A CONCISE TILING STRATEGY FOR PRESERVING SPATIAL CONTEXT IN EARTH OBSERVATION IMAGERY

Ellianna Abrahams, Tasha Snow, Matthew R. Siegfried, & Fernando Pérez
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA
{ellianna, fernando.perez}@berkeley.edu

Department of Geophysics
Colorado School of Mines
Golden, CO 80401, USA
{tsnow,siegfried}@mines.edu

ABSTRACT

We propose a new tiling strategy, **Flip-n-Slide**, which has been developed for specific use with large Earth observation satellite images when the location of objects-of-interest (OoI) is unknown and spatial context can be necessary for class disambiguation. Flip-n-Slide is a concise and minimalistic approach that allows OoI to be represented at multiple tile positions and orientations. This strategy introduces multiple views of spatio-contextual information, without introducing redundancies into the training set. By maintaining distinct transformation permutations for each tile overlap, we enhance the generalizability of the training set without misrepresenting the true data distribution. Our experiments validate the effectiveness of Flip-n-Slide in the task of semantic segmentation, a necessary data product in geophysical studies. We find that Flip-n-Slide outperforms the previous state-of-the-art augmentation routines for tiled data in all evaluation metrics. For underrepresented classes, Flip-n-Slide increases precision by as much as 15.8%.

1 INTRODUCTION

The volume of geospatial satellite imagery has grown rapidly in the past decade. Semantic segmentation presents a promising opportunity for rapidly parsing meaningful scientific understanding from these images. Despite the remarkable accomplishments of deep neural networks for such segmentation tasks (Ronneberger et al., 2015; Chen et al., 2019; Tan et al., 2020; Amara et al., 2022), these methods can underperform on data that have noisy or underrepresented labels (Shin et al., 2011; Guo et al., 2019) or when one set of data representations is used for a wider set of downstream tasks (Yang et al., 2018). These are common challenges in Earth observation imagery. To overcome these issues, data augmentation is a widely adopted technique for generalizing a model fit to make better predictions by expanding the size and distribution of training data through a set of transformations (Van Dyk & Meng, 2001; Hestness et al., 2017). In recent years, much focus has been given to upstream augmentation methods that address overfitting through data mixing (Zhang et al., 2017; Yun et al., 2019; Hong et al., 2021; Dabouei et al., 2021) or proxy-free augmentations (Cubuk et al., 2019; 2020; Reed et al., 2021; Li & Li, 2023)—strategic approaches that expand the training data, but also execute unrealistic data transformations. Furthermore, limited attention has been given to investigating upstream augmentation techniques in the realm of learning on tiled imagery, an approach often employed to parse large images into smaller tiles to overcome the intractable size of the overall image for the GPU memory (Pinckaers & Litjens, 2018; Huang et al., 2019).

Tiling techniques in scientific applications require intentional augmentation choices, as certain transformations are unphysical and therefore not useful to learn. Spatial context is needed to disambiguate between classes with similar channel outputs and surface textures (Pereira & dos Santos, 2021). This is particularly true in scientific use-cases where tasks like classification are employed to answer wider questions about rare or relatively unknown phenomena. Although most machine learning methods demonstrate robust performance with well-represented phenomena, this performance degrades when training data fails to accurately capture poorly represented features or the structure of the problem. Human experts are better able to identify rare and unknown phenomena by relying on domain context, such as the semantic or temporal proximity of other classes (Wang & Zhu, 2023); it stands to reason that machines could similarly use context for similar purposes.

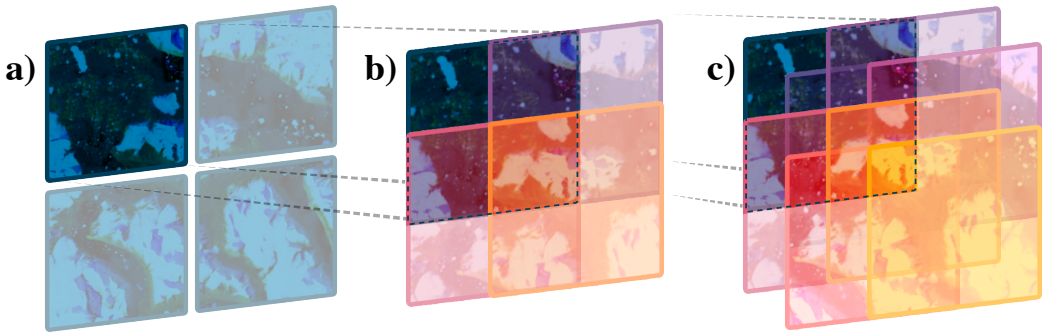

Figure 1: Flip-n-Slide's tile overlap strategy creates eight overlapping tiles for any image region more than a 75% tile threshold away from the overall image edge. Three tiling strategies, shown in false color to illustrate overlap, are visualized here. a) Tiles do not overlap. b) The conventional tile overlap strategy, shown at the recommended 50% overlap. c) Flip-n-Slide includes more tile overlaps, capturing more OoI tile position views for the training set.

To maintain access to context after tiling, the current state-of-the art employs tiling as a sliding window data augmentation, overlapping tiles by a significant percentage to extend dataset size and expose the model to multiple contextual views (Ronneberger et al., 2015; Ünel et al., 2019; Zeng & Zheng, 2019; Akyon et al., 2022) at training time, at test time, or during both. However, when applied upstream, this approach leads to redundancies within the training set, as many pixel windows are repeated more than once; for objects-of-interest (OoI) that are the size of the window slide, spatial context is limited beyond the singular sliding parameter (Zeng & Zheng, 2019; Reina et al., 2020). This leads to the question: **How can we best tile and augment large images with limited physically realistic transformations without losing important spatio-contextual information in the tiling process?** Further: **Can we achieve this without creating the redundancies that occur in simply overlapping tiles?**

To answer this question, we propose a concise augmentation strategy, **Flip-n-Slide**, built for use with large Earth observation images where: 1) tiling is necessary; 2) data transformations must be limited to rotations and reflections to be realistic; and 3) there is no prior knowledge of the pixel locations for which spatial context will be necessary. We argue that physically realistic transformations of the data can be implemented *alongside* the tiling overlap process, thereby removing redundancies when training convolutional neural networks (CNNs), in which orientation matters for learning (Ghosh et al., 2018; Szeliski, 2022). Our strategy allows us to create a larger set of samples without the superfluity of simply overlapping the tiles, enhancing downstream generalization. To achieve our goal, we slide through multiple overlap thresholds of the tiling window, exposing the model to more contextual views of each location (Figure 1), and distinctly permute any overlapping windows to eliminate redundancy with other tiles that share pixels. We run segmentation experiments using our routine on the Land Cover of Canada (LCC) dataset (Latifovic, 2020) to classify Landsat 8 satellite imagery (Earth Resources Observation And Science (EROS) Center, 2013; Roy et al., 2014). Our findings demonstrate that our strategy improves the segmentation performance of tiled data, especially for underrepresented classes, when compared to the conventional method (Figure 2c).

2 Tiling and Augmentation Method

We present the Flip-n-Slide algorithm, a data preprocessing strategy that tiles a large input image using a sliding window, providing eight sliding overlaps for every tile. Redundancies on overlap are eliminated using distinct, physically realistic transformations to permute each overlapping tile.

2.1 Mathematical Formalism for the Flip-n-Slide Algorithm

Flip-n-Slide follows a concise mathematical formulation to concurrently tile and transform a large image. In Algorithm 1 we outline the formal process for implementing Flip-n-Slide on an input

image, I. The algorithm is implemented in two stages. First a sliding window captures overlapping tiles, $t \in T$, at specified boundary ranges, $i \times j$, in the 2D image plane of I. The boundary ranges are given in S. This operation is specified in Line 2. The tile boundaries are defined such that any given $x \times y$ pixel location sufficiently away from the edge of the overall image I overlaps with eight tiles. The length of i is equal to the length of j, creating a square. Each of these overlapping tiles is permuted correspondingly from a specified set of eight distinct transformations, F. This operation is specified in Line 3. As augmentation is implemented concurrently with tiling, Flip-n-Slide must be implemented before the data streaming stage. This results in the constraint of CPU storage capacity required to accommodate the complete extended dataset. The overlap and transformation choices are explained in greater detail in Section 2.2. We release this algorithm as a versioned, documented, open-source `Python` package named `flipnslide`, available on PyPI and GitHub.

2.2 Hyperparameter Selection for Tiling and Transformation

We justify the fixed parameter choices for Flip-n-Slide to meet the goal of preserving spatial context for OoI, while eliminating informational redundancies that have been introduced in past tiling strategies. Our approach has the added benefit of expanding the initial training distribution with physically realistic transformations, adding further downstream generalization to the model fit.

Tile Size: Similar to past methods (Ronneberger et al., 2015; Ünel et al., 2019), Flip-n-Slide employs a fixed tile size to enable ease-of-scale when processing petapixel datasets. The algorithmic method is independent of tile size as needs will vary depending on the size of OoI, allowing users to pick a size that is most appropriate for a specific use-case. However, we limit the algorithm to square tiles to make the data more efficient for use with CNNs.

Tile Overlap : Previous methods employ a single overlap threshold, generally recommended to be 50%. Flip-n-Slide uses a strategy with multiple overlap thresholds, where a tile window slides along to capture a 0%, 25%, 50%, and 75% threshold on both spatial axes, leading to eight overlaps for any given $x \times y$ area that is more than a 75% tile threshold away from the edge of the original large image. This method provides the model with multiple tile views of contextual information for OoI, leading to greater downstream tile-position-invariance. Additionally, this method reduces the overhead present in dynamic methods by removing the need for an upstream check that small objects are not sliced during augmentation; in Flip-n-Slide sliced objects that are smaller than the tile size will have unsliced representations in other overlapping tiles.

Transformation Permutations : Rotations and reflections provide alternative views of spatial relationships to the convolutional kernel in CNNs (Van Dyk & Meng, 2001). For each of the eight tile overlaps, Flip-n-Slide employs a distinct rotation and/or reflection transformation from the following set of permutations $\{0°, 90°, 180°, 270°, (0°, \rightarrow), (0°, \uparrow), (90°, \rightarrow), (90°, \uparrow)\}$, shown in Figure 4. This set avoids commutative and indistinct transformation compositions to reduce redundancies for tiles that overlap in some way on the same initial image position. Additionally, it is limited to physically realistic transformations of the original image bypassing any pixel-level effects that can be introduced when rotating square matrices to angles indivisible by $90°$.

3 Experimental Setup

We evaluate our strategy's performance on a classification task with a benchmarked semantic segmentation approach and compare it with the current recommended tiling convention as a control. Across all experiments, we argue that our augmentation strategy exposes the model to enough semantic context to remove the need for augmentation and label averaging at training time. To test this hypothesis, we only tile the test image and do not add any further augmentations at test time.

We perform our experiment on data from Ellesmere Island in the Canadian Arctic to ensure that our tiling strategy performs well on datasets in wilderness regions, which are an understudied focus of inquiry in established machine learning datasets, (Figure 2a). We use land cover classifications from the publicly available Land Cover of Canada (LCC) dataset (Latifovic, 2020) and the corresponding images from NASA's Landsat 8 imagery (Earth Resources Observation And Science (EROS) Center, 2013; Roy et al., 2014) since the LCC was derived from Landsat images. The data are classified by seven labeled classes (Snow & Ice [39.36% of the overall dataset], Barren Rock [33.57%], Water [17.98%], Polar Moss & Lichen [4.97%], Polar Grassland [4.23%], Urban Development [0.0003%],

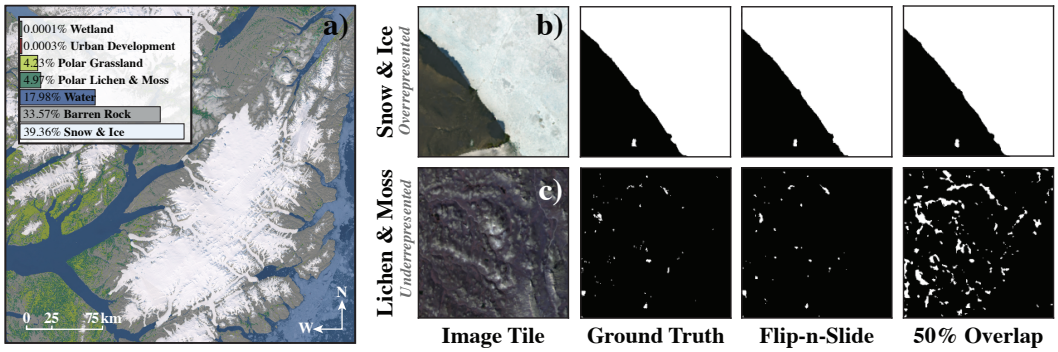

Figure 2: In Earth observation, classes can be extremely imbalanced as is shown here for Ellesmere Island, Nunavut, CA. a) Labels from the Land Cover of Canada dataset (Latifovic, 2020) overlain on the corresponding Landsat satellite imagery (Earth Resources Observation And Science (EROS) Center, 2013). The legend shows the relative class distribution. Image tiles showing (b) an over-represented class (snow and ice) and (c) an under-represented class (lichen and moss) with the binary class masks from the ground truth data set, segmentation after training using the Flip-n-Slide tiling algorithm, and segmentation after training using the conventional tiling algorithm (50% overlap). Although both algorithms perform well for the over-represented class case, Flip-n-Slide is more precise, by up to 15.8%, than the conventional strategy (50% tile overlap).

Wetland [0.0001%]). There is high class imbalance in this data as is common in geophysical land cover segmentation problems, where often the OoI are among the least represented classes.

We evaluate the performance of Flip-n-Slide on a semantic segmentation task, generating 12,800 256×256 tiles from the dataset. All training experiments are conducted using the benchmarked UNet architecture (Ronneberger et al., 2015), employed with the ADAM optimizer (Kingma & Ba, 2017) at a learning rate of 0.001 and an unweighted cross-entropy loss. We train additional models on the two control tiling strategies to test the comparative performance, employing the same model architecture and optimization. The models undergo 300 training epochs with a batch size of 32.

4 Experimental Results

We follow the standard evaluation setup for semantic segmentation tasks using a UNet (Ronneberger et al., 2015). We train a first model on tiles generated by employing the Flip-n-Slide strategy. As a comparative control, the second model is trained on non-overlapping tiles, and the third is trained on tiles generated using the conventional 50% overlap strategy (Ünel et al., 2019; Zeng & Zheng, 2019; Reina et al., 2020; Akyon et al., 2022). We test all three trained models on three random draws of non-overlapping tiles from the same geographic location, and share the performance results, averaged across the tests, in Table 1. In each experiment, Flip-n-Slide improves the model performance across all evaluation metrics when compared to conventional methods. These results position Flip-n-Slide as an effective augmentation strategy for tiling large scientific images, particularly in use-cases with high class-imbalance.

4.1 Performance Improvement for Underrepresented Classes

Incorporating the Flip-n-Slide strategy enhances model performance on underrepresented classes, even without tailoring the model architecture or loss function to addressing class imbalance (Table 1). Excluding the two classes in this investigation that exhibit extreme underrepresentation (<0.0004% of the overall data) and are therefore subject to model noise[1], the next two smallest classes each make up less than 5% of the overall dataset. Flip-n-Slide improves prediction precision by 13%, on average, in underrepresented classes, and outperforms other approaches in every met-

[1]Flip-n-Slide also outperforms other tiling methods on the two extremely underrepresented classes, but the performance improvement is within error, so we exclude those results here.

Performance Comparison Between Strategies					Performance On Underrepresented Class		
	Evaluated on All Classes				Moss & Lichen *(4.97% of Data)*		
Method	IoU	Precision	F1 Score	mAP	IoU	Precision	F1 Score
No Tile Overlap	70.7%	71.7%	82.8%	71.3	42.7%	44.4%	59.9%
50% Overlap	85.4%	87.1% %	92.1%	85.5%	58.3%	62.6%	73.6%
Flip-n-Slide *Ours*	**87.6%**	**90.1%**	**93.4%**	**87.8%**	**69.7%**	**78.4%**	**82.2%**

Table 1: Performance results for three tiling strategies across all class predictions. Flip-n-Slide outperforms other strategies in all metrics. Flip-n-Slide also improves performance results for underrepresented classes even with a basic loss function and model architecture.

ric tested. These results highlight the improved performance for underrepresented classes, with no changes to architecture, loss function, optimization, or any other model parameters.

Predicted masks for a withheld test set are shown for two classes in Figure 2. Although all tiling methods succeed at predicting the boundaries for a well-represented class (Snow & Ice, 39.36% of the data), only the model trained on Flip-n-Slide tiling reasonably estimates the underrepresented class (Lichen & Moss, 4.97%). Flip-n-Slide achieves this without altering the underlying class distribution while providing the model with more contextual views of all classes. In CNNs, classes can be distinguished by their channel spectrum or surface texture, but spatial context is also important for separating between classes (Wang & Zhu, 2023). This is especially true in scientific use-cases where the morphometry of the scene is dictated by physical processes and can aid in class separation for rare classes. Our results confirm that the Flip-n-Slide method sufficiently generalizes a model fit to include a fuller understanding of the training distribution, especially for underrepresented classes.

5 Conclusions

In this paper, we addressed a problem with earlier tiling augmentation strategies—namely that their overlap strategy caused data redundancies, which ultimately reduce a model's ability to generalize effectively. However, tile overlapping is necessary to preserve spatial context, which is important in segmentation tasks for Earth observation use-cases. To solve this, we argued that tile overlap when combined with distinct permutations of the data may not only eliminate this redundancy but also expose the model to an expanded training dataset with more spatio-contextual views for OoI. To maintain the fidelity of this context, we emphasized the necessity of realistic augmentation choices for use in scientific segmentation tasks. We introduced a new preprocessing strategy, Flip-n-Slide, built for tiling large Earth observation images. Flip-n-Slide maintains physically realistic transformations of the input data and does not degrade the spatio-contextual information available for small OoI in the overall training set. We show that Flip-n-Slide outperforms the previous benchmarked approach at scale for tiled augmentations in all evaluation metrics, especially in cases of class imbalance, improving the detection of rare phenomena even in large imagery.

Acknowledgements The majority of this work was conducted on the unceded territories of the xučyun, the ancestral land of the Chochenyo speaking Muwekma Ohlone people. We have benefited, and continue to benefit, from the use of this land. We recognize the importance of taking actions in support of American Indian and Indigenous peoples who are living, flourishing members of our communities today. Testing and development of this paper was done on the CryoCloud cloud-based JupyterHub (Snow et al., 2023) that is funded by the NASA Transform to Open Science Program and ICESat-2 Science Team (grant numbers 80NSSC23K0002 and 80NSSC22K1877), and on the Jupyter Meets the Earth (JMTE) cloud hub, an NSF EarthCube funded project (grant numbers 1928406 and 1928374). EA gratefully acknowledges support from a Two Sigma PhD Fellowship. We also gratefully acknowledge funding support from the NASA Cryospheric Science Program (grant number 80NSSC22K0385).

REFERENCES

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 966–970, October 2022. doi: 10.1109/ICIP46576.2022.9897990.

Kahina Amara, Ali Aouf, Hoceine Kennouche, A. Oualid Djekoune, Nadia Zenati, Oussama Kerdjidj, and Farid Ferguene. COVIR: A virtual rendering of a novel NN architecture O-Net for COVID-19 Ct-scan automatic lung lesions segmentation. *Comput Graph*, 104:11–23, May 2022. ISSN 0097-8493. doi: 10.1016/j.cag.2022.03.003.

Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollar. TensorMask: A Foundation for Dense Object Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2061–2069, October 2019. doi: 10.1109/ICCV.2019.00215.

Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Strategies From Data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, June 2019. doi: 10.1109/CVPR.2019.00020.

Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc Le. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, 2020.

Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, and Nasser M. Nasrabadi. SuperMix: Supervising the Mixing Data Augmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13789–13798, June 2021. doi: 10.1109/CVPR46437.2021.01358.

Earth Resources Observation And Science (EROS) Center. Collection-2 Landsat 8-9 OLI (Operational Land Imager) and TIRS (Thermal Infrared Sensor) Level-1 Data Products, 2013.

Arthita Ghosh, Max Ehrlich, Sohil Shah, Larry Davis, and Rama Chellappa. Stacked U-Nets for Ground Material Segmentation in Remote Sensing Imagery. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 252–2524, June 2018. doi: 10.1109/CVPRW.2018.00047.

Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R. Scott, and Serge Belongie. The iMaterialist Fashion Attribute Dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3113–3116, October 2019. doi: 10.1109/ICCVW.2019.00377.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically. *arXiv*, arXiv:1712.00409, December 2017. doi: 10.48550/arXiv.1712.00409.

Minui Hong, Jinwoo Choi, and Gunhee Kim. StyleMix: Separating Content and Style for Enhanced Data Augmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14857–14865, June 2021. doi: 10.1109/CVPR46437.2021.01462.

Bohao Huang, Daniel Reichman, Leslie M. Collins, Kyle Bradbury, and Jordan M. Malof. Tiling and Stitching Segmentation Output for Remote Sensing: Basic Challenges and Recommendations. *arXiv*, arXiv:1805.12219, February 2019. doi: 10.48550/arXiv.1805.12219.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv*, arXiv:1412.6980, January 2017. doi: 10.48550/arXiv.1412.6980.

Rasim Latifovic. 2020 Land Cover of Canada - Open Government Portal. https://open.canada.ca/data/en/dataset/ee1580ab-a23d-4f86-a09b-79763677eb47, 2020.

Lujun Li and Anggeng Li. A2-Aug: Adaptive Automated Data Augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2267–2274, June 2023. doi: 10.1109/CVPRW59228.2023.00221.

Matheus Barros Pereira and Jefersson Alex dos Santos. ChessMix: Spatial Context Data Augmentation for Remote Sensing Semantic Segmentation. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 278–285, October 2021. doi: 10.1109/SIBGRAPI54419.2021.00045.

Hans Pinckaers and Geert Litjens. Training convolutional neural networks with megapixel images. *arXiv*, arXiv:1804.05712, April 2018. doi: 10.48550/arXiv.1804.05712.

Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. SelfAugment: Automatic Augmentation Policies for Self-Supervised Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2673–2682, June 2021. doi: 10.1109/CVPR46437.2021.00270.

G. Anthony Reina, Ravi Panchumarthy, Siddhesh Pravin Thakur, Alexei Bastidas, and Spyridon Bakas. Systematic Evaluation of Image Tiling Adverse Effects on Deep Learning Semantic Segmentation. *Front Neurosci*, 14:65, February 2020. ISSN 1662-4548. doi: 10.3389/fnins.2020.00065.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

D. P. Roy, M. A. Wulder, T. R. Loveland, Woodcock C.e., R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy, T. A. Scambos, C. B. Schaaf, J. R. Schott, Y. Sheng, E. F. Vermote, A. S. Belward, R. Bindschadler, W. B. Cohen, F. Gao, J. D. Hipple, P. Hostert, J. Huntington, C. O. Justice, A. Kilic, V. Kovalskyy, Z. P. Lee, L. Lymburner, J. G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R. H. Wynne, and Z. Zhu. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145: 154–172, April 2014. ISSN 0034-4257. doi: 10.1016/j.rse.2014.02.001.

Hoo-Chang Shin, Matthew Orton, David J. Collins, Simon Doran, and Martin O. Leach. Autoencoder in Time-Series Analysis for Unsupervised Tissues Characterisation in a Large Unlabelled Medical Image Dataset. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 1, pp. 259–264, December 2011. doi: 10.1109/ICMLA.2011.38.

Tasha Snow, Joanna Millstein, Jessica Scheick, Wilson Sauthoff, Wei Ji Leong, James Colliander, Fernando Pérez, James Munroe, Denis Felikson, Tyler Sutterley, and Matthew Siegfried. CryoCloud JupyterBook. Zenodo, January 2023.

Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer International Publishing, Cham, 2022. ISBN 978-3-030-34371-2 978-3-030-34372-9. doi: 10.1007/978-3-030-34372-9.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10778–10787, June 2020. doi: 10.1109/CVPR42600.2020.01079.

F. Özge Ünel, Burak O. Özkalayci, and Cevahir Çiğla. The Power of Tiling for Small Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 582–591, June 2019. doi: 10.1109/CVPRW.2019.00084.

David A Van Dyk and Xiao-Li Meng. The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, March 2001. ISSN 1061-8600, 1537-2715. doi: 10.1198/10618600152418584.

Xuan Wang and Zhigang Zhu. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding*, 229:103646, March 2023. ISSN 1077-3142. doi: 10.1016/j.cviu.2023.103646.

Zhengyuan Yang, Yixuan Zhang, Jerry Yu, Junjie Cai, and Jiebo Luo. End-to-end Multi-Modal Multi-Task Vehicle Control for Self-Driving Cars with Visual Perceptions. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2289–2294, August 2018. doi: 10.1109/ICPR.2018.8546189.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, October 2019. doi: 10.1109/ICCV.2019.00612.

Guodong Zeng and Guoyan Zheng. Holistic decomposition convolution for effective semantic segmentation of medical volume images. *Medical Image Analysis*, 57:149–164, October 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.07.003.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. *CoRR*, abs/1710.09412, 2017.

A Appendix

A.1 The Flip-n-Slide Algorithm

Algorithm 1 The Flip-n-Slide Algorithm.

Inputs: I: Large Image $(x \times y)$
S: Set of sliding tile windows
$(i \times j) \in (x \times y)$
F: Set of corresponding transformations
$(i \times j) \in (x \times y)$
Output: T: Set of augmented image tiles, $t \in T$

for each $(s_{i\times j} \in S) \bigcap_{i\times j} (f_{i\times j} \in F)$ **do**:
    $T_t = s_{ii\times j}(I)$
    $T_t = f_{i\times j}(s_{i\times j})$
end for
return T

A.2 Computational Cost Analysis

In contrast to the previously accepted convention of overlap tiling, Flip-n-slide simultaneously augments and tiles the input image at the preprocessing stage, introducing a larger computational cost at this phase of the machine learning pipeline. The timing differences between the three tiling approaches used in this paper are detailed in Table 2. To find the run time, we run the 10240×10240 pixel image that we used for our segmentation task seven times through each of the three tiling strategies and take the average of these runs. We find that in both overlap approaches the 256×256 tiles are most efficient.

Although the Flip-n-Slide approach incurs a higher computational cost upfront, it streamlines downstream processes and enhances model performance during inference, particularly for underrepresented classes (Figure 2). By integrating data augmentation at the tiling stage of the pipeline, Flip-n-Slide also eliminates the need for additional augmentation at train time, thereby reducing computational costs at the training stage. Furthermore, many approaches that employ the previous overlap convention at training time recommend using the same overlapping tiling approach for prediction averaging at each pixel during test time Ünel et al. (2019); Zeng & Zheng (2019); Reina et al. (2020); Akyon et al. (2022). However, this necessitates tracking pixel locations from tile to tile and complicates the process of reconstructing test tiles back into the overall input image, incurring further computational overhead at the inference stage of the pipeline. Therefore, despite the increased computational cost associated with the Flip-n-Slide strategy during preprocessing, its performance improvements for underrepresented classes justify the upstream computational investment. Depending on the choices that a user makes throughout the rest of the pipeline, both at train and test time, these initial costs may result in reduced overhead later on.

Tiling Runtime for 10240×10240 Input Image				
Method	64×64 Tiles	128×128 Tiles	256×256 Tiles	512×512 Tiles
No Overlap	4.9s	4.3s	4.2s	4.0s
50% Overlap	16.4s	14.9s	14.9s	16.3s
Flip-n-Slide	71s	63s	59.3s	61s

Table 2: Computational cost for simultaneously tiling and augmenting a large input image using the Flip-n-Slide approach. Although Flip-n-Slide has an increased upstream cost, its performance improvements for underrepresented classes justify the upstream computational investment and these initial costs may result in reduced overhead later on.

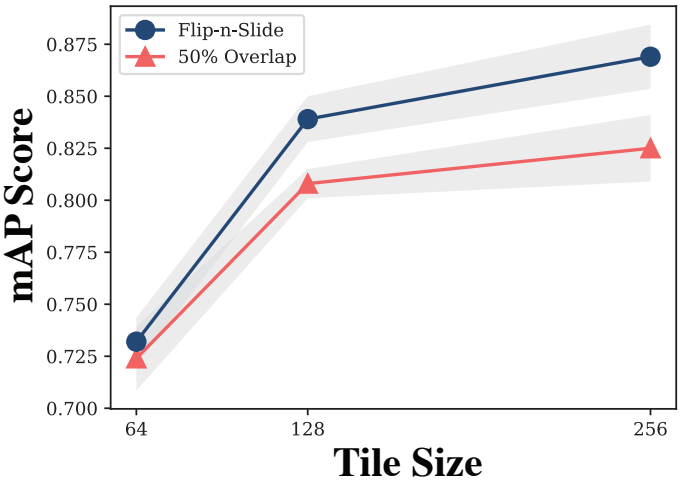

Figure 3: Ablation studies on varying tile size confirm that our strategy leads to expected model behavior. Here we show mAP for each study, averaged over three test runs with the standard deviation shown in grey. Flip-n-Slide outperforms the conventional 50% overlap strategy even at smaller input tile sizes.

A.3 Ablation Studies

Ablation studies are necessary to disambiguate between the role of various fixed decisions on the model outcomes. We explore the impacts of varying tile size choice on the model performance. Previous studies show that larger tiles are more effective for segmentation tasks under traditional augmentation strategies Zeng & Zheng (2019); Reina et al. (2020). We test the performance of the Flip-n-Slide strategy across a range of two additional tile sizes: 64×64, 128×128. Due to GPU size limitations we could not test tiles of 512×512 at the same batch size, so we do not include it here. We reproduce the results of previous studies, finding that larger tile size does lead to better performance. However, we find that even at smaller tile sizes, our approach outperform previous strategies even when they are implemented at larger sizes. When 128×128 tile sizes are generated with Flip-n-Slide they perform at a better score to 256×256 tiles generate by a 50% overlap strategy (Figure 3).

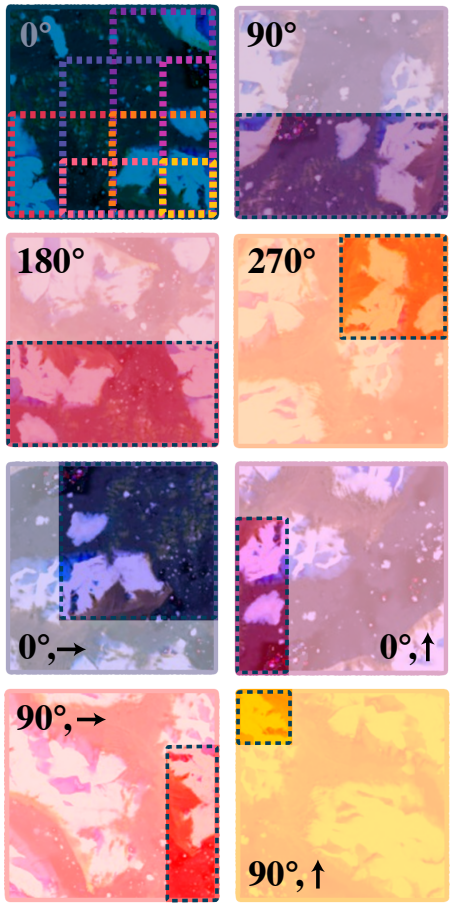

Figure 4: To minimize redundancy in the Flip-n-Slide strategy, each overlapping tile is uniquely permuted with a distinct, physically-realistic transformation, as shown here. Previous strategies have not employed overlap-specific transformations; any augmentations have been applied across all tiles or at random. Tiles are shown in false color to illustrate overlapping areas. Transparency indicates areas that do not overlap with the blue tile shown here. They overlap with neighboring blue tiles instead.