EFFECT OF LABEL NOISE IN SEMANTIC SEGMENTATION OF HIGH RESOLUTION AERIAL IMAGES AND HEIGHT DATA

A. Maiti[1,*], S.J. Oude Elberink[1], G. Vosselman[1]

[1] Department of Earth Observation Science, Faculty ITC, University of Twente, The Netherlands – (a.maiti, s.j.oudeelberink, george.vosselman)@utwente.nl

Commission II, WG II/6

KEY WORDS: Deep Learning, Semantic Segmentation, Label Noise, Very High Resolution

ABSTRACT:

The performance of deep learning models in semantic segmentation is dependent on the availability of a large amount of labeled data. However, the influence of label noise, in the form of incorrect annotations, on the performance is significant and mostly ignored. This is a big concern in remote sensing applications, wherein acquired datasets are spatially limited, labeling is done by domain experts with possible sources of high inter-and intra-observer variability leading to erroneous predictions. In this paper, we first simulate the label noise while conducting experiments on two different datasets with very high-resolution aerial images, height data, and inaccurate labels, responsible for the training of deep learning models. We then focus on the effect of these noises on the model performance. Different classes respond differently to the label noise. The typical size of an object belonging to a class is a crucial factor regarding the class-specific performance of the model trained with erroneous labels. Errors caused by relative shifts of labels are the most influential label errors. The model is generally more tolerant of the random label noise than other label errors. It has been observed that the accuracy gets reduced by at least 3% while 5% of label pixels are erroneous. In this regard, our study provides a new perspective of evaluating and quantifying the propagation of label noise in the model performance that is indeed important for adopting reliable semantic segmentation practices.

1. INTRODUCTION

Deep learning algorithms outperform traditional algorithms in semantic segmentation (Zhang et al., 2020a). However, they rely on the need for good and a large amount of training data. The quality of training labels is crucial for urban scene segmentation due to the specific characteristics of the classes such as object size, higher inter-class correlation, higher intra-class variability, etc. Thus, label errors act as incorrect examples that affect the model's learning process to correctly recognize the objects of different classes present in the scene. Despite the significant progress in the state-of-the-art, the impact of label noise on semantic segmentation did not get much attention. In particular, the performance of deep learning models influenced by the label noise during the training process has shown enormous differences compared to those without it (Jiang et al., 2020). This is especially concerning for remote sensing applications, wherein geo-information is limited, and labeling needs to be done by a domain expert. The latter is potentially biased due to high inter-/intra-observer variability and erroneous predictions. In the case of unsupervised image classification, it is not easy to obtain satisfactory results attributed to the unavailability of prior information such as labeled data (Laban et al., 2020). In comparison, supervised image classification approaches are generally considered to be more reliable (Congalton, 1991).

The current research trend in semantic segmentation primarily focuses on supervised deep learning-based feature learning. These approaches demonstrably work well and reliably maintain very high accuracy (Diakogiannis et al., 2020). There are various approaches like image augmentation to increase the volume of data synthetically. Furthermore, most deep learning methods are, in general, computationally costly. For these reasons, these models are typically developed in a very specific and expensive infrastructure (Diakogiannis et al., 2020). Although they are capable of providing very high accuracy, the extremely high computational cost does not always justify the relatively small improvement in the validation or test metrics (Huang et al., 2017; Hazırbaş et al., 2015; Li et al., 2021).

In addition to this, modern supervised semantic segmentation methods are sensitive to the accuracy of the training data. Inaccuracies in the data hamper the learning process and negatively affect the convergence and performance of the model. The availability of highly accurate controlled training data is minimal, limiting the applicability of these models to domains other than research. Some progress has been made in the field of computer vision to address this challenge where the self-supervised or semi-supervised model recursively refines the coarse or inaccurate training labels (Xu et al., 2015). These models have the potential advantage in terms of amount of data required for training and they are they are often designed to be more error tolerant (Hendrycks et al., 2019; Wang et al., 2021). However, standalone performances of these models are generally inferior compared to well trained supervised models. Devising improved methods semantic segmentation robust to noisy data is an active area of research. Although, the modern deep learning based models are excellent at solving certain tasks like segmentation, it is extremely hard to scientifically explain their behaviors with respect to changes in the inputs, such as how a semantic segmentation model behaves with changes in the quality of training labels. This kind of challenges important for XAI and need more attention (Gohel et al., 2021).

In this regard, our prime focus is to assess the impact of various types of labeling errors that commonly occur in geo-information

* Corresponding author

(due to inaccurate training labels) on semantic segmentation performance. In addition, we also analyze the extent of individual classes affected by label noise. One of the common approaches for generating labeled data is to use existing maps. However, maps are also prone to different kinds of errors such as co-registration error, human error, etc. Unlike the noise in the data, these types of label noises significantly impact the semantic segmentation process. In this study, we simulated 3 types of significant labeling errors, segment errors, relative shift of segment boundaries, and random labeling errors. We first conducted experiments with very high spatial resolution aerial images and height data alongside available geo-information, containing inaccurate labels for the training of deep learning models. We also investigated how varying amounts of these noises influence the model performance. Furthermore, quantitative and qualitative assessments have been presented, observing the results.

2. RELATED WORKS

2.1 Semantic Segmentation

The traditional methods of image classification are considered to be inadequate for the classification of very high-resolution imagery as they fail to capture complex feature space, and they generally do not consider local context very well (Richards, 2013). Furthermore, the recent emergence of deep learning in the field of computer vision has revolutionized the task of image segmentation and classification while addressing the drawbacks of conventional approaches. However, adapting the deep learning-based methods from the field of computer vision to fit the requirements of earth observation is not always trivial (Audebert et al., 2018). In the past few years, much has been achieved regarding the adaptation of deep learning-based methods into the field of earth observation, and those methods demonstrated to be considerably better than the traditional methods (Ma et al., 2019).

CNN and Semantic Segmentation: CNN based architectures are very efficient at capturing both the generalized context and local context when grouping the pixels for classifications compared to the pixel-based classification approaches. They also provide an end-to-end solution with minimal manual feature engineering and comparatively more tolerant of the noise present in the dataset (Girshick, 2015). The FCNs typically have an encoder-decoder architecture popularized by UNet proposed by Ronneberger et al. (2015). The encoder part of the architecture captures higher semantic features by successive convolution and pooling operations, whereas the decoder part gradually transforms the semantic features learned by the encoder into corresponding label features and recovers spatially. There are different CNN based approaches to further exploit the contextual information such as multi-scale inputs (Farabet et al., 2013), spatial pyramid pooling (Lazebnik et al., 2006) atrous spatial pooling (Papandreou et al., 2015) and dilated convolution (Yu and Koltun, 2016). While more complex networks with increasing numbers of parameters boosted the performance, progress has been made to optimize the networks without affecting their high accuracy using different convolution approaches such as depthwise convolution and separable convolution (Bello et al., 2021).

Current State-of-the-Art: The Inception architecture designed by Szegedy et al. (2015) has successfully shown that the decoupling cross channel correlation and spatial correlation can lead to better model performance with reduced computational cost. The Xception architecture by Chollet (2017) demonstrated further improvement and optimization of the Inception-based design. Furthermore, He et al. (2015) developed ResNet architecture and demonstrated the power of residual learning in semantic segmentation and image classification. However, despite advancements in network design, post-processing is often necessary to further refine the results generated by the models (Teichmann and Cipolla, 2018). Chen et al. (2017) was able to improve the state of the art performance of semantic segmentation using multi-scale atrous spatial pooling and the integration of CRF into their proposed network Deeplab. This network has been further improved into DeepLabV3+, adopting depthwise separable convolution, spatial pyramid pooling, and using earlier networks such as ResNet and Xception as its backbone (Chen et al., 2018). DeeplabV3+ achieved state-of-the-art performance without heavily increasing the baseline complexity compared to the existing state-of-the-art networks.

2.2 Model Complexity and Label Noise:

Although these models emerged from the field of computer vision and image recognition, they have been quickly adopted for photogrammetric applications (Song and Kim, 2020; Zhang et al., 2020b; Yuan et al., 2021). However, the complexity of the network and the requirement of massive, accurately labeled high-quality training data often pose challenges. While some studies focus on mitigating the problem of noise in the data (Klingner et al., 2020), the others are trying to reduce the impact of label noise in semantic segmentation (Patrini et al., 2017).

Label generation process in the domain of geo-information is prone to different types of errors affecting positional accuracy, attribute accuracy, consistency, and completeness (Oluseyi, 2002). Effects of these errors can be commonly observed in many of the geo-information data, such as thematic maps in the form of misaligned boundaries, relative shift, incorrect attribute, incorrect boundaries, etc. (Vargas-Munoz et al., 2021). In this study, we mainly focus on three types of errors:

Segment Error: Segment corruption generally occurs if the labels are auto-generated, typically from a thematic segmentation algorithm (Vargas-Munoz et al., 2021). An erroneous segment typically does not cover the entire underlying feature, and, as a result, a part of such feature is mislabeled.

Relative Shift of Object Boundaries: Relative shift of boundaries typically occurs due to co-registration error between image and corresponding label. This anomaly is very similar to boundary misalignment, which can also occur due to human error and bias (Vargas-Munoz et al., 2021).

Random Label Noise: Although this type of noise is relatively uncommon compared to other noise types, varying amounts of it can still be observed in the labels generated from pixel-wise classification methods (Su, 2016).

Although it is well known that any kind of noise in the data negatively impacts the model performance, very little research has studied the impact of label errors from a quantitative and analytical perspective. In our study, we focus on bridging this knowledge gap.

3. EXPERIMENTAL SETUP

3.1 Dataset Description

In this study we have used two ISPRS datasets, Potsdam dataset (ISPRS, 2016a) and Vaihingen dataset (ISPRS, 2016b). Both the dataset contains very high resolution airborne multi-spectral (RGB + NIR) images and DSM generated from LiDAR point clouds. The images have an ortho-rectified top view, and dense urban features typically dominate scenes. In the case of Potsdam, the semantic labels are available as images where each class is represented with a unique value. There are 24 tiles of 6000×6000 images with corresponding labels. We divided the dataset into 2 parts, 18 tiles have been randomly selected for training, and the rest of the 6 tiles have been selected randomly for validation. The test dataset contains 14 tiles of similar size for evaluating the model's performance.

In contrast, the Vaihingen dataset contains 16 tiles of different sizes for training and validation. Due to size variations of the tiles, 75% of the patches generated from all the tiles has been used for training, and the rest of the 25% patches are reserved for validation. The test set of the Vaihingen dataset consists of 17 tiles.

3.2 Simulation of Label Errors

We simulate the type and amount of errors in the training data to analyze the impact of specific errors. Our baseline is based on the original training labels from the dataset, which is assumed to be error-free. This allows us to compare our baseline to the current state-of-the-art and further analyze the impact of simulated label errors in the model performance. The errors are simulated as follows:

A. Random Label Noise:

1. Select random point from the label data.
2. Alter the label of that point to any of the other possible label.
3. Repeat step 1 to step 2 until expected noise threshold has been reached.

B. Segment Error:

1. Randomly pick a segment from the labeled data
2. Randomly pick a contiguous region inside the selected segment with size proportional to the noise threshold
3. Assign incorrect label to the selected region of the segment
4. Repeat step 1 to step 3 until expected noise threshold has been reached.

C. Relative Shift of Object Boundaries:

1. Spatially shift the labels in random direction with respect to the corresponding image.
2. Crop the image and the label restricting them to the intersecting area.

We set up experiments with the two noise thresholds for the three types of errors. In case of segment error and random label noise, we set the noise thresholds to be 5% and 10% of total pixels in the image. The thresholds for the shifts are 5 pixels and 10 pixels.

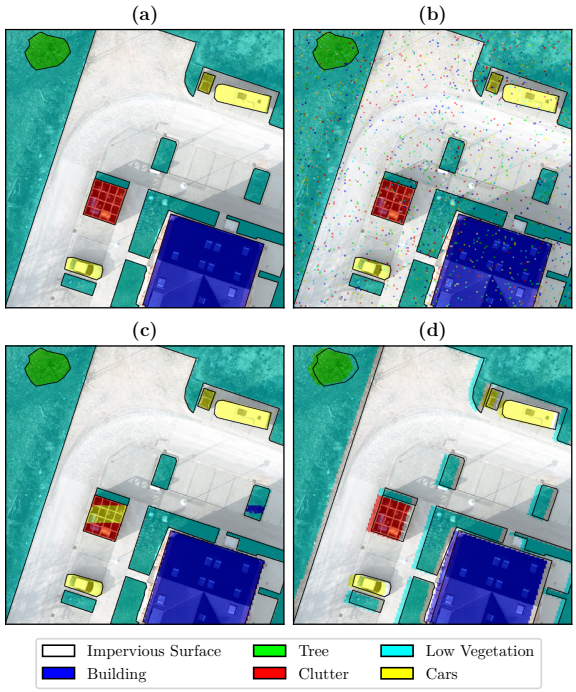

Figure 1. Labels before and after introducing different label noise. **(a)** original Label, **(b)** Label corrupted with random label (pixel) noise, **(c)** Label corrupted with segment error, **(d)** Label corrupted with segment shift

3.3 Prerequisites for the Experiments

Assumptions: For our experiments, we assume that original labels are ideal, i.e. these labels are perfectly delineated and absolutely error free. This assumption is necessary to be able to compare our observations with the original data. However, in reality, factors like the bias of the label experts, discrepancies due to data pre-processing such as ortho-rectification introduce additional errors in data as well as in labels. This inherent error can be crucial for the model while classifying difficult to recognize classes. It should also be noted that 2D labels and data representing a 3D scene is inherently prone to some fuzziness at the segment boundaries due to occlusion and highly complex true shape of the object. The tree class is particularly susceptible to this problem.

Although the types of label noises chosen to be simulated in this experiment are realistic, the simulation is performed at random, i.e. all segments have equal probability of getting attenuated. In reality however some classes can be relatively more susceptible to certain types of label noises compared to other classes.

Data Preprocessing: The original Potsdam dataset contains very high-resolution airborne imagery of 0.05m spatial resolution. In our experiments, we resampled the 6000×6000 image tiles to 1500×1500 dimension with 0.2m spatial resolution. The resolution of the imagery from Vaihingen dataset is 0.08m. These images also have been resampled to 0.2m resolution so that the combined dataset have uniform spatial resolution. This resolution is sufficient for our objects of interest, and it significantly reduces the computational time required for training. The images have been downscaled using nearest neighbor resampling to avoid interpolation of the DN values. Similarly, all the labels have also been downscaled using nearest neighbor resampling to preserve the integrity of the labels.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-275-2022 | © Author(s) 2022. CC BY 4.0 License.

Semantic Segmentation Model: For this study, we have used DeeplabV3+ (Chen et al., 2018) model to perform the experiments. The DeeplabV3+ model embraces the typical encoder-decoder structure for the network and uses depth-wise separable convolution to atrous spatial pyramid pooling. This helps the network identify the object boundaries more precisely while boosting comparatively faster computation. As part of the performance, the network demonstrates better results than the ResNet for semantic segmentation with relatively fewer parameters (Chen et al., 2018). The performance of DeeplabV3+ is considerably on a higher side when synergized with the Xception (Chollet, 2017) as its backbone (Chen et al., 2018). With a spatial pyramid pooling module, it is possible to encode multi-scale contextual information from input features, whereas an encoder-decoder network helps retrieve sharper object boundaries with precision. Xception, in this regard, leverages the advantages of both paradigms.

It is worth noting that some of the newer models, such as HRNet (Wang et al., 2020) may perform better than DeeplabV3+ in some cases. However, these models consist of many parameters in contrast to the performance gain. Moreover, over-parameterized complex models are susceptible to memorization instead of learning generalization (Zhang et al., 2021). However, the systematic and analytic design of deep learning networks is still in its infancy (Liang et al., 2019), thus choosing the suitable model depending upon the application still involves a few trade-offs. Hence, we have selected the popular DeeplabV3+ model with Xception as the backbone for this study which adequately fits our requirements.

Type of Error	Acc. (%)	F1	K
Rnd. Err. (5%)	95.00	0.8958	0.9322
Rnd. Err. (10%)	90.00	0.8178	0.8656
Seg. Err. (5%)	95.00	0.8909	0.9321
Seg. Err. (10%)	90.00	0.8251	0.8651
Rel. Shift (1m)	92.25	0.8734	0.8940
Rel. Shift (2m)	82.99	0.7403	0.7718

Table 1. Evaluation of simulated label errors in the combined training data (training set from Potsdam dataset + training set from Vaihingen dataset). The corresponding accuracy, Cohen's Kappa (K) coefficients and F1 scores are shown in the respective columns.

Preferred Hyper-parameters: Apart from the network architecture, model performance is also sensitive to the chosen hyperparameters (van Rijn and Hutter, 2018). The state-of-the-art machine learning algorithms require the manual selection of hyperparameters prior to the learning process. However, such an evaluation approach is time taking and leads to more complexity in terms of the dimensionality of the search space (Hinz et al., 2018). Sharma et al. (2019) identified some of the key hyperparameters for effective image segmentation. Some of these notable hyperparameters are learning rate, batch size, weight decay rate, number of epochs, momentum, etc. For achieving a reliable performance, it should be noted that even the least sensitive hyperparameters should be associated with an adequate value, as compared to the traditional frameworks (Sharma et al., 2019). We pre-select a few hyperparameters for our experiments based on initial experiments and recommendations from previous studies.

Training Strategy: We use 256×256 patch size for the experiments, which fits into our memory budget. Since CNN based semantic segmentation models are prone to boundary effect (Islam et al., 2021), we have used patches with approximately 32 pixels overlap to reduce that effect. In addition to this, we have used basic image augmentations techniques flip, reflection, and rotation to synthetically increase the size of the dataset. We also use gradient accumulation with a mini-batch size of 3 and a global batch size of 16 to mitigate the memory limitations. Finally, we have utilized an early stopping mechanism with the patience of 10 epoch to prevent overfitting. It is worth noting that the model in each experiment has been trained from scratch on their respective datasets.

4. RESULTS AND DISCUSSION

In this section, we present the results of our experiments with suitable metrics and respective analyses followed by qualitative assessment and interpretation.

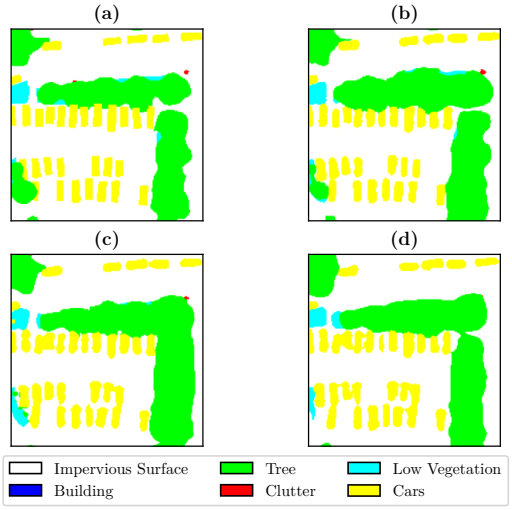

Figure 2. Label inference on the objects of the unseen test tile 6_10. **(a)** Ground truth, **(b)** Inference by the base line model, **(c)** Inference by the model trained on 5% segment errors, **(d)** Inference by the model trained on 1m relative shift.

4.1 Simulated Label Errors

Evaluation of each type of simulated label errors on Potsdam dataset are shown in Table 1. The accuracy of corresponding label error simulation in Table 1 shows that simulation of errors strictly followed the pixel-based error threshold in case of segment corruption and random label noise. The relative shift introduced relatively higher label errors compared to the others. Similar trends can be observed for both F1 and K scores and in the label errors simulated on the Vaihingen dataset.

4.2 Quantitative Analysis

First, we evaluate the performance of the baseline model trained on the training label without any simulated noise. We also compared our baseline performances in Table 3 and Table 4 with their expected performance on the corresponding dataset from the previous studies. Although our models are trained with down-sampled images, it performs comparably to the previous benchmarks of Deeplabv3+ on the ISPRS and Vaihingen dataset, presented in (Song and Kim, 2020, Table 3) and (Wang et al., 2018, Table 5) respectively.

We set the performance of the model trained on error-free labels as our baseline and compare the performance of the models trained on erroneous labels. In Table 3 and Table 4 it can be

Type of Error	Im. Surface		Building		Low Veg.		Tree		Car	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Rnd. Err. (5%)	84.97	0.8555	91.94	0.8965	75.71	0.7548	67.08	0.6658	59.84	0.6503
Rnd. Err. (10%)	82.80	0.8096	88.08	0.8666	74.54	0.7127	60.32	0.6395	37.96	0.4863
Seg. Err. (5%)	85.49	0.8491	91.17	0.8847	74.42	0.7372	68.61	0.6500	58.65	0.6465
Seg. Err. (10%)	81.65	0.8079	90.44	0.8619	68.33	0.6974	62.68	0.6315	40.69	0.4737
Rel. Shift (1m)	83.92	0.8364	94.52	0.9372	74.45	0.7462	70.45	0.7093	68.51	0.6797
Rel. Shift (2m)	79.76	0.7994	92.44	0.9163	71.95	0.7129	68.36	0.6849	54.33	0.5178
Baseline	89.50	0.8966	96.72	0.9678	85.51	0.8261	77.10	0.7824	89.69	0.8640

Table 2. Comparison of evaluation metrics (accuracy and F1 score for the test dataset) for the models trained with different type and amount of label errors introduced to Potsdam dataset. The last row of the table highlights the metrics for baseline model.

observed that the impact of relative shift is the highest among all 3 different label errors. The variation in error threshold also influences the performance. In all of the experiments, a higher label error resulted in lower accuracy. However, accuracy is most sensitive to variation relative shift and least sensitive to variation of random label noise. The F1 score shows a similar trend as the accuracy.

In the Table 2 the class-specific results have been shown for the Potsdam dataset. Overall, the models perform reasonably well for the building class among all the noises. However, among all the models, the model trained on labels having 10% random pixel noise performs the poorest in terms of accuracy of the building detection. The accuracy for this class is reduced in the range of approximately 2% to 8%. In the case of impervious surface, label error induced by 2m relative shift causes the poorest accuracy of that class. The accuracy hit for this class is in the range of approximately 4% to 10% for different label errors. The models appear to be very stable for identifying low vegetation class except for higher segment error or higher relative shift of the training labels. The performance hit for the low vegetation class is in the range of 10% to 17% in terms of accuracy. Regarding tree class, the models are fairly stable against label errors induced by relative shift; however, they perform comparatively poorly against segment errors and random label noise. The performance hit for this class is approximately in the range of 9% to 17% in terms of accuracy.

Type of Error	Acc. (%)	F1	K	MCC
Rnd. Err. (5%)	85.08	0.7462	0.7874	0.7881
Rnd. Err. (10%)	85.03	0.7406	0.7864	0.7874
Seg. Err. (5%)	84.54	0.7372	0.7801	0.7808
Seg. Err. (10%)	84.01	0.7268	0.7713	0.7728
Rel. Shift (1m)	83.05	0.7006	0.7590	0.7594
Rel. Shift (2m)	82.29	0.6982	0.7490	0.7492
Baseline	87.97	0.7968	0.8287	0.8291

Table 3. Evaluation metrics for Potsdam test set.

The models' performances concerning car class are most severely affected by the higher random label noises and higher segment corruption. Although the models comparatively perform better to identify the cars while training labels are corrupted with low label errors of any type, the overall performance penalty for this class is as high as approximately 21% to 52%. The clutter class appears to be the most challenging class to identify by the models, irrespective of the amount or type of noise. The performance penalty for this class is moderate at 8% to 24%. Variation of model performances in terms of F1 score with respect to variations in the error threshold in the labels are primarily consistent among all three error types. In general, label errors with lower thresholds caused a less detrimental effect on the F1 score. Overall, segment errors are most damaging, and random label errors are least damaging for the F1 score of the corresponding models.

The evaluation of overall performance on the test dataset is shown in Table 3 and Table 4 for Potsdam and Vaihingen dataset respectively. The accuracies and F1 scores reveal that the label errors caused by relative shift are the worst type of error among all the error types. The model appears to be more tolerant of the random label noise than the segment errors. We further verify this using $\boldsymbol{K}$ score and MCC. We observe that for 5% erroneous pixels in the training labels roughly decreases the test accuracy of the model by at least 3%.

4.3 Qualitative Assessment

The class-specific metrics in Table 2 show that, the baseline performance is competitive relative to the sate-of-the-art. The model performs best for the building class. A potential explanation for this is the availability of the height information in the input dataset set. The model also performs really well for the impervious surface class. Although height information does not have much influence, in this case, the model appears to be able to distinguish the mostly linearly shaped geometry and the texture associated with this class. In the case of car class, shape, size, and contextual association with roads (impervious surface) are the potential distinguishable features for the model. Thus the model performs well to identify the cars correctly. The NIR band helps the model to distinguish both tree and low vegetation classes. However, the model performs relatively poorly for the tree class compared to the low vegetation class, likely due to their irregularly shaped boundaries which is difficult for the model to delineate precisely. Evidently, the model performance is poorest for the clutter/background class, which matches the previous studies. This can be attributed to the vast intra-class variability of the clutter class in terms of shape, size, texture, and spatial context. Similar patterns have been observed for the Vaihingen dataset, strengthening the hypotheses stated above.

It has been observed that the tree class and car class are most affected by the label errors, irrespective of their types and amount. The footprints of the individual objects from both of these classes are relatively small compared to the objects of other classes, and this makes it difficult for the models to identify the individual objects from even smaller spatial contexts when additional label errors are introduced. The model is somewhat moderately tolerant to the lower amount of random label noise compared to other types of labeling errors. However, in a higher amount, it can be more damaging than others. A similar trend can be observed for the segment corruption. This can be attributed to the fact that in both the cases, models are exposed to a statistically similar amount of bad training examples in terms of erroneous pixels (see Table 1). The impact label errors caused by relative shift are mostly low compared to others, which complies with the translation equivariance property of the CNNs.

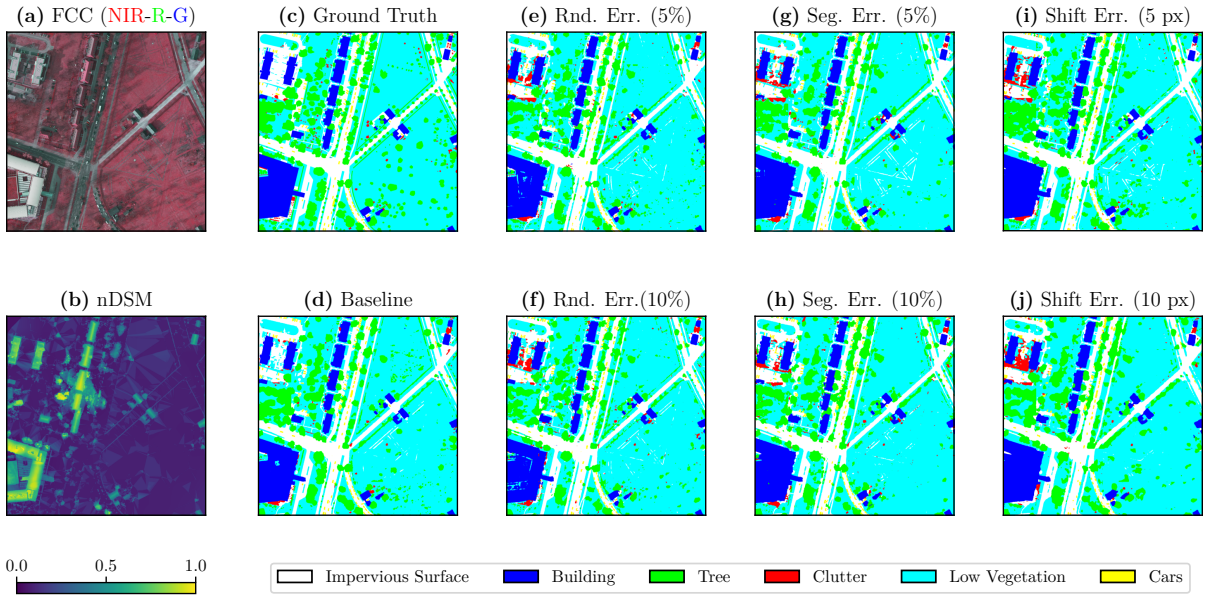

Figure 3. Illustration of predicted labels along with the overview of the scene. (a) False color composite of the scene, (b) Normalized DSM, (c) Ground Trth Label, (d) Baseline prediction, (e) - (j) predicted labels attributed to different noise type and threshold

Furthermore, visual analysis reveals that segment errors and relative shift cause dissolving edges among separate objects or result in less precise object boundaries. One example from the Potsdam dataset is shown in Figure 3. The baseline in Figure 3(b) closely resembles the ground truth in Figure 3(a). Most of the object edges are separate, and they are visually distinguishable. In both Figure 3(c) and Figure 3(d) the quality of object boundaries have been degraded, many objects of car class with distinct boundaries are no longer identifiable separately. However, these anomalies are more prevalent in Figure 3(d).

Type of Error	**Acc. (%)**	**F1**	$\boldsymbol{K}$	**MCC**
Rnd. Err. (5%)	81.33	0.7377	0.7222	0.7643
Rnd. Err. (10%)	80.57	0.7113	0.7195	0.7591
Seg. Err. (5%)	79.24	0.6804	0.7084	0.7433
Seg. Err. (10%)	77.92	0.6735	7050	0.7410
Rel. Shift (1m)	77.41	0.6632	0.6921	0.7089
Rel. Shift (2m)	77.06	0.6395	0.6914	0.6937
Baseline	83.32	0.7730	0.7736	0.7883

Table 4. Evaluation metrics for Vaihingen test set.

Overall, the label errors affect the model's ability to generalize and make it harder to infer the contextual patterns correctly. An example of this can be seen in Figure 4. In Figure 4(c), the small walkways are not delineated separately, although they are visually more similar to impervious surfaces. Ideally, the model should learn this from the contextual association. In the case of baseline inference in Figure 4(d), the model depicts this behavior quite well except for a few occurrences. However, as the generalization abilities of the models are affected by the label noise, the apparent misclassification of these walkways becomes more prevalent in Figure 4(e-g, i). This impact is hardly observable in Figure 4(j). This can be attributed to the inability to precisely detect narrow shapes and boundaries caused by the higher label of relative shift or segment errors.

5. CONCLUSION

In this work, we present a quantitative and qualitative analysis of the impact of various label errors in varying amounts on deep learning-based semantic segmentation using two well-known datasets. The errors we investigated commonly occur in the label preparation in the geo-information domain.

We used DeeplabV3+ as our chosen model for semantic segmentation. The baseline performance of this model is comparable to the state-of-the-art. Additional data such as DSM and NIR bands certainly helps the model to perform better than previous benchmarks such as (Song and Kim, 2020, Table 3). Among all three types of label errors, the errors caused by relative shift affect the model performance the most, followed by segment errors and random label noise. Relative shift and segment error prevent the model from learning the features of precise object boundaries and, to some extent, narrowly shaped small objects such as cars. Each class responds differently for each type and amount of label noise. However, in general, higher label errors in training labels result in lower model performance. The classes like clutter are inherently difficult to recognize for the model; additional label noise makes it even harder.

It is worth noting that, although the outcome of this experiment provides valuable insights into the impact of label noises on the model performance, these observations should be further validated on more datasets. In this experiment, the amount of label noise introduced to the data is relatively low. Thus, the model's behavior in case of severe label noise is yet to be observed. In the future, this experiment can be extended to include more models, more data, and more variations in the type and amount of noise. Furthermore, the changes in the model behaviors can be deeply investigated using saliency maps and gradient integration, aiming for the goals of XAI.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-275-2022 | © Author(s) 2022. CC BY 4.0 License.

ACKNOWLEDGMENT

This research is supported under 'Water4Change' project jointly funded by Department of Science and Technology (DST), Government of India, and the Dutch Research Council (NWO) project n. W 07.7019.103 — DST-1429-WRC.

ACRONYMS

CNN Convolutional Neural Network.

CRF Conditional Random Field.

DN Digital Number.

DSM Digital Surface Model.

FCN Fully Convolutional Network.

LiDAR Light Detection and Ranging.

MCC Matthews Correlation Coefficient.

NIR Near Infrared.

RGB Red - Green - Blue.

XAI Explainable Artificial Intelligence.

REFERENCES

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. 140, 20–32.

Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., Zoph, B., 2021. Revisiting ResNets: Improved Training and Scaling Strategies. http://arxiv.org/abs/2103.07579.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. http://arxiv.org/abs/1606.00915.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. http://arxiv.org/abs/1802.02611.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 1800–1807.

Congalton, R. G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. 37(1), 35–46.

Diakogiannis, F. I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. 162, 94–114.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning Hierarchical Features for Scene Labeling. 35(8), 1915–1929.

Girshick, R., 2015. Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 1440–1448.

Gohel, P., Singh, P., Mohanty, M., 2021. Explainable ai: current status and future directions.

Hazırbaş, C., Diebold, J., Cremers, D., 2015. Optimizing the relevance-redundancy tradeoff for efficient semantic segmentation. J.-F. Aujol, M. Nikolova, N. Papadakis (eds), *Scale Space and Variational Methods in Computer Vision*, 9087, Springer International Publishing, 243–255. Series Title: Lecture Notes in Computer Science.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385.

Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019. Using self-supervised learning can improve model robustness and uncertainty.

Hinz, T., Navarro-Guerrero, N., Magg, S., Wermter, S., 2018. Speeding up the Hyperparameter Optimization of Deep Convolutional Neural Networks. 17(2), 1850008.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K., 2017. Speed/accuracy trade-offs for modern convolutional object detectors. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3296–3297.

Islam, M. A., Kowal, M., Jia, S., Derpanis, K. G., Bruce, N. D. B., 2021. Position, Padding and Predictions: A Deeper Look at Position Information in CNNs. http://arxiv.org/abs/2101.12322.

ISPRS, 2016a. 2D semantic labeling dataset - Potsdam.

ISPRS, 2016b. 2D semantic labeling dataset - Vaihingen.

Jiang, Z., Kirby, M. S., He, W., Sainju, A. M., 2020. Deep Learning for Earth Image Segmentation based on Imperfect Polyline Labels with Annotation Errors. http://arxiv.org/abs/2010.00757.

Klingner, M., Bär, A., Fingscheidt, T., 2020. Improved Noise and Attack Robustness for Semantic Segmentation by Using Multi-Task Training with Self-Supervised Depth Estimation. http://arxiv.org/abs/2004.11072.

Laban, N., Abdellatif, B., Ebied, H. M., Shedeed, H. A., Tolba, M. F., 2020. Multiscale satellite image classification using deep learning approach. A. E. Hassanien, A. Darwish, H. El-Askary (eds), *Machine Learning and Data Mining in Aerospace Technology*, 836, Springer International Publishing, 165–186. Series Title: Studies in Computational Intelligence.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, 2, IEEE, 2169–2178.

Li, Y., Li, X., Xiao, C., Li, H., Zhang, W., 2021. EACNet: Enhanced Asymmetric Convolution for Real-Time Semantic Segmentation. 28, 234–238. https://ieeexplore.ieee.org/document/9325538/.

Liang, T., Poggio, T., Rakhlin, A., Stokes, J., 2019. Fisher-rao metric, geometry, and complexity of neural networks. K. Chaudhuri, M. Sugiyama (eds), *The 22nd International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 89, PMLR, 888–896.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. 152, 166–177.

Oluseyi, F. O., 2002. The effects of gis architecture, data models and data sources on the accuracy of digital maps: an example of digital terrain model of ibadan region. *ISPRS 2002 - 2002 ISPRS Commission IV Symposium*, ISPRS, 6191–6194.

Papandreou, G., Kokkinos, I., Savalle, P.-A., 2015. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 390–399.

Patrini, G., Rozza, A., Menon, A., Nock, R., Qu, L., 2017. Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach. http://arxiv.org/abs/1609.03683.

Richards, J. A., 2013. *Image Classification in Practice*. Springer Berlin Heidelberg, 381–435.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. http://arxiv.org/abs/1505.04597.

Sharma, A., van Rijn, J. N., Hutter, F., Müller, A., 2019. Hyperparameter importance for image classification by residual neural networks. P. Kralj Novak, T. Šmuc, S. Džeroski (eds), *Discovery Science*, 11828, Springer International Publishing, 112–126. Series Title: Lecture Notes in Computer Science.

Song, A., Kim, Y., 2020. Semantic Segmentation of Remote-Sensing Imagery Using Heterogeneous Big Data: International Society for Photogrammetry and Remote Sensing Potsdam and Cityscape Datasets. 9(10), 601.

Su, T.-C., 2016. A filter-based post-processing technique for improving homogeneity of pixel-wise classification data. 49(1), 531–552.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the Inception Architecture for Computer Vision. http://arxiv.org/abs/1512.00567.

Teichmann, M. T. T., Cipolla, R., 2018. Convolutional CRFs for Semantic Segmentation. http://arxiv.org/abs/1805.04777.

van Rijn, J. N., Hutter, F., 2018. Hyperparameter Importance Across Datasets. 2367–2376. http://arxiv.org/abs/1710.04725.

Vargas-Munoz, J. E., Srivastava, S., Tuia, D., Falcao, A. X., 2021. OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing. 9(1), 184–199. https://ieeexplore.ieee.org/document/9119753/.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2020. Deep High-Resolution Representation Learning for Visual Recognition. http://arxiv.org/abs/1908.07919.

Wang, Y., Liang, B., Ding, M., Li, J., 2018. Dense Semantic Labeling with Atrous Spatial Pyramid Pooling and Decoder for High-Resolution Remote Sensing Imagery. *Remote Sensing*, 11(1), 20. https://doi.org/10.3390/rs11010020.

Wang, Z., Li, Y.-L., Guo, Y., Wang, S., 2021. Combating noise: Semi-supervised learning by region uncertainty quantification. A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (eds), *Advances in Neural Information Processing Systems*.

Xu, B., Wang, N., Chen, T., Li, M., 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. http://arxiv.org/abs/1505.00853.

Yu, F., Koltun, V., 2016. Multi-Scale Context Aggregation by Dilated Convolutions. http://arxiv.org/abs/1511.07122.

Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. 169, 114417.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. 64(3), 107–115.

Zhang, X., Han, L., Han, L., Zhu, L., 2020a. How Well Do Deep Learning-Based Methods for Land Cover Classification and Object Detection Perform on High Resolution Remote Sensing Imagery? 12(3), 417.

Zhang, Z., Huang, J., Jiang, T., Sui, B., Pan, X., 2020b. Semantic segmentation of very high-resolution remote sensing image based on multiple band combinations and patchwise scene analysis. 14(1), 1.