A CLOSER LOOK AT SEGMENTATION UNCERTAINTY OF SCANNED HISTORICAL MAPS

Sidi Wu[1]*, Magnus Heitzler[1], Lorenz Hurni[1]

[1]Institute of Cartography and Geoinformation, ETH Zurich, Switzerland - (sidiwu, hmagnus, lhurni)@ethz.ch

Commission IV, WG IV/3

KEY WORDS: Deep Learning, Uncertainty Analysis, Historical Maps, Semantic Segmentation

ABSTRACT:

Before modern earth observation techniques came into being, historical maps are almost the exclusive source to retrieve geo-spatial information on Earth. In recent years, the use of deep learning for historical map processing has gained popularity to replace tedious manual labor. However, neural networks, often referred to as ”black boxes”, usually generate predictions not well calibrated for indicating if the predictions are trustworthy. Considering the diversity in designs and the graphic defects of scanned historical maps, uncertainty estimates can benefit us in deciding when and how to trust the extracted information. In this paper, we compare the effectiveness of different uncertainty indicators for segmenting hydrological features from scanned historical maps. Those uncertainty indicators can be categorized into two major types, namely aleatoric uncertainty (uncertainty in the observations) and epistemic uncertainty (uncertainty in the model). Specifically, we compare their effectiveness in indicating erroneous predictions, detecting noisy and out-of-distribution designs, and refining segmentation in a two-stage architecture.

1. INTRODUCTION

Historical maps serve as useful and almost unique resources to depict geo-spatial phenomena on Earth before modern earth observation techniques came into being. Tremendous scanned historical maps with diverse designs, scales, and different graphic qualities require automatic, generic, and robust methods. In recent years, deep learning methods have been leveraged to segment features from scanned historical maps (Uhl et al., 2017; Uhl et al., 2018, 2020; Heitzler and Hurni, 2020) and have shown promising results. However, neural networks, often referred to as ”black boxes”, usually generate predictions not well calibrated for indicating if and when the predictions are trustworthy. As historical maps inevitably have graphic defects and diversity in designs, uncertainty estimates can benefit researchers in deciding when and how to trust the extracted information. Two major types of uncertainty have been studied in the field of computer vision (Kendall and Gal, 2017), namely aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty captures noise inherent in the data (e.g., map noise from the original painting, aging effect, and the scanning process). This uncertainty is inevitable and cannot be reduced with more training data collected. Epistemic uncertainty, also known as model uncertainty, describes the imperfectness in model parameters, which can be reduced with more training data collected.

For historical map segmentation, Wu et al. (2022) incorporated the concepts of Bayesian deep learning to model uncertainty inherent in historical maps. In this paper, we have a closer look at both aleatoric and epistemic uncertainty through three types of uncertainty indicators, with hydrological features as the focus. Specifically, we test the effectiveness of those types of uncertainty in indicating erroneous predictions, capturing noisy and out-of-distribution designs, and improving segmentation results in a two-stage architecture. The method we propose is agnostic to specific deep learning architectures and the framework can be applied to other uncertainty estimation techniques. Also, despite that historical maps are our focus, our methodology is general enough for other applications.

2. RELATED WORK

With the advances of computer vision, deep-learning-based map processing methods are becoming increasingly popular for segmenting buidings (Uhl et al., 2017; Uhl et al., 2018, 2020; Heitzler and Hurni, 2020) and hydrological features (Wu et al., 2022). The network architectures alter from LeNet (Uhl et al., 2017; Uhl et al., 2018, 2020), to U-Net (Heitzler and Hurni, 2020), and to ASPP-integrated U-Net (Wu et al., 2022) which incorporates multi-scale contexts.

To provide a quality indicator to calibrate the network’s predictions, two types of segmentation uncertainty are modelled under the concept of Bayesian deep learning, namely aleatoric uncertainty (data-dependent) and epistemic uncertainty (model-dependent). Aleatoric uncertainty has been modelled by either a probabilistic distribution (Kendall and Gal, 2017; Cipolla et al., 2018; Gurevich and Stuke, 2018) to calibrate the prediction probability or an interpolation degree between the targeted and the predicted distributions (DeVries and Taylor, 2018a,b; Wu et al., 2022). Epistemic uncertainty has been modelled by either using Monte-Carlo dropout to sample the neural network parameters (Kendall and Gal, 2017; DeVries and Taylor, 2018b) or using an ensemble of neural networks trained independently with random initializations (Lang et al., 2022; Lakshminarayanan et al., 2016). The second option is computationally more efficient and tends to have better performances in practice (Lakshminarayanan et al., 2016). However, since both sampling and assembling models often are time-consuming and resource-intensive, modelling aleatoric uncertainty is more computationally efficient than modelling epistemic uncertainty in general.

In previous works, DeVries and Taylor (2018b) compared different types of uncertainty mentioned above for segmenting skin lesions and investigated their effectiveness in predicting the segmentation quality. Our work compares the effectiveness

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-189-2022

of different uncertainty indicators for segmenting scanned historical maps. We have a closer look at their effectiveness in three aspects both qualitatively and quantitatively. Instead of using uncertainty to predict the segmentation quality, we propose a novel two-stage architecture to use the estimated uncertainty to refine segmentation results.

3. METHOD

The proposed two-stage architecture is depicted in Figure 1. The first stage is a normal semantic segmentation task. The network takes an input image and outputs a segmentation map. We apply the ASPP-integrated U-Net proposed by Wu et al. (2022), which adds an atrous spatial pyramid pooling (ASPP) block (Chen et al., 2017) after the backbone of a normal U-Net (Ronneberger et al., 2015), to segment four hydrological classes - streams, wetlands, rivers and lakes. We use Sigmoid activation at the end to obtain segmentation probability. To produce uncertainty maps, we investigate three methods: prediction entropy, learned confidence estimates and ensemble variance. The first two captures aleatoric uncertainty while the last one indicates epistesmic uncertainty. In the second stage, the output segmentation map, uncertainty map together with the input image from the first stage are fed into a refinement network to generate a refined segmentation map. Both stages share the same network architecture.

3.1 Prediction Entropy

The first uncertainty estimation technique we investigate is the prediction entropy. As our model outputs a soft score (probability) between 0 and 1, we can directly calculate the entropy of the predicted probability. The entropy, as a common uncertainty indicator, can be obtained for free from any classification network without introducing additional parameters. Since our model allows multi-class prediction, we calculate entropy per class instead of averaging it across class dimensions:

$$c_{ij} = -p_{ij} log p_{ij} \tag{1}$$

for the prediction p_{ij} of class j at the pixel i.

3.2 Learned Confidence Estimates

The second technique we investigate is Learned Confidence Estimates (LCE) proposed by DeVries and Taylor (2018a,b), which is an effective method to generate pixel-wise confidence/uncertainty maps for segmentation. The network learns to produce an additional uncertainty map beside the segmentation map. The uncertainty estimation is encouraged as a calibration mask that controls the interpolation degree between the predicted and the targeted prediction:

$$p_{ij}^{'} = (1 - \sigma_{ij}) * p_{ij} + \sigma_{ij} y_{ij} \tag{2}$$

where σ_{ij} is the predicted uncertainty, p_{ij} is the segmentation probability and y_{ij} is the binary ground truth. When the network is highly uncertain about the results ($\sigma_{ij} \rightarrow 1$), it will receive the correct label ($p_{ij}^{'} \rightarrow y_{ij}$) to avoid penalizing the uncertain cases. A log penalty is added to prevent the model from always predicting high uncertainty scores.

3.3 Ensemble Variance

The third technique we investigate is the ensemble variance (Lang et al., 2022; Lakshminarayanan et al., 2016). We train several models independently with random initializations, regarding each model as a randomly-sampled distribution of the ensemble parameters. The variance of the ensemble can represent the uncertainty in model parameters:

$$\begin{aligned} \hat{p} &= \frac{1}{M} \sum_{m=1}^{M} p_m \\ var(\hat{p}) &= \frac{1}{M} \sum_{m=1}^{M} (p_m - \hat{p})^2 \end{aligned} \tag{3}$$

3.4 Effectiveness of Uncertainty Indicators

We test the effectiveness of the uncertainty indicators in the following three aspects:

- We test if the estimated uncertainty can indicate erroneous predictions based on the assumption that a well-calibrated uncertainty indicator should capture untrustworthy predictions. Specifically, we observe the change of accuracy after filtering out/dilating/eroding uncertain pixels. We believe that the remaining relatively more confident pixels will have higher prediction accuracy. For dilation, we treat all uncertain predictions as 1 while for erosion, we reduce all uncertain predictions to 0, to test if the uncertainty indicator is biased towards negative or positive predictions.
- We test if the estimated uncertainty can identify noisy and out-of-distribution inputs, assuming that a functional uncertainty indicator should identify peculiar situations. We compare the difference of uncertainty levels between normal map sheets (used for training) with the noisiest map sheets and different map designs.
- We test if the estimated uncertainty can aid the refinement network to improve the raw predictions. We suppose that the highly structured and meaningful uncertainty indicator can help the model make reasonable decisions regarding when and how to trust the raw predictions.

3.5 Evaluation Metrics

We evaluate prediction accuracy using four common metrics - dice coefficient, F1 score, precision, and recall. For line features, their metrics can be sensitive to small width changes of the thin structures. In our ground truth, we give streams (line vectors) a fixed buffer width on average before rasterizing them to pixel-level annotations. However, as the width of streams alters, a fixed-width can lead to mismatches between predictions and ground truth annotations. Therefore, we skeletonize both predicted and annotated streams into centerlines and calculate the metrics by relaxing the notions in a buffered distance similar to the approaches described by Mosinska et al. (2018) and Wegner et al. (2013) for road extraction. Predicted centerline pixels are regarded as true positives (TP) if they lie in a buffer of the ground-truth centerline and false positives (FP) otherwise. Ground-truth centerline pixels are regarded as false negatives (FN) if they don't lie in a buffer of the predicted centerline. The variant recall $\frac{TP}{TP+FN}$ and precision $\frac{TP}{TP+FP}$ are similar to the completeness and correctness in the works of Mosinska

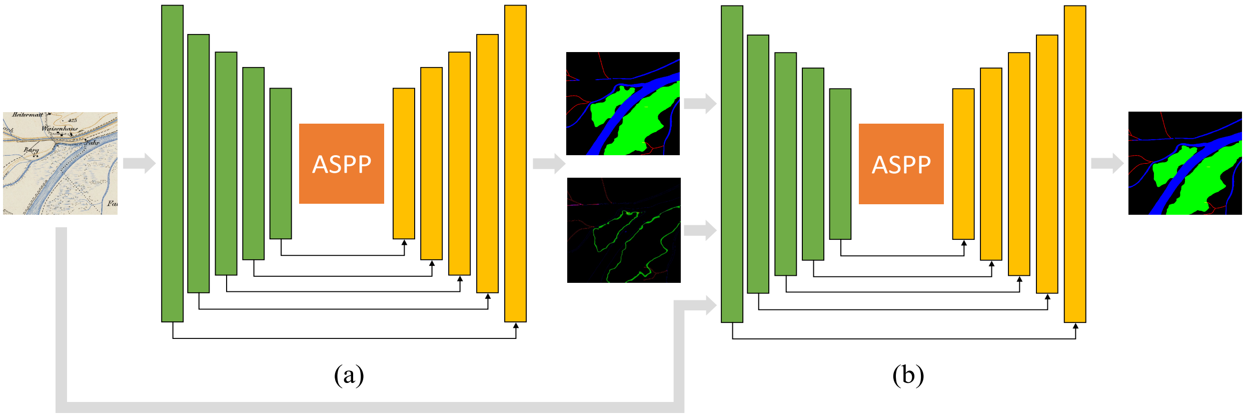


Figure 1. Our proposed two-stage architecture. In stage (a), a segmentation network (an ASPP-integrated U-Net) takes an input image and outputs a segmentation map. An uncertainty map is either output additionally by the network or calculated directly out of the segmentation map. In stage (b), a refinement network takes the segmentation map, uncertainty map and the input image from (a) to generate a refined prediction map. The network architecture stays the same.

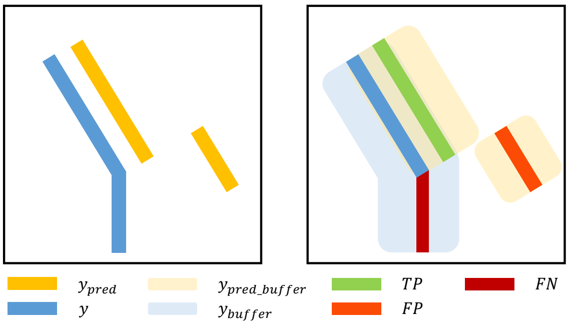


Figure 2. Calculating evaluation metrics for line features. We skeletonize both predictions and ground truth and calculate the metrics in a buffered distance. Predicted centerline pixels y_{pred} are regarded as true positives TP if they lie in a buffer of ground truth y and false positives FP otherwise. Ground-truth centerline pixels are regarded as false negatives FN if they don't lie in a buffer of the predicted centerline y_{pred}.

et al. (2018) and Wegner et al. (2013). Then we obtain the F1 score by $2 * \frac{precision*recall}{precision+recall}$. For the dice coefficient, we count intersections by the number of predicted centerline pixels y_{pred} that lie in a buffer y_{buffer} of the ground-truth centerline y: $2 * \frac{y_{pred} \cap y_{buffer}}{y_{pred}+y}$. We illustrate the process in Figure 2.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Experimental Data and Settings

In our experiment, we use the scanned Topographic Atlas of Switzerland (”Siegfried map”) that was published between 1870 and 1949, and then scanned and stored in a digital archive by the Federal Office of Topography (Swisstopo). Our training data contains 439 map sheets of scale 1:25000 around the year 1880, each of which is 7000 pixels wide and 4800 pixels high. We divide the map sheets into training (80%), validation (10%), and testing (10%) and sample them into small training tiles of 256 × 256 pixels.

For stage (a) in Figure 1, we trained seven models independently for 100 epochs with a patience of 25 epochs for early stopping. We use the Adam optimizer with an initial learning rate of 0.001 and decrease it by 10% after each epoch. For stage (b), we trained the models for only 50 epochs with the same optimizer and learning rate.

4.2 Uncertainty Maps

We visualize different uncertainty maps in Figure 3 for qualitative comparisons. As different uncertainty is yielded in different numerical ranges, for better visualization, we discretize uncertainty by their quantiles. Since the majority of maps does not contain hydrological features and tend to have uncertainty values approximately 0, we clip out those values first before we obtain the 60% (Q1), 80% (Q2) and 90% (Q3) quantiles of uncertainty. We visualize the resulting 10%, 20% and 40% most uncertain pixels. From Figure 3 we can see, all three types of uncertainty are able to indicate uncertainty in normal object boundaries (e.g. streams, rivers). This is within our expectation because the object borders tend to have less discriminative information and possibly aliasing effects from scanning. For features without explicit boundaries (i.e. wetlands), LCE and variance displays a wider band of uncertainty than entropy, including the blank area between two wetlands. We find that variance and LCE are better at identifying erroneous predictions influenced by input noise (e.g. noisy strokes (b), scanning artifacts (c)) than entropy. Variance is relatively more sensitive to input noise than LCE, showing high uncertainty regardless if the segmentation errs. All of three uncertainty indicators can recognize exotic/out-of-distribution inputs (d) that are from a totally different map series. Now we are going to check the difference between their effectiveness quantitatively.

4.3 Uncertainty Indicators for Erroneous Predictions

To quantitatively investigate the capacity of these indicators for erroneous predictions, we filter/dilate/erode the 10%, 20%, and 40% most uncertain pixels by each indicator and observe changes of accuracy. We regard every uncertain pixel as positive for dilation and negative for erosion to gain more insights into whether the uncertainty indicator is biased towards positive or negative predictions. Figure 4 shows an example of erosion and dilation using the ensemble variance.

From Table 1 we can see, after filtering out uncertain pixels, the accuracy metrics improve for all three uncertainty indicators,

		Dice				F1				Precision				Recall			
Operation	Uncertainty	-	10%	20%	40%	-	10%	20%	40%	-	10%	20%	40%	-	10%	20%	40%
Filter	Entropy	0.829	0.899	0.902	0.905	0.871	0.930	0.932	0.935	0.829	0.878	0.881	0.887	0.917	0.988	0.989	0.989
	Variance	0.829	0.892	0.926	0.926	0.871	0.918	0.938	0.938	0.829	0.911	0.952	0.968	0.917	0.924	0.925	0.910
	LCE	0.829	0.883	0.898	0.904	0.871	0.911	0.919	0.918	0.829	0.905	0.927	0.948	0.917	0.917	0.912	0.890
Dilation	Entropy	0.829	0.838	0.838	0.830	0.871	0.877	0.877	0.870	0.829	0.832	0.830	0.822	0.917	0.928	0.929	0.924
	Variance	0.829	0.820	0.804	0.795	0.871	0.864	0.853	0.847	0.829	0.799	0.778	0.764	0.917	0.940	0.945	0.950
	LCE	0.829	0.838	0.838	0.830	0.871	0.877	0.877	0.870	0.829	0.831	0.830	0.822	0.917	0.929	0.929	0.924
Erosion	Entropy	0.829	0.848	0.846	0.845	0.871	0.885	0.884	0.880	0.829	0.878	0.881	0.887	0.917	0.892	0.887	0.872
	Variance	0.829	0.852	0.875	0.846	0.871	0.886	0.897	0.871	0.829	0.911	0.952	0.968	0.917	0.862	0.848	0.792
	LCE	0.829	0.845	0.841	0.810	0.871	0.878	0.868	0.831	0.829	0.905	0.927	0.948	0.917	0.853	0.817	0.740

Table 1. Uncertainty indicators for erroneous predictions. We test the segmentation accuracy after filtering out the most 10%, 20% and 40% pixels and dilating/eroding these uncertain pixels.

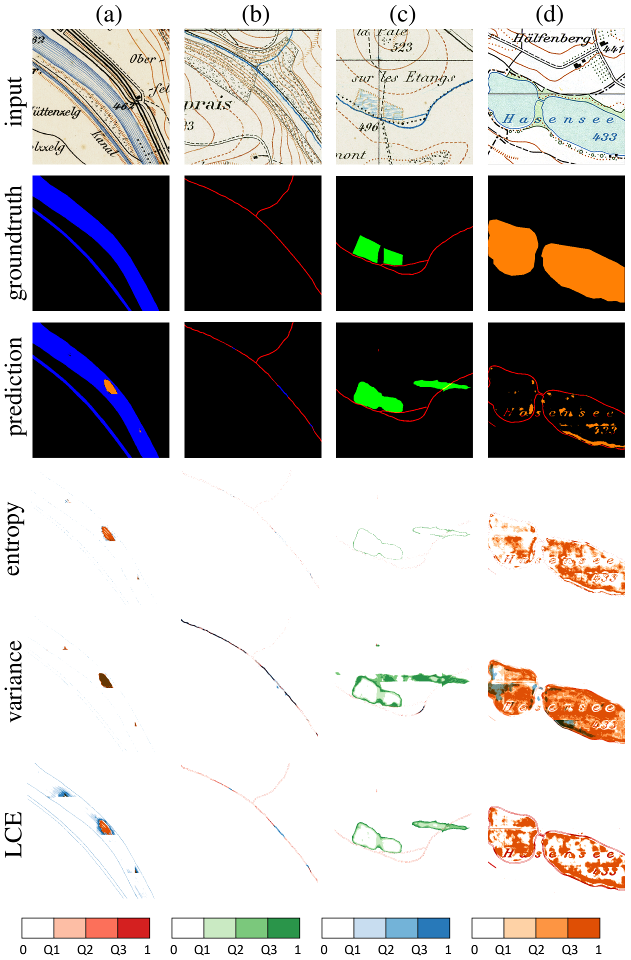


Figure 3. Comparison of different uncertainty indicators. Color bars are presented at the bottom. (a) shows an example where a river is partially misclassified as a lake. (b) and (c) present cases where the prediction accuracy suffers due to noisy strokes (b) for streams and scanning artifacts (c) for wetlands. (d) is an out-of-distribution example where the design varies from the training map sheets.

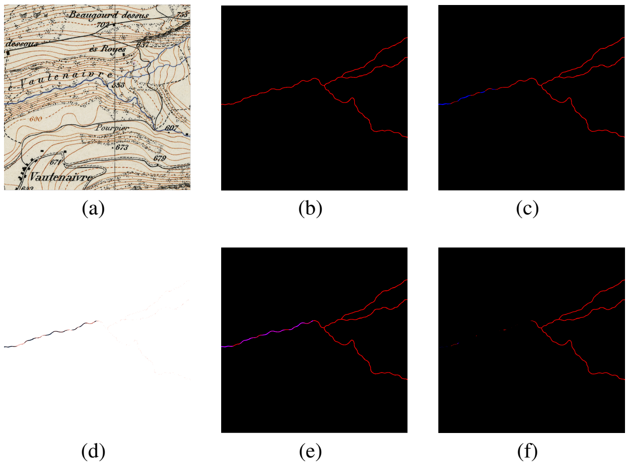


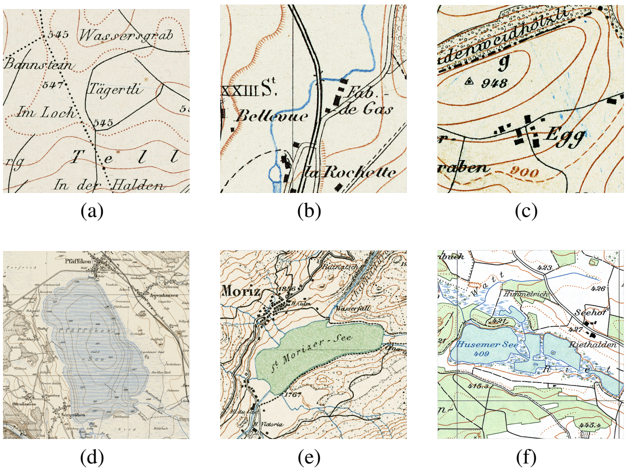
Figure 4. Examples of using the ensemble variance to erode and dilate features. From (a) to (f) are input, ground truth, prediction, variance, dilated prediction and eroded prediction. For each class, we regard uncertain pixels as negative for erosion and positive for dilation.

which implies their general correlation with erroneous predictions. Interestingly, the accuracy has merely a slight increase after filtering more than 20% top uncertain pixels for all these indicators. This implies that erroneous predictions are more likely to accumulate in the 20% most uncertain pixels (even 10% for entropy and LCE).

Accuracy increases after dilating uncertain pixels for entropy and variance while it decreases for variance. This indicates that high variance is likely to correspond more to true negatives than false negatives while the opposite is true for entropy and LCE. The improved accuracy after erosion demonstrates that all three indicators capture more false positives than true positives (except LCE for the top 20% – 40% uncertain pixels). The accuracy metrics alter more after erosion than dilation for all three indicators. This shows that all of them are rather biased towards positive predictions than negative ones.

4.4 Uncertainty Indicators for Noisy and Out-of-Distribution Data

As mentioned before, the graphic flaws and design diversity characterizing scanned historical maps can influence the extraction results. To test the effectiveness of those uncertainty indicators for noisy data, we select ten most noisy map sheets with scanning artifacts 5(a), noisy strokes 5(b) and painting flaws 5(c). For exotic(out-of-distribution) data, we select ten



(a) (b) (c)

(d) (e) (f)

Figure 5. Examples of noisy ad out-of-distribution data. Noisy examples include scanning artifacts (a), noisy strokes (b) and painting errors (c). (d) (e) (f) show examples of different designs for lakes of Siegfried maps (25k), Siegfried maps (50k), and old national maps (25k), respectively.

map sheets from other different map series - five from Siegfried maps of 1:50k and five from old national maps of 1:25k. While Siegfried maps of 1:25k mainly cover the Swiss plateau, Siegfried maps of 1:50k mainly cover the Swiss Alps and their designs can vary from 1:25k. Figure 5(e) shows an example of Siegfried of 1:50k where the lakes are painted green. Starting from the middle 20th century, national maps of Switzerland were implemented and replaced the previous map series (Dufour and Siegfried maps) with a new design schema. The maps we select are old national maps of 1:25k published between 1958 and 1990. Figure 5(f) shows an example of old national maps of 1:25k where the symbolization varies from Siegfried maps 5(a) – 5(d).

We randomly select ten sheets from Siegfried maps of 1:25k in the testing areas as ”normal” data of which the distribution is similar to the training data (i.e. produced at similar years with the same design schema). We produce both segmentation and uncertainty maps for those selected map sheets and calculate the ratio between the sum of uncertainty values and the sum of prediction probabilities to normalize the uncertainty w.r.t the object content per map sheet. A higher ratio represents a larger uncertainty level proportionally.

As we can see from Figure 6, numerically LCE is the largest while entropy is the smallest. Nevertheless, all of them present a clear distinction between the normal designs, noisy designs, and out-of-distribution designs. LCE can capture the most significant difference between the normal data and noisy/out-of-distribution data. Generally speaking, the uncertainty level of noisy in-distribution designs is higher than normal in-distribution designs but lower than out-of-distribution designs.

4.5 Uncertainty Indicators for Refining Segmentation

We assume that highly structured and meaningful uncertainty indicators can help the network to learn when and how to trust the prediction. We input either the calculated uncertainty (variance/entropy) or the learned uncertainty (LCE), together with the input image and the segmentation map, to the refinement network for a second-stage training. As we can see from Table 2, all three uncertainty indicators help to refine the predictions, especially improving the precision significantly. Among the three indicators, LCE aids in refining predictions to the largest degree. The epistemic uncertainty indicator (variance) does not lead to a greater improvement than the aleatoric indicators (LCE and entropy). This is probably because, in our case, uncertainty caused by the noise inherent in data has a more systematic influence on prediction than noise in model parameters. Originally, wetlands and lakes have relatively poorer segmentation. All three uncertainty indicators perform well in refining wetland segmentation, which is within our expectation as we note that the majority of false-positive wetlands are caused by map artifacts. However, the improvements for lakes are fairly limited. This implies that poor segmentation of lakes might not be simply explained by noise in data or model parameters. As pointed by Wu et al. (2022), inadequate contexts can inhibit a model from making consistent predictions, especially when similar symbolization exists for different feature classes, as shown in Figure 7 for rivers (c) and lakes (a) (b). As most studies about model-level uncertainty only focus on noise in model parameters, the context-level uncertainty should be further investigated.

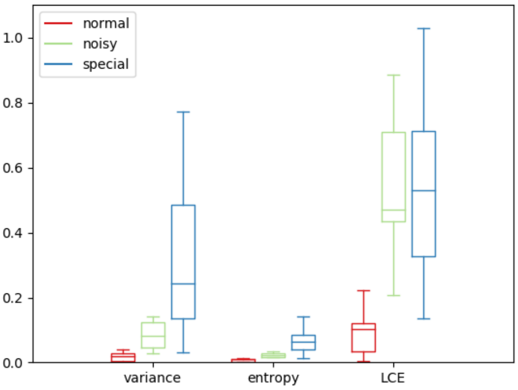
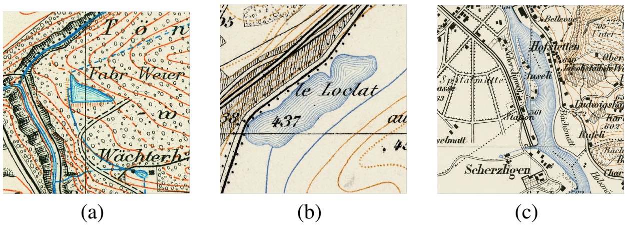


Figure 6. Comparison between indicated uncertainty levels. The uncertainty is normalized by the content of segmented features. The box plot shows the minimum, the maximum, quartiles and the median.



(a) (b) (c)

Figure 7. Examples of similar symbolization between rivers (c) and lakes (a) (b).

4.6 Conclusions

In this work, we have investigated three types of uncertainty indicators that capture aleatoric and epistemic uncertainty in historical map segmentation — prediction entropy, model variance, and LCE. We found that these indicators generally correspond to erroneous predictions and are biased towards positive predictions (rather false positives) than negative ones. All three indicators are able to distinguish noisy and out-of-distribution

	Dice					F1					Precision					Recall				
Refine	avg	stream	wetland	river	lake	avg	stream	wetland	river	lake	avg	stream	wetland	river	lake	avg	stream	wetland	river	lake
-	0.829	0.942	0.752	0.906	0.718	0.871	0.954	0.830	0.912	0.783	0.829	0.940	0.747	0.895	0.734	**0.917**	0.968	0.933	0.929	0.838
Entropy	0.843	0.965	0.789	0.893	0.726	0.878	0.963	0.853	0.899	0.792	**0.872**	0.967	0.808	0.873	0.841	0.884	0.958	0.902	0.928	0.748
Variance	0.843	0.965	0.793	0.890	0.721	0.876	0.963	0.857	0.897	0.786	0.850	0.969	0.810	0.866	0.757	0.903	0.956	0.910	0.930	0.816
LCE	**0.858**	0.965	0.840	0.894	0.731	**0.886**	0.963	0.883	0.900	0.795	0.864	0.967	0.855	0.876	0.759	0.908	0.958	0.913	0.925	0.834

Table 2. Refinement using different uncertainty indicators. We calculate the dice, F1, precision, and recall to evaluate the accuracy in our testing areas for four classes - stream, wetland, river, and lake, and their average accuracy.

designs clearly from normal designs. To test if the yielded uncertainty is highly structured and meaningful, we have proposed a two-stage network to use the uncertainty to refine raw predictions (the second stage) from a normal segmentation network (the first stage). The aid of ensemble variance that captures epistemic uncertainty does not lead to a greater improvement than the other two indicators for aleatoric uncertainty, possibly because noise in data dominates predictions more systematically in our case. Among all three indicators, LCE shows the greatest performance in the refinement task. However, none of the indicators is able to capture the uncertainty in context. In the future, we are going to investigate the impact of context-level uncertainty on model performance.

References

Chen, L., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv:1706.0558.*

Cipolla, R., Gal, Y., Kendall, A., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7482–7491.

DeVries, T., Taylor, G. W., 2018a. Learning Confidence for Out-of-Distribution Detection in Neural Networks. *arXiv:1802.04865.*

DeVries, T., Taylor, G. W., 2018b. Leveraging Uncertainty Estimates for Predicting Segmentation Quality. *arXiv:1807.00502.*

Gurevich, P., Stuke, H., 2018. Pairing an arbitrary regressor with an artificial neural network estimating aleatoric uncertainty. *arXiv:1707.07287.*

Heitzler, M., Hurni, L., 2020. Cartographic reconstruction of building footprints from historical maps: A study on the Swiss Siegfried map. *Transactions in GIS*, 24(2), 442–461.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 5580–5590.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv:1612.01474.*

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J. D., 2022. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment*, 268, 112760.

Mosinska, A., Marquez-Neila, P., Koziński, M., Fua, P., 2018. Beyond the pixel-wise loss for topology-aware delineation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3136–3145.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Uhl, J. H., Leyk, S., Chiang, Y., Duan, W., Knoblock, C. A., 2017. Extracting human settlement footprint from historical topographic map series using context-based machine learning. *Proceedings of the 8th International Conference of Pattern Recognition Systems (ICPRS)*, 1–6.

Uhl, J. H., Leyk, S., Chiang, Y. Y., Duan, W., Knoblock, C. A., 2018. Spatialising uncertainty in image segmentation using weakly supervised convolutional neural networks: A case study from historical map processing. *IET Image Processing*, 12(11), 2084–2091.

Uhl, J. H., Leyk, S., Chiang, Y. Y., Duan, W., Knoblock, C. A., 2020. Automated extraction of human settlement patterns from historical topographic map series using weakly supervised convolutional neural networks. *IEEE Access*, 8, 6978–6996.

Wegner, J. D., Montoya-Zegarra, J. A., Schindler, K., 2013. A higher-order crf model for road network extraction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1698–1705.

Wu, S., Heitzler, M., Hurni, L., 2022. Leveraging uncertainty estimation and spatial pyramid pooling for extracting hydrological features from scanned historical topographic maps. *GIScience & Remote Sensing*, 59(1), 200-214.