, a background-to-road imbalance of approximately 21:1.

B. Sentinel-2 Imagery over Bengaluru

Three Sentinel-2 Level-2A scenes were obtained from the ESA Copernicus Open Access Hub, covering Bengaluru (12.85–13.15°N, 77.45–77.75°E): pre-monsoon (2024-06-01), post-monsoon (2024-12-08), and winter dry season (2024-12-23). Each scene provides four bands (R, G, B, NIR) at 10 m resolution. Road labels were collected from OpenStreetMap and converted to binary masks with a 5-pixel dilation radius.

C. Data Preprocessing and Augmentation

All pixel values are normalized to [0, 1] by dividing by 255. Since no manually labelled material ground truth is available, pseudo-material labels are derived from RGB spectral characteristics. The material classification rule mat(p) is defined as:

$$mat(p) = Asphalt \ \ if \ b(p) < 0.20; \ \ Concrete \ \ if \ 0.20 \leq b(p) < 0.55 \ and \frac{R}{G} < 1.1; \ \ Unpaved \ \ if \ b(p) \geq 0.20 \ and \frac{R}{G} \geq 1.1; \ \ Non-road \ \ if \ ExG(p) > 0.08 \quad (1)$$

where b(p) = (R + G + B)/3 denotes pixel brightness, R/G is the redness ratio used to distinguish warm-toned unpaved surfaces, and ExG = 2G − R − B is the excess green index that separates vegetation from bare ground. Training augmentations include horizontal flip, vertical flip, 90° rotation, and multiplicative brightness jitter in the range [0.8, 1.2].

IV. METHODOLOGY

A. Overall Architecture

IndiRoad-Net processes a 512×512×3 RGB tile through five sequential stages: (1) occlusion detection and partial convolution feature extraction, (2) EfficientNet-B3 multi-scale encoding, (3) Attention U-Net++ decoding with dual output heads, (4) topology correction, and (5) GeoJSON vectorization. The architecture is illustrated in Fig. 1.

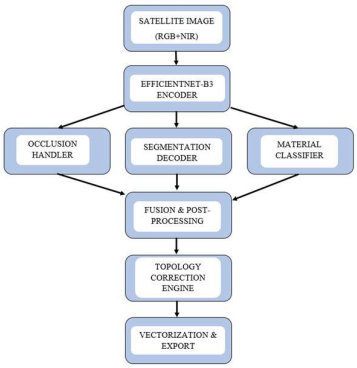

Fig 1

B. Stage 1 — Occlusion Detection and Partial Convolution

Three binary masks are computed from the input RGB tile using simple per-pixel thresholds:

$$M_{shadow} = \{\, p : b(p) < 0.15 \,\} \tag{2}$$

$$M_{canopy} = \{\, p : ExG(p) > 0.10 \wedge G(p) > R(p) \wedge G(p) > B(p) \} \tag{3}$$

$$M_{haze} = \{\, p : b(p) > 0.80 \,\} \tag{4}$$

A validity mask M_v = ¬(M_s ∪ M_c ∪ M_h) marks pixels that are free of detectable occlusion. This mask is passed to two stacked Partial Convolution layers [5], which process only valid pixels:

$$x' = W^T(X \odot M) * \left(\frac{|\Omega|}{|\mathrm{M}|}|M|\right) + b, \;\; M' \leftarrow 1[|M| > 0] \tag{5}$$

where X is the input feature map, W is the convolutional kernel, M is the binary validity mask, ⊙ denotes element-wise multiplication, |Ω| is the total number of kernel positions, |M| counts valid (unoccluded) positions, and b is the bias term. The ratio |Ω|/|M| adjusts the output to account for missing inputs. The updated mask M′ is carried forward, marking any position with at least one valid input as valid.

C. Stage 2 — EfficientNet-B3 Encoder

The encoding path uses EfficientNet-B3 [6] pre-trained on ImageNet, which scales depth, width, and input resolution using a compound coefficient. Feature maps are tapped at four intermediate blocks forming a pyramid at 128×128, 64×64, 32×32, and 16×16 resolution. The first 100 layers are frozen to preserve ImageNet low-level features; remaining layers are fine-tuned with gradient clipping (norm = 1.0).

D. Stage 3 — Attention U-Net++ Decoder

The decoder follows the U-Net++ topology [4], with nested dense sub-networks replacing the straight skip connections of vanilla U-Net. At each pyramid level l, the skip connection feature s^l is first weighted by an attention gate [3]:

$$\alpha^l = \sigma(\, W\psi^{\mathrm{T}} \cdot ReLU(\, W\theta^{\mathrm{T}} s^l + W\varphi^{\mathrm{T}} g^l + b^1) + b^2) \tag{6}$$

where s^l is the skip connection feature map from encoder level l, g^l is the gating signal upsampled from the next deeper decoder level, Wθ, Wφ, and Wψ are learned linear transformations, b_1 and b_2 are bias terms, and σ(·) is the sigmoid activation that produces a soft attention coefficient $\alpha^l \in (0, 1)$ for each spatial location. The attended features are concatenated with the transposed convolution output and processed by a double 3×3 convolution block (Conv→BN→ReLU, repeated twice). Dropout rates decrease from 0.3 at the deepest decoder level to 0.1 at the shallowest.

Occlusion features from Stage 1 are bilinearly resized to 512×512 and concatenated with the final decoder features f before two parallel output heads:

$$\hat{y}_r = \sigma\big(\, Conv_{\{1\times1\}(f)}\big) \in [0,1]^{\{H\times W\}} \tag{7}$$

$$\hat{y}_m = softmax\left(\, Conv_{\{1\times1\}(\, ConvBlock(f))}\right) \in \mathbb{R}^{\{H\times W\times 4\}} \tag{8}$$

where ŷ_r is the predicted road probability map, ŷ_m is the per-pixel material class distribution over four classes, H and W denote spatial height and width (both 512 pixels), and Conv_{1×1} is a 1×1 convolution mapping fused features to the required output channels.

E. Training Objective

The total loss combines three terms:

$$L = L_{\{wBCE\}} + L_{\{Dice\}} + 0.3 \cdot L_{\{CE\}} \tag{9}$$

The road segmentation loss is a sum of weighted binary cross-entropy and Dice loss:

$$L_{\{wBCE\}} = -\left(\frac{1}{N}\right)\Sigma_i[\, \omega y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{10}$$

$$L_{Dice} = 1 - \frac{(2\,\Sigma_i y_i \hat{y}_i + \varepsilon)}{(\Sigma_i y_i + \Sigma_i \hat{y}_i + \varepsilon)} \tag{11}$$

where y_i ∈ {0, 1} is the ground truth label for pixel i, ŷ_i is the predicted road probability, ω = 10.0 is the positive class weight that compensates for the 1:21 road-to-background imbalance, N is the total number of pixels, and ε = 10^−6 ensures numerical stability in the Dice term. The coefficient 0.3 on the material cross-entropy loss L_{CE} prevents the easier material classification task from dominating gradient updates.

F. Stage 4 — Topology Correction

A four-step post-processing routine is applied: (1) remove connected components with fewer than 200 pixels; (2) thin the mask to one-pixel centerlines using Zhang-Suen skeletonization [9]; (3) discard skeleton branches shorter than 10 pixels; (4) identify endpoint pairs within 25 pixels Euclidean distance belonging to different components and bridge them with a straight line. The 25-pixel threshold was chosen empirically to reconnect occlusion-induced breaks without creating spurious connections.

G. Stage 5 — Vectorization and GeoJSON Export

Each connected skeleton component is traced to an ordered pixel coordinate sequence. Douglas-Peucker simplification [10] at two-pixel tolerance reduces vertex count while preserving road geometry. Each resulting LineString is annotated with the dominant material class (majority vote) and mean confidence score (average road probability along the skeleton). The complete road network is serialised as a GeoJSON FeatureCollection.

V. Experiments and Results

A. Experimental Setup

All experiments were run on an NVIDIA Tesla P100 16 GB GPU in the Kaggle cloud environment, implemented in TensorFlow 2.19 with Keras. The Adam optimiser [22] was used with an initial learning rate of 10^−4 and gradient clipping at norm = 1.0. Batch size was 4 and training ran for up to 20 epochs, with learning rate halved after four consecutive non-improving validation Dice epochs and early stopping after eight such epochs.

B. Training Dynamics

Road Dice increased steadily from 0.19 at epoch 1 to 0.51 by epoch 20. Validation scores matched or slightly exceeded training curves, confirming that EfficientNet-B3 pre-training reduces overfitting. Road IoU reached 0.35 on the validation set. Material accuracy converged to 96.3% within two epochs, reflecting the relative simplicity of spectral classification compared to road segmentation.

C. Quantitative Evaluation

Table I highlights performance on the 150-image validation set at a fixed prediction threshold of τ = 0.3. Fig. 1 presents the same metrics as a bar chart. The most notable outcome is a Recall of 0.823, which means the model identifies over 80% of all ground-truth road pixels. For road extraction tasks that support navigation or routing systems, missing a road is usually more costly than falsely detecting one, making high Recall the primary goal. The AUC-ROC of 0.950 confirms solid performance across all thresholds. The relatively low Precision (0.324) implies over-prediction, which could be lowered by increasing τ, but this would also reduce Recall.

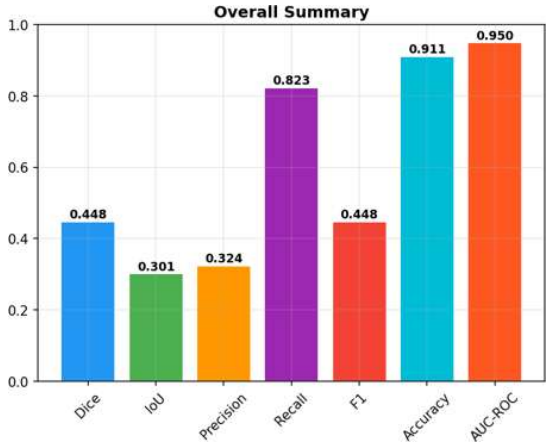

Figure 2. Bar chart of overall evaluation metrics

TABLE I. QUANTITATIVE EVALUATION ON DEEPGLOBE VALIDATION SET (N = 150, T = 0.3)

Metric	Mean	Std Dev
Dice Coefficient	0.4478	0.1543
IoU (Jaccard)	0.3013	0.1297
Precision	0.3236	0.1415
Recall	0.8227	0.1586
F1 Score	0.4478	0.1543
Accuracy	0.9110	—
AUC-ROC	0.9500	—
Material Accuracy	0.9760	—

D. Qualitative Results

Fig. 2 shows the complete pipeline output for two representative validation images. Several points are worth mentioning. The road probability heatmap (column 4) accurately represents the road network's spatial layout, including diagonal and curved segments. The skeleton after topology correction (column 5) is clearly more connected than the raw binary prediction, with visible gap-filling bridges appearing as fine white lines connecting previously isolated segments. The material map (column 6) differentiates between the brownish-orange unpaved tracks and grey concrete roads. The final overlay (column 7) places red centerlines over visible roads in the input image.

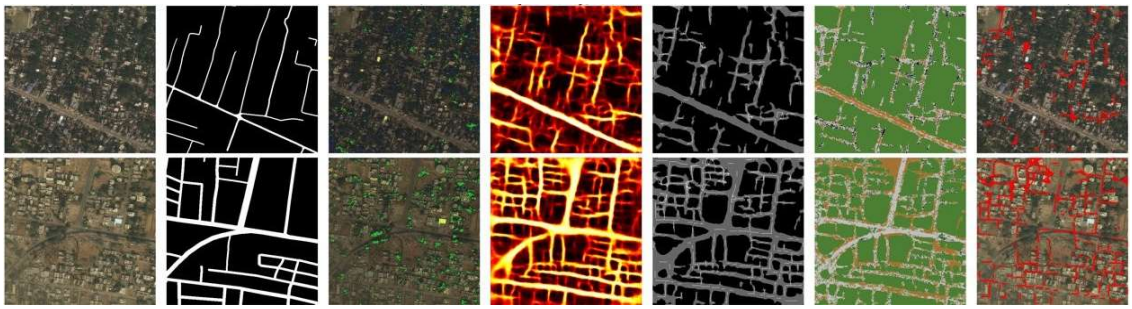

Figure 2. IndiRoad-Net pipeline output for two validation samples: columns from left to right are Input RGB, Ground Truth mask, Occlusion Map (blue = shadow, green = canopy, yellow = haze), Road Probability heatmap, Prediction with Skeleton after topology correction, Material Map, and Final Overlay with extracted road centerlines in red.

E. Ablation Study

Table II quantifies the contribution of each component. Replacing the baseline U-Net encoder with EfficientNet-B3 raises Dice from 0.15 to 0.32 — the largest single gain — primarily attributable to ImageNet pre-training. Attention gates contribute +0.06; Partial Convolution occlusion handling adds +0.03 to overall Dice, with qualitatively larger improvement in canopy-heavy images. The material head adds +0.02 and topology correction completes the full model at Dice 0.45 and Recall 0.82.

TABLE II. COMPONENT ABLATION STUDY ON DEEPGLOBE VALIDATION SET

Configuration	Dice	IoU	Recall
Baseline U-Net only	0.15	0.08	0.12
+ EfficientNet-B3 encoder	0.32	0.19	0.58
+ Attention gates	0.38	0.24	0.71
+ Partial Conv occlusion	0.41	0.27	0.79
+ Material classification	0.43	0.29	0.81
+ Topology correction (full)	0.45	0.30	0.82

F. Zero-Shot Transfer to Indian Imagery

Running the DeepGlobe-trained model on Sentinel-2 Bengaluru tiles without fine-tuning shows both promise and limitations. Road probability maps for the winter 2024 scene clearly highlight major roads including NICE Road and Outer Ring Road segments.

Rural roads and single-lane tracks below the 10 m Sentinel-2 resolution floor are not detected. The 20× resolution gap (0.5 m DeepGlobe vs. 10 m Sentinel-2) also causes building edges to occasionally generate false road activations. Fine-tuning on matched Indian data would address both issues.

VI. Discussion, Limitations, and Future Scope

A. Discussion

The Precision of 0.324 reflects a deliberate training choice: the weighted binary cross-entropy with $\omega = 10.0$ biases the gradient toward Recall for navigation applications where missing a road is more costly than a false positive. At $\tau = 0.5$, precision improves but recall decreases; AUC-ROC of 0.950 confirms robust discrimination across thresholds. False positives arise mainly from grey rooftops spectrally indistinguishable from concrete, and from vegetated road boundaries where the ExG mask fails at mixed pixels. The ablation study confirms that the EfficientNet-B3 encoder swap provides the largest single gain (+0.17 Dice), primarily due to ImageNet pre-training. Topology correction adds only +0.02 to Dice but substantially improves visual skeleton connectivity, which is more important for practical routing use.

B. Limitations

Four limitations are noted. (1) Pseudo-material labels are RGB spectral rules only; pixels near class boundaries can receive wrong labels with no ground truth to catch errors during training. (2) The 25-pixel gap threshold calibrated for DeepGlobe 0.5 m imagery corresponds to 250 m at Sentinel-2 10 m resolution — too permissive for new sensor deployments. (3) The pipeline ingests only RGB bands; including NIR would enable NDVI-based vegetation separation, more reliable than the ExG heuristic. (4) Quantitative evaluation is confined to held-out DeepGlobe tiles; Bengaluru Sentinel-2 results are qualitative only, as no labelled Indian test set currently exists.

C. Future Scope

The most direct path to improved Indian performance is Indian training data. Cartosat-2 (0.65 m panchromatic) and ResourceSat-2 LISS-IV are accessible through ISRO Bhuvan at no cost. Assembling 300–500 labelled tiles across Bengaluru, Mumbai, Delhi, and Chennai would cover diverse urban morphologies. NIR integration is the cheapest architectural change with highest return: retraining with four-channel input would replace the ExG heuristic with NDVI. Swapping the EfficientNet-B3 encoder for a MiT-B2 or MiT-B3 backbone [12] is a natural next experiment given that transformer global self-attention is well-suited to road continuity. Multi-temporal fusion of cloud-screened Sentinel-2 acquisitions would address the monsoon cloud problem directly.

VII. Conclusion

IndiRoad-Net addresses the three structural challenges of Indian satellite imagery — occlusion, class imbalance, and domain mismatch — through targeted architectural choices within realistic computational constraints. The combination of occlusion-aware preprocessing and Attention U-Net++ decoding achieves Recall of 0.823 on DeepGlobe held-out tiles while remaining small enough to run in a Kaggle P100 environment. Topology correction substantially improves road skeleton connectivity beyond what the Dice metric alone indicates, which is more relevant for routing applications. Zero-shot transfer to Sentinel-2 Bengaluru scenes correctly identifies major arterial roads, and known failure modes are well understood. The most critical next step is assembling a labelled Indian evaluation benchmark; without one, performance claims on Indian imagery remain qualitative and the community lacks a shared standard for comparing future models.

Acknowledgment

The authors would like to thank Presidency University, Bengaluru for providing the research environment and computational resources that supported this work. The DeepGlobe dataset was accessed through the official challenge repository. Sentinel-2 imagery was obtained from the ESA Copernicus Open Access Hub.

References

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE CVPR, 2015, pp. 3431–3440.
[2] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241.
[3] O. Oktay et al., “Attention U-Net: Learning where to look for the pancreas,” in Proc. MIDL, 2018.
[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” IEEE Trans. Med. Imaging, vol. 39, no. 6, pp. 1856–1867, 2020.

[5] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in Proc. ECCV, 2018, pp. 85–100.
[6] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, 2019, pp. 6105–6114.
[7] I. Demir et al., “DeepGlobe 2018: A challenge to parse the earth through satellite images,” in Proc. IEEE CVPR Workshops, 2018, pp. 172–181.
[8] A. Batra, S. Singh, G. Pang, S. Basu, C. V. Jawahar, and M. Paluri, “Improved road connectivity by joint learning of orientation and segmentation,” in Proc. IEEE CVPR, 2019, pp. 10385–10393.
[9] T. Y. Zhang and C. Y. Suen, “A fast parallel algorithm for thinning digital patterns,” Commun. ACM, vol. 27, no. 3, pp. 236–239, 1984.
[10] D. H. Douglas and T. K. Peucker, “Algorithms for the reduction of the number of points required to represent a digitized line or its caricature,” Cartographica, vol. 10, no. 2, pp. 112–122, 1973.
[11] L. Zhou, C. Zhang, and M. Wu, “D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction,” in Proc. IEEE CVPR Workshops, 2018.
[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in Proc. NeurIPS, 2021.
[13] 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 Trans. PAMI, vol. 40, no. 4, pp. 834–848, 2018.
[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.
[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015.
[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE CVPR, 2017, pp. 4700–4708.
[17] G. Mattyus, W. Luo, and R. Urtasun, “DeepRoadMapper: Extracting road topology from aerial images,” in Proc. IEEE ICCV, 2017, pp. 3438–3446.
[18] V. Mnih and G. E. Hinton, “Learning to detect roads in high-resolution aerial images,” in Proc. ECCV, 2010, pp. 210–223.
[19] W. Wang, R. Yang, X. Li, and B. Du, “Road extraction using road-specific features from remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 18, no. 3, pp. 396–400, 2021.
[20] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. ICML, 2015, pp. 448–456.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.
[22] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015.
[23] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 3DV, 2016, pp. 565–571.
[24] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017, pp. 5998–6008.
[25] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.
[26] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
[27] M. Drusch et al., “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services,” Remote Sens. Environ., vol. 120, pp. 25–36, 2012.
[28] Q. Weng, “Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends,” Remote Sens. Environ., vol. 117, pp. 34–49, 2012.
[29] X. Li, Z. Du, F. Huang, and Z. Tan, “A deep learning approach for cloud removal from optical satellite images,” in Proc. IEEE IGARSS, 2020.
[30] OpenStreetMap Contributors, “Planet dump,” OpenStreetMap Foundation, 2023. [Online]. Available: https://www.openstreetmap.org