COMPOSITE AUGMENTATIONS FOR SEMANTIC SEGMENTATION IN AERIAL IMAGES WITH FEW SAMPLES

Pranav Chandramouli & Ian Stavness
Department of Computer Science
University of Saskatchewan
{prc051,stavness}@cs.usask.ca

Philip D. McLoughlin & Branden Neufeld
Department of Biology
University of Saskatchewan
philip.mcloughlin@usask.ca

ABSTRACT

ML4RS has the potential to enable comprehensive population monitoring to inform wildlife and biodiversity conservation. However, annotated datasets of wildlife in-situ are often difficult, expensive, and time-consuming to procure. This paper proposes a computational and data efficient method to synthesize composite images to supplement real-world data in data-sparse environments with few positive samples. Our method showed up to a 3% increase in target-class IoU scores on three aerial remote sensing datasets. We aim to use this method with a novel aerial dataset of the Boreal forest for ungulate monitoring.

1 INTRODUCTION

Remote sensing of wildlife habitat at scale via aerial imaging and deep-learning (DL) based detection/segmentation methods has potential to dramatically improve our understanding of remote ecozones. Detailed and comprehensive aerial imaging to count and track animals can be used to inform our understanding of fragile population dynamics, monitor and track progress on efforts to improve biodiversity (e.g. the reintroduction of native species), and improve the security of traditional food sources for Indigenous communities (e.g. Caribou in the Northern Boreal forest). Building DL models for this problem is challenging due to variability in vegetation and animal appearance across ecozones and seasons, and is particularly challenging for sparsely populated ecozones, such as large mammals in forested environments, where images with positive samples are rare. We have encountered these issues while curating a novel aerial image dataset of the Canadian Boreal Plains ecozone: of 612 distinct image sets captured over 5 flights and an area greater than 1000 km^2, we found only 24 images with positive samples (127 animal instances across 8 target classes).

The USask-Wilds dataset is a novel multispectral wildlife dataset that aims to develop our understanding of the food web dynamics at scale in the Canadian Boreal Plains ecozone. The dataset aims to provide data to allow for modeling beyond the reach of ecologists due to a need for more data on the densities of interacting species, especially for species that are costly to monitor, like large mammals in forested environments. New approaches to wildlife detection tackle the fundamental, yet intractable, problem of cost-effectively obtaining accurate, precise, and simultaneous data on multiple wildlife populations at scale to monitor complex population dynamics.

Synthetic data has been shown to effectively supplement real-world data in training DL models in cases of sparse positive samples or unbalanced class distributions (Nikolenko, 2019; Kim et al., 2022). Some of the most common methods use generative adversarial networks (GANs, Goodfellow et al., 2014), variational autoencoders (Kingma & Welling, 2022), diffusion models (Ho et al., 2020; Li et al., 2023), and computer graphics (Ubbens et al., 2018) to generate high-fidelity synthetic images from masks or label references. These approaches have shown promise in alleviating the challenges posed by unbalanced datasets. However, the successful application of these generative approaches requires high computational costs and large datasets, e.g. reduced training data sizes results in drastic performance degradation for GANs Zhao et al. (2020). An alternative light-weight approach for synthetic data augmentation is image compositing: taking the foreground objects of one image and combining it with the background from another image (Chen & Kae, 2019; Zhang et al., 2020). Compositing has been used previously for remote sensing of weeds in field images, where weed plant instances were very sparse compared to instances of crop plants Gao et al. (2020).

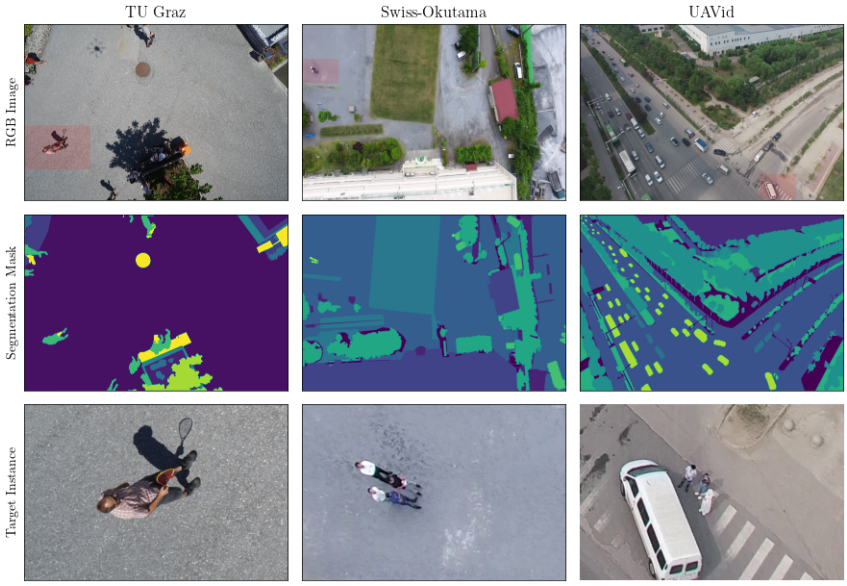

Figure 1: Example images from the TUGraz (left), Swiss-Okutama (middle), and UAVid (right) datasets, including the original RGB images (top row, red box denotes target instance), the corresponding segmentation masks (middle row), and a zoomed-in view of a target instance (bottom row).

Dataset	# Images	# Objects	# Targets	% Targets	% Target Pixels
TUGraz	398	141,508	1,566	1.106	1.052
Swiss	281	108,326	136	0.125	0.004
UAVid	250	98,067	5,750	5.863	0.906

Table 1: Summary of datasets used, including target class frequency by count and pixel count.

In this paper, we propose a novel image compositing approach tailored to aerial images with limited positive samples. We evaluate its use in data augmentation for training semantic segmentation models for *person* detection in three public aerial image datasets. Our results show a 3% increase target-class IoU and tighter IoU prediction spreads. In future, we plan to apply this approach to animal detection in our Boreal aerial image dataset, which is currently in progress.

2 METHODS

2.1 DATASETS

As our new wildlife aerial dataset is under development, we used three public remote sensing datasets with similar characteristics (drone-captured, sparse target class instances) to evaluate image-composite based data augmentation in this study (see examples in Figure 1, summary in Table 1).

The **TUGraz Semantic Drone Dataset** (TUGraz, 2019) was curated by the Institute of Computer Graphics and Vision at the Graz University of Technology (TU Graz) with autonomous flight safety and landing procedures in mind. The dataset focuses on a semantic understanding of urban scenes, with 20 houses from a nadir (bird's eye) view acquired 5 to 30 meters above the ground.

The **Swiss-Okutama Drone Datasets** (Speth et al., 2022) is comprised of the Swiss Dataset — 100 images taken around Cheseaux-sur-Lausanne in Switzerland, at a flight height of around 80 meters and the Okutama Drone Dataset, with 91 images taken around Okutama, west of Tokyo, Japan, flying at a height of around 90 meters. All images represent are taken from a nadir, and are hand-labeled with pixel-wise semantic segmentation annotations.

The **UAVid Dataset** (Lyu et al., 2020) is a high-resolution UAV semantic segmentation dataset, which includes large-scale variation and moving objects. The dataset consists of 30 video sequences capturing high-resolution images in oblique views.

2.2 Data Augmentation with Composite Images

Image composites aim to modify the distribution of target class objects (classes of interest) by placing these target objects onto a different part of the image. The new location is determined using constraints on the background pixels to ensure reasonable realism in the composite images. The process was designed to be as unrestrictive as possible — allowing the script to randomize location, maximizing variation, and only restricting the background pixels to ensure a realistic spatial association with the background to reflect pragmatic areas in which the target class would naturally be located. The target masks are shared across images, allowing the final target pixels in composites to originate from another image. This was especially useful in cases with cut-off target objects or images with no target objects. Our image composites are based on target objects in the pictures, and we aim to use the composites to supplement images with low-quality target objects or no target objects. We used the class representing human objects in the images as the target class for this study.

2.3 Model and Evaluation

We used DeepLabV3+ (Chen et al., 2018) as the baseline model, because it is still considered the leading CNN approach to semantic segmentation. DeepLabV3+ adds an encoder-decoder to DeepLabV3 (Chen et al., 2017), which, itself, was novel in removing the `DenseCRF` post-processing layer and adding atrous convolution layers to improve performance and computational efficiency. We used Iakubovskii (2019)'s DeepLabV3+ implementation and a ResNeXt (Xie et al., 2017) encoder with ImageNet (Deng et al., 2009) weights, yielding 86M trainable parameters.

We used the Lovász-Softmax (Berman et al., 2018) loss for this application, as it is a direct optimization of Jaccard-Index (see Sec 2.4). It was shown to perform better with respect to Jaccard losses than the traditionally used cross-entropy loss. We can provide an optional argument weight — a 1D Tensor assigning weight to each class, which is particularly useful when working with an unbalanced training set. We elected not to do this for this study as the real and synthetic datasets have completely different distributions, which would require two different sets of weights. While this would make the model perform well within the chosen homogenous data for the experiment, it would hamper performance in other experiments with different distributions. We used the `OneCycleLR` learning rate scheduler, which takes advantage of a phenomenon called "super-convergence" (Smith & Topin, 2018), where neural networks can be trained an order of magnitude faster than with standard training methods. This was especially useful in this application considering that Smith & Topin (2018) found that super-convergence provides a more significant performance boost than standard training when the amount of labeled training data is limited.

2.4 Metric — Mean Intersection-over-Union (Jaccard-Index)

To measure model performance, we use the canonical semantic segmenation metric of Intersection over Union (IoU), also known as the Jaccard Index. IoU is the area of the intersection over the union of the predicted segmentation and the ground truth segmentation.

The Mean Intersection over Union, mIoU, is the average IoU over all classes in the image. It reflects the model's performance at predicting, on average, across all object classes in the image. We use equal weights for the different classes, so no penalty or reward is applied for different class types. For a ground-truth mask GT and a predicted mask P, with k classes, the mIoU can be calculated as:

$$mIoU = \frac{1}{n} \sum_{i=1}^{k} \frac{GT_i \cap P_i}{GT_i \cup P_i} \tag{1}$$

In addition to the mean IoU, we also use a class IoU that is often used in instance-segmentation applications to judge the performance of our approach on the target class. This metric differs slightly, because we only consider the masks and predictions with respect to the target class (T), as follows:

$$IoU_T = \frac{GT_T \cap P_T}{GT_T \cup P_T} \tag{2}$$

We also report mIoU Kurtosis $(\alpha 4)$, which is a measure of the "tailedness" of the probability distribution (Kallner, 2018). The standard normal distribution has a kurtosis of 3, and is considered mesokurtic. $(\alpha 4) > 3$ distributions, also known as leptokurtic distributions can be visualized as a thin bell curve - with a high peak, while $(\alpha 4) < 3$ distributions are called playtkurtic, and have broad peaks. Kurtosis of IoU distributions has been used to evaluate the performance of a model (Taran et al., 2018), which compared the performance of models to a normal distribution.

3 RESULTS

We evaluate our method on the real-world test sets. We ran two separate experiments, a baseline performed on the raw dataset (*Real*) and a hybrid experiment, which includes a 1:1 ratio of synthetic and real training images (*Hybrid*). Both experiments were evaluated on real-world test images. We ran over 20 iterations of the compositing and evaluation process to calculate these results over multiple composite data.

Dataset	Experiment	$\mu\ IoU \uparrow$	$\mu\ IoU_T \uparrow$	$\alpha 4\ IoU$	$\alpha 4\ IoU_T$
TU Graz	Real	0.565	0.667	1.091	6.029
	Hybrid	0.579	0.695	1.106	1.387
Swiss	Real	0.590	0.108	-0.764	-1.082
	Hybrid	0.613	0.108	-0.604	-0.742
UAVid	Real	0.514	0.204	-0.405	-0.581
	Hybrid	0.519	0.235	-0.146	-0.663

Table 2: Mean (μ) and Kurtosis $(\alpha 4)$ calculated from predicted masks on the real test sets. Higher mean IoU values indicate better predictions. A Kurtosis value of 3 suggests a normal distribution, high values indicate pronounced peaks, and low values suggest flattened peaks.

The real and hybrid experiments are similar in terms of their overall performance; the hybrid experiment outperforms the real experiment, albeit with minor performance improvements — which could be attributed partly to the larger dataset size in the synthetic dataset. However, the hybrid dataset outperforms the real data in target mask inference, with $0.02 - 0.03$ improvements in IoU. UAVid shows a more significant increase, which indicates that the shifted oblique point-of-view lends to more robust inferences with synthetic data.

The overall mIoU Kurtosis once again improved from real data to hybrid data. The improvement in mIoU kurtosis is more pronounced than the change in mean metrics. The Class IoU of the target classes shows a more intriguing trend. UAVid, an urban dataset showing city scenes with roads, shows a marginal decrease in kurtosis. TUGraz had an extremely high Kurtosis of 6, which overcorrected to a kurtosis of 1.3 on the hybrid data, owing to the variation in open urban areas — cemented/paved areas, grass, and urban stairs. The Swiss-Okutama dataset has a combination of more open areas and a varied mix of urban concentration and also showed a more reasonable correction of kurtosis, closer to the normal distribution.

4 DISCUSSION

The objective of this study was to assess whether including composite images of a target class is an effective data augmentation strategy for aerial image datasets. We found that our image composite augmentations resulted in a 3% increase in target-class IoU, and tighter IoU prediction spreads that are closer to the normal distribution. While the work generated positive results, there are multiple avenues for future work worth exploring. We are interested in studying the effects of distribution shifts in model performance and robustness to distribution shifts which are common in aerial image datasets (different locations, different time-of-year, et al.). We are also interested in applying this method to animal detection in our new aerial image dataset of ungulates in the Canadian Boreal Plains ecozone, which we will report on our progress at the ML4RS workshop.

REFERENCES

Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, 2018.

Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8407–8416, 2019. doi: 10.1109/CVPR.2019.00861.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pp. 248–255, 06 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848/.

Junfeng Gao, Andrew P. French, Michael P. Pound, Yong He, Tony P. Pridmore, and Jan G. Pieters. Deep convolutional neural networks for image-based convolvulus sepium detection in sugar beet fields. *Plant Methods*, 16(1):29, Mar 2020. ISSN 1746-4811. doi: 10.1186/s13007-020-00570-z. URL https://doi.org/10.1186/s13007-020-00570-z.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.

Anders Kallner. Laboratory statistics (second edition). In Anders Kallner (ed.), *Laboratory Statistics (Second Edition)*, pp. iv. Elsevier, second edition edition, 2018. ISBN 978-0-12-814348-3. doi: https://doi.org/10.1016/B978-0-12-814348-3.00003-4. URL https://www.sciencedirect.com/science/article/pii/B9780128143483000034.

Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35710–35723. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e8507db80464ced5658d16b49bd458b9-Paper-Datasets_and_Benchmarks.pdf.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023.

Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2020.05.009. URL http://www.sciencedirect.com/science/article/pii/S0924271620301295.

Sergey I. Nikolenko. Synthetic data for deep learning, 2019.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates, 2018.

Simon Speth, Artur Gonçalves, Bastien Rigault, Satoshi Suzuki, Mondher Bouazizi, Yutaka Matsuo, and Helmut Prendinger. Deep learning with rgb and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6):840–868, 2022. doi: https://doi.org/10.1002/rob.22082. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22082.

Vlad Taran, Nikita Gordienko, Yuriy Kochura, Yuri Gordienko, Alexandr Rokovyi, Oleg Alienin, and Sergii Stirenko. Performance evaluation of deep learning networks for semantic segmentation of traffic stereo-pair images. In *Proceedings of the 19th International Conference on Computer Systems and Technologies*, CompSysTech'18. ACM, September 2018. doi: 10.1145/3274005.3274032. URL http://dx.doi.org/10.1145/3274005.3274032.

ICG TUGraz. Semantic drone dataset. http://dronedataset.icg.tugraz.at/, 2019.

Jordan Ubbens, Mikolaj Cieslak, Przemyslaw Prusinkiewicz, and Ian Stavness. The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant methods*, 14: 1–10, 2018.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M. Patel. Deep image compositing, 2020.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. *CoRR*, abs/2006.10738, 2020. URL https://arxiv.org/abs/2006.10738.