Exploration of Optimized Semantic Segmentation Architectures for edge-Deployment on Drones

Vivek Parmar1,2, Narayani Bhatia1, Shubham Negi1,2, and Manan Suri1,2,∗
1 Indian Institute of Technology Delhi, New Delhi, India; 2 CYRAN AI Solutions; ∗ manansuri@ee.iitd.ac.in

Abstract—In this paper, we present an analysis on the impact of network parameters for semantic segmentation architectures in context of UAV data processing. We present the analysis on the DroneDeploy Segmentation benchmark. Based on the comparative analysis we identify the optimal network architecture to be FPN-EfficientNetB3 with pretrained encoder backbones based on Imagenet Dataset. The network achieves IoU score of 0.65 and F1-score of 0.71 over the validation dataset. We also compare the various architectures in terms of their memory footprint and inference latency with further exploration of the impact of TensorRT based optimizations. We achieve memory savings of ∼4.1x and latency improvement of 10% compared to Model: FPN and Backbone: InceptionResNetV2.

Index Terms—Remote Sensing; Semantic Segmentation; UAV; TensorRT.

I. INTRODUCTION

Remote sensing applications have gained immense traction in recent years owing to the advent of high quality acquisition systems, sophisticated processing algorithms and increasingly accurate detection and classification methods enabled by deep learning. Owing to it’s rich features, some of the most popular remote sensing tasks rely on semantic segmentation to assign class-wise labels to each pixel in the frame. Semantic segmentation algorithms are used for a plethora of applications such as anomaly detection, event detection, land use cover change, etc [1]. Unmanned Aerial Vehicles (UAVs) have enabled capture of ultra-high resolution data [1] due to characteristics like low-cost, flexibility and low-flying altitude thus leading to increasing interest in the field.

In the context of UAV-centric deep learning applications, there has been some work around autonomous navigation [2], object tracking [3], change detection [4], semantic segmentation [5]. In recent times, deep learning implementations have been employed for real-time recognition and tracking on drones [6] with cloud-based processing. With resource-hungry Deep Convolutional Neural Networks (DCNNs) breaking accuracy ceilings, a significant aspect of the viability relies on reduction of costs in terms of network communication and computational energy. Hence UAV systems warrant use of new-age edge AI (artificial intelligence) hardware accelerators such as edge-GPU (Graphics Processing Units) [7], edge-TPU (Tensor Processing Units) [8], specialized ASIC (Application specific Integrated Circuits) [9], [10], which help achieve faster inference time, lightweight deployment and low-power localized processing. Studies proposing embedded deployment in the remote sensing domain have been limited to object detection [11], [12], scene classification [13], or semantic segmentation for satellite deployment [14].

To this purpose, we present a first-of-its-kind intensive algorithmic-hardware exploration for performing segmentation using UAV images based multi-class dataset in a resource-efficient manner for future-ready edge-AI deployments. Key contributions of the paper are listed below: (i) Detailed performance benchmarking of standard segmentation models on a new multi-class UAV segmentation dataset (DroneDeploy [15]). (ii) First demonstration of EfficientNet based Semantic Segmentation in context of UAV images on edge-AI accelerators. (iii) First hardware-software co-optimization study for semantic segmentation in context of UAV images.

II. MATERIALS AND METHODS

Dataset Description: For the purpose of this study, we have used DroneDeploy Dataset [15], comprising of 55 RGB images, along with single-channel elevation maps and label maps. The label maps are annotated with 7 classes - namely Building, Clutter, Vegetation, Water, Ground, Car and ‘Ignore’ - the last class referring to missing pixels/ boundaries. The ground resolution is 10 cm/pixel. For this study, we have only used raw RGB TIFFs, in order to demonstrate generalized capability without the need of additional channels such as elevation (as in the case of this dataset) or hyper-spectral bands (as in the case of other UAV datasets) due to relatively high costs of lightweight multispectral cameras [1].

Semantic Segmentation Architectures: In this work, we experiment with a variety of models, backbones, hyper-parameters and training variations before arriving at an accurate model and further try to optimize the architecture for embedded implementation. Segmentation models and encoder-backbones investigated for this study are listed in Table I.

Five encoder-backbones have been considered for this study: (i) EfficientNetB3 (ENB3) [16], (ii) InceptionResNetV2 (IRNv2) [17], (iii) MobileNetV2 [18] (MNv2) (iv) MobileNetV3 (MNv3) [19] (v) ShuffleNetv2 (SNv2) [20]. The prime motivation for the choice of backbones was to ensure an exhaustive exploration. For analyzing trade-off between

| Models          | UNet, FPN, LinkNet, PSPNet, UNet++ |
|-----------------|-----------------------------------|
| Encoder Backbone| MNv2, MNv3, ENB3, IRNv2, SNv2      |

TABLE I NETWORK ARCHITECTURES USED IN THIS STUDY
memory and accuracy over the complete design space, we have
selected the five backbones as representative workloads.

III. RESULTS

A. Memory Profiling

Real world edge-deployment on any UAV platform will
require DNN to have low memory footprint, low latency and
low power. To achieve lightweight implementation in terms of
memory footprint, there are two strategies: (i) Network archi-
tecture selection to reduce model weights, (ii) Quantization of
model weights and computation graph. Both these strategies
are discussed further.

Model Weights: Here, we compare the memory consumption
of different models’ weights as shown in Table II in order to
select the architecture with the best trade-off between model
weight size and validation accuracy.

Quantization based Memory Optimization: In order to
reduce memory footprint and inference latency, we analyze the
impact of quantizing floating-point-32 (FP32) precision net-
work to a more efficient data-type i.e. floating-point-16 (FP16).
FP16 quantization is performed based on NVIDIA TensorRT
(TRT) [21]. Fig. 1 demonstrates the prediction performance for
the baseline as well as TRT optimized models. We observe
that TRT FP16 model performance matches closely with
baseline FP32 model, and thus making it a valid candidate for
further latency analysis. Fig. 2 shows prediction performance
comparison of FP16-architectures with FPN model using all 3
encoder-backbones. As per expectations, MNv2 shows noisy
predictions whereas ENB3 and IRNv2 are at par with each
other.

B. Latency Profiling

For estimating edge-performance, using size of a complete
image from the dataset might not be accurate. We present
inference latency for image size corresponding to a typical
drone sensor (960×480) [22] in Table III. Out of the edge-
devices referred to in the study, the best performance is
obtained for NVIDIA Jetson Nano-Tensorflow framework.

IV. DISCUSSION

FPN model with encoder-backbone as ENB3 has the highest
validation IoU score when compared to all other architectures,
thus being the preferred choice. The choice of ENB3 over
IRNv2 stems from the marginally higher IoU score, along
with memory footprint considerations as shown in Table II.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Following are some important observations based on trends shown in Section III-A: (i) All models with MNv2 as backbone are lightweight, however, highly underperform when compared to other backbones’ Validation IoU. (ii) All IRNv2 models are very memory intensive, however perform better or at par when compared to other backbones for the same model. (iii) Out of all models, PSPNet has least memory consumption, however, it’s performance is also considerably lower than all other models. (iv) If models with best validation IoU are compared in terms of memory, IRNv2 is 4× heavier than ENB3 making it a suboptimal choice for lightweight on-the-edge deployment.

The above trends further explain motivation behind choosing the investigated architectures, since they provide a more complete picture of the trade-off between accuracy and memory footprint. MNv2 is highly resource efficient for deployment, whereas IRNv2 performs consistently and accurately, however ENB3 depicts the tradeoff between accuracy and memory consumption. Fig. 1 shows performance of FP 16 optimized models to be at par with FP32 models in terms of accuracy. This observation, compounded with memory footprint savings, motivates the need for investigating the latency performance for the TF-TRT models. Section III-B shows lower inference times for TF-TRT optimized models (improvements in the range of ~ 1.05 × to ~ 1.4 × except in the case of ENB3 performed on NVIDIA RTX 2080 Ti, possibly owing to a memory bottleneck issue. NVIDIA Jetson Nano emerges as the preferable choice as based on edge device estimates described in Table III.

## V. CONCLUSIONS

In this paper, we present a detailed benchmarking study of semantic segmentation models in context of UAV applications on the DroneDeploy dataset. We also present the first demonstration of semantic segmentation based on EfficientNet architectures for remote sensing applications. Based on extensive exploration, the best configuration is found to be: Model: FPN, Backbone: ENB3, Pretraining: ImageNet weights, Optimizer: AdamW with Learning rate scheduler and complete training which achieves IoU score of 0.65 and F1-score of 0.71 over the validation dataset. We also profile memory usage and latency for each model and optimize them for inference based on TensorRT with FP16 precision. Based on this we achieve memory savings of ~ 4.1× and latency improvement of ~ 10% compared to Model: FPN and Backbone: IRNv2.

## REFERENCES

[1] Huang Yao, Rongjun Qin, and Xiaoyu Chen, “Unmanned aerial vehicle for remote sensing applications—a review,” Remote Sensing, vol. 11, no. 12, pp. 1443, jun 2019.
[2] Paula Fraga-Lamas, Lucía Ramos, Víctor Mondejar-Guerra, and Tiago M. Fernández-Caramés, “A review on iOT deep learning UAV systems for autonomous obstacle detection and collision avoidance,” Remote Sensing, vol. 11, no. 18, pp. 2144, sep 2019.
[3] Wilbert G. Aguilar, Marco A. Luna, Julio F. Moya, Vanessa Abad, Humberto Parra, and Hugo Ruiz, “Pedestrian detection for UAVs using cascade classifiers with meanshift,” in 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017, IEEE.
[4] Danilo Avola, Luigi Cinque, Gian Luca Foresti, Niki Martinel, Daniele Pannone, and Claudia Pesarelli, “A UAV-based approach for monitoring and change detection from low-altitude flights,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 50, no. 6, pp. 2139–2149, jun 2020.
[5] Ming-Der Yang, Hsin-Hung Tseng, Yu-Chun Hsu, and Hui Ping Tsai, “Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date UAV visible images,” Remote Sensing, vol. 12, no. 4, pp. 633, feb 2020.
[6] Neurala, “Neurala Drones.” https://www.neurala.com/solutions/drones.
[7] Stephen Cass, “Nvidia makes it easy to embed AI: The jetson nano packs a lot of machine-learning power into DIY projects - [hands on],” IEEE Spectrum, vol. 57, no. 7, pp. 14–16, jul 2020.
[8] Google, “Edge TPU,” https://coral.ai/docs/.
[9] Luca Puglia, Mircea Ionica, Giancarlo Raiconi, and David Moloney, “Passive dense stereo vision on the myriad2 VPU,” in 2016 IEEE Hot Chips 28 Symposium (HCS), aug 2016, IEEE.
[10] Vivek Parmar, Jung-Ho Ahn, and Manan Suri, “Hyperspectral image classification for remote sensing using low-power neuromorphic hardware,” in 2019 International Joint Conference on Neural Networks (IJCNN), jul 2019, IEEE.
[11] Luis Barba-Guaman, José Eugenio Naranjo, and Anthony Ortiz, “Deep learning framework for vehicle and pedestrian detection in rural roads on an embedded gpu,” Electronics, vol. 9, no. 4, pp. 580, 2020.
[12] Christos Kyrkou, George Plastiras, Theocharis Theocharides, Stylianos I. Venieris, and Christos-Savvas Bougianis, “DroNet: Efficient convolutional neural network detector for real-time UAV applications,” in 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), mar 2018, IEEE.
[13] Christos Kyrkou and Theocharis Theocharides, “Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), jun 2019, IEEE.
[14] Junxing Hu, Ling Li, Yijun Lin, Fengge Wu, and Junsu Zhao, “A comparison and strategy of semantic segmentation on remote sensing images,” in The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Springer, 2019, pp. 21–29.
[15] DroneDeploy, “DroneDeploy Segmentation Benchmark,” https://github.com/dronedeploy/dd-ml-segmentation-benchmark.
[16] Mingxing Tan and Quoc V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, Kamalika Chaudhuri and Ruslan Salakhutdinov, Eds. 2019, vol. 97 of Proceedings of Machine Learning Research, pp. 6105–6114, PMLR.
[17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Thirty-first AAAI conference on artificial intelligence, 2017.
[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, jun 2018, IEEE.
[19] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le, “Searching for MobileNetV3,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV). oct 2019, IEEE.

[20] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, “ShuffleNet v2: Practical guidelines for efficient CNN architecture design,” in Computer Vision – ECCV 2018, pp. 122–138. Springer International Publishing, 2018.

[21] Han Vanholder, “Efficient Inference with TensorRT,” 2016.

[22] DJI, “DJI MATRICE 300 Specs,” https://www.dji.com/matrice-300/specs.