Grid-Shift: An Image Preprocessing Approach to Reduce Overfitting in AI Training

Kevin Kocon
Fraunhofer IGD
TU Darmstadt
Darmstadt, Germany
kevin.kocon@igd.fraunhofer.de

Michel Krämer
Fraunhofer IGD
TU Darmstadt
Darmstadt, Germany
michel.kraemer@igd.fraunhofer.de

Lina Emilie Budde
Fraunhofer IGD
TU Darmstadt
Darmstadt, Germany
lina.emilie.budde@igd.fraunhofer.de

***Abstract*—We present Grid-Shift, a lightweight image preprocessing approach to counteract overfitting when training Convolutional Neural Networks. Grid-Shift solves the problem that tiling large images for training disrupts coherent features (i.e. an object may be split at the edge of a sub-image) and thus leads to information loss. Existing augmentation methods that reduce overfitting do not solve this problem explicitly. In our case study of Land Use and Land Cover Classification, Grid-Shift outperforms all other approaches tested (a raw UNet, a UNet with Batch Normalization, and various augmentation methods). Grid-Shift achieves a Categorical Accuracy of 95%, which is almost 20% better than a raw UNet and still 4% better than the best augmentation approach tested.**

***Index Terms*—Convolutional Neural Networks, Computer Vision, Data Augmentation, Image Processing, Remote Sensing Data**

I. Introduction

Artificial Intelligence (AI) technology has become increasingly popular in recent years. A McKinsey Global Survey on AI shows that in 2022, 50% of the surveyed companies integrate AI technologies in at least one application area. In 2017, it was just 20% [1].

AI can be found in many areas of our daily lives. Important applications include the analysis of medical image data [2], environmental recognition for self-driving cars [3], or the analysis of aerial images for agriculture [4]. Deep Neural Networks (Deep NNs) are particularly popular, as they are very powerful and make it possible to recognise well-trained facts automatically and reliably.

Despite its popularity, a major problem with supervised learning is the availability of training data [5]. Firstly, ground truth data is usually rare, but a large amount of it is needed to obtain good results. At the same time, manual labelling is time-consuming and expensive. Further, images typically have to be converted to a resolution specified by the used network (and hardware), which is usually done by tiling the images into smaller *sub-images*. This causes the problem that objects at the border are sometimes split, which means information is lost.

For this reason, our paper deals with the question of how to best use the limited data available for training. The goal is to reduce overfitting. Our focus is on Earth Observation data and the investigated use case is Land Use and Land Cover Classification (LULC), although our approach can be applied to all computer vision tasks where tiling of the original images is required.

In this paper, we describe and evaluate an image preprocessing approach, which we named *Grid-Shift*. The approach is intended to counteract the loss of information at the border of training samples. Grid-Shift is a lightweight scheme, in which adjacent samples are combined to create a new training sample. Although we have seen the approach of overlapping tiles for training AI applications in several papers ([6], [7], [8], [9], [10]), none of them explicitly focus on it. To the best of our knowledge, a standardized terminology for this method and a detailed evaluation are missing in literature. In the same manner, the Grid-Shift approach is never explicitly described as an augmentation strategy. We close this gap with our work.

This paper is structured as follows: We first discuss relevant work aiming to avoid overfitting. We then explain the basic idea of Grid-Shift and describe the implementation as well as the evaluation setup. Subsequently, we compare the accuracy of Grid-Shift to existing approaches. The paper finishes with a discussion and conclusion.

II. Related Work

A common problem in training neural networks is the small amount of labelled data [5]. Given the good generalization ability of Deep NNs, lack of data can lead to overfitting. As summarized by Hao et al. [11], there are currently two overarching approaches to counteract this problem. One is a network-based approach adapting the model architecture. The other is a data-based approach through augmenting images.

A. Model Architecture Adaptation

One way to avoid overfitting is to utilize network-based approaches. Here, the network structure is stabilized in order to better deal with noise and non-linearities. The best known methods are Dropout [12] and Batch Normalization [13]. Dropout randomly removes neurons—including all inputs and outputs. This way, the network learns more robust features, which prevents overfitting. Batch Normalization normalizes the layer inputs. This allows for better handling of non-linearities. According to Ioffe and Szegedy, this often eliminates the need for Dropout [13].

With the goal of counteracting the loss of information at the border of image subsamples, several works have focused on exploiting neighbourhood information [14], [15]. Fu et al. [15] extended the DeepLabV3+ architecture [16] with an extra context attention module that combines atrous convolutions and a spatial attention module based on a non-local block [17] on all eight neighbouring samples to produce a feature map. Niloy et al. [14] used a spatial attention mechanism that is not based on a non-local block. Their proposed mechanism captures dependencies between neighbouring samples through a series of convolutions and combinations with the image that is currently segmented. Both works achieved minor improvements in their respective metrics.

B. Data Augmentation

Another common way to avoid overfitting is to apply data augmentation. In recent years, there has been a lot of work on the topic of image data augmentation [18], [19], [11]. According to Hao et al. [11], data based augmentation methods can be separated into One-Sample Transformation, Multi-Sample Synthesis, Deep Generative Models, and Virtual Sample Generation. One-Sample Transformation has been used extensively so far [20]. The reason for its popularity is its simple nature, as it only consists of geometric transformations, sharpness transformations, noise disturbances, and random erase methods. Often, multiple One-Sample Transformations are used in conjunction [21], [22], [23]. Additionally, Wang et al. have shown that multiple transformations achieve the best network accuracy for object detection tasks (e.g. the “dog vs. cat” classification) [24]. Taylor and Nitschke have compared different One-Sample Transformations and noticed that cropping gives the best improvement [25]. Since cropping uses only a subsample of the (sub-)sample, further information is lost and the training image resolution changes. Therefore, we focus on a modified approach that we have named Grid-Shift, which is actually more comparable to Multi-Sample Synthesis methods.

Multi-Sample Synthesis methods generate new data by combining multiple samples. Known algorithms are Mixup, [26] where samples are interpolated to obtain a new sample, and BC [27], where a neural network is trained to output a mixture ratio for a given mixture of samples with different labels. In addition, in CutMix [28] new samples are created by removing a part of the image and replacing it with a part from another image. These algorithms belong to linear stacking methods.

Furthermore, there are so-called non-linear blending methods in many variations [11]. Here, a coefficient λ influences how samples are combined. In the case of Vertical or Horizontal Concat, λ is a ratio determining how much of a given sample is used when creating a new sample while $1 - \lambda$ determines how much of the second sample is used. Takahashi et al. have already shown that non-linear blending methods are superior to linear stacking methods [29]. Nevertheless, Hao et al. remark that non-linear blending methods “lack interpretability” [11]. Grid-Shift uses neighbouring samples, which alleviates this problem by creating realistic and interpretable images. At the same time, the lost information at the border of image subsamples is recovered.

This general idea of Grid-Shift is already known. However, to the best of our knowledge, the approach has not been given a name before and has not been evaluated in detail. Reina et al. [10] address the tiling of images for deep semantic segmentation. They compare the accuracy of different tile sizes. They also mention that overlapping tiles are commonly used, but do not elaborate on this. Huang et al. [9], on the other hand, deal with the merging of tiles towards the overall image. Here, as well, the overlapping tiles are named as a tiling strategy, but it is also not discussed in detail.

Cira et al. [8] compare the influence of different tile sizes and overlapping tiles in training. They perform a comparison of the accuracy of training without overlapping tiles and with an overlap of 12.5% and find that overlapping tiles provide better results. But an exact comparison of different overlaps and classical augmentation methods is not available.

Other similar works are by An et al. [6] and Bullinger et al. [7]. Both papers deal with the complete pipeline of tiling and merging. The lack of standardized terminology becomes clear in the paper of Bullinger et al. [7], as they do not explicitly speak of an overlap in Δ-percent, but of a ”stride size x-times < the tile size”. They show that this generates a larger amount of training data, resulting in better accuracy. Nevertheless, a detailed comparison (e.g. different overlap proportions, different number of new tiles, detailed comparison to classical augmentation) on this approach is also missing. An et al. [6] compare training with 50% overlapping tiles with no overlapping tiles. They show, that the results with overlap are better, but comparisons of different overlaps and classic augmentation methods are also missing here.

Summarizing, this shows the potential of Grid-Shift, but also the need for standardized terminology and detailed evaluation. In addition, the Grid-Shift approach is not explicitly described as an augmentation strategy anywhere in literature. We aim to close this gap with our paper.

III. Material and Methods

Grid-Shift is a data-based approach, which generates new samples by combining neighbouring samples. The approach as well as the test environment were implemented with Keras [30] and the TensorFlow [31] backend. We used the classic UNet [32] as the underlying CNN, since it achieved the best results in our previous experiments when segmenting Earth Observation data [33].

A. General Idea

Grid-Shift is motivated by the fact that large images are often tiled into smaller samples for training to match the image resolution required by the used network. Since labelled training data is rare, it is particularly important to maximally

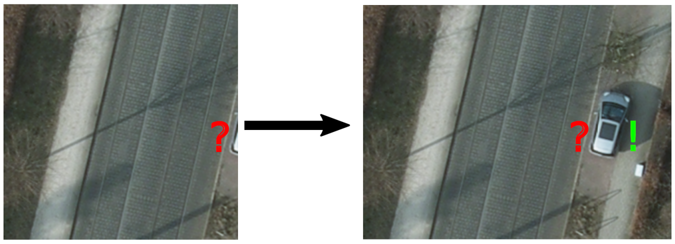


Fig. 1: The general motivation for Grid-Shift

exploit the information from the available data. However, image tiling can lead to information loss. This can be seen in Figure 1, where a car has been split into two parts at the border of the image sample. In the image on the left, it is not visible what kind of object it is. Therefore, it is very difficult for the network to extract meaningful features. If only a part of the neighbouring image is added, as it has been done in the image on the right, it becomes obvious that the object is a car. Exactly this is the basic idea behind Grid-Shift. By shifting the grid cells, information is restored to the edges of the image that was previously lost during tilling.

There are existing non-linear blending methods, where partial images are put together to a new one. Nevertheless, these are not explicitly spatially neighbouring images but arbitrary ones, which can lead to random results. For this reason, Hao et al. also criticizes them [11]. Grid-Shift, on the other hand, only uses spatially neighbouring images.

B. Grid-Shift

The name *Grid-Shift* explains the functionality of our approach very well. Tiling the original large image into smaller ones creates a regular grid of image samples. Grid-Shift generates additional samples by using further regular grids that are shifted in relation to the initial grid. The grid can be moved by

$$\frac{i}{g} \cdot \lambda_x \quad \text{and/or} \quad \frac{i}{g} \cdot \lambda_y$$
$$g \in \mathbb{N} \quad \text{and} \quad i \in 1, ..., g-1$$

where λ_x and λ_y correspond to the cell size of the grid (i.e. the size of a sample) in the x and y direction, $g-1$ is the number of shifts that are performed per tile and i is the id of the current shift. Thus, a new sample always consists of a fraction

$$\frac{i}{g} \cdot \lambda_{x,y} \quad \text{and} \quad (1-\frac{i}{g}) \cdot \lambda_{x,y}$$

in each x and y direction of two adjacent initial ones. This is illustrated in Figure 2. In Section IV, we compare $g \in 1, 2, 3, 4$.

Existing non-linear blending methods can lead to unrealistic and uninterpretable merged images [11]. By moving the whole regular grid, on the other hand, the new subsamples are still parts of the actual image. Furthermore, the samples generated with this method regain the information that was lost during tiling.

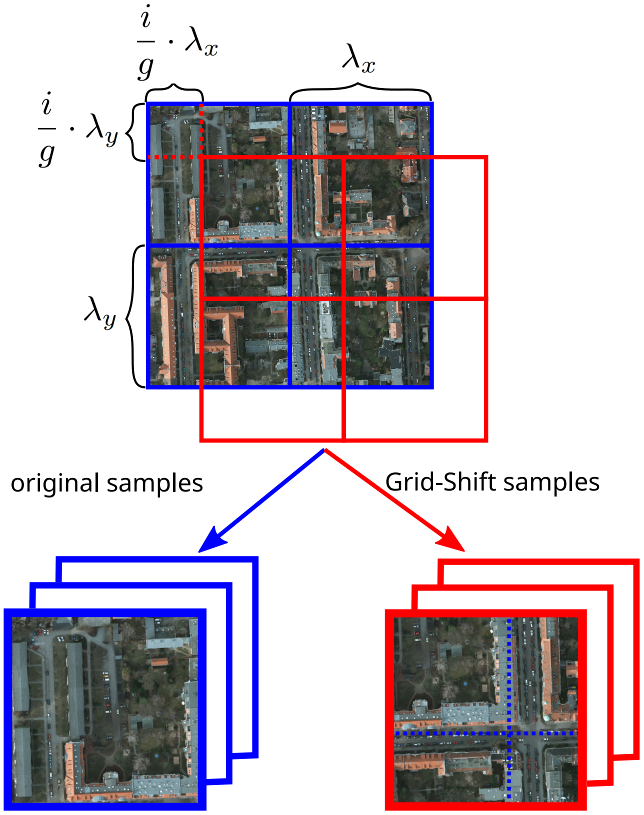


Fig. 2: A visual representation of Grid-Shift: The generation of new samples (red) from the initial samples (blue).

Also, the major drawback of classic cropping is eliminated. Instead of a further loss of information by cropping the border areas of the image out of the visible field, lost information is added to the database for training. However, it has to be noted that this approach is limited to spatially contiguous samples, i.e. when samples are generated from a large image, as is often the case when using remote sensing data.

C. Evaluation Setup

Besides a raw UNet, we also compare our approach to Batch Normalization and data augmentation, currently the most popular ways to counteract overfitting. For the augmentation, we use One-Sample Transformation methods. Here, we generate up to 15 new samples per sample in the different test cases. We randomly applied a random number of 1-3 of the following geometric transformations.

- Horizontal and Vertical Flip
- Rotation
- X and Y Scale
- Crop

Each configuration of the study was trained and evaluated with the ISPRS Potsdam dataset [34]. This dataset can be used freely for research purposes. It consists of orthophotos with R, G, B, and IR channels, as well as matching labelled ground truth images. The sampling distance of one pixel is 5 centimetres.

We created samples with a resolution of 224×224 pixels. This resulted in 25 688 samples. Additionally, the colour values of all image samples were normalized. The ratio between training and evaluation data was $\frac{2}{3}$:$\frac{1}{3}$. We classified the following six classes:

- impervious surfaces
- building
- low vegetation
- tree
- car
- clutter/background

As evaluation metrics, we used Categorical Accuracy as well as one hot mean Intersection over Union (IoU).

To avoid distortions of the results due to unfavourable hyperparameters, we performed hyperparameter tuning for each configuration. This applies to Grid-Shift as well as all other state-of-the-art methods we compare to. For this, we used KerasTuner with the Hyperband algorithm [35]. The search space was defined as follows:

- Learning rate: 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001
- Loss: Categorical CrossEntropy, Sigmoid Focal CrossEntropy, Jaccard Distance
- Start filter size: 8, 16
- Optimizer: Adam, Gradient Descent

IV. Comparison of Results

In this section, we compare the quality of Grid-Shift with that of the current state of the art. The results are shown in Table I. It can be seen that a UNet with Batch Normalization (BN) gives better results than a raw UNet (UNet). In addition, data-based approaches achieve significantly better results than Batch Normalization.

First, we look at the results of the augmentation aug_l, where l denotes how much larger the data set is compared to the original data set. The best accuracy of 0.91 and an IoU of 0.79 was achieved with aug_{16}, i.e. with 15 additional augmented images per sample.

The results of Grid-Shift are labelled $gs_{g,d}$, where g denotes by how much the grid is shifted (see Figure 2). d indicates whether only horizontal (h), vertical (v), or both (a) directions were shifted. Thus, $gs_{4,a}$ means that the grid was shifted three times by $\frac{1}{4}$ in each direction (i.e. horizontally and vertically). This setting produced the best results with a Categorical Accuracy of 0.95 and an IoU of 0.87.

Compared to a raw UNet, the IoU is almost 0.2 better. When comparing to classical augmentation methods, it is additionally noticeable that Grid-Shift already provides very good results with significantly less generated samples. If we compare $gs_{2,a}$ (four times the number of original samples) with aug_{16} (sixteen times the number), we see the accuracies are around 0.9, but Grid-Shift needs 4 times less data and correspondingly less time for training to achieve this result. With the same number of data, it is also visible that Grid-Shift always performs better than augmentation.

TABLE I: Evaluation results. The acronyms mean the following: UNet = raw UNet; BN = UNet with Batch Normalization; aug_l = UNet trained with augmented dataset, where the new dataset is l times larger; $gs_{g,d}$ = UNet trained with Grid-Shift dataset, where g denotes how much ($\frac{1}{g}$) the grid was moved and how often ($g-1$), and d denotes the direction (v = vertical, h = horizontal, a = vertical and horizontal). The bold entry highlights the best result, namely Grid-Shift with $g = 4$, moved in vertical and horizontal direction.

Modell	Acc.	IoU
$UNet$	0.84	0.68
BN	0.85	0.69
aug_2	0.85	0.69
aug_3	0.86	0.70
aug_4	0.86	0.70
aug_9	0.89	0.75
aug_{16}	0.91	0.79
$gs_{2,a}$	0.90	0.79
$gs_{2,v}$	0.87	0.72
$gs_{2,h}$	0.87	0.72
$gs_{3,a}$	0.92	0.83
$gs_{3,v}$	0.89	0.75
$gs_{3,h}$	0.89	0.75
$gs_{4,a}$	**0.95**	**0.87**
$gs_{4,v}$	0.91	0.80
$gs_{4,h}$	0.91	0.80

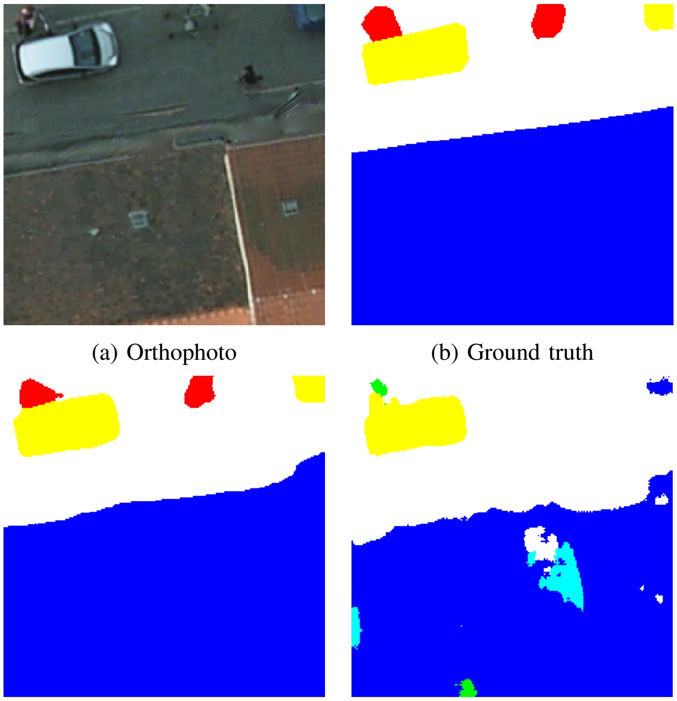
Visual examples of the predictions can be seen in Figures 3 and 4. Subfigures (a) and (b) show tiled orthophotos and the corresponding ground truth images respectively. In (c), the UNet trained with Grid-Shift ($gs_{4,a}$) was applied, while (d) shows the UNet with augmented data (aug_{16}). In both examples, it can be seen that Grid-Shift leads to improved results. The edges of individual classes are detected much more sharply. Also, the confidence in class detection is higher, and the classification produces less false positives, as is the case in (d).

V. Discussion and Conclusion

With this paper, we were addressing the problem of overfitting during the training of CNNs. For this purpose, we introduced a data-based approach, which we named Grid-Shift. It is an alternative way to the classical augmentation. Here, for large images, the samples used to train the CNN are not simply generated by tiling the original image in a regular grid but by additional shifting (as seen in Figure 2). This approach is designed to recover information lost during image tiling.

We compared Grid-Shift to state-of-the-art approaches aiming to counteract overfitting: image augmentation as well as a network with and without Batch Normalization. The results clearly show that Grid-Shift outperforms all other approaches. A Categorical Accuracy of 0.95 was achieved for a Land Use and Land Cover Classification of the ISPRS Potsdam dataset [34]. According to “Papers with Code” [36], the best accuracy ever achieved on this dataset has been 0.94 so far.

However, we must also add two limitations of Grid-Shift at this point. The approach can only be used for spatially contiguous images or for large images that have to be tiled into samples. But if the spatial information is given, there are



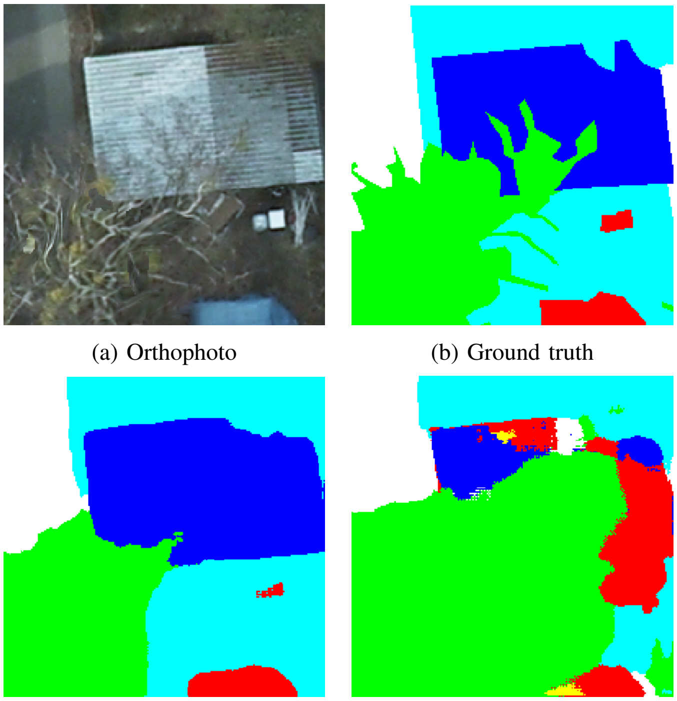
(a) Orthophoto

(b) Ground truth

(c) Predicted image (Grid-Shift $gs_{4,a}$)

(d) Predicted image (Data Augmentation aug_{16})

Fig. 3: First visual comparison of Grid-Shift with ground-truth data and data augmentation



(a) Orthophoto

(b) Ground truth

(c) Predicted image (Grid-Shift $gs_{4,a}$)

(d) Predicted image (Data Augmentation aug_{16})

Fig. 4: Second visual comparison of Grid-Shift with ground-truth data and data augmentation

no restrictions to the used machine learning approach or use cases.

Due to the required spatial relationship between the images, there is also a limitation in pre-processing. Grid-Shift must be applied to all data before training. That means, in addition to the samples created by tiling the original image, all samples created by Grid-Shift must also be stored. If only the original samples should be saved, a spatial index is necessary to be able to find spatially neighbouring samples. Nevertheless, since Grid-Shift and augmentation methods are mainly used if you do not have large amounts of data in the first place, and since memory is not a limiting factor any longer today, this is not a significant limitation for us. Additionally, we have shown in Section IV that Grid-Shift can work with less data than other state-of-the-art methods while maintaining or even improving quality.

Acknowledgment

This work has been partially funded by the European Union (GA 101137787).

References

[1] McKinsey & Company, "McKinsey Global Survey on AI," (last accessed: 2023-07-19). [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

[2] R. Altman, "Artificial intelligence (AI) systems for interpreting complex medical datasets," *Clinical Pharmacology & Therapeutics*, vol. 101, no. 5, pp. 585–586, 2017.

[3] A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," *Array*, vol. 10, p. 100057, 2021.

[4] A. Dakir, F. Barramou, and O. B. Alami, "Opportunities for artificial intelligence in precision agriculture using satellite remote sensing," *Geospatial Intelligence: Applications and Future Trends*, pp. 107–117, 2022.

[5] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE, 2018, pp. 117–122.

[6] Y. An, Q. Ye, J. Guo, and R. Dong, "Overlap training to mitigate inconsistencies caused by image tiling in cnns," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 2020, pp. 35–48.

[7] S. Bullinger, F. Fevers, C. Bodensteiner, and M. Arens, "Geo-tiles for semantic segmentation of earth observation imagery," *arXiv preprint arXiv:2306.00823*, 2023.

[8] C.-I. Cira, M.-Á. Manso-Callejo, N. Yokoya, T. Sălăgean, and A.-C. Badea, "Impact of tile size and tile overlap on the prediction performance of convolutional neural networks trained for road classification," *Remote Sensing*, vol. 16, no. 15, p. 2818, 2024.

[9] B. Huang, D. Reichman, L. M. Collins, K. Bradbury, and J. M. Malof, "Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations," *arXiv preprint arXiv:1805.12219*, 2018.

[10] G. A. Reina, R. Panchumarthy, S. P. Thakur, A. Bastidas, and S. Bakas, "Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation," *Frontiers in neuroscience*, vol. 14, p. 65, 2020.

[11] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, and X. Li, "A review of data augmentation methods of remote sensing image target recognition," *Remote Sensing*, vol. 15, no. 3, p. 827, 2023.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[14] F. F. Niloy, M. A. Amin, A. A. Ali, and A. M. Rahman, "Attention toward neighbors: A context aware framework for high resolution image segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2279–2283.

[15] W. Fu, Q. Peng, Y. Gong, M. Xie, S. Wang, and F. Li, "Semantic segmentation of high resolution remote sensing images with extra context attention mechanism," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*. IEEE, 2020, pp. 1372–1376.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[19] K. Maharana, S. Mondal, and B. Nemade, "A review: Data preprocessing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[20] D. Ma, P. Tang, L. Zhao, and Z. Zhang, "21. review of data augmentation for image in deep learning," *Journal of Image and Graphics*, vol. 26, no. 03, pp. 0487–0502.

[21] Y. Yan, Z. Tan, and N. Su, "A data augmentation strategy based on simulated samples for ship detection in rgb remote sensing images," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 276, 2019.

[22] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang, "Sar target detection based on ssd with data augmentation and transfer learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 150–154, 2018.

[23] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 4, pp. 549–553, 2017.

[24] J. Wang, L. Perez *et al.*, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. 2017, pp. 1–8, 2017.

[25] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2018, pp. 1542–1547.

[26] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk management," in *Proc. 6th Int. Conf. Learn. Represent.(ICLR)*, 2018, pp. 1–13.

[27] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5486–5494.

[28] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[29] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for deep CNNs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2917–2931, 2019.

[30] "Keras," (last accessed: 2023-07-21). [Online]. Available: https://keras.io/

[31] "Tensorflow," (last accessed: 2023-07-21). [Online]. Available: https://www.tensorflow.org/

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[33] Anonymized, "Anonymized title," *Anonymized*, vol. 3, p. 42, 2022.

[34] The International Society for Photogrammetry and Remote Sensing, "2D semantic labeling contest - Potsdam," (last accessed: 2023-07-20). [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[35] "Hyperband tuner," (last accessed: 2023-08-29). [Online]. Available: https://keras.io/api/keras_tuner/tuners/hyperband/

[36] Papers with Code, "Semantic segmentation on ISPRS Potsdam," (last accessed: 2023-08-29). [Online]. Available: https://paperswithcode.com/sota/semantic-segmentation-on-isprs-potsdam