UAVPal: A New Dataset for Semantic Segmentation in Complex Urban Landscape With Efficient Multiscale Segmentation

Abhisek Maiti, Sander Oude Elberink, and George Vosselman

***Abstract*—Semantic segmentation has recently emerged as a prominent area of interest in Earth observation. Several semantic segmentation datasets already exist, facilitating comparisons among different methods in complex urban scenes. However, most open high-resolution urban datasets are geographically skewed toward Europe and North America, while coverage of Southeast Asia is very limited. The considerable variation in city designs worldwide presents an obstacle to the applicability of computer vision models, especially when the training dataset lacks significant diversity. On the other hand, naively applying computationally expensive models leads to inefficacies and sometimes poor performance. To tackle the lack of data diversity, we introduce a new UAVPal dataset of complex urban scenes from the city of Bhopal, India. We complement this by introducing a novel dense predictor head and demonstrate that a well-designed head can efficiently take advantage of the multiscale features to enhance the benefits of a strong feature extractor backbone. We design our segmentation head to learn the importance of features at various scales for each individual class and refine the final dense prediction accordingly. We tested our proposed head with a state-of-the-art backbone on multiple UAV datasets and a high-resolution satellite image dataset for LULC classification. We observed improved intersection over union (IoU) in various classes and up to 2% better mean IoU. Apart from the performance improvements, we also observed nearly 50% reduction in computing operations required when using the proposed head compared to the traditional segmentation head.**

***Index Terms*—CNN, deep learning, Earth observstion (EO), semantic segmentation, vision transformer.**

I. Background

LAND use land cover (LULC) classification, a crucial task in Earth Observation (EO) utilizing images from airborne or space-borne platforms. Its significance lies in applications, such as monitoring, planning, and decision-making. It can be considered as a specific application of semantic segmentation. Semantic segmentation is a widely studied area in computer vision that involves assigning pixels in an image to specific categories. In recent years, significant advancements have been made in deep learning based semantic segmentation approaches [1]. These advancements encompass improvements in architectures, training techniques, datasets, and loss functions. However, evaluating the overall progress can sometimes be challenging due to variations in training methods, datasets, and the overall training process [2]. Moreover, the dominance of the increasingly complex and computationally expensive architectures in the benchmarks creates a skewed perspective that achieving state-of-the-art (SOTA) results requires ever-increasing model complexity and computational resources [3]. Apart from the improvements with respect to models, the volume and quality of the training data play a vital role in achieving the desired performance in semantic segmentation. When it comes to remote sensing platforms, the number of datasets available for semantic segmentation is much smaller. This hinders the generalization capability of the models and makes transfer learning less effective.

In the field of EO, recent research has focused on understanding different aspects of visual scenes through diverse datasets, aiming to develop improved methods for specific tasks [4]. However, several challenging domain-specific issues persist, including variations in object scales, overlapping class distributions, class imbalance, and texture heterogeneity within the same class. These challenges often lead to biased decisions and high uncertainty in predictions, even with the best-performing models [5]. Addressing these problems is an active area of research. However, most efforts concentrate on designing more complex and intricate backbones aiming to have more powerful feature extractors with better generalization capability and improved robustness to the aforementioned data-related issues [6].

In a typical design, the backbone serves as a deep feature extractor, resembling an image classifier. The subsequent part of the architecture, known as the head or decoder, is responsible for dense prediction using the coarse features generated by the backbone [7]. The evolution of backbone architectures, is becoming deeper and more complex since the early days of VGGs [8], the model named after Oxford's Visual Geometry Group. It has been a major driving factor behind improved performance in both image classification and semantic segmentation [9]. In addition to enhancing backbone architectures, improving the segmentation head also plays a vital role in boosting the overall performance of semantic segmentation subject to good quality training data [3].

Manuscript received 29 August 2023; revised 18 October 2023 and 1 November 2023; accepted 2 November 2023. Date of publication 6 November 2023; date of current version 23 November 2023. This work was supported by the "Water4Change" project jointly funded by the Department of Science and Technology (DST), the Government of India, and the Dutch Research Council (NWO) project W 07.7019.103 | DST-1429-WRC. *(Corresponding author: Abhisek Maiti.)*

The authors are with the Faculty ITC, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: a.maiti@utwente.nl; s.j.oudeelberink@utwente.nl; george.vosselman@utwente.nl).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2023.3330758, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2023.3330758

The current availability of public semantic segmentation datasets primarily focuses on common objects found in natural images captured from the ground. Popular datasets like MS-COCO [10], Pascal VOC [11], ADE20 K dataset [12] offer semantic segmentation tasks for recognizing common objects in typical scenes. These datasets are commonly used for common object recognition and semantic segmentation tasks. In contrast, when it comes to remote sensing, the choice of available datasets for semantic segmentation is very limited. The images are generally expected to be captured in the nadir view and usually show only the top of objects. Acquiring very high-resolution imagery conforming to such criteria from satellites or UAVs over a sufficiently large area is usually very expensive. Moreover, preparing the corresponding training labels is a time-consuming process. In this respect, the ISPRS 2D semantic labeling benchmarks [13] provide Vaihingen and the Potsdam semantic segmentation datasets containing very high-resolution imagery from UAVs. The LoveDA dataset [14] provides a semantic segmentation dataset from China containing very high-resolution Google Earth imagery. However, open semantic segmentation datasets covering highly populated and diverse urban regions of Southeast Asian cities are difficult to find. Due to high variation and distinct spatial patterns, models trained on images from European or North American regions perform poorly, and the lack of region-specific data bars the potential of transfer learning.

By introducing urban scene semantic segmentation tasks for the UAV platform, researchers can gain valuable insights into visual understanding in UAV scenes, forming the basis for more advanced smart applications. In this regard, our contributions encompass two aspects.

1) A new UAVPal dataset capturing complex urban scenes from the city of Bhopal, Madhya Pradesh, India.
2) An efficient multiscale segmentation head that reduces computational load without compromising performance.

The dataset is openly available.[1] The associated implementations of this research are available on Github.[2]

II. Related Work and Motivation

Popular semantic segmentation networks employ network trunks with a low output stride. This design choice allows the networks to achieve better resolution for capturing details at a moderate spatial scale. However, it has the drawback of reducing the receptive field, making it challenging for the networks to predict relatively large objects in a given scene accurately. Similarly, very small objects often consist of intricate details and fine structures that are not preserved in the feature representations learned by CNNs. To address this issue, pyramid pooling is utilized to assemble multiscale context. PSPNet [12] incorporates a spatial pyramid pooling module, which combines features from multiple scales obtained through pooling and convolution operations on the final layer of the network trunk. DeepLab [15] adopts Atrous Spatial Pyramid Pooling (ASPP), which employs atrous convolutions with various dilation levels to create denser features than PSPNet. Overall, these architectures utilize intermediate features along with the final layer features of the network trunk to create a comprehensive multiscale context.

Pyramid pooling techniques focus on fixed, square context regions due to the symmetric application of pooling and dilation. In addition, these techniques are usually static and not learned. On the other hand, relational context methods such as CFNet [16] establish context by considering the relationships between pixels and are not restricted to square regions. Moreover, relational context methods can adaptively learn and construct context based on the composition of the image. However, average pooling may not be optimal as it equally weights output from different scales. To overcome this, attention mechanisms have been proposed by Chen et al. [17], where attention heads are trained across multiple scales simultaneously to improve contextual information. Such methods, however, require a fixed set of scales during training and inference.

Tao et al. [18] introduced a novel attention-based head for combining multiscale predictions without limiting the number of scales. We take this idea further by introducing a novel head that enhances the capabilities of a SOTA backbone with better efficiency. This enhancement is achieved through the utilization of visual-cognition based attention and a more efficient multiscale feature fusion mechanism. Mehta et al. [19] highlighted the importance of the segmentation head of the model by demonstrating that a simple but optimally designed head not only reduces compute cost, but also potentially improves the model performance. The conceptual underpinnings of our work are rooted in the research conducted by [18] and [19]. However, unlike [17], our method does not rely on extra scale-specific supervision.

Apart from the development of better models, several new EO-related vision datasets have been released over the years. Many of these datasets focus on various specific tasks, ranging from object detection [20], and different variants of segmentations [21] to change detection [22]. Apart from computer vision datasets of common scenes and terrestrial images, datasets like UAVid [23], and SpaceNet 4 [24] provide datasets with oblique or off-nadir imagery. The quintessential Potsdam and Vaihingen datasets released by ISPRS [13] contain ortho-rectified high-resolution imagery and DSM. Similarly MiniFrance dataset [25] contains high-resolution aerial imagery from various cities in France, where a considerable portion of the images are annotated, enabling both supervised and unsupervised segmentation tasks. While building detection datasets like [26] covers various cities in Africa and FloodNet [27] and iSAID [28] datasets covering parts of China, high-resolution EO datasets from southeast Asia, including India, is hard to find. This motivates us to openly release the high-resolution dataset acquired through drones, with a focus on India to address the aforementioned data availability issue in the region. A comprehensive comparison of UAVPal with other popular segmentation datasets in terms of data characteristics, annotation quality, geographic coverage, and other relevant factors is presented in Table I.

[1] [Online]. Available: https://doi.org/10.17026/dans-z55-6gt4
[2] [Online]. Available: https://github.com/digital-idiot/SemSeg

TABLE I
COMPARATIVE OVERVIEW OF DATASET SPECIFICATIONS

Dataset	Classes	Annotation Type	Annotated Area (km^2)	Acquisition Platform	Orthorectified	Spectral Bands	Height Data	Region Covered
UAVid[23]	8	Fine	-	Airborne	✗	RGB	✗	China (30 cities)
Potsdam[29]	6	Fine	3.4*	Airborne	✓	IR-RGB	Stereo	Germany (1 city)
Vaihingen[30]	6	Fine	1.1*	Airborne	✓	IR-RG	LiDAR	Germany (1 city)
MiniFrance[25]	15	Coarse	10000.0*	Airborne	✓	RGB	✗	France (16 cities)
FloodNet[27]	10	Fine	1.2*	Airborne	✗	RGB	✗	USA (2 cities)
LoveDA[14]	7	Fine	536.2	Spaceborne	✓	RGB	✗	China (3 cities)
UAVPal (ours)	5	Fine	1.0	Airborne	✓	RGB	Stereo	India (1 city)

The areas estimated approximately are * marked. An annotation is considered fine if more granular objects such as individual buildings, and cars are delineated.

III. UAVPAL DATASET

Our approach to obtaining and labeling data is specifically tailored for semantic segmentation in complex urban scenes from relatively small cities in India. Civilian use of UAVs is highly restricted in India [31] therefore, it is difficult to acquire data over a large area in the Indian cities. In our project, we obtained the opportunity to fly a DJI Phantom 4 over the city of Bhopal, located in the state of Madhya Pradesh in the central part of India. In order to avoid motion blur, the drone has been flown with an average speed of approximately 9 ± 1 m/s at a height of approximately 90 m above ground. The survey was carried out around noon to minimize the effect of shadows from buildings and other vertically large objects.

A. Dataset Properties

In order to acquire a diverse dataset with good-quality labels, we considered the following aspects:

Resolution: Due to relatively low flying height, the images are acquired at very high spatial resolution. Each image out of 1642 raw images is of size 5472×3648. The pixel size of these images is approximately 1.9 cm on average. The objects of interest are visually clear and distinguishable with a lot of details. Spectrally, the images have three optical RGB channels with unsigned 8 b discrete intensity levels. The spectral distribution of the image channels is shown in Fig. 1(a).

Spatial Coverage: The survey has been carried out over the central part of the city. Before, preprocessing, the spatial coverage of the survey is approximately 1.12×1.09 km^2. The area is very densely populated and dominated by man-made structures.

Ortho-rectification: The raw images are preprocessed using a photogrammetric pipeline to generate a high-quality ortho-mosaic. This preserves the relative distance between the objects and corrects distortions due to camera perspective, camera orientation, and lens properties. Ortho-rectification of the stack of raw images results into a high-resolution ortho-mosaic with 2.2 cm spatial resolution with coverage of 1.01 km^2.

Height Map: The height map of the entire area of interest is generated using a stereo-matching pipeline. First, the images are processed through Pix4D Mapper [32] generating point clouds through dense matching. The point cloud from dense matching is then rasterized to obtain a high-quality digital surface model (DSM) of 8.5 cm spatial resolution. The height distribution of DSM is shown in Fig. 1(b).

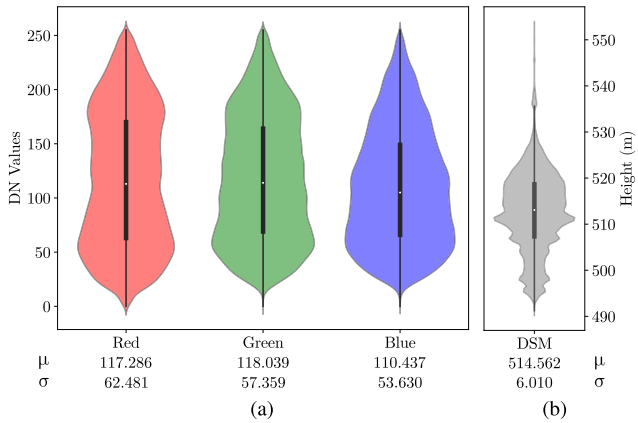


Fig. 1. Violin plot of (a) RGB channels of the images and (b) DSM. The horizontal span represents the probability density, the hollow dot shows the mean (μ) value, and the thick black line represents quartiles on either side of the mean. The corresponding mean and standard deviation (σ) values are reported underneath the figures.(a) Spectral Stat. (b) DSM Stat.

Scene Complexity: It is qualitatively evident that our UAVPal dataset has much higher scene complexity than the other existing UAV semantic segmentation dataset. An example of such complexity of the texture in the images is shown in Fig. 2. Quantitatively, the standard deviation of the pixel values of the buildings in our dataset is approximately 7 times higher than the buildings in the Vaihingen dataset [30]. The roads in our dataset contain various visually distinct objects other than labeled cars. The background class also has immense textural variability, and many cases have a similar textural pattern to other classes making the semantic segmentation task relatively more challenging.

Annotation: The variability of the texture and nuance present across different classes make the common automated annotation tools ineffective. Considering the complexities present in the scene, all the annotations have been generated manually without any automated tools. To further ensure the accuracy and reliability of the labels, an additional independent quality check has been performed. The annotations are first created in a vector format using GIS software. For training, these vector annotations are later converted into indexed raster labels matching the spatial

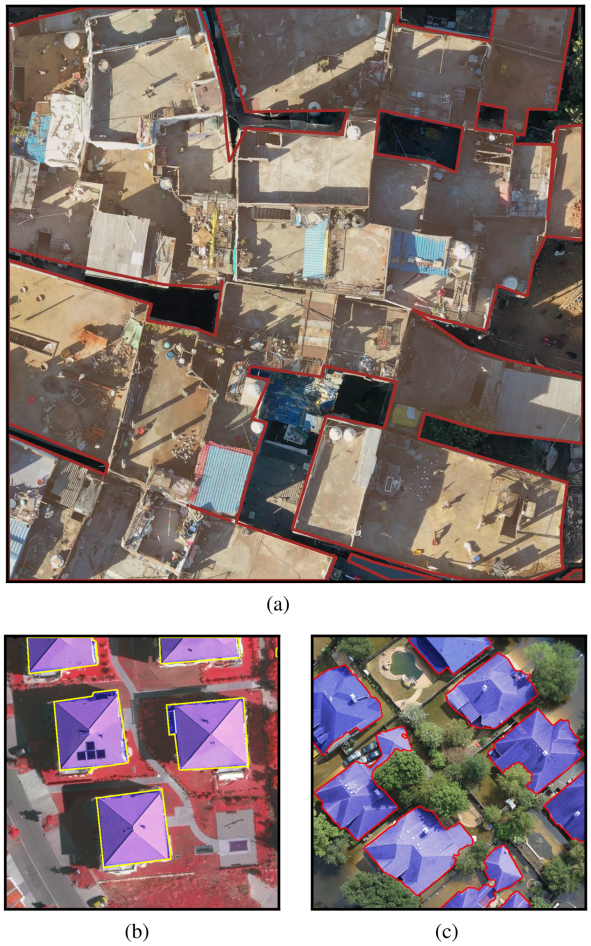


Fig. 2. Visual complexity of rooftops in the UAVPal dataset: Highlighting the intricate roof patterns contrasting the developed western regions. (a) A typical scene roofs from UAVPal dataset. (b) Typical roofs in Vaihingen. (c) Typical roofs in FloodNet.

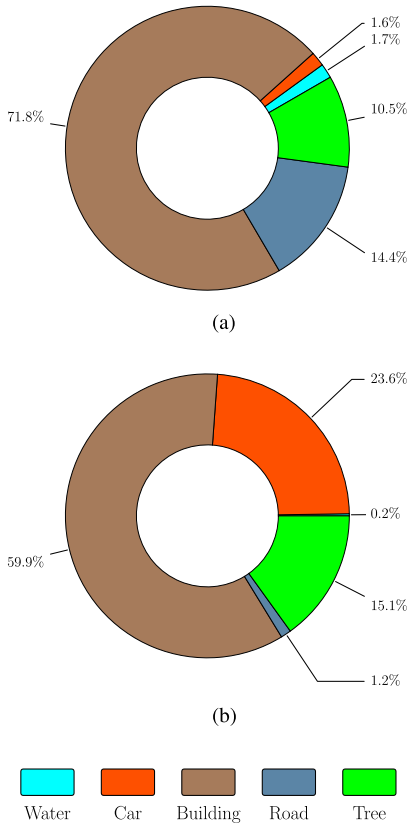


Fig. 3. Distribution of (a) pixels and (b) segments among different designated classes. The highest share of both pixels and segments belongs to the building class. (a) Pixel Distribution. (b) Segment Distribution.

resolution of the corresponding images. The annotations consist of five distinct classes, namely water, road, car, building, and tree. Pixels not belonging to any of the aforementioned classes are tagged into an extra background class. The distributions of the classes are presented in terms of the share of the pixels and the number of distinct segments in the vector annotation belonging to each class in Fig. 3(a) and (b), respectively.

B. Dataset Preparation

Following the preprocessing, the dataset is prepared to make it suitable for model training and inference. First, the DSM mosaic is resampled to match the resolution of the image mosaic. Both the image and the DSM are clipped to remove the distorted regions around the edges of the scene. Finally, the mosaics and the rasterized annotation are split into 2048×2048 tiles. The overlap between neighboring tiles is adjusted to 60 pixels to avoid tiles slacking beyond the scene extent. There is a total of 529 tiles, each tile spatially covers 45×45 m^2 area. Among the 529 samples, 159 are randomly picked and reserved for testing, and the rest of the 70% tiles are available for training. The spatial distribution of the train and test tiles is visually presented in Fig. 2 of the supplementary material.

IV. Method

Apart from simplicity and efficiency, the main guiding principle of our novel head is that harmonization of the receptive fields at different scales leads to better object detection and dense prediction. In a typical multihead segmentation backbone, each head has different effective receptive fields.

A. Designing Segmentation Head

In a typical backbone, the multiscale features are generated at different spatial sizes. Therefore, these features are transformed into the same scale using a scale injection module. In this module, all features are interpolated into the target spatial size. Next, the lower-resolution features are injected into subsequent high-resolution features as shown in (1). The global and the local semantics are individually passed through 1×1 convolutions. Afterward, global semantics are further transformed into semantic weights by applying a subsequent sigmoid activation. These semantic weights are multiplied with the transformed local token and added to the global semantic transformed by a separate 1×1 convolution to generate the scale-specific feature ($\tilde{\chi}$). In (1), $f_{1\times1}(\cdot)$ is 1×1 convolution, $[\cdot]_{\mathcal{N}}$ represents batch normalization, $\varphi(\cdot)$ is sigmoid activation, and $\circ$ represents Hadamard

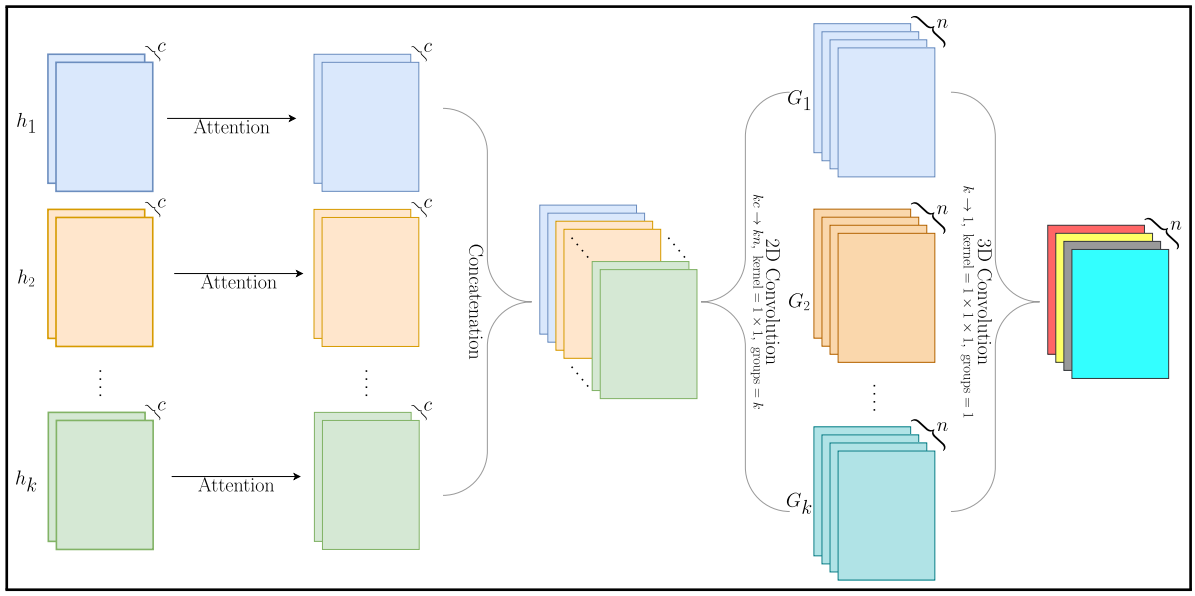


Fig. 4. Architecture of our proposed head. First, the features of each head are refined using attention. The features are brought to the target spatial scale using bilinear interpolation. 2D Group convolution is applied to get class scores from each scale and stacked along a new dimension. Finally, class scores from each scale are fused together using a 3D convolution operation to predict final class scores reducing the added extra dimension.

product. $\boldsymbol{\tau}$ and $\boldsymbol{\chi}$ are corresponding local tokens and global semantics, respectively. This scale injection module helps bridge the semantic gap between two consecutive scales compared to just interpolation for scaling the features

$$\tilde{\boldsymbol{\chi}} = [f_{1\times1}(\boldsymbol{\chi})]_{\mathcal{N}} \circ \varphi([f_{1\times1}(\boldsymbol{\tau})]_{\mathcal{N}}) + [f_{1\times1}(\boldsymbol{\chi})]_{\mathcal{N}}. \quad (1)$$

The design of our proposed head architecture is shown in Fig. 4. We apply attention to features at each scale to refine the features separately at their respective scale. We first aim to predict dense class scores at each scale and then generate the weighted average of all these groups of class scores to predict the final class scores, where the weights are learnable. We do this by bringing the features to one uniform resolution and then using group convolution on the concatenated features to predict class scores for each scale. If there are k heads and c channels at each head, the concatenated feature will have kc channels. Assuming the number of possible classes is n, We efficiently transform this to features with kn channels using a convolution of 1×1 kernel and k groups. Each group of adjacent c channels transforms into subsequent n channels in the resultant feature without cross-influence. We reshape the output features where each group is stacked along a new dimension. We apply a 3D convolution to get the final class scores. The $k \times 1 \times 1 \times 1$ kernel weights of the 3D convolution are responsible for learning the appropriate weights across the scale-specific groups for each of the n classes.

B. Attention Mechanism

Our proposed head uses two convolution blocks without additional computational overhead to the backbone. To keep the minimal overhead, we adopt the attention mechanism proposed by [33]. The objective here is to minimize the linear separability among the neurons so that the target neuron appears more visually distinctive from the surrounding neurons. A closed-form solution of the related energy function has been obtained, assuming each channel follows a single distribution. The minimal energy function for deriving attention is given by (2) [33]. Here, e_t^* is the minimal energy of channel t, with a closed-form solution as a function of t, it's channel-wise mean μ_t and channel-wise standard deviation σ_t. The term λ is a free parameter representing potential bias in the solution. When the energy e_t^* is reduced, neuron t becomes more distinguishable from the neighboring neurons thus the importance of each neuron is defined by $1/e_t^*$

$$e_t^* = \frac{4({\sigma_t}^2 + \lambda)}{(t - {\mu_t}^2) + 2{\sigma_t}^2 + 2\lambda}. \quad (2)$$

In [33], the optimum λ is found using cross-validation search, which is a time-consuming additional process. We mitigate this issue by considering λ as a learnable parameter.

V. Experiment Setup

Here, we describe how we set up our experiments to test the performance of our proposed head compared to the default segmentation heads. We also provide some key details regarding implementation. We further analyze the effectiveness of the head through ablation studies.

Apart from UAVPal, we selected four more well-known datasets to evaluate our method. All the datasets contain very high-resolution images, but datasets are acquired from either space-borne (LoveDA) or airborne (FloodNet, Potsdam, Vaihingen) platforms. The details of the datasets are provided in Table II.

For the FloodNet dataset, we resize the images to 1024×768 for training but perform prediction at full resolution. The pixel size of the resized training images is approximately 12 cm.

TABLE II
DETAILS OF DATASETS USED IN THE EXPERIMENTS

Dataset	Train	Val.	Test	Size	Resolution (m)	Channels
FloodNet[27]	1445	450	448	4000×3000	0.03	RGB
LoveDA[14]	2522	1669	1796	1024×1024	1.2	RGB
Potsdam[29]	24	–	14	6000×6000	0.05	IR-RGB+D
Vaihingen[30]	16	–	17	–	0.08	IR-RG+D
UAVPal (Ours)	370	–	159	2048×2048	0.02	RGB+D

TABLE III
RESULTS OBTAINED ON THE TEST SET OF FLOODNET

Model	BCG	BFL	BNF	RDF	RNF	WTR	TRE	VCL	POL	GRS	mIoU
TF	92.3	83.2	**93.3**	79.2	83.2	72.0	77.0	87.2	84.3	84.4	83.6
TF + MSF	**93.2**	**84.2**	93.2	**82.3**	**86.3**	**76.4**	**77.2**	**87.4**	**85.1**	**87.4**	**85.3**
UF	92.4	83.3	94.4	80.0	83.6	73.1	77.2	88.3	84.1	84.6	84.1
UF + MSF	**93.0**	**83.7**	**94.6**	**80.2**	**84.2**	**73.2**	**77.8**	**88.9**	**84.5**	**85.2**	**84.5**

IoUs are reported in percentages. The best values are marked in bold. BCG: background, BFL: building flooded, BNF: building non-flooded, RDF: road flooded, RNF: road non-flooded, WTR: water, TRE: tree, VCL: vehicle, POL: pool, GRS: grass.
The best values are marked in bold.

Resizing the images makes it possible to input a batch of entire images into the network. We downsample instead of cropping mainly because cropping the images into smaller tiles reduces the effective receptive field [34]. The down-sampling of the image may reduce unnecessary details, such as transmission lines, clutter, etc. We use these images from the rest of the datasets without scaling. Large scenes in Potsdam and Vaihingen datasets are sliced into 1024×1024 tiles to facilitate training. For each experiment, we train the network from scratch. Evaluation of the test performance on the LoveDA experiment is done by submitting the inferred labels to the official portal. In our experiment, we use the whole training set to train the model instead of training and fine-tuning the model for rural and urban scenes separately. When a dataset lacks a designated validation set, for training, we employ k-fold cross-validation with $k = 5$, which entails an $80 : 20$ ratio of training to validation data.

Two of the SOTA backbones have been used to test the effectiveness of our proposed head. Both TopFormer [35] and UNetFormer [36] share an overall encoder–decoder architecture with the option for multiscale prediction. While TopFormer applies a transformer block at the bottleneck, UNetFormer applies specialized transformers at each scale of the decoder. While TopFormer utilizes a lightweight transformer block to refine high-level features, UnetFormer exploits global-local transformer blocks for future fusion at each scale.

In all of our experimental endeavors, the well-regarded categorical cross-entropy has been employed as the loss function. To rectify class imbalances, we have utilized inverse frequency weighting to ascertain class weights, subsequently incorporating these within the loss computation. The Rectified Adam optimizer (RAdam) [37], a refined version of the conventional Adam optimizer [38], has been selected for the experiments, owing to its superior performance characteristics. We used the one-cycle learning rate policy [39] to optimize the learning rate dynamically, facilitating more rapid and enhanced convergence. To attain superior generalization capabilities, stochastic weight averaging (SWA) [40] has been deployed during the training phase. The strategy of weight averaging serves to smooth the loss landscape, preventing the optimization procedure from ensnaring within local minima, which in succession, cultivates enhanced model performance. The batch size of 24 has been determined using a grid search subject to available memory and gradient accumulation strategy.

For the evaluation of performance, we use IoU. We compare both class-specific IoUs and overall mean IoU (mIoU). In order to minimize the effect of numerical instability and stochasticity, we report mean values of five trials for each experiment. The IoU is a metric used to evaluate the overlap between two segmented areas. Its definition is shown as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{FP} + \text{TP} + \text{FN}}. \tag{3}$$

In the context of semantic segmentation, IoU offers a more unbiased assessment of model performance compared to accuracy. While accuracy can be misleadingly high if a class dominates the image, apart from true positives (TP) IoU takes into account both false positives (FP) and false negatives (FN) in its calculation. Thus, it ensures that even if a segment is a small portion of the image, its accurate prediction is crucial for a high IoU score, promoting balanced model performance across classes [41].

VI. RESULTS

In this section, we present the quantitative and qualitative results of our experiments and analyze them. Finally, we show the results of a few essential ablation experiments. We will denote the *T*op*F*ormer with its original head as TF and the network with our proposed *M*ulti-*S*cale *F*usion as MSF. Similarly, UF represents *U*Net*F*ormer with its original head. We also measured the performances of a few well-known semantic segmentation models on our UAVPal dataset, these measurements are available in Table I of the supplementary material.

A. Experiments on FloodNet Dataset

Table III shows the IoUs of the experiments on the FloodNet dataset. It can be readily observed that the network with our proposed head (MSF) performs equally, if not better, for each class, along with better mIoU compared to the default heads of both networks. Scores of the experiment with our proposed method for the classes with large object sizes, such as grass and background have improved, along with classes containing small objects such as pools. In addition, we observe better detection of linearly shaped roads and flooded buildings.

In Fig. 5(a)–(d), we can observe a test scene with its associated ground truth labels and the corresponding inferences from each model. Visually comparing Fig. 5(c) and (d), we observe that the inference from MSF is less spurious, and the shapes and boundaries of the objects seem more precise. For a more detailed visual comparison, we refer to Fig. 5(e)–(h) focusing on a subset area of a test scene. Here, notable differences can be observed by comparing the areas highlighted with white rectangles. Both TF and MSF misclassify the building at the bottom left. However, the latter correctly labels the building immediately on the right. On the top right portion, TF fails to detect one of the two vehicles, but MSF detects both vehicles correctly.

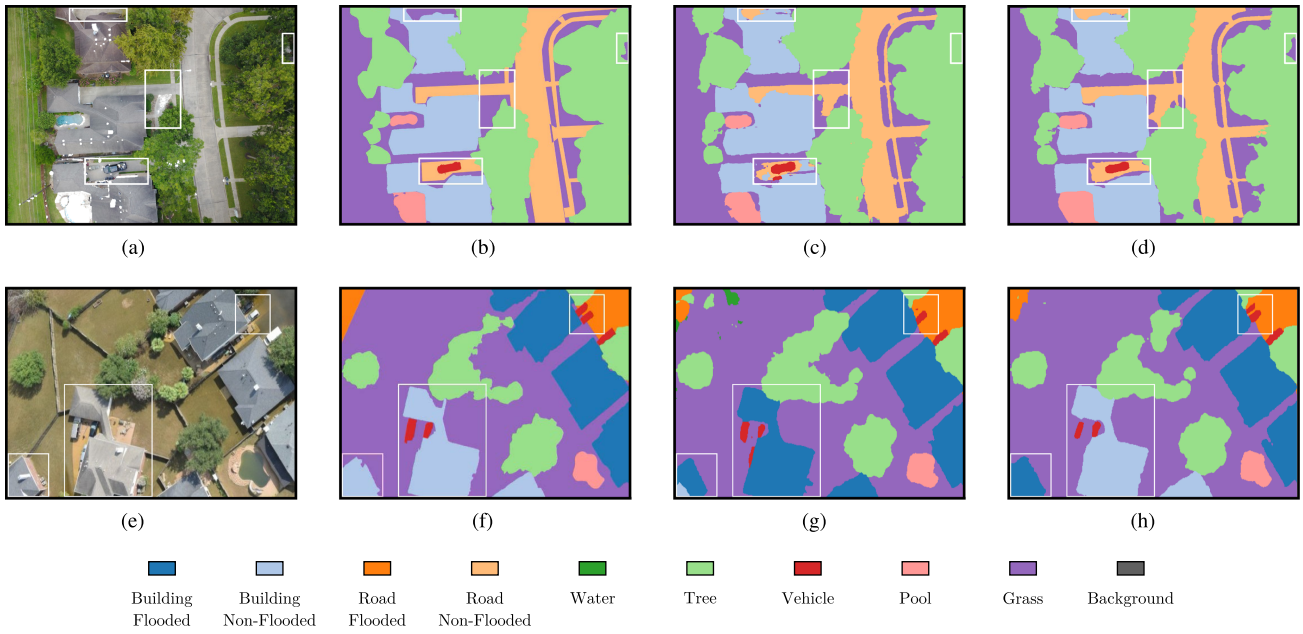


Fig. 5. Visual comparison of inferences with respect to the ground truth for a test image from the FloodNet dataset. * indicates zoomed scene. (a) Image: 10813. (b) Ground Truth: 10813. (c) TF: 10813. (d) TF + MSF (ours): 10813. (e) Image*: 7323. (f) Ground Truth*: 7323. (g) TF*: 7323. (h) TF + MSF* (ours): 7323.

TABLE IV
RESULTS OBTAINED ON THE TEST SET OF LOVEDA

Model	BG	BD	RD	WR	BR	FR	AG	mIoU
TF	39.1	**53.4**	55.1	74.2	**12.2**	38.3	**56.2**	46.9
TF + MSF	**41.1**	53.1	**56.3**	**77.5**	11.4	**41.3**	54.4	**47.9**
UF	44.7	58.8	54.9	79.6	20.1	46.0	**62.5**	52.4
UF + MSF	**45.2**	**60.1**	**55.0**	**80.4**	**20.3**	**46.5**	62.3	**52.9**

The IoUs are reported in percentages. The best values are marked in bold. BG: background, BD: building, RD: road, WR: water, BR: barren, FR: forest, AG: agriculture.
The best values are marked in bold.

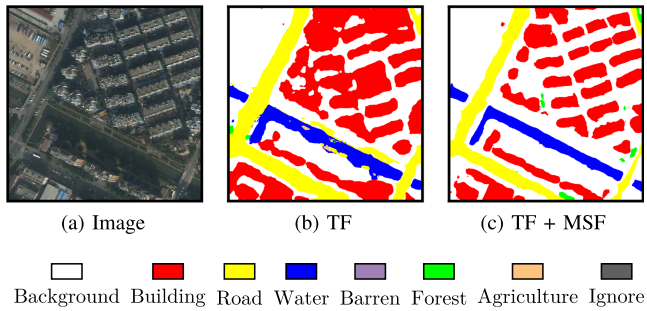


Fig. 6. Visual comparison of inferences for a test image from LoveDA dataset. Ignore represents no data region. (a) Image. (b) TF. (c) TF + MSF.

B. Experiments on LoveDA Dataset

The test scores for the LovDA dataset have been obtained by submitting the predicted results to the official leaderboard [14] for evaluation and are shown in Table IV. We observe that MSF can outperform the base models for many classes. However, here we also observe that MSF slightly underperformed with respect to TF for barren and agriculture classes. Nevertheless, MSF achieves better mIoU than TF, which agrees with the FloodNet results. For qualitative and visual analysis, we compare the predictions in Fig. 6(b) and (c).

Although we cannot qualitatively compare the predictions of the test set due to the unavailability of the respective ground truths, we can still intuitively compare the visual quality among the predictions. In the context of shadows, MSF can produce more coherent predictions than TF. Shadows of the buildings are harder to detect due to their low illumination, irregular shape, size, and orientation. Shadow detection requires more contextual information. Thus, a larger receptive field compared to what is required to detect the associated object. In both TF and MSF, multi-scale features from the backbone help in this regard. However, MSF learns the importance of features at different scales for each class. Therefore, the likelihood of MSF resolving shadows is higher than TF. We also observe that the shapes and boundaries of the objects are better delineated by MSF. Moreover, MSF seems less likely to introduce spurious noise in the prediction than TF. We encountered similar patterns and equivalent observations throughout the predictions for the test set.

C. Experiment on Potsdam and Vaihingen Dataset

In the field of Earth observation, the ISPRS Vaihingen and Potsdam datasets are extensively utilized for benchmarking semantic segmentation tasks. In Table V, we present the performance metrics comparing MSF to the baselines with respect to both of these datasets. Similar to the FloodNet and LoveDA experiments, we observe IoU improvements in several classes due to MSF.

The improvements in mIoU can also be observed in most cases. In the case of the Vaihingen dataset, the MSF could not

TABLE V
METRICS FROM THE POTSDAM AND VAIHINGEN EXPERIMENT

Dataset	Model	Surf.	Bld	Veg.	Tree	Car	mIoU
Potsdam	TF	91.9	94.2	85.4	**88.1**	92.3	84.7
	TF + MSF	**92.2**	**94.9**	**86.2**	88.0	**93.1**	**85.4**
	UF	93.6	97.2	**87.7**	88.9	96.5	86.8
	UF + MSF	**94.0**	**97.7**	87.5	**89.2**	**96.6**	**87.1**
Vaihingen	TF	**91.3**	93.8	84.6	88.1	**89.1**	**82.9**
	TF + MSF	90.9	**93.9**	**84.8**	**88.4**	88.8	**82.9**
	UF	92.7	95.3	84.9	90.6	88.5	82.7
	UF + MSF	**92.9**	**95.7**	**85.4**	**90.7**	**89.0**	**83.8**

The IoUs are reported in percentages. The best values are marked in bold. Surf.: impervious surface, Bld: building, Veg.: low vegetation.
The best values are marked in bold.

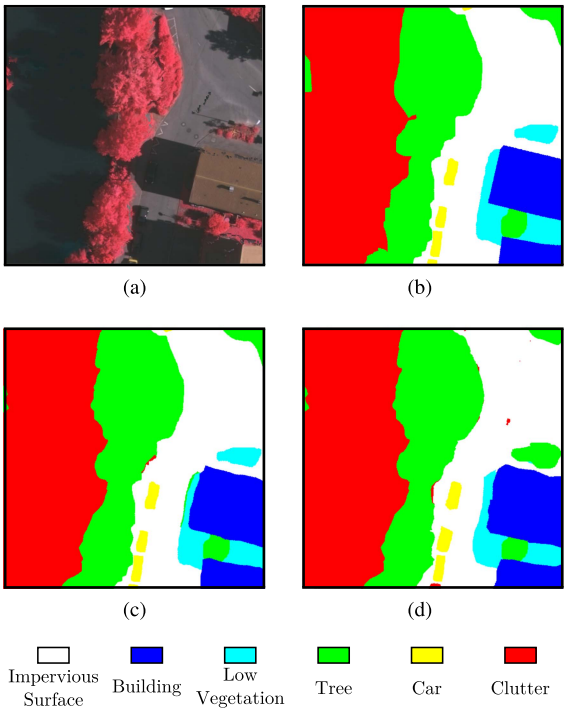


Fig. 7. Visual comparison of inferences for a test image from the Vaihingen dataset. The clutter class is considered as no data and ignored in metrics. (a) Image. (b) Ground Truth. (c) UF. (d) UF + MSF.

improve overall IoU with the TF backbone. However, in the case of the UF backbone, the MSF improves the mIoU by more than 1%. In Fig. 7, we observe that with MSF the spurious prediction is reduced and the shapes of objects are relatively better delineated. Similar trends have also been observed for the Potsdam dataset.

D. Experiment on UAVPal Dataset

Fig. 8 shows the qualitative comparison of the performance of our proposed model on the UAVPal dataset with the ground truth. We found significant improvements in the accuracy of individual classes, particularly in the building class. It can be seen that the borders of the buildings are very accurately delineated from other classes. There are a lot of missing road segments in Fig. 8(d) compared to that of our proposed model results.

TABLE VI
RESULTS OBTAINED ON THE TEST SET OF UAVPAL

Model	WR	RD	CR	BD	TR	mIoU
TF	93.1	81.9	62.4	93.8	84.1	83.1
TF + MSF	**95.5**	**85.1**	**65.7**	**96.9**	**85.4**	**85.7**
UF	83.9	80.9	61.3	91.9	82.5	80.1
UF + MSF	**87.1**	**83.8**	**62.4**	**94.6**	**82.9**	**82.1**

The IoUs are reported in percentages. The best values are marked in bold. WR: water, RD: road, CR: car, BD: building, TR: tree.
The best values are marked in bold.

Interestingly, a car on the road in the top left of the image in Fig. 8(e) has shown a proper distinction which is not captured in UF. Despite the complexity of the UAVPal dataset in terms of closely clustered buildings and uneven distribution of different classes, our proposed model outperforms TF and is found to be strongly aligned with the ground truth.

In Table VI, we see that the IoU of 96.9% is associated with the building class, which proves to be the highest compared to other classes. This can also be demonstrated by qualitative inspection in Fig. 8. Moreover, we observe an increase in IoU by 3.2% for road class over TF. This is a significant increase, which is reflected in Fig. 8(d) and (e). Recognition of the water class is also increased by 2.4% and 3.2%, respectively, over both TF and UF by using our MSF head. Furthermore, in the car class, IoU has an overall increase of 3.3% and 1.1% compared to baseline TF and UF, respectively. A proper demarcation of cars on the road in Fig. 8(e) is a perfect example of these quantitative improvements. In the end, we demonstrate that despite the spatial heterogeneities and imbalanced class distribution, the models perform well on the UAVPal dataset with the potential to perform even better with our proposed MSF head.

It is worth noting that, despite the increased scene complexities compared to the Potsdam and Vaihingen datasets, the overall performance of the models on the UAVPal is comparable to the performance on the Potsdam and Vaihingen datasets. We suspect that in the label rasterization, the nearby buildings with insufficient separation get merged into a single segment, which makes it easier for the model to detect a cluster of buildings instead of delineating well-separated individual buildings in other datasets. In addition, in the UAVPal, waterbody segments are relatively large, easy to detect, and very limited, which improves the mIoU.

E. Ablation Study

To investigate the impact of attention mechanisms used in our proposed segmentation head, we train a separate network without applying the attention mechanism and observe its behavior to examine the effect of attention mechanisms used in our proposed segmentation head. Comparing the loss curves of the models in Fig. 9, we notice that attention induces faster convergence and slightly lower expected loss.

Moreover, we observed that without attention, mIoU decreases by approximately $0.5\% - 1\%$. These observations are

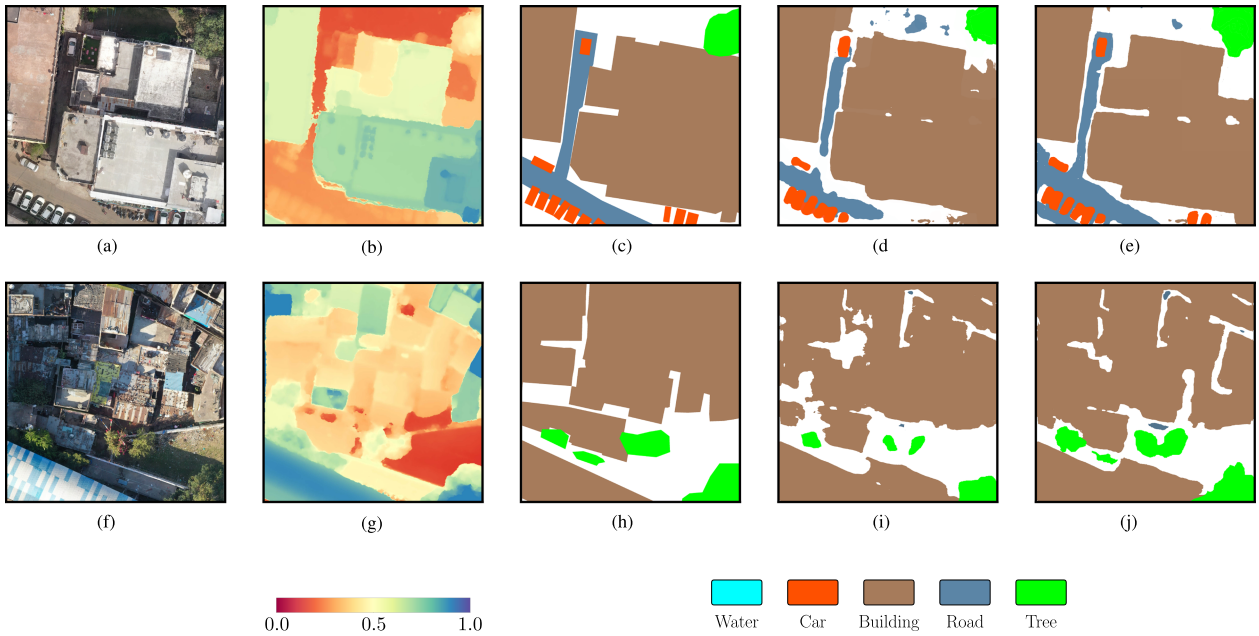


Fig. 8. Visual comparison of inferences with respect to the ground truth for a test image from the UAVPal dataset. Sample 14_00 and 07_09 from the test set are shown in the first and second rows, respectively. The DSM * refers to the DSM where the Z-scores of DSM scaled into $[0, 1]$ range. (a) Image14 00. (b) DSM*: 14 00. (c) Ground Truth: 14 00. (d) TF: 14 00. (e) TF + MSF (ours): 14 00. (f) Image: 07 09. (g) DSM*: 07 09. (h) Ground Truth: 07 09. (i) UF: 07 09. (j) UF + MSF (ours): 07 09.

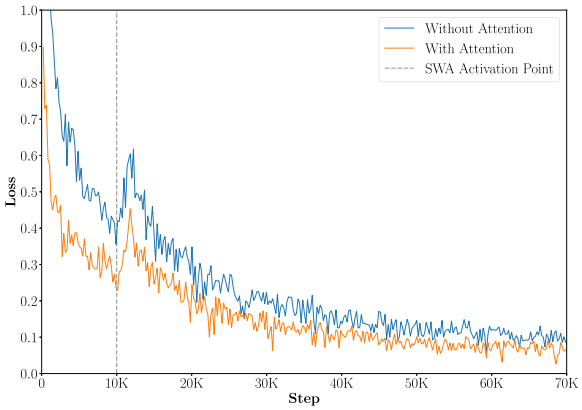


Fig. 9. Change of loss with respect to progression of training steps. A comparison of both curves highlights the effect of attention. SWA: SWA.

TABLE VII
EFFECT OF ATTENTION

	Attention	Potsdam	Vaihingen	UAVPal
HMSA[18]	✓	85.0	82.3	84.1
MSF	✗	84.8	82.1	84.1
	✓	**85.4**	**82.9**	**85.7**

The best values are marked in bold.

presented in Table VII. In addition, we compare the proposed head with the hierarchical multi-scale attention (HMSA) head [18] and discover that MSF outperforms the latter.

We also compare the TF with TF + MSF in terms of the number of trainable parameters, the estimated size of a single pass during training, and the number of floating point operations per second (FLOPS), representing the complexity and efficiency of the models. We use 1024×1024 input images with three channels and a batch having a single image for these estimations. In our implementations, TF has 4.47 million trainable parameters requiring 50.75 GFLOPS for a single forward pass, and the estimated size of a single pass is 2.12 GB. When combined with HMSA the TF has 4.56 million trainable parameters with 2.14 GB pass and 42.71 GFLOPs. In contrast, TF + MSF has 4.41 million trainable parameters with one single pass having an estimated size of 2.04 GB, and a forward pass requires 25.99 GFLOPS.

Although the number of trainable parameters and the size of a single pass for TF and MSF are not drastically different, the FLOPS required for MSF is almost half of what is needed for TF. We observed a similar trend in these metrics for UF and UF + MSF as well. This improvement in efficiency is mainly due to the efficiency of the group convolution operation and the subsequent reduction of learnable parameters attributed to the reduced channel depth of the feature groups. The implication of this faster training and inference time for MSF, compared to TF, is subject to the computing capability of the hardware. Yet, setting the parameter λ as learnable does not directly enhance the model's efficiency, but it eliminates the necessity for further hyperparameter tuning for λ.

F. Impact on Inference Time

To provide a clearer understanding of the computational benefits of our proposed approach, we present a direct comparison of inference times between our method and other relevant methods. Table VIII showcases the inference times, highlighting the efficiency gains achieved by our approach. Evidently, our MSF head

TABLE VIII
IMPACT OF PROPOSED MSF ON THE INFERENCE SPEED OF THE MODEL

Method	MSF	Latency (ms)
TF	✗	8.2
	✓	6.7
UF	✗	11.3
	✓	7.9

significantly reduces the inference time. In the case of both TF and UF, the inference latency decreased by approximately 20% when used with the proposed MSF. The inference times were measured with batch size 24 of 1024×1024 images having four channels using a single Nvidia A40 GPU. For each experiment, the latency is measured for 1000 iterations and the mean value is reported in Table VIII. These measurements have been carried out in FP32 mode. Depending upon the hardware, a combination of lower precision modes such as FP16 and adequate model quantization further optimizes the speed of inference.

VII. DISCUSSION

In this section, we explore the broader implications and context of our study, emphasizing the nuances and potential areas of interest for future research.

UAVPal in Broader Context: The introduction of the UAVPal dataset is a step toward rectifying the geographical imbalance in available datasets for urban scene understanding. Its focus on Bhopal, India, brings to light the unique urban landscapes of Southeast Asia, which have been underrepresented in previous datasets. This dataset's significance lies not just in its geographical focus but also in the potential it offers for region-specific research.

Balancing Efficiency and Performance: Our research underscores the possibility of achieving computational efficiency without sacrificing model performance. The design choices in our multiscale segmentation head highlight this balance, which is pivotal for applications where computational resources might be constrained.

Navigating Challenges: The UAVPal dataset, while valuable, comes with its set of challenges, including limited spatial coverage. In addition, the regulatory landscape around UAV usage in regions like India poses challenges for extensive data acquisition. One of the primary challenges is the manual annotation of images, which is not only time-consuming and cumbersome, but also expensive. Furthermore, manual annotations are prone to human errors affecting the quality and reliability. Minimizing these errors require stringent quality check, which further adds to the aforementioned drawbacks.

Looking Ahead: The groundwork laid by this study opens avenues for further exploration. Whether it is expanding the dataset's coverage, delving deeper into computational techniques, or applying the insights to other vision tasks. Moreover, The recent surge in popularity of interactive segmentation models such as segment anything [42] potentially reduces the effort required for manual annotation. This reduced necessity for manual intervention, while still retaining the option for manual refinement can significantly improve both the efficiency and accuracy of image annotation workflows in the future.

VIII. CONCLUSION

In our study, we focus on providing a new UAVPal dataset, which expands the overall diversity of the high-resolution open datasets for urban scene understanding. We describe the data collection procedure and subsequent preprocessing performed on the data to prepare it for model training and inference. The dataset contains very high-resolution images acquired from a low altitude, making it ideal for understanding the complexities present in a typical urban scene from a densely populated Indian city. The precisely annotated labels available with the dataset are adequate for computer vision tasks like semantic segmentation, as we have demonstrated through our experiments. However, the UAVPal dataset has limited spatial coverage. In the future, we would like to expand the dataset by expanding the spatial coverage and covering more cities, making it more challenging and useful.

Furthermore, a new multiscale segmentation head is implemented to reduce computational overhead with the potential to improve performance. We derive semantics from multiple available scales separately and fuse them according to scale-specific inferred importance. The proposed head has been thoroughly tested on two modern backbones (TopFormer and UNetFormer) and five high-resolution datasets with different properties showcasing the applicability of the segmentation head, on images with different resolutions, object size variabilities, viewing conditions, viewing perspectives, etc. Our experiments show that the proposed head has improved the IoU of several classes and the overall mIoU. While a better backbone largely improves overall feature extraction, our proposed head exploits the extracted feature better, which results in more stable and better inference. Moreover, the proposed head has fewer trainable parameters in the order of 6×10^4 and uses nearly 50% fewer compute operations, resulting in better training and inference times. These empirical findings underscore the effectiveness of our method and demonstrate its potential on our newly introduced UAVPal dataset.

VIII. ACKNOWLEDGMENT

The authors would like to thank the funding agencies for their support. In addition, the authors would like to extend our appreciation to the Accionland team for their assistance in conducting the survey for data acquisition. We declare that there are no known conflicting personal or financial interests that could have potentially influenced the research presented in this article.

REFERENCES

[1] X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, and X. Zhang, "A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet," *Sci. Rep.*, vol. 13, no. 1, May 2023, Art. no. 7600. [Online]. Available: https://doi.org/10.1038/s41598-023-34379-2

[2] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[3] D. Mehta et al., "Simple and efficient architectures for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 2628–2636.

[4] Z. Chen et al., "Vision transformer adapter for dense predictions," 2022, *arXiv:2205.08534*.

[5] X. Deng, Y. Zhu, Y. Tian, and S. Newsam, "Scale aware adaptation for land-cover classification in remote sensing imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2159–2168.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[9] 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. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[10] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer, 2014, pp. 740–755.

[11] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jun. 2014. [Online]. Available: https://doi.org/10.1007/s11263-014-0733-5

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 6230–6239.

[13] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 93, pp. 256–271, Jul. 2014. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2013.10.004

[14] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021, doi: 10.5281/zenodo.5706578.

[15] 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. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018, doi: 10.1109/TPAMI.2017.2699184.

[16] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 548–557. [Online]. Available: https://doi.org/10.1109/cvpr.2019.00064

[17] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[18] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," 2020, *arXiv:2005.10821*.

[19] M. Dushyant et al., "Simple and efficient architectures for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 2627–2635. [Online]. Available: https://doi.org/10.1109/cvprw56347.2022.00296

[20] D. Lam et al., "xView: Objects in context in overhead imagery," 2018. [Online]. Available: https://arxiv.org/abs/1802.07856

[21] J. Chen, Y. Xu, S. Lu, R. Liang, and L. Nan, "3-D instance segmentation of MVS buildings," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5704014.

[22] R. Gupta et al., "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 10–17.

[23] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020. [Online]. Available: https://doi.org/10.1016%2Fj.isprsjprs.2020.05.009

[24] CosmiQ Works, "SpaceNet 4: Off-Nadir buildings — SpaceNet.ai," 2019. Accessed: Jul. 27, 2023. [Online]. Available: https://spacenet.ai/off-nadir-building-detection/

[25] J. Castillo Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre, "MiniFrance," IEEE DataPort, 2020. [Online]. Available: https://dx.doi.org/10.21227/b9pt-8x03

[26] Global Facility for Disaster Reduction and Recovery (GFDRR) Labs, "Open cities AI challenge dataset," 2020. [Online]. Available: https://registry.mlhub.earth/10.34911/rdnt.f94cxb

[27] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[28] S. W. Zamir et al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 28–37.

[29] ISPRS, "2D semantic labeling dataset - Potsdam," 2016. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[30] ISPRS, "2D semantic labeling dataset - Vaihingen," 2016. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[31] S. Srivastava, S. Gupta, O. Dikshit, and S. Nair, "A review of UAV regulations and policies in India," in *Proc. Int. Conf. Unmanned Aerial Syst. Geomatics*, Springer, 2020, pp. 315–325. [Online]. Available: https://doi.org/10.1007/978-3-030-37393-1_27

[32] Pix4D, "Pix4D mapper," 2023. Accessed: Oct. 11, 2023. [Online]. Available: https://www.pix4d.com/product/pix4dmapper-photogrammetry-software

[33] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.

[34] Y. Liu, J. Yu, and Y. Han, "Understanding the effective receptive field in semantic image segmentation," *Multimedia Tools Appl.*, vol. 77, no. 17, pp. 22159–22171, Jan. 2018. [Online]. Available: https://doi.org/10.1007/s11042-018-5704-3

[35] W. Zhang et al., "TopFormer: Token pyramid transformer for mobile semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12073–12083.

[36] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[37] L. Liu et al., "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.

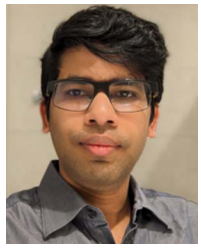
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[39] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Proc. Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl.*, Tien Pham, Eds., 2017, doi: 10.1117/12.2520589.

[40] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Proc. 34th Conf. Uncertainty Artif. Intell.*, 2018, pp. 876–885.

[41] B. Cheng, R. Girshick, P. Dollar, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15329–15337. [Online]. Available: https://doi.org/10.1109/cvpr46437.2021.01508

[42] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.



Abhisek Maiti received the M.Sc. degree in geo-information science and Earth observation from the University of Twente, The Netherlands, in 2019.

Since 2020, he has been working as Ph.D. Researcher with the Faculty ITC, University of Twente, The Netherlands. His main research interests are earth observation, photogrammetry, computer vision, and deep learning.

Sander Oude Elberink received the M.Sc. degree in geodetic engineering from the Delft University of Technology, the Netherlands, in 2000, and the Ph.D. degree in Geo-informatics from the University of Twente, The Netherlands, in 2010.

Since 2009, he has been an Assistant/Associate Professor with the Department of Earth Observation Science, ITC, University of Twente. His main research interests are automated information extraction from airborne and mobile laser scanner point clouds and 3D city and landscape modeling.

Dr. Oude Elberink was a recipient of the ISPRS Giuseppi Inghilleri award in 2016. He is the member of the International Science Advisory Committee from ISPRS.

George Vosselman received the M.Sc. degree in geodetic engineering from the Delft University of Technology, the Netherlands, in 1986, and the Ph.D. degree in photogrammetry from the Rheinische Friedrich Wilhelms University of Bonn, Germany, in 1991.

He worked as a Researcher with the Institute of Photogrammetry of the Stuttgart University, Germany, until 1992. After a year as a visiting scientist with the University of Washington, Seattle, WA, USA, he was appointed as a Professor of photogrammetry and remote sensing with the Delft University of Technology, the Netherlands, in 1993. In 2004, he joined ITC, University of Twente, Enschede, the Netherlands. From 2005 until 2012, he was an Editor-in-Chief of the ISPRS Journal of Photogrammetry and Remote Sensing. Currently, he is an Editor-in-Chief of the *ISPRS Open Journal of Photogrammetry and Remote S*ensing. He has authored or coauthored more than 300 journal and conference papers on photogrammetry and laser scanning. His research interests include the extraction of geo-information from imagery and point clouds.

Dr. Vosselman was a recipient of the ISPRS Karl Kraus medal (2012), ISPRS Schwidesfky Medal (2012), and ASPRS Photogrammetric Award (2015). In 2020, he was elected an ISPRS Fellow.