Unsicherheitsauswertung von semantischer Segmentierung mittels Neuronaler Netze

LINA E. BUDDE[1,2], STEFAN SCHMOHL[1] & UWE SÖRGEL[1]

Zusammenfassung: Künstliche Neuronale Netze haben sich in weiten Teilen der maschinellen Dateninterpretation als Stand der Technik etabliert. Jedoch ist die Qualitätseinschätzung ihrer Klassifikationsergebnisse ohne gegebene Ground Truth noch ein offenes Forschungsfeld. Von zunehmendem Interesse ist dabei das sogenannte Monte-Carlo Dropout, mit welchem eine Verteilung der prädizierten Pseudo-Wahrscheinlichkeiten geschätzt werden kann. In dieser Arbeit werden verschiedene daraus abgeleitete Unsicherheitsmaße sowohl quantitativ als auch visuell miteinander verglichen. Als Studienobjekt dient dafür der ISPRS Potsdam 2D Semantic Labeling Benchmark Datensatz zur Luftbildklassifikation. Es zeigt sich, dass insbesondere mittels der Shannon Entropie der Pseudo-Wahrscheinlichkeiten unsichere bzw. fehlerhafte Pixelklassifikationen identifiziert werden können. Diese identifizierten Pixel befinden sich in erster Linie an Objekträndern und ermöglichen beispielsweise durch die Entfernung der unsichersten 10 % der Pixel eine Steigerung der Gesamtgenauigkeit von 82,6 % auf 86,5 %.

1 Einleitung

Derzeit genießt maschinelles Lernen, im Speziellen Deep Learning, große Aufmerksamkeit in der Wissenschaft und zunehmend auch in der praktischen Anwendung. Dabei gibt es keine Beschränkung auf eine spezielle Disziplin, sondern eine Vielzahl von Anwendungsmöglichkeiten, zum Beispiel in den Bereichen Physik, Biologie und Fertigung. Eine häufige Aufgabe in der Bildanalyse im Allgemeinen und in der Fernerkundung im Besonderen ist die sogenannte semantische Segmentierung von Bildern, bei der jedem Pixel eine Klasse zugeordnet wird. Aus Fernerkundungsdaten können so qualitativ hochwertige und aktuelle Landbedeckungskarten erstellt werden, die zum Beispiel für die Land- und Forstwirtschaft, Stadtentwicklung und Umweltüberwachung genutzt werden.

Die damit gewonnene, pixelweise Klassenzugehörigkeit ist jedoch als Information nicht immer ausreichend. Solche automatisch erzeugten Klassifikationen sind grundsätzlich fehlerbehaftet. Zusätzliche Informationen zu ihrer Qualität sind daher ebenfalls wünschenswert.

Die Bedeutung solcher Qualitätsmaße im Bereich der Fernerkundung ergibt sich dadurch, dass gerade die Landbedeckungskarten eine wichtige Grundlage sowohl politischer als auch wirtschaftlicher Entscheidungen bilden. Indikatoren für mögliche Fehlzuordnungen können außerdem verwendet werden, um effizient manuell nachzubearbeitende Bereiche auszuwählen, sollte eine sehr hohe Klassifikationsgüte erforderlich sein.

[1] Universität Stuttgart, Institut für Photogrammetrie, Geschwister-Scholl-Str. 24D, D-70174 Stuttgart, E-Mail: st115986@stud.uni-stuttgart.de, [stefan.schmohl, uwe.soergel]@ifp.uni-stuttgart.de

[2] Technische Universität Darmstadt, Institut für Geodäsie, Franziska-Braun-Straße 7, D-64287 Darmstadt, E-Mail: lina.budde@tu-darmstadt.de

Die bei Klassifikationen üblichen Genauigkeitsmaße zur Beurteilung der Qualität benötigen allerdings Referenzdaten als direkte Vergleichswerte, im Fall der Fernerkundung z.B. ausgewählte Referenzflächen im Testgebiet. Solche Referenzdaten mit bekannten Klassenzuordnungen sind aufwendig in der Erstellung und daher leider nur begrenzt verfügbar. Eine Möglichkeit diesem Problem entgegenzuwirken, wird von KÖLLE et. al. (2020) aufgezeigt. Dort wird die sogenannte Vote Entropie eingesetzt, um schrittweise mittels Active Learning die Verbesserung eines Random Forest Klassifikators zu erzielen. Die Quantifizierung der Klassifikations-Unsicherheit stellt somit eine Alternative dar, bei der Unsicherheiten unmittelbar vom Klassifikationsmodell mitgeliefert werden. Jedoch können die meisten Deep Learning Modelle solche Unsicherheiten nicht direkt ausgeben. Stattdessen werden häufig die prädizierten Softmax-Scores als Wahrscheinlichkeiten und damit fälschlich als Modellkonfidenz interpretiert (GAL 2016).
Um dennoch ein fundiertes Maß für die Unsicherheit zu erhalten, wird in der vorliegenden Arbeit das von GAL & GHAHRAMANI (2015) entwickelte Monte-Carlo Dropout eingesetzt, um Klassifikations-Ergebnisse von Luftbildern zu beurteilen. Der Einsatz von Monte-Carlo Dropout im Kontext semantischer Segmentierung wurde erstmals von KENDALL et al. (2015) vorgeschlagen. KAMPFFMEYER et al. (2016) zeigten, dass aus Monte-Carlo Dropout erzeugten Standardabweichungen als Unsicherheitsmaße per Schwellwert Fehlklassifikationen in Luftbildern herausgefiltert werden können. In dieser Arbeit präsentieren wir eine detaillierte Analyse dreier aus Monte-Carlo Dropout abgeleiteter Unsicherheitsmaße (Standardabweichung, Shannon Entropie und Mutual Information) und untersuchen diese hinsichtlich ihrer Tauglichkeit, falsch klassifizierte Pixel zu identifizieren.

2 Methodik

2.1 Monte-Carlo Dropout

Beim Dropout handelt es sich um eine übliche Regularisierungstechnik für neuronale Netze. Es basiert auf der Idee des Ensemble-Lernens, wobei mehrere schwache Klassifikatoren zusammen einen starken ergeben. Statt mehrere Netze separat zu trainieren, werden beim Dropout Subnetze innerhalb des eigentlichen Netzes simuliert. Hierfür wird beim Training pro Iteration ein anderer Teil der Neuronen deaktiviert, üblicherweise in einer der letzten neuronalen Schichten. Bei der Inferenz sind dagegen wieder alle Neuronen aktiv.
Die spezielle Variante des Monte-Carlo Dropouts ermöglicht die effiziente Simulation Bayesscher Netze (GAL, 2016). Statt Dropout bei der Inferenz zu deaktivieren, wird es nun genutzt, um bei mehreren Klassifikationsdurchläufen der Daten durch das neuronale Netz variierende Pseudo-Wahrscheinlichkeiten pro Punkt und Klasse zu erzielen, deren Verteilungsparameter hier als Unsicherheitsparameter dienen. Als prädiziert gilt jene Klasse mit den im mittel höchsten Pseudo-Wahrscheinlichkeiten. Gegenüber der Verwendung von Bayesschen Netzen zur Unsicherheitsauswertung besteht der wesentliche Vorteil dieses Verfahrens darin, dass es in jedem neuronalen Netz mit Dropout eingesetzt werden kann, ohne zusätzliche Parameter aufwendig im Optimierungsprozess bestimmen zu müssen. Auch wird kein a-priori Vorwissen über Randverteilungen benötigt.

2.2 Unsicherheitsmaße

Aus den variierenden Pseudo-Wahrscheinlichkeiten beim Monte-Carlo Dropout können pro Punkt verschiedene Werte abgeleitet werden, die als Unsicherheitsmaße verwendet werden. Unter anderem ergibt sich über die Verteilungen der Pseudo-Wahrscheinlichkeiten pro Klasse eine Standardabweichung σ_c, welche über alle Klassen gemittelt werden (KAMPFFMEYER et al., 2016):

$$\sigma_c = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(p_c^t - \bar{p}_c)^2} \tag{1}$$

Dabei steht p_c für den Softmax-Output für die Klasse c am jeweiligen Punkt und T für die Anzahl an Stichproben. Die gemittelte Standardabweichung über alle Klassen stellt ein Maß für die gesamte Modellunsicherheit dar, welches unabhängig von den einzelnen Klassen ist. Als weiteres Maß dient die normierte Shannon Entropie H aus den gemittelten Stichproben pro Punkt:

$$H = -\frac{1}{\log_2(C)}\sum_{c=1}^{C}\bar{p_c}\cdot\log_2(\bar{p_c}) \tag{2}$$

Die Shannon Entropie ermöglicht Aussagen zum Informationsgehalt pro Punkt. Dabei ist der Informationsgehalt dann größer, wenn die Unsicherheit größer ist. Das dritte Unsicherheitsmaß ist schließlich die Mutual Information (GAL 2016):

$$I_{MC} = \frac{1}{\log_2(C)}\left[-\sum_{c=1}^{C}\left(\sum_{t=1}^{T}p_c^t\cdot\log_2\sum_{t=1}^{T}p_c^t\right)+\frac{1}{T}\sum_{c,t=1}^{C,T}p_c^t\cdot\log_2(p_c^t)\right] \tag{3}$$

Über die Mutual Information wird der Unterschied zwischen der Entropie der Prädiktion und der A-posteriori Wahrscheinlichkeit der Parameter angegeben.

3 Versuchsaufbau

3.1 Semantische Segmentierung

Für die Untersuchung der Klassifikations-Unsicherheiten wird in dieser Studie der ISPRS Potsdam 2D Semantic Labeling Benchmark[3] verwendet, bestehend aus 38 6000×6000 Pixel großen Multispektralbildern mit 5 cm GSD (GERKE et al. 2014). Für die semantische Segmentierung stehen somit zum einen True Orthophotos der Kanäle RGB & IR, zum anderen ein dazugehöriges nDSM zur Verfügung. Das nDSM leitet sich aus einem DSM ab, welches mittels dense image matching erzeugt wurde.

Abweichend zum ursprünglichen Benchmark wurde in dieser Arbeit eine andere Verteilung von Test- und Trainingsbildern verwendet. Der neue Testsatz besteht aus nur noch acht statt vierzehn

[3] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

Bildern und ist über das gesamte Gebiet verteilt. Auf diese Weise konnte u.a. der bisher ausschließlich in den Trainingsdaten vorhandene See berücksichtigt werden. Auch wurden so leicht höhere Genauigkeiten erzielt.
Aufgrund der begrenzten Rechenleistung der verwendeten Hardware wird eine Auflösungsreduktion um den Faktor zwei für diese Daten durchgeführt. Die semantische Segmentierung erfolgt mittels eines einfachen U-Net (RONNEBERGER et al. 2015) und folgt somit einer Encoder-Decoder Struktur. Jedoch werden zu dem von RONNEBERGER et al. (2015) veröffentlichten U-Net Ergänzungen, wie Zero-Padding, Batch-Normalisierung und Dropout hinzugefügt. Das so aufgebaute Netz wird mit Bildausschnitten von jeweils 300×300 Pixel unter der Verwendung der fünf zu Verfügung stehenden Kanäle RGB, IR & nDSM trainiert. Im Gegensatz dazu wird für die Testphase ein Bild in vier Teile zerteilt, die aus 1.500×1.500 Pixeln ($\widehat{=} 150\ \mathrm{m} \times 150\ \mathrm{m}$ auf dem Boden) bestehen.
Die für das Training notwendigen Referenzdaten beinhalten sechs verschiedene Klassen: „versiegelte Oberfläche“, „Gebäude“, „niedrige Vegetation“, „Bäume“, „Autos“ und alle weiteren Objekte zusammengefasst in der Klasse „Sonstiges“. Für das Training selbst wird zusätzlich eine künstliche Datenerweiterung durchgeführt. Durch Rotation, Spiegelung und additivem Rauschen ergibt sich eine größere Variation der Trainingsdaten, sodass die Gefahr der Überanpassung verringert werden kann. Als Verlustfunktion wird eine gewichtete Kreuzentropie mit weight decay verwendet. Dadurch kann die ungleiche Auftrittshäufigkeit der verschiedenen Klassen berücksichtigt werden. Insgesamt beträgt die Dauer der Trainingsphase 50 Epochen. Durch das Aufspalten des Datensatzes in Trainings-, Validierungs- und Testdaten, kann bereits während der Trainingsphase der Trainingserfolg mit den Validierungsdaten überprüft werden. Nach der Trainingsphase werden für die Auswertung, die bisher für das Netz unbekannten, Testdaten verwendet.

3.2 Unsicherheitsauswertung

Aus den Pseudo-Wahrscheinlichkeiten von insgesamt 20 Monte-Carlo Durchläufen je Bild werden die drei Unsicherheitsmaße Standardabweichung, Shannon Entropie und Mutual Information gebildet. Dazu werden die in den Gleichungen (1), (2) und (3) angegeben Formeln genutzt. Zusätzlich wird die Shannon Entropie auch ohne Monte-Carlo Dropout bestimmt. Dadurch kann verglichen werden, inwieweit das Monte-Carlo Dropout für eine Unsicherheitsauswertung notwendig ist.
Mithilfe von Schwellwerten können aus den so erhaltenen Unsicherheitsmaßen anschließend die unsichersten Klassifikationen herausgefiltert werden, um einen möglichst großen Teil der Fehlklassifikationen zu entfernen. Dabei können sowohl absolute als auch relative Schwellwerte vorgegeben werden. Bei einem absoluten Schwellwert darf das jeweilige Pixel diesen Schwellwert nicht überschreiten, um als sicher eingestuft zu werden. Wird dagegen ein relativer Schwellwert angegeben, wird dadurch die Anzahl der unsicheren beziehungsweise sicheren Pixel festgelegt, die auf Grundlage ihres Unsicherheitswertes aussortiert beziehungsweise beibehalten werden.

4 Ergebnisse

4.1 Semantische Segmentierung

Aus den 20 Inferenzdurchläufen mit aktiviertem Monte-Carlo-Dropout gilt pro Pixel diejenige Klasse als prädiziert, welche den höchsten gemittelten Softmax-Score besitzt. Das Ergebnis einer solchen semantischen Segmentierung zeigt beispielhaft Abb. 1 für einen kleineren Testbildausschnitt.

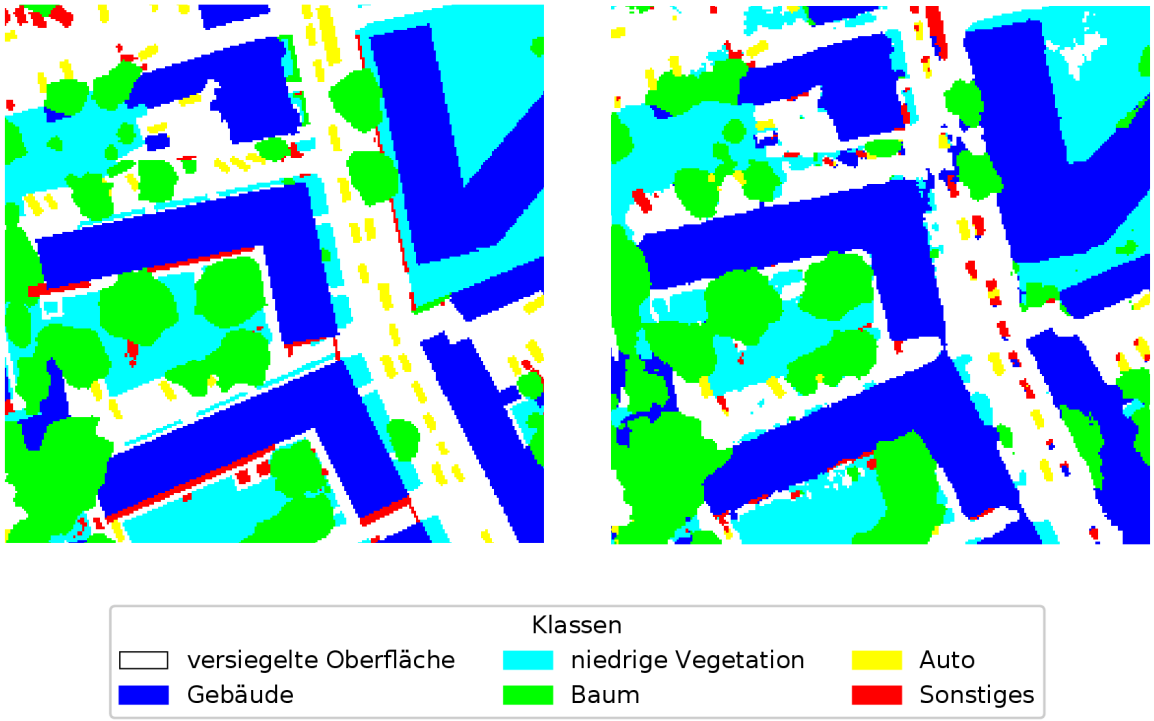

Abb. 1: Testbildausschnitt. Links: Referenz, rechts: prädizierte Label

Mit Hilfe der Referenzdaten ergeben sich aus den acht Testbildern die in Abb. 2 dargestellte Confusion Matrix und die in Tab. 1 enthaltenen F1-Werte. Es zeigt sich, dass besonders die Klasse „Gebäude“ sehr gut erkannt wird. Dagegen kann für die Klasse „Sonstiges“ nur eine geringe Genauigkeit erreicht werden. Dies liegt wahrscheinlich hauptsächlich in der Inhomogenität der Klasse „Sonstiges“. Starke Verwechslungen existieren zwischen den beiden Vegetationsklassen „Baum“ und „niedrige Vegetation“. Zudem konnte eine große Abhängigkeit der Klassifikation von der Qualität des gegeben nDSM festgestellt werden, da in Bereichen von fehlerhaften nDSM-Daten die Klassifikation ebenfalls fehlerhaft war. Insgesamt ergibt sich mit dem verwendeten Netzwerk über alle Testbilder eine Gesamtgenauigkeit von 82,6 %.
Neben den Fehlklassifikationen bei den Klassen „Sonstiges“ und „Auto“ werden in Abb. 1 die teils ausgefransten Objektkanten sichtbar. Die Identifikation solcher Fehlklassifikationen ist bei der nachfolgenden Unsicherheitsauswertung wünschenswert.

Tab. 1: F1-Werte der einzelnen Klassen aus den Testdaten

	Versiegelte Oberfläche	Gebäude	Niedrige Vegetation	Baum	Auto	Sonstiges
F1 [%]	87,7	90,3	79,6	72,2	65,7	23,7

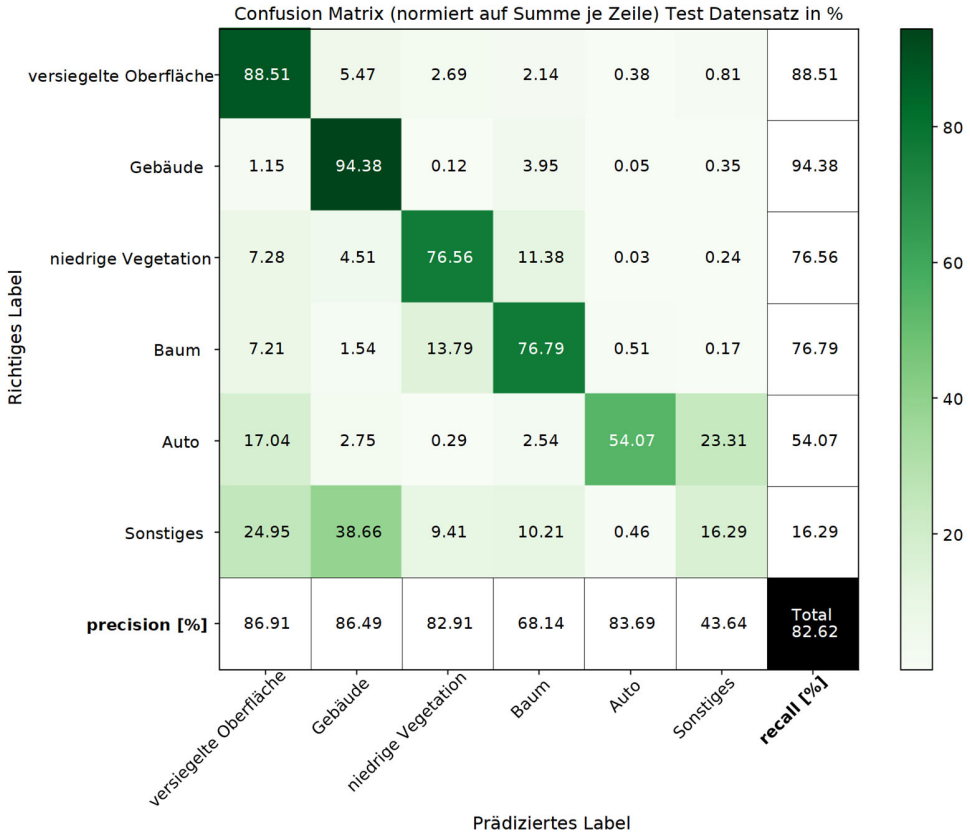

Abb. 2: Normierte Confusion Matrix in % aus den Testdaten

4.2 Unsicherheitsmaße

Zusätzlich zu Abb. 1 enthalten die Abb. 3 bis Abb. 6 die gemittelten Softmax-Scores sowie die drei Unsicherheitsmaße aus Monte-Carlo-Dropout für denselben Testbildausschnitt. Allen Abbildungen ist gemeinsam, dass Objektstrukturen eine homogene innere Färbung besitzen. Deutlich ist ein sich durchziehender senkrechter Bruch zu erkennen. Dies hängt mit der Zerteilung der einzelnen Testbilder in jeweils vier Bildausschnitte für die Prozessierung zusammen. Durch die iterative Klassifikation der einzelnen Bildausschnitte durch das Netz basieren die einzelnen Pseudo-Wahrscheinlichkeiten bei jedem Bildausschnitt durch das Monte-Carlo Dropout auf unterschiedlichen Sub-Netzen. Dies betrifft nicht nur die gemittelten Softmax-Scores der prädizierten Klasse, sondern auch alle aus dem Monte-Carlo Dropout abgeleiteten Größen. Diese Variationen der einzelnen Bildausschnitte sind ebenfalls ein Zeichen für die Unsicherheit des Modells, da ohne Unsicherheiten jedes der Sub-Netze ähnliche Ergebnisse liefern sollte. Jedoch haben diese Variationen hier keinen Einfluss auf die prädizierte Klasse. Wie in Abb. 1 zu sehen ist.

Die mittlere Standardabweichung und Mutual Information erzeugen bei der optischen Betrachtung ähnliche Ergebnisse (Abb. 3b & Abb. 4a). Die Unsicherheitswerte sind bei Objekten innerhalb einer Klasse konsistent. So weisen Bäume im Großteil höhere Werte auf als niedrigere Objekte. Die geringsten Werte besitzen Gebäude. Im Ausschnitt rechts unten sind die Verhältnisse teilweise invertiert. Die Unsicherheiten scheinen hier daher in erster Linie mit der Klassenart zu korrelieren und weniger mit Fehlklassifikationen.
Im Gegensatz dazu treten bei der Shannon Entropie die höchsten Werte klar als linienförmige Strukturen auf (Abb. 4b). In diesen Bereichen sind die Unsicherheitswerte am stärksten und befinden sich entlang der Klassen- bzw. Objektgrenzen, was auch für einen großen Teil der Fehlklassifikationen zutrifft (vgl. Abb. 6). Ein ähnliches Muster zeigt sich bei den oft als Unsicherheitsmaß herangezogenen, hier mittleren, Softmax-Scores aus Abb. 3a.

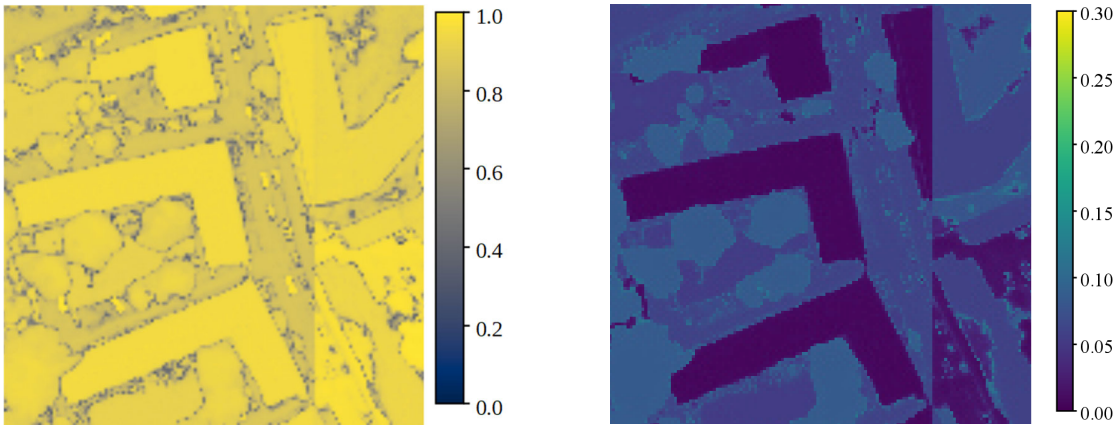

Abb. 1: Testbildausschnitt. (a): mittlere Softmax-Scores (Pseudo-Wahrscheinlichkeiten) der jeweils stärksten Klasse, (b): über alle Klassen gemittelte Standardabweichung

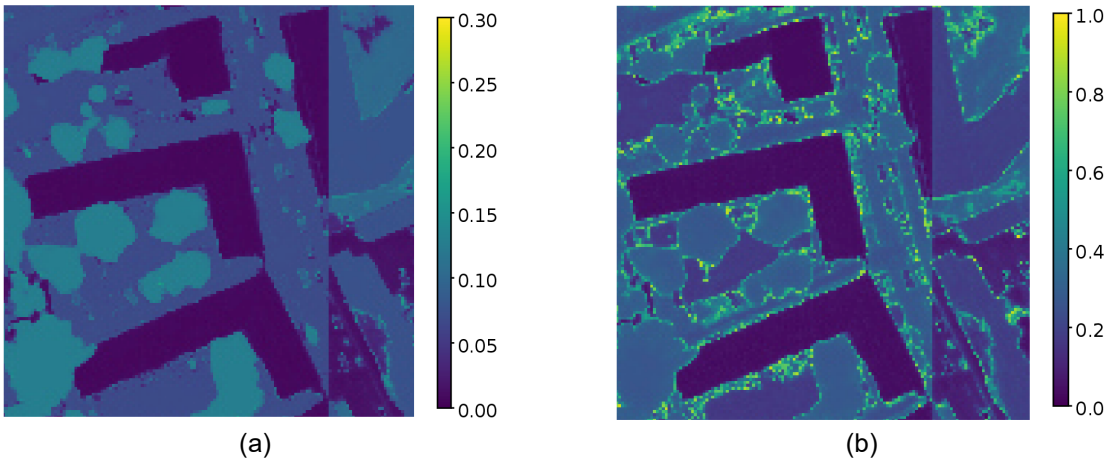

Abb. 2: Testbildausschnitt. (a): Mutual Information, (b): Shannon Entropie mit Monte-Carlo Dropout

Um die Unsicherheitsmaße besser vergleichen zu können, werden anhand verschiedener Schwellwerte Precision-Recall Kurven erstellt (Abb. 5). Dafür werden die Pixel mit Unsicherheitswerten,

die größer als der jeweilige Schwellwert sind, verworfen. Es zeigt sich, dass die Shannon Entropie mit Monte-Carlo Dropout das geeignetste Verfahren zur Bestimmung der Zuverlässigkeit ist. Dicht gefolgt von der Precision-Recall Kurve der gemittelten Softmax-Scores der prädizierten Klasse. Deutlich schlechter schneiden die Standardabweichung und die Mutual Information ab. Da sich die Shannon-Entropie auch ohne Monte-Carlo-Dropout berechnen ließe, ist zum Vergleich auch solch eine Kurve dargestellt. Diese ist besonders auffällig, da die maximal erreichte Precision im Bereich von 84 % liegt.

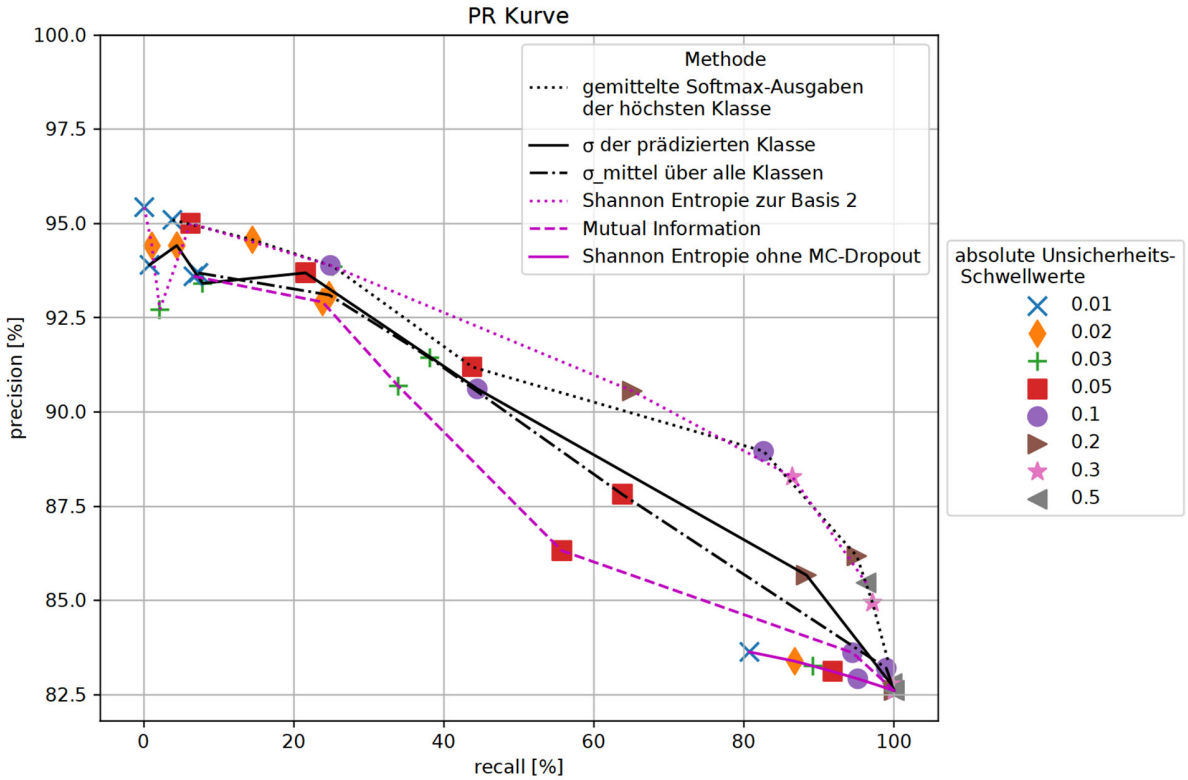

Abb. 5: Precision-Recall Kurven zum Vergleich verschiedener Unsicherheitsmaße mittels unterschiedlicher absoluter Schwellwerte

In der Praxis kann es praktischer sein, statt fester Unsicherheitsschwellwerte, relative Schwellwert anhand eines prozentualen Anteils an zu verwerfenden Pixeln festzulegen. Nach dem Herausfiltern von 10 % der Pixel mit den größten Entropiewerten, verbessert sich die Gesamtgenauigkeit von 82,6 % auf 86,5 %. Nach der Entfernung von 20 % liegt die Genauigkeit bereits bei 88,2 %. Die Ergebnisse solcher Filterungen sind in Abb. 6 & 7 zu sehen. Dabei kann unterschieden werden, ob ein Pixel richtig klassifiziert und gleichzeitig als unsicher identifiziert wurde (weiße Pixel) oder sowohl als falsch klassifiziert sowie zudem als unsicher (schwarze Pixel) anzusehen ist. In erster Linie werden die Pixel an den Objekträndern herausgefiltert. Der Schwellwert, nach dem die Pixel als sicher beziehungsweise unsicher kategorisiert werden, kann entsprechend den gewünschten Anforderungen festgelegt werden.

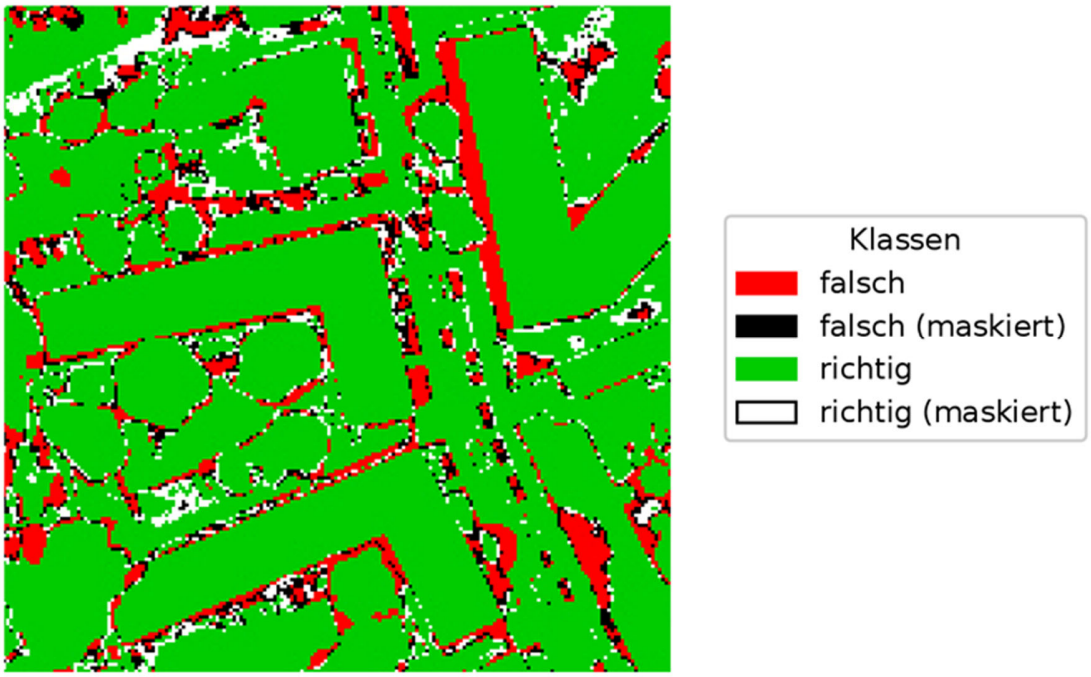

Abb. 6: Testbildausschnitt, bei dem die unsichersten 10 % der Pixel aus der Shannon Entropie maskiert wurden.

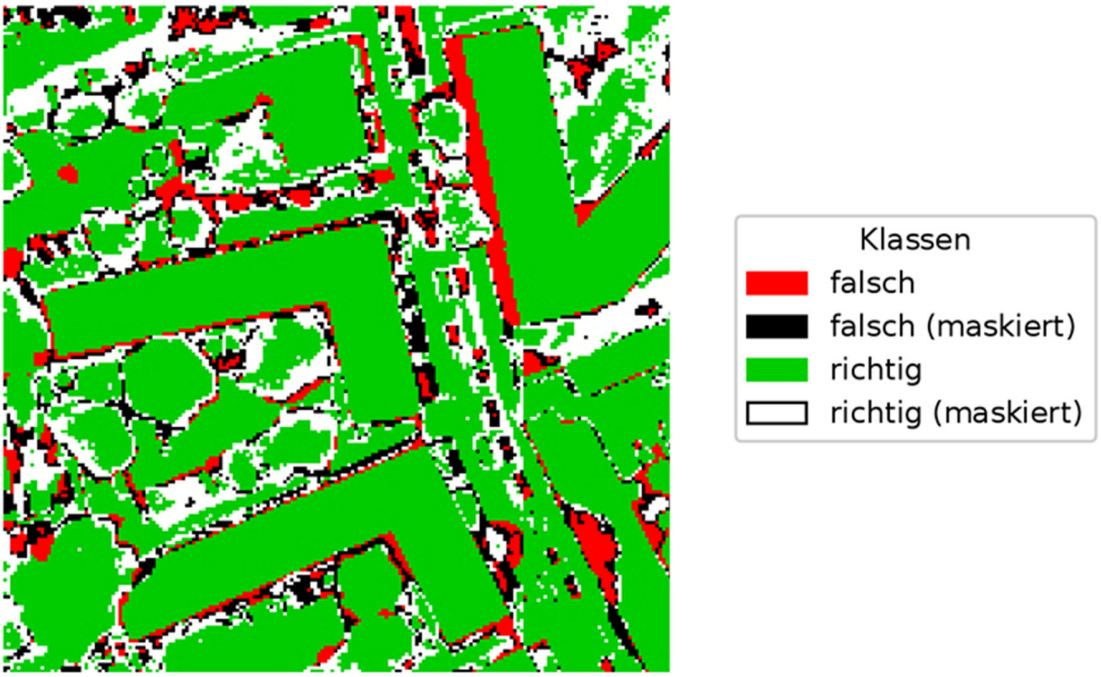

Abb. 7: Testbildausschnitt, bei dem die unsichersten 20 % der Pixel aus der Shannon Entropie maskiert wurden

5 Fazit & Ausblick

In dieser Arbeit konnte gezeigt werden, dass Monte-Carlo-Dropout robustere Maße für die Modellunsicherheit liefert als die Auswertung der Pseudo-Wahrscheinlichkeiten einer einzelnen Inferenz.
Die Precision-Recall Kurve aus Abb. 5 unterstützt den optischen Eindruck, dass die Shannon Entropie dabei das geeignetste Verfahren zur Bestimmung der Unsicherheit ist. Die Bereiche an den Objektgrenzen werden dabei als am unsichersten identifiziert. Dort sind auch viele der Fehlklassifikationen zu finden, was die Genauigkeitssteigerung durch das Herausfiltern von Pixeln mit besonders hohen Entropiewerten erlaubt. Je höher die Anforderung an die Precision ist, desto mehr Pixel müssen gegebenenfalls nachbearbeitet werden. Es ist allerdings mit diesem Ansatz nicht möglich Fehlklassifikationen zu erkennen, die sich aufgrund mangelnder Qualität der Eingabedaten auf ganze Objekte beziehen. Daher wäre es interessant ein gegebenenfalls vorhandenes Vorwissen zur Qualität der Daten mitberücksichtigen zu können.
Ein zusätzlicher Anwendungsbereich bildet das Active Learning. Mittels der zusätzlichen Informationen zu den Unsicherheiten jedes Pixels kann beispielsweise anstelle eines Random Forest wie bei KÖLLE et. al. auch ein neuronales Netz eingesetzt werden. Dabei können, durch gezieltes Labeln unsicherer, bisher nicht mit ground truth versehener Pixel, mit einem geringeren Aufwand neue und für die Klassifikation wertvolle Trainingsdaten geschaffen werden, anstatt größere Datenmengen vollständig manuell zu labeln.

6 Literaturverzeichnis

GAL, Y., 2016: Uncertainty in Deep Learning. Dissertation. University of Cambridge.

GAL, Y. & GHAHRAMANI, Z., 2015: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. CoRR (arXiv: 1506.02142).

GERKE, M., ROTTENSTEINER, F., WEGNER, J. & SOHN, G., 2014: ISPRS Semantic Labeling Contest. PCV - Photogrammetric Computer Vision.

KAMPFFMEYER, M. SALBERG, A. & JENSSEN, R., 2016: Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 680-688.

KENDALL, A. BADRINARAYANAN, V. & CIPOLLA, R., 2015: Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. CoRR (arXiv: 1511.02680).

KÖLLE, M., WALTER, V., SCHMOHL, S. & SÖRGEL, U., 2020: Evaluierung der Leistungsfähigkeit der Crowd für das Labeln von 3D-Punktwolken im Kontext von Active Learning. Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V., Band **29**, 299-311.

RONNEBERGER, O., FISCHER, P. & BROX, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS **9351**, 234-241.