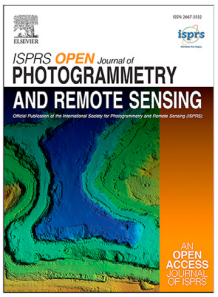
Journal Pre-proof



Integrating uncertainty quantification into transformer models for large-scale 3D semantic segmentation of urban areas

Anis Farshian, Steven Landgraf, Markus Ulrich, Martin Weinmann

PII: S2667-3932(26)00043-8
DOI: https://doi.org/10.1016/j.ophoto.2026.100157
Reference: OPHOTO 100157

To appear in: *ISPRS Open Journal of Photogrammetry and Remote Sensing*

Received date : 12 March 2026
Revised date : 17 August 2026
Accepted date : 21 August 2026

Please cite this article as: A. Farshian, S. Landgraf, M. Ulrich et al., Integrating uncertainty quantification into transformer models for large-scale 3D semantic segmentation of urban areas. *ISPRS Open Journal of Photogrammetry and Remote Sensing* (2026), doi: https://doi.org/10.1016/j.ophoto.2026.100157.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 Published by Elsevier B.V. on behalf of International Society of Photogrammetry and Remote Sensing.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Revised Manuscript (Clean Version) Journal Pre-proof Click here to view linked References

Integrating Uncertainty Quantification into Transformer Models for Large-scale 3D Semantic Segmentation of Urban Areas

Anis Farshian Steven Landgraf Markus Ulrich Martin Weinmann
Institute of Photogrammetry and Remote Sensing (IPF)
Karlsruhe Institute of Technology
{anis.farshian, steven.landgraf, markus.ulrich, martin.weinmann}@kit.edu

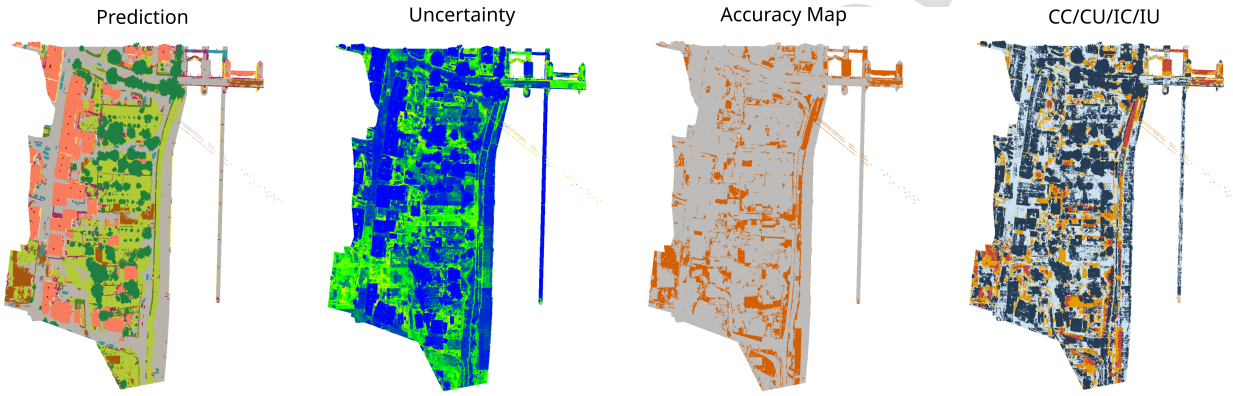


Figure 1. Visualization of results achieved by a specific configuration of our framework (DSE-FE-TTA) on the Hessigheim 3D dataset in bird's-eye view (BEV): predicted labels, uncertainty map, accuracy map, and CC/CU/IC/IU representation.

Abstract

Large-scale 3D point cloud semantic segmentation is a key capability for remote sensing applications, yet modern deep models often remain overconfident in challenging and ambiguous regions. This work studies uncertainty-aware large-scale 3D semantic segmentation by integrating complementary uncertainty mechanisms into a strong Point Transformer V3 backbone. Specifically, we consider Monte Carlo Dropout (MCD), Deep Sub-Ensembles (DSE) variants, and Test-Time Augmentation (TTA), within a unified training and inference pipeline. We evaluate these methods, together with the baseline model, on two benchmark datasets with distinct characteristics (Hessigheim 3D and DALES) using segmentation metrics, calibration errors, and uncertainty measures, complemented by qualitative visualizations of predicted labels, accuracy maps, uncertainty maps, and accuracy-uncertainty breakdowns. Compared to the baseline, incorporating uncertainty awareness yields gains on the Hessigheim 3D dataset of up to 2.3% in mIoU, 2.3% in mF1, and 2.8% in mAcc. On DALES, the corresponding improvements reach up to 1.9% in mIoU, 1.5% in mF1, and 0.5% in mAcc. Based on the analysis of the final results, the influence of TTA is more evident on segmentation performance and has subtler effects on uncertainty metrics. Furthermore, the MCD variants can improve calibration and uncertainty usefulness as the dropout rate increases, at the cost of segmentation accuracy. Finally, DSE variants yield more structured and reliable uncertainty estimates. In particular, we find that the DSE variant with frozen encoder combined with TTA (DSE-FE-TTA) reduces overconfident errors and offers the most robust overall compromise across both datasets, with clearer benefits on the more challenging Hessigheim 3D benchmark in terms of segmentation, calibration, and uncertainty metrics.

1. Introduction

Large-scale 3D point clouds acquired by LiDAR and photogrammetry have become a core data modality in remote sensing for mapping, inspection, and scene understanding. A fundamental task on such data is 3D semantic segmentation. Instead of labeling pixels as in semantic seg-

Journal Pre-proof

mentation with 2D images, it assigns a semantic class label to each 3D point in a point cloud [94]. Unlike 2D images arranged on regular grids, point clouds are irregular sets of points in continuous space, with non-uniform sampling density and no fixed neighborhood structure. These characteristics make it non-trivial to apply standard convolutional networks directly and motivate architectures that can operate natively on unordered sets.

Transformers offer a principled foundation for point cloud learning because self-attention [83] aggregates information over sets without relying on grid-aligned convolutions and can naturally handle variable-sized inputs. Therefore, because self-attention is invariant to the order or number of input elements, transformer architectures have emerged as a promising direction in 3D vision [19]. Point Transformer models [92, 93, 97] leverage this property to learn expressive point-wise representations by combining attention-based feature aggregation with geometric cues, enabling deep networks that operate directly on raw point sets. In this work, we follow our previous work [20] and build on Point Transformer V3 [92] as a strong baseline for large-scale 3D semantic segmentation, as it provides a practical balance between representation capacity and scalability in modern point cloud processing pipelines.

Despite strong predictive performance, standard deep segmentation models often tend to be overconfident [22,26]. They can be very confident in assigning labels to incorrect predictions, particularly in challenging and ambiguous regions. This behavior is problematic in decision-sensitive deployments, where downstream actions should depend not only on predicted labels but also on the reliability of those predictions. Uncertainty quantification (UQ) provides an operational remedy by estimating predictive uncertainty alongside class probabilities, allowing uncertainty-aware postprocessing, risk-aware decision making, and more informative error diagnosis. However, uncertainty mechanisms can also introduce additional variance or reduce the model's effective capacity. Therefore, their practical value must be assessed jointly with segmentation performance and the calibration of the model's predicted softmax probabilities (i.e., how closely predicted probabilities match the empirical frequency of correct predictions [22, 26, 61]).

This paper studies uncertainty-aware large-scale 3D semantic segmentation within a unified framework that integrates three complementary uncertainty mechanisms into a Point Transformer V3 architecture: (i) Monte Carlo Dropout (MCD) [21] to approximate uncertainty through stochastic inference, (ii) Deep Sub-Ensembles (DSE) [80] to induce predictive diversity via multiple decoder heads, and (iii) Test-Time Augmentation (TTA) [2, 22, 70, 91] to derive uncertainty from the dispersion of predictions under geometric perturbations, while aggregating them for a more consistent final output. We evaluate these methods on two benchmark datasets with distinct characteristics (Hessigheim 3D [37] and DALES [81]), using segmentation metrics, calibration errors, and uncertainty measures that quantify whether the model is correct when confident, and uncertain when wrong.

We address two central research questions in this work. First, we investigate how integrating uncertainty estimation influences the segmentation performance and calibration of a strong backbone such as Point Transformer V3 in large-scale point cloud settings (cf. Figure 1). Second, we examine which uncertainty mechanism delivers the most reliable segmentation results together with informative uncertainty estimates across different datasets.

In summary, the main contributions of this paper are:

- To improve prediction reliability, we address overly confident predictions by incorporating uncertainty quantification methods into a strong Point Transformer V3 backbone for large-scale 3D point cloud semantic segmentation.
- We integrate and compare complementary uncertainty mechanisms (MCD, DSE variants, and TTA) under a unified training and inference pipeline.
- We evaluate the proposed framework quantitatively and qualitatively on two large-scale benchmarks with different characteristics, analyzing segmentation results, calibration errors, and uncertainty measures, and highlighting how these aspects relate across methods. In particular, we show that some of the uncertainty-aware methods can reduce overconfident errors and, on the more challenging dataset, can also yield a meaningful improvement in segmentation performance relative to the baseline.

This paper is organized as follows. Section 2 provides a brief overview of related work on 3D semantic segmentation and uncertainty quantification. Section 3 presents the fundamentals of the proposed framework for uncertainty-aware large-scale 3D semantic segmentation. Section 4 describes implementation details as well as the datasets and evaluation metrics. Furthermore, quantitative and qualitative results are presented in this section. Section 5 discusses the findings, and Section 6 concludes the paper and outlines directions for future work.

2. Related Work

In this section, we briefly summarize related work on 3D semantic segmentation (Section 2.1) and uncertainty quantification (Section 2.2).

2.1. 3D Semantic Segmentation

Traditional approaches for 3D semantic segmentation rely on the use of handcrafted features extracted for each 3D

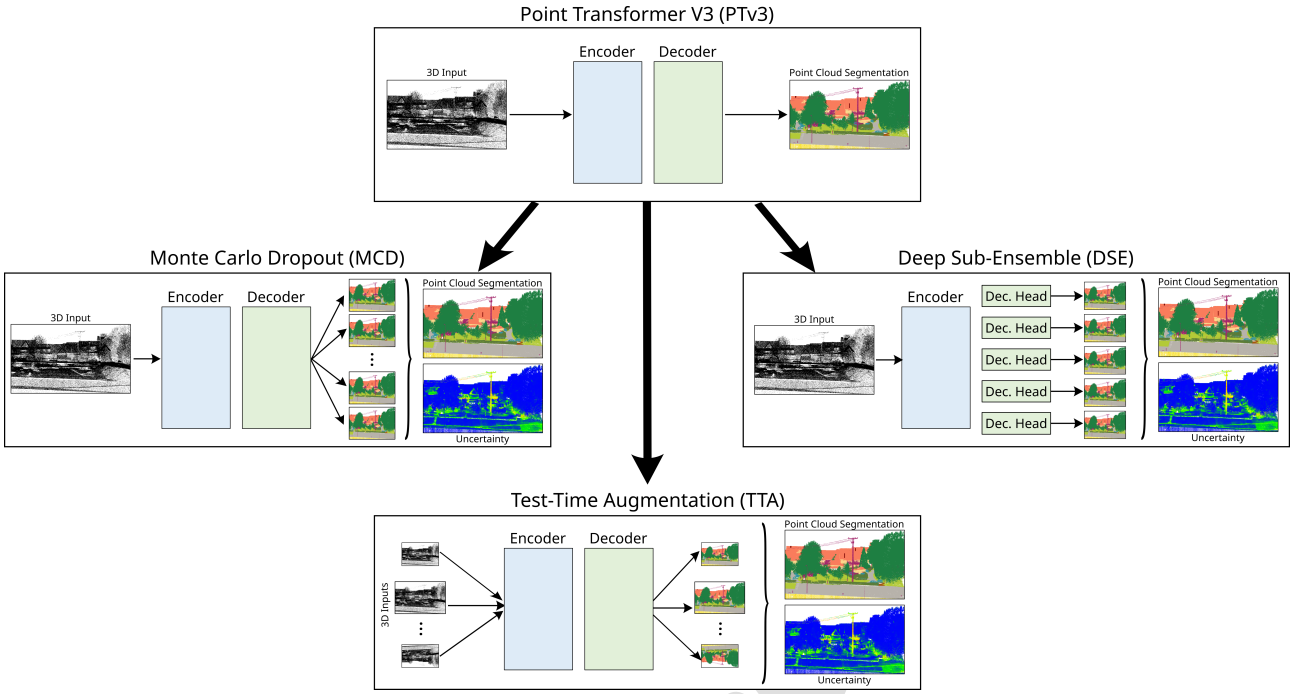


Figure 2. Schematic overview of integrating various uncertainty quantification mechanisms into the Point Transformer V3 (PTv3) baseline.

point and concatenated into a feature vector, which is then used as input to a shallow classifier (e.g., Support Vector Machine (SVM) [16] or Random Forest (RF) [11]) to predict the corresponding class label [30, 77, 89]. Since each 3D point is treated independently based only on its feature vector, the resulting predictions are often noisy, whereas a spatially regular labeling is typically desirable.

To account for possibly correlated class labels of nearby 3D points, contextual classification approaches have been proposed. Here, the classification relies not only on the respective feature vector for labeling an individual 3D point, but also on the relationships within its local neighborhood, which are inferred from the training data through the modeling of interactions between neighboring points. For this modeling of relationships, numerous approaches [44,60,74] rely on the use of a Conditional Random Field (CRF) [39]. A more general formulation to model relationships within local neighborhoods can be achieved in the form of structured regularization [45]. Although contextual classification approaches typically ensure spatially regular labeling, their dependence on handcrafted features makes them less suitable for scenarios with many and highly similar classes.

For scenarios in which it is difficult to define features that enable a clear separation between different classes, modern deep learning approaches show great potential. Such approaches focus on learning both the features and the classification in an end-to-end fashion, given sufficient labeled training data. In general, deep learning approaches for 3D semantic segmentation can be categorized into three different groups [1, 8, 27, 31, 94]:

- Voxel-based approaches focus on the extension of the convolution operation from 2D to 3D, since the learning of 2D filter masks applied via convolution has shown great success for numerous image analysis tasks. More specifically, the 3D space or point cloud is represented in the form of a voxel grid (e.g., a voxel occupancy grid) and then 3D convolutions can be applied, which is the basis for different 3D-CNN architectures [29, 34, 76].

- Projection-based approaches leverage the success of image analysis techniques by transferring the 3D semantic segmentation task into three successive steps: (i) rendering the 3D point cloud from different perspectives to get corresponding 2D imagery, (ii) applying a standard 2D-CNN architecture to derive a per-pixel labeling for the given 2D imagery, and (iii) back-projecting these predictions onto the 3D points followed by refinement to obtain a consistent labeling of the original point cloud [10, 47, 75].

- Point-based approaches focus on the direct processing of 3D point clouds, and they can be further categorized into Multi-Layer Perceptron (MLP)-based approaches (e.g., PointNet [13], PointNet++ [65] or PointNeXt [66]), point-convolution-based approaches (e.g.,

Journal Pre-proof

KPConv [78], KPConvX [79], LCA-KPConv [5], or RandLA-Net [33]), graph-convolution-based approaches (e.g., DGCNN [88], RFFS-Net [51] or SPG [46]) and transformer-based approaches (e.g., PointASNL [95], Stratified Transformer [40] or approaches of the Point Transformer family [92, 93, 97]).

Beyond conventional closed-set supervised segmentation, recent work has also begun to address more flexible urban scene understanding settings. For instance, Wang et al. proposed OpenUrban3D [85], a label-free open-vocabulary framework for large-scale urban point clouds that combines multiview rendering, vision-language feature extraction, and distillation into a 3D backbone. Related work has also considered open-world 3D segmentation from a different perspective. Riz et al. [67] introduced SNOPS, a method for novel class discovery for 3D semantic segmentation, where previously unlabeled classes are discovered from partially labeled training data using online clustering, uncertainty estimation, and semantic distillation. Besides work that focuses on the discovery of novel classes, several works particularly focused on underrepresented objects and imbalanced classes [3, 4], for instance by data augmentation in terms of inserting additional objects into the point cloud used for training.

In our work, we learn task-specific features end-to-end. The proposed uncertainty integration is therefore largely model-agnostic and could in principle be combined with other 3D segmentation backbones. We nevertheless adopt a recent transformer-based model, Point Transformer V3 (PTv3) [92], because its scale-driven design enables efficient processing of large-scale point clouds, while maintaining strong segmentation performance. Its attention-based feature aggregation captures both local geometry and broader context, while its design enables efficient processing of high point counts typical for remote sensing and urban scenes. Using a single, competitive backbone allows us to isolate the effects of the uncertainty mechanisms and avoids mixing UQ behavior with architectural differences.

2.2. Uncertainty Quantification

Traditional machine learning approaches that use handcrafted features as input to a shallow model, can incorporate confidence in several ways. One option is to use any discriminative classifier that delivers a soft labeling in the form of a probabilistic output, i.e., the prediction of a probability distribution over the defined classes instead of the prediction of just a single label. Typical examples are represented by probabilistic modeling (e.g., using a Naive Bayes classifier [35]), Ensembling (e.g., a Random Forest [11]), Neural Networks (e.g., MLP with a final softmax activation layer), or SVMs (e.g., with an additional sigmoid function to map the SVM outputs into probabilities [64]).

A derived soft labeling, in turn, can be used as unary potential within a graphical model for contextual classification to achieve a spatially regular labeling. Regarding 3D semantic segmentation, this has been addressed by the use of a Conditional Random Field (CRF) [38, 39] and linking its association potential to the soft labeling of a discriminative classifier [44, 60, 90]. Thus, the association potential indicates how a specific 3D point would be labeled if only its corresponding feature vector is considered, whereas the other term within the CRF formulation, i.e., the interaction potential, represents a pairwise potential that addresses spatial smoothness of the labeling based on the feature vectors and label of 3D points within a local neighborhood. A more versatile alternative to the standard graphical model approach is represented by the structured regularization framework proposed by Landrieu et al. [45], which allows for the choice between a wide range of fidelity functions and regularizers. This also includes configurations under which the spatially regular labeling retains a confidence associated with each label.

While the aforementioned traditional approaches provide a straightforward measure of label confidence, they rely on a single, deterministic point-estimate (e.g., a softmax probability). Consequently, these methods often suffer from overconfidence and cannot easily distinguish between inherent data noise and a fundamental lack of model knowledge. To achieve true uncertainty awareness, modern approaches shift from interpreting single-output confidence to quantifying the dispersion of a predictive distribution. Rather than relying on one deterministic forward pass, techniques such as Monte Carlo Dropout (MCD) [21], Deep Ensembles (DE) [41], and Test-Time Augmentation (TTA) [2] simulate a distribution by generating a diverse set of predictions for the same input (either by perturbing the network's weights, training multiple independent models, or by augmenting the input data). By measuring the disagreement across these multiple predictions, these methods can explicitly capture uncertainty arising from data ambiguity or model ignorance, providing a far more robust assessment of reliability than traditional soft labels.

Beyond traditional machine learning approaches, deep learning approaches have achieved remarkable success across a wide range of applications. Their ability to learn complex, high-dimensional representations from large datasets has led to outstanding performance in many tasks. However, standard deep learning models typically do not provide a well-calibrated measure of confidence, limiting their reliability in safety-critical and decision-sensitive settings. This limitation can be addressed with uncertainty quantification in deep neural networks [22]. Conceptually, it is often distinguished between uncertainty originating from inherent noise in the data (aleatoric uncertainty) and uncertainty stemming from limited knowledge of model parameters or structure (epistemic uncertainty) [18]. For 3D

Journal Pre-proof

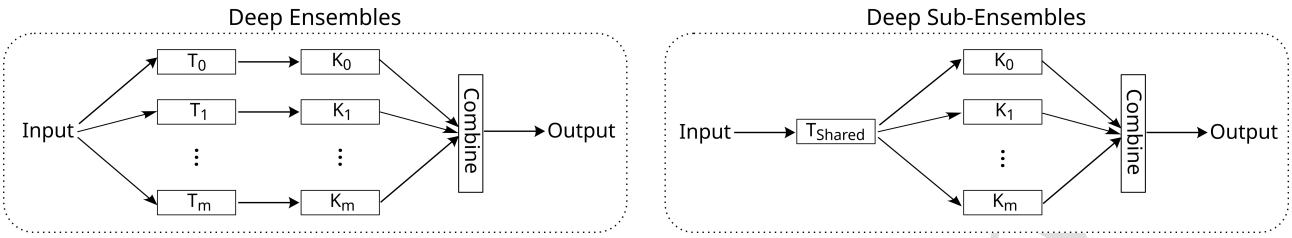


Figure 3. Deep Ensembles (DE) vs. Deep Sub-Ensembles (DSE) with m ensemble members. Unlike DE, DSE relies on one shared trunk T, while training multiple independently initialized task heads K, whose outputs are combined by averaging at inference, resulting in ensemble-like uncertainty at substantially lower computational cost.

point clouds, aleatoric uncertainty captures, for instance, noise inherent to the 3D measurement, such as LiDAR range errors, density variations, or missing returns. This type of uncertainty remains irreducible even with additional data. In contrast, epistemic uncertainty reflects gaps in what the model has learned and can typically be reduced with more diverse training data [22, 36]. While aleatoric uncertainty can be learned directly from the data itself and requires only a single forward pass, epistemic uncertainty necessitates more sophisticated approaches such as Bayesian Neural Networks [59], Bayesian approximations like Monte Carlo Dropout [21], or Ensembling [41, 80]. These techniques differ in their computational requirements and their ability to represent different sources of uncertainty.

Regarding 3D semantic segmentation based on deep learning, only few studies have explicitly investigated uncertainty in point cloud predictions. Cortinhal et al. [17] introduced SalsaNext, an uncertainty-aware LiDAR segmentation network that employs a Bayesian treatment to estimate per-point epistemic and aleatoric uncertainty for autonomous driving settings. More specifically, epistemic uncertainty is assessed by employing dropout layers and a Monte Carlo sampling during inference, whereas aleatoric uncertainty is derived by replacing the network's activations with probability distributions. Similarly, Petschnigg et al. [62, 63] proposed Bayesian neural networks to estimate epistemic and aleatoric uncertainty. They showed that their approach clearly surpasses a deterministic baseline and that leveraging its uncertainty leads to more accurate results in an industrial manufacturing plant. A further approach for uncertainty-aware 3D semantic segmentation has been presented by Vassilev et al. [82] for infrastructure projects and is based on Bayesian neural networks that employ Monte Carlo Dropout and Variational Inference to conduct inference with a KPConv backbone. Moreover, Bhandary et al. [9] performed a systematic comparison of multiple uncertainty quantification methods [21, 41, 53] and found that Deep Ensembles consistently outperform the other methods in their experimental evaluation on the SemanticKITTI dataset [6] comprising sequential scans of a moving car. In a class-incremental learning context, Yang et al. [96] proposed a geometry and uncertainty-aware framework for 3D point cloud semantic segmentation, addressing catastrophic forgetting through uncertainty-aware pseudo-labeling and geometric feature relation transfer. Additionally, Cai et al. [12] introduced Cross-Point Embeddings (CUE) for uncertainty estimation in 3D dense prediction tasks, including semantic segmentation, by leveraging cross-point relations to quantify both epistemic and aleatoric uncertainties in a generic manner. Finally, some works [49, 69] proposed to fit a Gaussian mixture model to the network's feature space to obtain uncertainty-aware predictions, while outperforming the traditional baselines. A different concept based on Test-Time Augmentation [2] has been proposed by Vu et al. [84] as a systematic way for data augmentation at test time. Their approach (i) transforms the input point cloud via data augmentation, (ii) performs predictions on the augmented versions of the input point cloud, and (iii) combines the prediction results.

In our work, we focus on the use of different options for assessing uncertainty (in terms of method and representation) on a recent powerful 3D semantic segmentation model for analyzing large-scale remote sensing data and urban scenes. More specifically, we involve Monte Carlo Dropout [21] and Deep Ensembles [41]. For the latter, we take into account the high computational burden typically arising for large-scale 3D semantic segmentation and thus involve Deep Sub-Ensembles [80]. Finally, we also involve Test-Time Augmentation [2].

3. Methodology

In this section, we provide an overview of our framework for uncertainty-aware large-scale 3D semantic segmentation. We first describe our baseline model for which we use a Point Transformer V3 (PTv3) as the backbone [92] (Section 3.1). Then, we explain how we incorporate three complementary uncertainty mechanisms: Monte Carlo Dropout (MCD) [21] (Section 3.2), Deep Sub-Ensembles (DSE) [80] (Section 3.3), and Test-Time Augmentation (TTA) [22, 70, 91] (Section 3.4). Figure 2 provides a schematic overview of the framework and indicates how each mechanism is in-

Journal Pre-proof

tegrated into the Point Transformer V3.

3.1. Baseline Model

The Point Transformer family [92, 93, 97] comprises three successive point-based transformer backbones that progressively trade architectural complexity for efficiency and scalability while maintaining strong performance on 3D point cloud tasks.

Earlier Point Transformer variants (the original Point Transformer (PTv1) [97] and Point Transformer V2 (PTv2) [93]) introduce local vector attention and improve efficiency through grouped attention and grid-based pooling. The third variant in this family, Point Transformer V3 (PTv3) [14, 92] adopts a different design philosophy and moves toward greater simplicity and scalability by focusing on overcoming the inherent tension between accuracy and efficiency in large-scale point cloud processing. Instead of introducing a more complex attention mechanism, PTv3 updates the pipeline to allow significantly larger neighborhoods and faster computation, which is specifically suitable for our use case.

The core idea is the use of point cloud serialization. The 3D space is traversed with a space-filling curve (e.g., Z-order [54], Hilbert curve [32], and their transposed variants) to impose a structured ordering on points. Neighbors in this ordering tend to be spatially close. PTv3 uses this by replacing explicit kNN search with an efficient serialized neighbor mapping, where points attend to other points that are nearby in the sorted order. This strategy preserves locality but is far cheaper to compute than a fresh kNN search at each layer. It enables PTv3 to massively expand the attention receptive field (from 16 points in PTv2 up to 1024 points per neighborhood in PTv3) while simultaneously increasing speed and reducing memory usage. Within the resulting serialized sequence, the model then applies patch-based attention similar to windowed attention in image transformers. The ordered point list is divided into fixed-size patches, and self-attention is performed within each patch to capture local interactions efficiently. To compensate for the rigid patch partitioning, PTv3 introduces patch interaction schemes, such as shifted patches and shuffled orders, which enables information exchange between patches across layers.

On the other hand, PTv3 replaces costly RPE with an enhanced conditional positional encoding (xCPE). xCPE is implemented as a sparse convolutional layer (with a skip connection) placed before the attention block. By using xCPE, PTv3 eliminates the need to compute and lookup relative coordinate differences inside the attention module for every pair of points, substantially improving efficiency.

Overall, PTv3 retains the same high-level U-Net [68] architecture and even the grid pooling approach of PTv2, but its attention blocks are redesigned for scale. By removing explicit kNN queries and RPE and exploiting serialization and patch-based attention, PTv3 can process more points per layer than PTv2. Therefore, it expands the effective receptive field from 16 to 1024 points while being about 3× faster and over 10× more memory-efficient than PTv2. The main trade-off is that locality is enforced through serialized neighborhoods rather than exact kNN sets, i.e., strict permutation invariance and precise neighbor selection are deliberately relaxed in favor of scalability.

Training Criteria. During training, the PTv3 model is optimized using a combined objective comprising the standard cross-entropy loss [24] and the Lovász-Softmax loss [7]. This formulation promotes accurate per-point classification while additionally encouraging higher region overlap via an IoU-aligned surrogate.

Taking p as the number of points in a point cloud or minibatch, we train a multi-class classifier that assigns each point i a softmax probability $f_i(c)$ over classes $c \in C$. Using one-hot targets $y_{i,c}$, the standard objective is the cross-entropy loss [7, 24],

$$\mathcal{L}_{\text{CE}}(f) = -\frac{1}{p}\sum_{i=1}^{p}\sum_{c\in C} y_{i,c}\log f_i(c) = -\frac{1}{p}\sum_{i=1}^{p}\log f_i(y_i^*) \tag{1}$$

where y_i^* is the ground truth class of point i.

For multi-class semantic segmentation, the Lovasz-Softmax loss [7] is a differentiable surrogate designed to better align training with the Jaccard index (IoU metric) than per-point cross-entropy. For each class $c \in C$, it forms a per-point error vector $m(c) \in [0,1]^p$ from the predicted softmax probabilities $f_i(c)$ as

$$m_i(c) = \begin{cases} 1 - f_i(c), & \text{if } c = y_i^*, \\ f_i(c), & \text{otherwise.} \end{cases} \tag{2}$$

Consequently, errors increase when confidence is low on the true class or high on an incorrect class.

The class-wise loss is then computed via the Lovasz extension $\bar{\Delta}$ of the Jaccard set loss, evaluated on $m(c) : \mathcal{L}_c = \bar{\Delta}_{J_c}(m(c))$. Concretely, the Lovasz extension is a weighted sum of the sorted errors,

$$\begin{aligned} \bar{\Delta}(m) &= \sum_{i=1}^{p} m_{\pi_i}\, g_i(m), \\ g_i(m) &= \Delta(\{\pi_1,\ldots,\pi_i\}) - \Delta(\{\pi_1,\ldots,\pi_{i-1}\}), \end{aligned} \tag{3}$$

where π ranks the components of m in decreasing order.

To align with mIoU, the final Lovasz-Softmax objective averages the class-wise surrogates over classes:

$$\mathcal{L}_{\text{Lovasz-Softmax}}(f) = \frac{1}{|C|}\sum_{c\in C}\bar{\Delta}_{J_c}\big(m(c)\big) \tag{4}$$

Journal Pre-proof

Uncertainty Measure. To quantify uncertainty in semantic segmentation, we use entropy computed from the model's predicted softmax probabilities $f_i(c)$ [36],

$$H(f_i) = -\sum_{c \in C} f_i(c)\ \log(f_i(c)), \tag{5}$$

where high entropy indicates a less confident prediction. In contrast to variance or standard deviation, which require samples from multiple stochastic forward passes, entropy can be obtained directly from a single forward pass of the baseline model. For uncertainty quantification methods that generate multiple samples, we use the average of the predicted softmax probabilities across all samples to compute the entropy, enabling a consistent comparison between the baseline and uncertainty quantification methods.

3.2. Monte Carlo Dropout

Monte Carlo Dropout (MCD) [21, 73] is a widely used uncertainty quantification technique that reinterprets standard dropout (primarily used as a regularization technique for fine-tuning machine learning models) as approximate Bayesian inference in deep neural networks. In this framework, training a network with dropout layers corresponds to performing variational inference that approximates a stochastic Gaussian process. To obtain predictive uncertainty, dropout is applied not only during training, but is also kept active at test time, so that each forward pass uses a different random dropout mask. By running multiple stochastic forward passes (N) for the same input, MCD generates an ensemble of predictions from a single trained model that approximates samples from the posterior predictive distribution. The mean of these predictions is taken as the final output, and used to compute model's final uncertainty. Although these multiple forward passes increase computational cost, MCD is straightforward to implement and has become one of the commonly applied choices for uncertainty quantification.

Monte Carlo Dropout's behavior depends on several design choices: (i) how many dropout layers are used and where they are located in the network, (ii) the dropout rate (the probability of randomly deactivating different sets of neurons), and (iii) the number of stochastic forward passes N used to estimate predictive mean and entropy. In PTv3, dropout layers are used in various parts through its architecture, including dropout on attention weights (i.e., randomly masking a fraction of attention links), and projection and MLP dropouts (i.e., feature-level dropout in both attention output and feed-forward network). We use these implemented layers and keep them active during both training and inference. Moreover, we consider dropout rates of 10%, 20%, and 50%, and perform ten stochastic forward passes to compute the final prediction and its associated uncertainty [21, 28, 42].

3.3. Deep Sub-Ensemble

Deep Ensembles [41] are among the most well-known and powerful methods for estimating a model's uncertainty. They train and evaluate multiple completely independent networks, which is computationally demanding. In order to reduce the high training and inference cost, Deep Sub-Ensembles (DSE) [80] are introduced as a simplification of Deep Ensembles, as depicted in Figure 3. Rather than replicating an entire network, the model is split into two sub-networks: a shared "trunk" network T and a "task" network K. Thus, one full instance is trained end-to-end, the trunk is then frozen, and $M-1$ additional independently initialized task heads are trained on the same data (M is the number of ensemble members). During inference, the trunk runs once for input, each head produces a prediction, and the outputs are combined by averaging, yielding uncertainty estimates similar to Deep Ensembles but with substantially lower computation.

We consider the encoder of our baseline model as the shared trunk network, and its decoder as a task network. We then implement the Deep Sub-Ensemble method by attaching five independently initialized decoder heads to the shared backbone. During training, we cycle through the decoder heads to encourage diversity across them while reducing training time. Therefore, in each minibatch, only a single decoder head is activated and updated, while the remaining heads stay inactive for that iteration.

We involve three different configurations: (i) Deep Sub-Ensemble (DSE) in which the encoder (trunk) and all decoder heads are trained end-to-end, (ii) Deep Sub-Ensemble - frozen encoder (DSE-FE) in which we first train a single-head baseline model to convergence, then freeze the encoder and continue training the multiple decoder heads on top of the fixed features, and (iii) DSE-FE-TTA in which we apply Test-Time Augmentation, which will be explained in Section 3.4, in addition to DSE-FE.

During inference, each decoder head produces a separate prediction from the shared encoder features. Consequently, we combine the head outputs by computing their mean and quantify uncertainty using predictive entropy of this mean distribution, yielding both the final segmentation and an associated uncertainty estimate.

3.4. Test-Time Augmentation

Test-Time Augmentation (TTA) [2, 22, 70, 91] is a comparatively simple strategy for estimating predictive uncertainty, which operates entirely at inference time. Instead of evaluating each test sample only once, multiple augmented variants of the same input are generated using standard data augmentation operations (e.g., intensity or geometric perturbations, flips, rotations, etc.), and the model is run on all of them to obtain a set of predictions. All these predictions give us a distribution of possible outputs, which we can then

Journal Pre-proof

use to quantify how uncertain the model is. By exposing the model to different "views" of the same underlying sample, TTA captures uncertainty, tests the stability of its outputs, and, in practice, often leads to better-calibrated probability estimates.

The approach has been particularly popular in medical image analysis [2, 55, 86], where data augmentation is already a core component of training pipelines and annotated data is expensive to obtain. This makes it more convenient to reuse these transformations at test time to quantify uncertainty. Remote sensing and earth observation is another domain which TTA is effectively employed [58].

Overall, TTA is appealing because it leaves the underlying network architecture and weights unchanged and does not require additional labeled data. It can be implemented straightforwardly using standard augmentation techniques to generate multiple predictions, which are then used to assess how certain the final prediction is.

4. Experimental Results

In this section, we describe the experimental setup used to evaluate the proposed methods. We first present the datasets (Section 4.1) and the data augmentation strategies applied during training and inference (Section 4.2). We then outline the implementation details (Section 4.3), including network configuration and training procedure. Subsequently, we present the evaluation metrics used to assess both predictive performance and uncertainty estimates (Section 4.4). Finally, we report and analyze the experimental results based on these metrics (Sections 4.5- 4.6), and we perform an ablation regarding the most critical design choices (Section 4.7).

4.1. Datasets

We conduct the experiments on two LiDAR datasets, Hessigheim 3D [37] and DALES [81]. The Hessigheim 3D (H3D) benchmark targets high-resolution semantic segmentation of geospatial 3D data over the village of Hessigheim, Germany. Data was collected across four epochs (March 2016, March and November 2018, March 2019). H3D consists of 11 classes: low vegetation, impervious surfaces, vehicles, urban furniture, roofs, facades, shrubs, trees, soil/gravel, vertical surfaces, and chimneys. For our experiments, we use the LiDAR data from the March 2018 epoch. The official split comprises approximately 59.4 M, 14.5 M, and 51.7 M points for training, validation, and test, respectively.

The Dayton Annotated LiDAR Earth Scan (DALES) dataset is another large-scale, high-density airborne LiDAR benchmark designed for 3D semantic segmentation in complex urban scenes. DALES is partitioned into 40 non-overlapping tiles (each approximately 12 M points, 0.5 km^2, 50 points/m^2). Points are annotated with eight classes: ground, vegetation, cars, trucks, poles, power lines, fences, and buildings. Two dataset variants are provided: the original DALES (semantic labels) and DALES Objects [71], which adds instance labels and intensity. In our experiments, we use the original DALES format and follow the standard split: 29 scenes for training and 11 for testing, reserving 5 of the training scenes for validation.

For both datasets, we only consider spatial 3D coordinates as input to our framework.

4.2. Data Augmentations

During training, we apply a standard 3D data augmentation pipeline to improve generalization and introduce controlled randomness, following [92]:

- Random rotation (RandomRotate): with probability 0.5, the point cloud is rotated around the z-axis, reducing sensitivity to small orientation changes.
- Random scaling (RandomScale): point coordinates are uniformly scaled by a factor sampled from [0.9,1.1], improving robustness to moderate size variations.
- Random flipping (RandomFlip): with probability 0.5, the point cloud is randomly mirrored along the x and/or y axes, helping the model handle symmetric variations.
- Random jittering (RandomJitter): Gaussian noise with $\sigma = 0.005$, clipped at ± 0.02, is added to the coordinates to simulate local perturbations and measurement noise.

Together, these augmentations improve the model's robustness to changes in orientation, scale, symmetry, and local noise, leading to better generalization.

During inference, we follow the TTA setting suggested by the paper [92] and apply five fixed scalings ($s \in \{0.9, 0.95, 1.0, 1.05, 1.1\}$) as well as the same scalings followed by a deterministic flip of both x and y.

4.3. Setting/Implementation Details

In all experimental settings, we use a batch size of 12 and train for 500 epochs to ensure a fair comparison between the baseline and all uncertainty estimation variants. Based on the training setup of Point Transformer V3 [92], we optimize the network using AdamW [50] with an initial learning rate of 0.002 and a weight decay of 0.005, while transformer block parameters are assigned a ten-times smaller learning rate. The learning rate is further controlled by a OneCycleLR scheduler [72] with cosine annealing. Across all the experiments, the model takes only 3D coordinates as input features, without using any color or intensity information. To handle the very large input point clouds, we partition them into smaller chunks for memory efficiency. In

Journal Pre-proof

Method	mIoU ↑	mF1 ↑	mAcc ↑	ECE ↓	ACE ↓	$p(acc\|cer)$ ↑	$p(unc\|inacc)$ ↑	Latency (s/scene)
Baseline	0.629	0.749	0.731	0.123	0.189	0.879	0.692	0.224
Baseline-TTA	0.646	0.763	0.744	0.117	0.195	0.882	0.699	4.54
MCD (10%)	0.630	0.745	0.733	0.124	0.188	0.879	0.680	2.12
MCD (10%)-TTA	0.641	0.756	0.745	0.111	0.187	0.888	0.694	44.76
MCD (20%)	0.618	0.741	0.724	0.108	0.187	0.885	0.684	2.15
MCD (50%)	0.616	0.736	0.725	0.092	0.159	0.888	0.722	2.16
DSE	0.615	0.738	0.729	0.096	0.174	0.905	**0.737**	0.430
DSE-FE	0.643	0.764	0.751	0.054	0.116	**0.906**	0.727	0.431
DSE-FE-TTA	**0.652**	**0.772**	**0.759**	**0.050**	**0.115**	0.904	0.733	8.56

Table 1. Results of various approaches for 3D semantic segmentation and uncertainty quantification on the Hessigheim 3D dataset.

Method	LVeg	ISurface	Vehicle	UFurn.	Roof	Facade	Shrub	Tree	Soil	VSurface	Chimney	mF1
Baseline	0.821	0.776	0.723	0.588	0.934	0.802	0.723	0.971	0.276	0.738	0.889	0.749
Baseline-TTA	0.835	0.814	0.733	0.597	0.937	0.812	**0.730**	**0.972**	0.308	0.762	**0.899**	0.763
MCD (10%)	0.830	0.787	**0.757**	0.595	0.938	0.814	0.705	**0.972**	0.167	0.739	0.887	0.745
MCD (10%)-TTA	0.834	0.808	0.749	0.599	0.938	**0.823**	0.712	**0.972**	0.216	0.765	0.896	0.756
MCD (20%)	**0.841**	0.798	0.662	0.575	0.931	0.813	0.715	0.970	0.293	0.709	0.844	0.741
MCD (50%)	0.815	0.790	0.683	0.566	0.933	**0.823**	0.689	0.968	0.236	0.711	0.889	0.736
DSE	0.823	**0.826**	0.679	0.545	0.933	0.794	0.665	0.970	0.303	0.755	0.830	0.738
DSE-FE	0.839	0.814	0.694	0.610	**0.943**	0.813	0.716	0.971	0.379	0.770	0.857	0.764
DSE-FE-TTA	0.833	0.822	0.701	**0.617**	0.938	0.820	0.728	**0.972**	**0.400**	**0.779**	0.880	**0.772**

Table 2. Class-wise segmentation results for the Hessigheim 3D dataset in terms of F1 scores. LVeg. refers to low vegetation, ISurface to impervious surface, UFurn. to urban furniture, and VSurface to vertical surface.

addition, grid sampling is applied as a voxel-based downsampling step with a grid size of 0.05 m. The point coordinates are discretized into occupied grid cells, reducing point density and computational cost while preserving the overall spatial structure of the point cloud. During training, one representative point is randomly selected from each occupied cell, whereas during test time, the grid cells are used to generate structured point partitions for inference. All experiments are carried out on four NVIDIA A100 GPUs, and CloudCompare [23] is employed for visualization throughout the study.

4.4. Metrics

For quantitative evaluation of semantic segmentation, we use the mean Intersection-over-Union (mIoU; IoU is also known as the Jaccard Index), the mean F1 (mF1) score, and the mean accuracy (mAcc) as our primary metrics. In addition, we report the per-class F1 scores. To assess the calibration of the predicted softmax probabilities, we compute the expected calibration error (ECE) [26, 57] and adaptive calibration error (ACE) [61]. ECE reports the sample-weighted average of the per-bin mismatch between mean confidence and empirical accuracy over B equally-spaced confidence bins, whereas ACE adaptively partitions predictions into R bins with variable widths but an equal number of samples per bin. Moreover, we adopt the uncertainty evaluation measures proposed by Mukhoti and Gal [56]:

1. $p(acc|cer)$: the probability that a prediction is correct given that its uncertainty lies below a chosen threshold.

2. $p(unc|inacc)$: the probability that the model's uncertainty exceeds a chosen threshold given that the prediction is incorrect.

To simplify the setup and better reflect a practical deployment scenario, we set the uncertainty threshold in all experiments to the median uncertainty across all valid points in each point set [43]. Importantly, this threshold is used only to convert continuous uncertainty values into a binary partition of "certain" and "uncertain" predictions for the computation of $p(acc|cer)$ and $p(unc|inacc)$. It does not modify the predicted labels or softmax probabilities and therefore does not affect mIoU, mF1, mAcc, ECE, or ACE. Among commonly used thresholding choices such as the mean and median [43], we use the median uncertainty as the decision threshold for uncertainty evaluation metrics because it provides a robust and balanced split between lower and higher uncertainty predictions without requiring a manually tuned and method-specific cutoff. In practice, uncertainty distributions in deep neural networks are often right-skewed, meaning that many predictions exhibit low uncertainty while a smaller subset shows very high

Journal Pre-proof

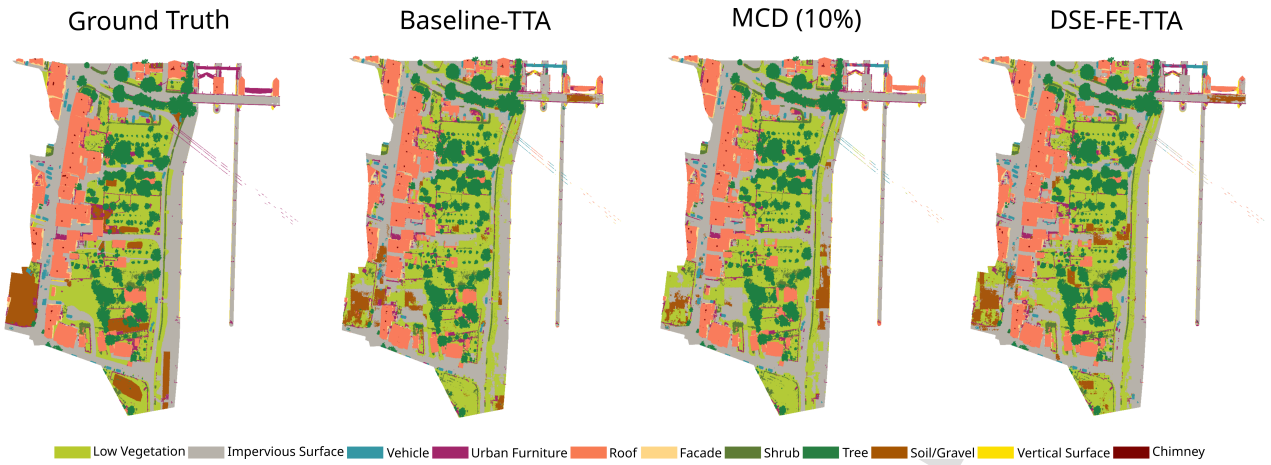


Figure 4. Labels for the H3D dataset in BEV.

uncertainty. As a result, the mean can be shifted by relatively few highly uncertain predictions. This leads to a more permissive threshold, under which moderately uncertain errors may still be counted as "certain", thereby increasing the risk of overconfident mistakes. In contrast, the median is less sensitive to such outliers and therefore provides a more stable and fair basis for comparing uncertainty-aware methods.

In addition to segmentation, calibration, and uncertainty metrics, we report inference latency as an efficiency-related measure. Latency is reported in s/scene and corresponds to the average forward-only inference time per test scene. Because each scene is processed fragment-wise, this value includes the summed forward time over all fragments required to produce the final scene prediction. Measurements were performed with batch size 1 on a single GPU. For TTA-based methods, the reported latency includes all augmented forward passes.

4.5. Quantitative Results

For our experiments, we involve nine configurations of our framework that correspond to three method categories: baseline variants (with/without TTA), MCD variants (10%, 20%, 50%, 10%-TTA), and Deep Sub-Ensembles variants (DSE, DSE-FE, DSE-FE-TTA). These configurations are applied on both the H3D and DALES datasets and are evaluated using segmentation metrics (mIoU, mF1, mAcc), calibration metrics (ECE, ACE), and uncertainty evaluation measures ($p(acc|cer)$, $p(unc|inacc)$).

H3D. The results shown in Table 1 confirm that adding TTA consistently improves all segmentation metrics (mIoU, mF1, mAcc) and reduces ECE compared to the same method evaluated without TTA, as well as the baseline. The strongest overall performance in terms of segmentation and calibration is achieved by DSE-FE-TTA (highest segmentation scores and lowest calibration error). Table 2 summarizes the per-class F1 scores for H3D.

Regarding uncertainty measures, TTA generally improves $p(unc|inacc)$ compared to the corresponding no-TTA setting. DSE variants outperform both baseline variants and MCD variants and improve the important metric $p(unc|inacc)$ by about 3 to 4 percentage points relative to the baseline. Specifically, the DSE method attains the highest $p(unc|inacc)$ (0.737), i.e., it is comparatively better at being uncertain when wrong. In contrast, MCD exhibits a clear sensitivity to the dropout rate. Increasing dropout rate from 10% to 50% improves uncertainty ($p(acc|cer)$: 0.879→0.888, $p(unc|inacc)$: 0.680→0.722) and reduces calibration error (ECE: 0.124→0.092, ACE: 0.188→0.159), but degrades segmentation performance (mIoU: 0.630→0.616, mF1: 0.745→0.736).

Overall, the DSE-FE-TTA method shows the most consistent gains over the baseline on H3D, improving segmentation (mIoU: 0.629→0.652, mF1: 0.749→0.772, mAcc: 0.731→0.759), while reducing calibration error (ECE: 0.123→0.050, ACE: 0.189→0.115). It also improves uncertainty metrics, increasing ($p(acc|cer)$ by about 2.5 percentage points and $p(unc|inacc)$ by about 4 percentage points.

DALES. Table 3 shows that introducing TTA consistently increases segmentation metrics (mIoU, mF1) and keeps ECE almost untouched relative to the corresponding no-TTA setting and the baseline. The baseline-TTA achieves the strongest segmentation results on DALES, with the highest mIoU, mF1, and mAcc, while keeping ECE essentially unchanged relative to the baseline. This indicates that when the primary objective is segmentation quality, baseline-TTA is the most effective choice on DALES. In

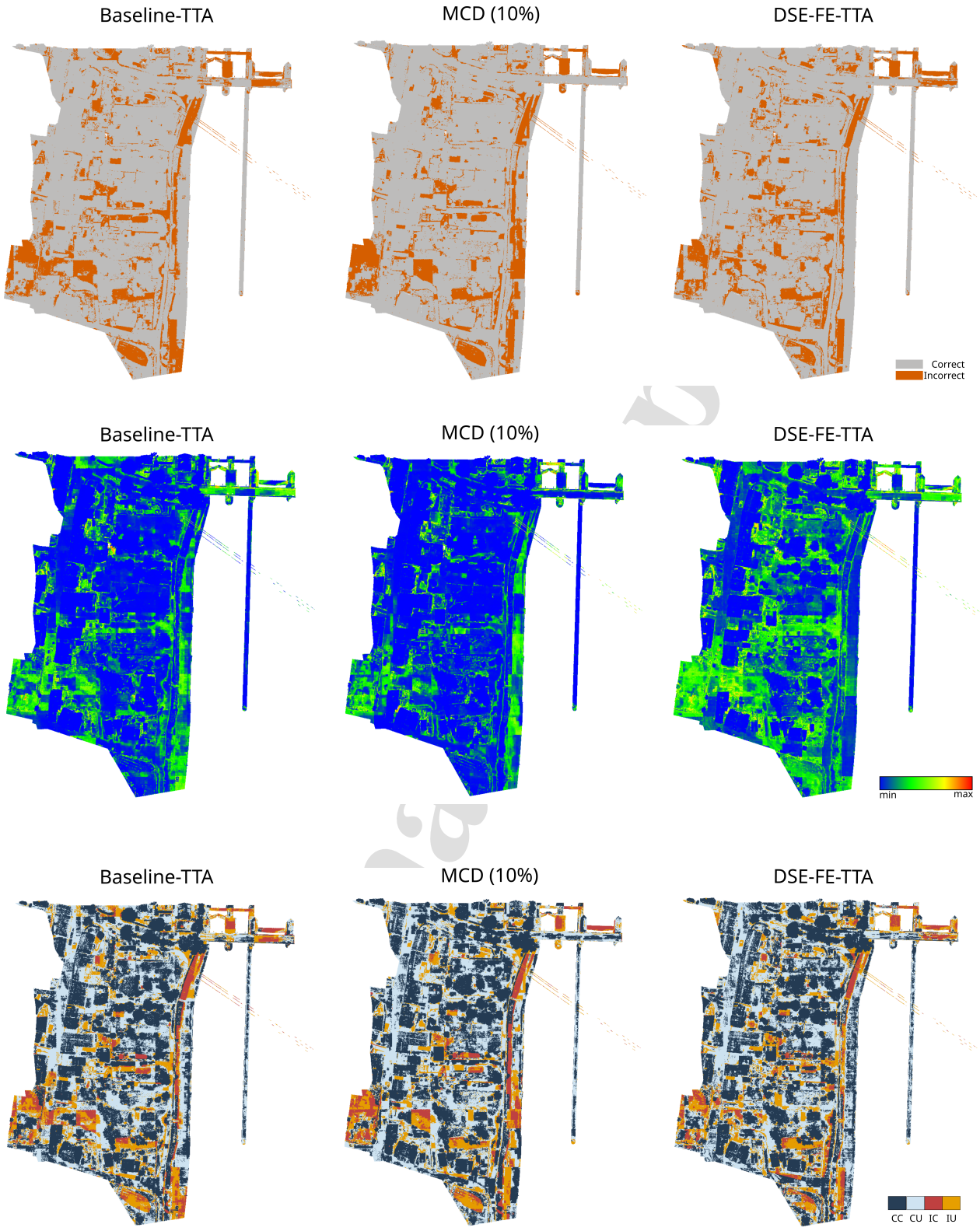


Figure 5. Accuracy map (first row), uncertainty map (second row), and CC/CU/IC/IU representation (third row) in BEV for the H3D dataset.

Journal Pre-proof

Method	mIoU ↑	mF1 ↑	mAcc ↑	ECE ↓	ACE ↓	$p(acc\|cer)$ ↑	$p(unc\|inacc)$ ↑	Latency (s/scene)
Baseline	0.656	0.769	0.747	0.021	0.198	0.992	0.935	0.334
Baseline-TTA	**0.675**	**0.784**	**0.752**	0.021	0.205	0.992	0.935	3.45
MCD (10%)	0.659	0.770	0.733	0.020	0.201	0.991	0.928	3.63
MCD (10%)-TTA	0.674	0.782	0.739	0.021	0.208	0.990	0.922	35.90
MCD (20%)	0.657	0.768	0.736	0.020	0.198	0.991	0.927	3.50
MCD (50%)	0.653	0.766	0.729	0.017	**0.195**	0.991	0.932	3.53
DSE	0.651	0.762	0.727	0.020	0.212	**0.993**	0.946	0.702
DSE-FE	0.647	0.760	0.724	**0.016**	0.197	**0.993**	**0.948**	0.690
DSE-FE-TTA	0.665	0.775	0.732	**0.016**	0.203	**0.993**	0.945	7.03

Table 3. Quantitative comparison of different approaches for 3D semantic segmentation and uncertainty quantification on the DALES dataset.

Method	Ground	Vegetation	Cars	Trucks	Powerlines	Fences	Poles	Buildings	mF1
Baseline	0.959	0.933	0.842	0.422	0.812	0.697	0.574	0.912	0.769
Baseline-TTA	**0.963**	0.938	0.858	**0.430**	0.832	0.715	0.614	**0.921**	**0.784**
MCD (10%)	0.958	0.934	0.851	0.383	0.820	0.702	0.601	0.909	0.770
MCD (10%)-TTA	0.962	**0.939**	**0.863**	0.404	**0.833**	**0.719**	**0.616**	0.918	0.782
MCD (20%)	0.957	0.933	0.850	0.388	0.818	0.700	0.593	0.906	0.768
MCD (50%)	0.957	0.932	0.850	0.410	0.810	0.698	0.564	0.906	0.766
DSE	0.958	0.933	0.848	0.342	0.810	0.697	0.595	0.909	0.762
DSE-FE	0.957	0.932	0.845	0.397	0.803	0.692	0.549	0.904	0.760
DSE-FE-TTA	0.961	0.937	0.859	0.420	0.820	0.710	0.577	0.915	0.775

Table 4. Class-wise segmentation results (F1 scores) for the DALES dataset.

contrast, MCD variants degrade segmentation performance as the dropout rate increases, while improving calibration errors. DSE-FE-TTA improves mIoU and mF1 compared to the baseline approach. Table 4 presents per-class segmentation metric results (F1) for the DALES dataset.

Considering uncertainty evaluation measures, adding TTA typically leaves uncertainty metrics in a similar range compared to the corresponding no-TTA setting. The MCD variants are sensitive to the chosen dropout rate, and increasing the dropout rate typically raises some uncertainty metrics such as $p(unc|inacc)$. The DSE variants provide the most consistent uncertainty improvements, increasing $p(acc|cer)$ and $p(unc|inacc)$ while also achieving the lowest ECE, likely because decoder head diversity produces more meaningful disagreement in hard regions. Overall, DSE-FE-TTA offers a decent combination of segmentation quality and uncertainty estimates on DALES, by slightly improving $p(unc|inacc)$ and $p(acc|cer)$ relative to the baseline while also increasing mIoU and mF1.

Latency. From a computational perspective, the latency results across both datasets highlight the trade-off of uncertainty-aware inference. The baseline model is the fastest configuration, while DSE-based variants remain comparatively efficient, with only a moderate increase in latency relative to the baseline. In contrast, MCD leads to a substantially higher inference cost because it relies on multiple stochastic forward passes, whereas varying the dropout rate has only a negligible effect on runtime. The highest latency is observed for TTA-based variants, particularly when combined with MCD, since the final prediction is obtained from many repeated augmented evaluations. Overall, these results indicate that the improved uncertainty behavior of MCD and TTA-based methods comes at a clear computational cost, whereas DSE-based variants offer a more favorable trade-off between uncertainty quality and inference efficiency.

4.6. Qualitative Results

For qualitative comparison, we select one representative method from each category (baseline, MCD, and sub-ensemble), namely baseline-TTA, MCD (10%), and DSE-FE-TTA. We visualize the final results in terms of predicted labels, accuracy maps, uncertainty maps, and combined accuracy-uncertainty representations (CC/CU/IC/IU). CC, CU, IC, and IU represent correctly predicted and certain, correctly predicted but uncertain, incorrectly predicted but certain, and incorrectly predicted and uncertain point labels, respectively.

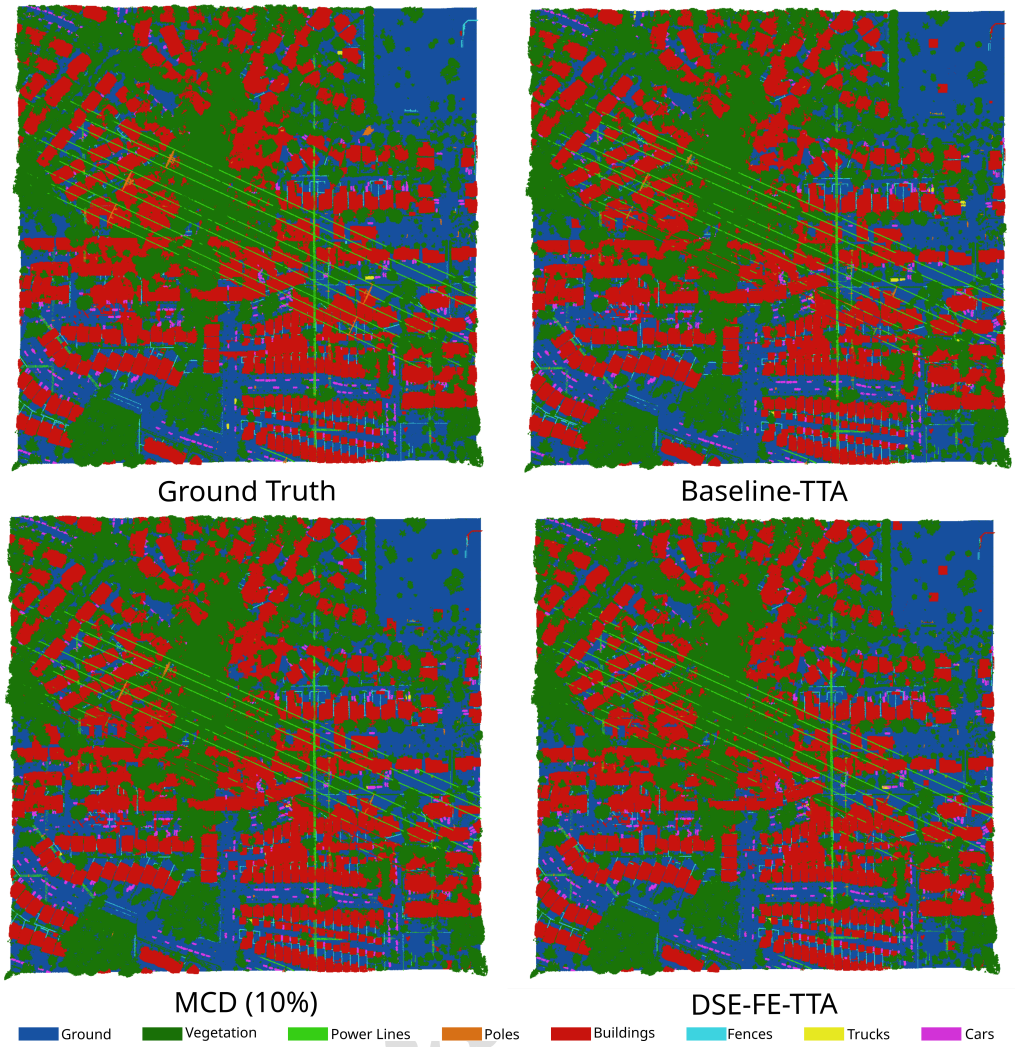


Figure 6. Labels for the DALES dataset in BEV.

H3D. Figure 4 and Figure 8 in the Appendix present final predicted labels for different methods on H3D, while incorrect labels and different measures of uncertainty are visualized in Figure 5 and Figures 9-11 in the Appendix. All three approaches could recover the global scene layout reasonably well. Dominant structures such as building blocks (roof/facade), tree-covered areas, and major ground regions (impervious surface and low vegetation) are spatially consistent with the ground truth. The remaining errors concentrate in failure modes expected for geometry-only input data. In particular, confusions occur between geometrically similar ground classes (impervious surface/low vegetation/soil or gravel), at class boundaries where multiple materials meet, and for thin/rare objects (e.g., urban furniture, vehicles, vertical surfaces). Moreover, regions with vertical overlap of vegetation (low vegetation/shrub/tree) increase label ambiguity because points from different semantic classes coexist within a small horizontal footprint but at different heights. Overall, DSE-FE-TTA seems to improve spatial consistency, as expected from the effect of averaging multiple decoder heads on top of a shared, high-quality encoder representation.

The uncertainty visualizations in Figure 5 and Figure 10 in the Appendix report predictive entropy (blue=low uncertainty, yellow/red=high uncertainty). Baseline-TTA reveals predominantly low-entropy over most of the scene, suggesting that the model is often highly confident, including in heterogeneous regions where class confusion is likely. By comparison, DSE-FE-TTA produces a more structured and spatially aligned uncertainty. Low entropy is preserved on clearly separable regions, while uncertainty increases over challenging ground-like areas and along

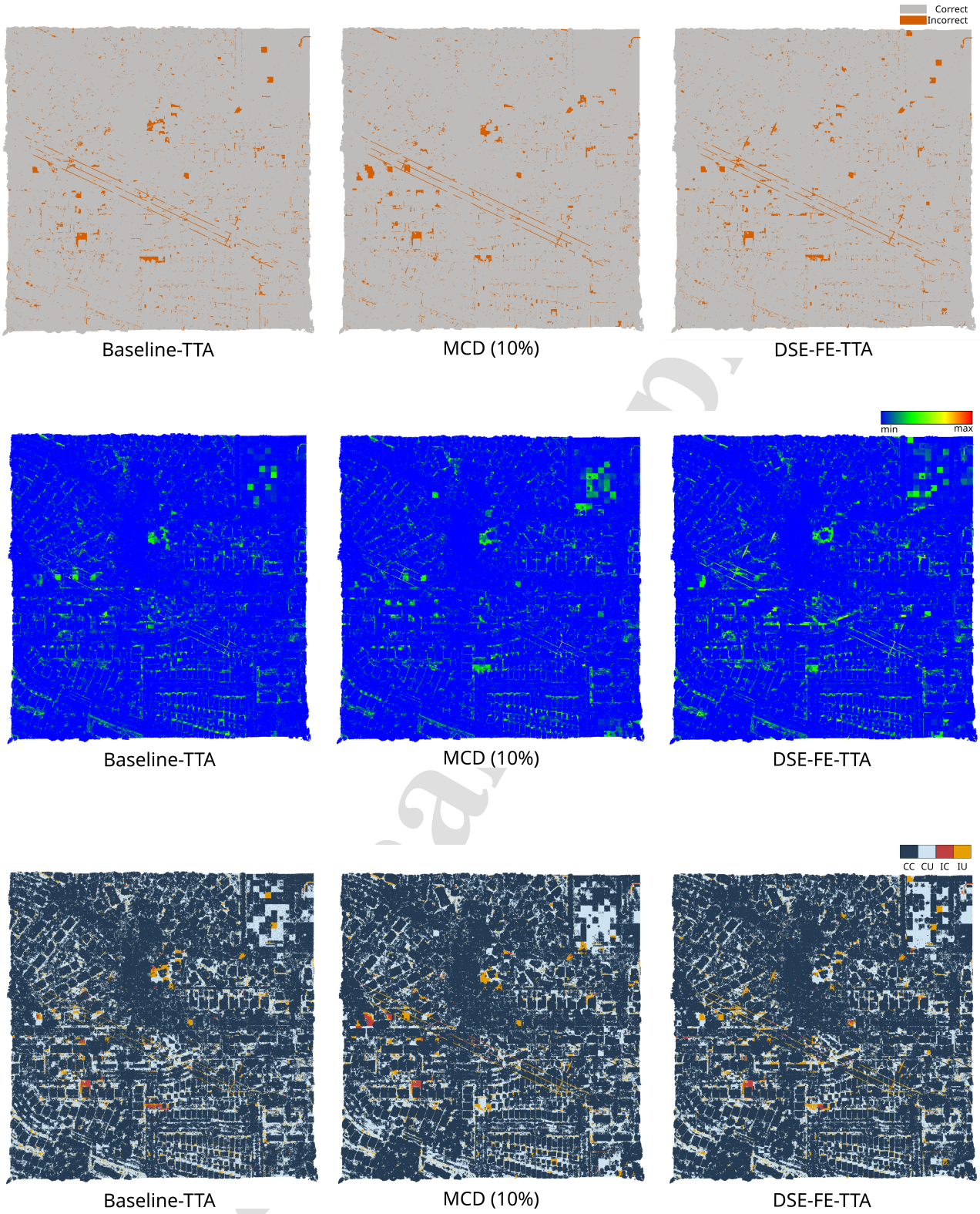


Figure 7. Accuracy map (first row), uncertainty map (second row), and CC/CU/IC/IU representation (third row) in BEV for the DALES dataset.

Journal Pre-proof

complex interfaces (e.g., impervious surface/vegetation/soil transitions and mixed vegetation), indicating reduced overconfidence on such difficult points. This qualitative behavior matches the quantitative trends on H3D, where DSE-FE-TTA achieves relatively better calibration and uncertainty evaluation measures while also improving segmentation performance.

Finally, the accuracy-uncertainty breakdown in Figure 5 and Figure 11 in the Appendix further supports this interpretation. DSE-FE-TTA shows relatively fewer inaccurate but certain predictions (IC, highlighted in red color) than baseline-TTA and MCD, meaning that its mistakes are more often accompanied by higher uncertainty. This is precisely the behavior desired for reliable uncertainty-aware deployment. Accuracy maps for different approaches are also shown in Figure 5 and Figure 9 in the Appendix. An accuracy map indicates, for each point, whether the predicted label matches the ground truth (correct vs. incorrect).

DALES. In Figure 6 and Figure 16 in the Appendix, the qualitative predictions of all three methods are largely consistent with the ground truth at the scene level, correctly capturing the main spatial extent of the dominant classes (buildings, vegetation, and ground). Most visible discrepancies arise in the typical DALES challenging regions, including thin linear structures (power lines), and small or rare classes (poles, fences, trucks). Among the compared methods, baseline-TTA appears visually closest to the ground truth overall, with slightly cleaner region boundaries and fewer fragmented patches, which is consistent with its best quantitative segmentation scores on DALES. DSE-FE-TTA produces similarly coherent outputs for the major classes and remains competitive on challenging structures (e.g., fences and power lines). Overall, the differences are subtle in terms of segmentation quality, matching the relatively small performance gaps reported in the DALES quantitative results.

The DALES uncertainty maps in Figure 7 and Figure 18 in the Appendix are dominated by low uncertainty (blue=low uncertainty, yellow/red=high uncertainty) across all three methods, indicating that the models are generally confident on the major, well-separated classes. Elevated entropy concentrates in the expected challenging regions (such as thin structures, small/rare classes, and zones where classes meet and local geometry becomes ambiguous). Baseline-TTA stays the most uniformly "blue", suggesting very confident predictions almost everywhere. This is plausible on DALES because most of the scene is composed of large, easily separable regions, but it also risks being too confident on fine structures and boundary noise. MCD(10%) introduces more spatially scattered moderate uncertainty. DSE-FE-TTA produces the most structured uncertainty, with clearer uncertainty traces along the difficult regions and linear corridors, which aligns with its more consistent uncertainty usefulness behavior on DALES even though the qualitative differences between methods remain relatively subtle.

Figure 7 and Figure 19 in the Appendix complement the quantitative results by showing that all three methods behave similarly at a global scale, which is expected given the small numerical gaps on DALES. Among the compared approaches, DSE-FE-TTA exhibits slightly fewer red IC regions and a correspondingly higher tendency to flag mistakes with uncertainty, consistent with its marginally higher uncertainty reliability (about 1 percentage point improvement in $p(unc|inacc)$ relative to the baseline variants). However, the overall differences remain modest, reflecting that DALES is dominated by large, easy regions where all methods are already confident and accurate. Therefore, uncertainty quantification gains primarily manifest as small reductions in confident errors on rare or challenging structures rather than large shifts in aggregate segmentation metrics. Finally, Figure 7 and Figure 17 in the Appendix show the corresponding accuracy maps for the different approaches on DALES.

4.7. Backbone Comparison and Baseline Selection

To clarify the choice of Point Transformer V3 (PTv3) as the backbone for the uncertainty analysis, in our first ablation study we additionally compare several representative segmentation architectures under the same experimental protocol described in Sections 4.2-4.4, namely SparseUNet [15, 25], OctFormer [87], KPConvX [79], and PTv3 [92]. As shown in Table 5, PTv3 achieves the strongest overall performance on the Hessigheim 3D dataset in our setup, yielding the highest mF1 and mIoU among the compared backbones, while also providing competitive or leading class-wise results for several categories. KPConvX remains competitive on some classes. However, in our experiments, it requires substantially more architecture-specific hyperparameter tuning and shows stronger GPU memory limitations, restricting the number of points that could be processed per sample. SparseUNet and OctFormer provide useful reference points, but yield lower overall performance under the same data split, input features, augmentation strategy, and optimization protocol. Based on this comparison, we selected PTv3 as a strong and scalable baseline so that the effect of the uncertainty quantification methods could be analyzed without conflating it with differences in backbone architecture. This choice should therefore be understood as a pragmatic and experimentally supported baseline selection for uncertainty analysis within our setup, rather than as a claim of universal superiority across all architectures or settings.

In the second ablation study, we investigate the effect of different loss functions on the performance of PTv3 for

Journal Pre-proof

Method	LVeg	ISurface	Vehicle	UFurn.	Roof	Facade	Shrub	Tree	Soil	VSurface	Chimney	mF1	mIoU
SparseUNet [25]	0.821	0.784	0.710	0.504	0.956	0.817	0.649	0.965	0.309	0.740	0.833	0.735	0.612
OctFormer [87]	0.811	0.843	0.620	0.473	0.860	0.720	0.643	0.957	0.300	0.577	0.775	0.689	0.549
KPConvX [79]	0.829	0.847	0.670	0.493	0.957	0.808	0.621	0.962	0.381	0.810	0.799	0.743	0.621
PTv3 [92]	0.821	0.776	0.723	0.588	0.934	0.802	0.723	0.971	0.276	0.738	0.889	0.749	0.629

Table 5. Comparison of class-wise F1-scores for 3D semantic segmentation achieved by various backbones on the Hessigheim 3D dataset, along with the corresponding mF1 and mIoU.

	Validation		Test	
Loss Combination	mIoU	mF1	mIoU	mF1
Focal	0.503	0.614	0.535	0.652
Focal+Dice	0.457	0.563	0.483	0.588
Focal+CE+Lovasz	0.597	0.713	0.605	0.730
Focal+Dice+CE+Lovasz	0.532	0.639	0.543	0.659
CE+Lovasz	0.592	0.718	0.622	0.751

Table 6. Effect of different loss functions on the performance of 3D semantic segmentation using PTv3 on the Hessigheim 3D dataset. CE denotes cross-entropy loss.

3D semantic segmentation. To address the class imbalance issue in the datasets, we additionally evaluate several loss formulations, including focal loss [48], which is often beneficial in imbalanced settings because it down-weights easy examples and emphasizes harder or underrepresented classes, as well as dice loss [52] applied individually and in combination with cross-entropy and Lovasz losses. This comparison is intended as a practical loss ablation under a fixed training setup rather than an exhaustive optimization of all candidate loss formulations and their relative weighting. Table 6 shows that these alternatives do not lead to improved semantic segmentation performance compared to the combination of cross-entropy and Lovasz losses used in our study. This finding is consistent with the original Point Transformer V3 paper [92], where the combination of cross-entropy and Lovasz loss is shown to be effective across multiple datasets with varying characteristics. We therefore retain this loss formulation in the final model.

5. Discussion

The results highlight that uncertainty-aware methods can reduce overconfident errors and improve uncertainty reliability and calibration, while also delivering segmentation gains over the baseline, although the magnitude varies across mechanisms.

TTA can be useful for deriving uncertainty from the dispersion of predictions under various augmentations, as well as improving robustness by averaging predictions across geometric perturbations. The impact of TTA is more noticeable in increasing segmentation performance and has smaller effects on uncertainty metrics. Moreover, On H3D, adding TTA reduces ECE relative to the no-TTA setting, while on DALES, it remains almost unchanged. The main drawback of TTA is its inference overhead, i.e., prediction time scales approximately with the number of augmentations.

MCD injects stochasticity at inference time and aggregates multiple forward passes, and its behavior and performance are highly dependent on its dropout rate. Among MCD variants, increasing the dropout rate can strengthen uncertainty usefulness and reduce calibration error, but this comes with a clear accuracy penalty and degradation of the segmentation results.

DSE variants use a shared encoder and employ multiple decoder heads to produce diverse predictions. The results show that the frozen-encoder variant with TTA (DSE-FE-TTA) is consistently very reliable in terms of uncertainty behavior, especially on the H3D dataset where ambiguity is higher due to geometry-only input, more fine-grained classes, and stronger class overlap. Qualitatively, predictions of DSE-FE-TTA are more spatially consistent, and their uncertainty maps are more structured, i.e., uncertainty increases where it should (impervious surface/vegetation/soil transitions and mixed vegetation), rather than remaining uniformly low. This behavior is also reflected in the CC/CU/IC/IU visualizations, where the DSE-FE-TTA method reduces the most undesirable failure mode (IC, inaccurate but certain) predictions, relative to baseline-TTA and MCD. This is exactly the behavior required for reliable uncertainty-aware deployment. On DALES, although the baseline-TTA wins the highest segmentation performance, DSE-FE-TTA can still slightly improve segmentation quality relative to the baseline (no-TTA) while yielding more useful uncertainty estimates. Overall, DSE-FE-TTA provides the most robust overall compromise and gains for the H3D and DALES dataset compared to the baseline. From a practical deployment perspective, the latency results indicate that repeated inference strategies, particularly MCD and MCD-TTA, impose a substantial computational burden, whereas DSE-based variants offer a more favorable balance between uncertainty quality and inference efficiency.

A key takeaway is that on the H3D dataset, uncertainty-

Journal Pre-proof

aware methods yield clearer improvements in terms of calibration and uncertainty estimates because the dataset contains more challenging regions in the available input features (e.g., similar ground-related classes, class overlap, boundary clutter, and vegetation stratification) and provides less complementary information (e.g., missing RGB). DALES, in contrast, is dominated by large, easily separable regions (ground/vegetation/buildings), so all methods appear confident over most points and the remaining uncertainty improvements mainly concentrate on thin structures and small or rare classes. As a result, quantitative gaps in uncertainty metrics are smaller on DALES and qualitative differences are more subtle, even when incorrect-certain (IC) regions are slightly reduced. Overall, these results suggest that the practical benefit of uncertainty modeling depends on dataset characteristics and the frequency of inherently challenging and ambiguous regions.

Finally, it is important to note that this study has only investigated the reliability uncertainty estimates from predictive entropy for in-domain settings, which does not necessarily generalize to out-of-distribution behavior. Future work could improve reliability through training strategies that better handle class imbalance and rare categories, and through evaluation that emphasizes thin and rare objects, where uncertainty is most practically useful. Further investigation of loss functions specifically designed to address class imbalance is also encouraged. In addition, the effect of data chunking/tiling on uncertainty maps and calibration remains an important topic for investigation.

A further open question concerns Test-Time Augmentation. Both the type of augmentation and the number of augmented samples can influence uncertainty quality [22]. Simple transformations (e.g., reflections) may not capture meaningful variability, whereas task or domain-specific perturbations could reveal additional uncertainty, but at increased computational cost. This is an important consideration for global-scale earth observation. For Deep Sub-Ensembles, performance similarly depends on design choices. The trunk-task split location and the number of decoder heads are hyperparameters that can influence the quality of uncertainty estimates [80].

6. Conclusion

This paper investigated uncertainty-aware large-scale 3D point cloud semantic segmentation by integrating uncertainty quantification into a strong Point Transformer V3 backbone. We consider three complementary mechanisms within this framework: Monte Carlo Dropout (MCD), Deep Sub-Ensembles (DSE) variants, and Test-Time Augmentation (TTA). Using the proposed framework, we analyze their effects on segmentation performance, calibration error, and uncertainty measures.

Across two benchmarks with distinct characteristics (Hessigheim 3D and DALES), the results show that uncertainty-aware inference can reduce overconfident errors and improve reliability, with effects that depend on both the selected mechanism and dataset complexity. Compared to the baseline, the uncertainty-aware counterpart achieves gains of up to 2.3% in mIoU, 2.3% in mF1, 2.8% in mAcc, 2.7% in $p(acc|cer)$, and 4.5% in $p(unc|inacc)$ for the Hessigheim 3D dataset and gains of up to 1.9% in mIoU, 1.5% in mF1, 0.5% in mAcc, 0.1% in $p(acc|cer)$, and 1.3% in $p(unc|inacc)$ for the DALES dataset. In particular, TTA primarily strengthens segmentation quality, with subtler effects on uncertainty measures. MCD is highly sensitive to the dropout rate, providing improved calibration and uncertainty usefulness at the expense of segmentation accuracy when stochasticity is increased. Among the evaluated approaches, the DSE variants yield the most structured and informative uncertainty maps and reduce inaccurate-but-certain predictions. The frozen-encoder variant combined with TTA (DSE-FE-TTA) provides the most robust overall gains over the baseline, particularly on the more challenging H3D dataset, where limited input cues and challenging regions make reliability more critical.

While these findings highlight the practical value of uncertainty modeling for reliable segmentation, our analysis is restricted to in-domain reliability based on predictive entropy. Future work should extend evaluation to dataset shift and out-of-distribution settings, refine uncertainty assessment for thin and rare classes, and further study the influence of data chunking on uncertainty quality. In addition, more informative and task-relevant uncertainty evaluation metrics should be developed and adopted, especially for 3D data. Finally, systematic exploration of augmentation design and sub-ensemble hyperparameters (e.g., trunk-task split location and number of decoder heads) may further improve uncertainty quality and robustness in practical large-scale remote sensing settings.

7. Acknowledgment

We gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG). Furthermore, we acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology. We also acknowledge the use of generative AI tools (ChatGPT, OpenAI) during the preparation of this manuscript, primarily to assist with text structuring and improving clarity. After using these tools, we reviewed and

revised all AI-assisted content and take full responsibility for the final manuscript.

References

[1] Adnan Anouzla, Mohamed Bakali El Mohamadi, Nabila Zrira, and Khadija Ouazzani-Touhami. 3d semantic segmentation from lidar point clouds: A review. In Brahim El Bhiri, Amir Hussain, and Yassine Maleh, editors, *AI and Digital Transformation: Innovations in Supply Chain, Education, and Energy Systems*, pages 93–100, Cham, 2025. Springer Nature Switzerland. 3

[2] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018. 2, 4, 5, 7, 8

[3] O. C. Bayrak, Z. Ma, E. M. Farella, F. Remondino, and M. Uzar. ESTATE: A large dataset of under-represented urban objects for 3d point cloud classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2-2024:25–32, 2024. 4

[4] Onur Can Bayrak, Zhenyu Ma, Elisa Mariarosaria Farella, Fabio Remondino, and Melis Uzar. Combining 3d urban objects from all around the world to improve object classification and semantic segmentation. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 64(4501917):1–17, 2026. 4

[5] Onur Can Bayrak and Melis Uzar. Local contextual attention for enhancing kernel point convolution in 3d point cloud semantic segmentation. *Applied Sciences*, 15(17):9503, 2025. 4

[6] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9296–9306, 2019. 5

[7] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 6

[8] Thodoris Betsas, Andreas Georgopoulos, Anastasios Doulamis, and Pierre Grussenmeyer. Deep learning on 3d semantic segmentation: A detailed review. *Remote Sensing*, 17(2), 2025. 3

[9] Swaroop Bhandary, Nico Hochgeschwender, Paul Plöger, Frank Kirchner, and Matias Valdenegro-Toro. Evaluating uncertainty estimation methods on 3d semantic segmentation of point clouds. *arXiv preprint arXiv:2007.01787*, 2020. 5

[10] Alexandre Boulch, Joris Guerry, Bertrand Le Saux, and Nicolas Audebert. SnapNet: 3d point cloud semantic labeling with 2d deep segmentation networks. *Computers & Graphics*, 71:189–198, 2018. 3

[11] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 3, 4

[12] Kaiwen Cai, Chris Xiaoxuan Lu, and Xiaowei Huang. Uncertainty estimation for 3d dense prediction via cross-point embeddings. *IEEE Robotics and Automation Letters*, 8(5):2558–2565, 2023. 5

[13] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 77–85, 2017. 3

[14] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. https://github.com/Pointcept/Pointcept, 2023. 6

[15] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022. 15

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 3

[17] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, pages 207–222. Springer, 2020. 5

[18] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 4

[19] Anis Farshian, Markus Götz, Gabriele Cavallaro, Charlotte Debus, Matthias Nießner, Jón Atli Benediktsson, and Achim Streit. Deep-learning-based 3-d surface reconstruction—a survey. *Proceedings of the IEEE*, 111(11):1464–1501, 2023. 2

[20] Anis Farshian Abbasi and Martin Weinmann. Transfer Learning for 3D Semantic Segmentation using Transformers. *Publikationen der DGPF, Band 33*, pages 374–388, 2025. 2

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 2, 4, 5, 7

[22] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. 2, 4, 5, 7, 17

[23] Daniel Girardeau-Montaut. CloudCompare - 3d point cloud and mesh processing software. Open Source Project, 2024. Version 2.13.x [GPL software]. 9

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning, 2016. 6

[25] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 15, 16

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 2, 9

Journal Pre-proof

Journal Pre-proof

[27] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2021. 3

[28] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319, 2020. 7

[29] Timo Hackel, Nicolay Savinov, Lubor Ladicky, Jan Dirk Wegner, Konrad Schindler, and Marc Pollefeys. SEMANTIC3D.NET: a new large-scale point cloud classification benchmark. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1/W1:91–98, 2017. 3

[30] Timo Hackel, Jan Dirk Wegner, and Konrad Schindler. Fast semantic segmentation of 3d point clouds with strongly varying density. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-3:177–184, 2016. 3

[31] Yong He, Hongshan Yu, Xiaoyan Liu, Zhengeng Yang, Wei Sun, Saeed Anwar, and Ajmal Mian. Deep learning based 3d segmentation in computer vision: A survey. *Information Fusion*, 115:102722, 2025. 3

[32] David Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. In *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes: Nebst Einer Lebensgeschichte*, pages 1–2. Springer, 1935. 6

[33] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11105–11114, 2020. 4

[34] Jing Huang and Suya You. Point cloud labeling using 3d convolutional neural network. In *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2670–2675, 2016. 3

[35] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 4

[36] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 5, 7

[37] Michael Kölle, Dominik Laupheimer, Stefan Schmohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. The hessigheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1:100001, 2021. 2, 8

[38] Sanjiv Kumar and Martial Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, 2006. 4

[39] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 3, 4

[40] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8490–8499, 2022. 4

[41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 4, 5, 7

[42] Steven Landgraf, Markus Hillemann, Theodor Kapler, and Markus Ulrich. Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation. In *DAGM German Conference on Pattern Recognition*, pages 348–364. Springer, 2024. 7

[43] Steven Landgraf, Markus Hillemann, Theodor Kapler, and Markus Ulrich. A comparative study on multi-task uncertainty quantification in semantic segmentation and monocular depth estimation. *tm-Technisches Messen*, 92(7-8):298–310, 2025. 9

[44] Loic Landrieu, Clement Mallet, and Martin Weinmann. Comparison of belief propagation and graph-cut approaches for contextual classification of 3d lidar point cloud data. In *Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium*, pages 2768–2771, 2017. 3, 4

[45] Loic Landrieu, Hugo Raguet, Bruno Vallet, Clement Mallet, and Martin Weinmann. A structured regularization framework for spatially smoothing semantic labelings of 3D point clouds. *ISPRS J. Photogramm. Remote Sens.*, 132:102–118, 2017. 3, 4

[46] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4567, 2018. 4

[47] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Deep projective 3d semantic segmentation. In Michael Felsberg, Anders Heyden, and Norbert Krüger, editors, *Computer Analysis of Images and Patterns*, pages 95–107, Cham, 2017. Springer International Publishing. 3

[48] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 16

[49] Pengfei Liu, Guizhen Yu, Zhangyu Wang, Bin Zhou, Ruotong Ming, and Chunhua Jin. Uncertainty-aware point-cloud semantic segmentation for unstructured roads. *IEEE Sensors Journal*, 23(13):15071–15080, 2023. 5

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8

[51] Yongqiang Mao, Kaiqiang Chen, Wenhui Diao, Xian Sun, Xiaonan Lu, Kun Fu, and Martin Weinmann. Beyond single receptive field: A receptive field fusion-and-stratification network for airborne laser scanning point cloud classifica-

Journal Pre-proof

tion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:45–61, 2022. 4

[52] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 16

[53] Aryan Mobiny, Hien Van Nguyen, Supratik Moulik, Naveen Garg, and Carol C. Wu. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *arXiv preprint arXiv:1906.04569*, 2019. 5

[54] Guy M Morton. *A computer oriented geodetic data base and a new technique in file sequencing*. International Business Machines Company, 1966. 6

[55] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports*, 10(1):5068, 2020. 8

[56] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 9

[57] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015. 9

[58] Jakub Nalepa, Michal Myller, and Michal Kawulok. Training-and test-time data augmentation for hyperspectral image segmentation. *IEEE Geoscience and Remote Sensing Letters*, 17(2):292–296, 2019. 8

[59] Radford M Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012. 5

[60] Joachim Niemeyer, Franz Rottensteiner, and Uwe Soergel. Contextual classification of lidar data and building object detection in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:152–165, 2014. 3, 4

[61] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019. 2, 9

[62] Christina Petschnigg and Jürgen Pilz. Uncertainty estimation in deep neural networks for point cloud segmentation in factory planning. *Modelling*, 2(1):1–17, 2021. 5

[63] Christina Petschnigg, Markus Spitzner, Lucas Weitzendorf, and Jürgen Pilz. From a point cloud to a simulation model—bayesian segmentation and entropy based uncertainty estimation for 3d modelling. *Entropy*, 23(3):301, 2021. 5

[64] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999. 4

[65] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5105–5114, 2017. 3

[66] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: revisiting PointNet++ with improved training and scaling strategies. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022. 3

[67] Luigi Riz, Cristiano Saltori, Yiming Wang, Elisa Ricci, and Fabio Poiesi. Novel class discovery meets foundation models for 3d semantic segmentation. *International Journal of Computer Vision*, 133(2):527–548, 2025. 4

[68] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[69] Hanieh Shojaei Miandashti, Qianqian Zou, and Max Mehltretter. Uncertainty estimation and out-of-distribution detection for lidar scene semantic segmentation. In *European Conference on Computer Vision*, pages 116–131. Springer, 2024. 5

[70] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 2, 5, 7

[71] Nina M Singer and Vijayan K Asari. Dales objects: A large scale benchmark dataset for instance segmentation in aerial lidar. *IEEE Access*, 9:97495–97504, 2021. 8

[72] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019. 8

[73] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 7

[74] Moritz Steinsiek, Przemyslaw Polewski, Wei Yao, and Peter Krzystek. Semantische Analyse von ALS- und MLS-Daten in urbanen Gebieten mittels Conditional Random Fields. *Publikationen der DGPF, Band 26*, pages 521–531, 2017. 3

[75] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 945–953, 2015. 3

[76] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. SEGCloud: Semantic segmentation of 3d point clouds. In *Proceedings of the 2017 International Conference on 3D Vision*, pages 537–547, 2017. 3

[77] Hugues Thomas, François Goulette, Jean-Emmanuel Deschaud, Beatriz Marcotegui, and Yann LeGall. Semantic classification of 3d point clouds with multiscale spherical neighborhoods. In *Proceedings of the 2018 International Conference on 3D Vision*, pages 390–398, 2018. 3

[78] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the 2019 IEEE/CVF Inter-*

Journal Pre-proof

national Conference on Computer Vision, pages 6410–6419, 2019. 4

[79] Hugues Thomas, Yao-Hung Hubert Tsai, Timothy D. Barfoot, and Jian Zhang. Kpconvx: Modernizing kernel point convolution with kernel attention. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5535, 2024. 4, 15, 16

[80] Matias Valdenegro-Toro. Sub-ensembles for fast uncertainty estimation in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4119–4127, 2023. 2, 5, 7, 17

[81] Nina Varney, Vijayan K Asari, and Quinn Graehling. Dales: A large-scale aerial lidar data set for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 186–187, 2020. 2, 8

[82] Hristo Vassilev, Marius Laska, and Jörg Blankenbach. Uncertainty-aware point cloud segmentation for infrastructure projects using bayesian deep learning. *Automation in Construction*, 164:105419, 2024. 5

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[84] Tuan-Anh Vu, Srinjay Sarkar, Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. Test-time augmentation for 3d point cloud classification and segmentation. In *2024 International Conference on 3D Vision (3DV)*, pages 1543–1553, 2024. 5

[85] Chongyu Wang, Kunlei Jing, Jihua Zhu, and Di Wang. Openurban3d: Label-free open-vocabulary semantic segmentation of large-scale urban point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 2026. 4

[86] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *International MICCAI Brainlesion Workshop*, pages 61–72. Springer, 2018. 8

[87] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023. 15, 16

[88] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 38(5), 2019. 4

[89] Martin Weinmann, Boris Jutzi, Stefan Hinz, and Clément Mallet. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:286–304, 2015. 3

[90] Martin Weinmann, Alena Schmidt, Clement Mallet, Stefan Hinz, Franz Rottensteiner, and Boris Jutzi. Contextual classification of point cloud data by exploiting individual 3d neighborhoods. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4:271–278, 2015. 4

[91] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020. 2, 5, 7

[92] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024. 2, 4, 5, 6, 8, 15, 16

[93] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 2, 4, 6

[94] Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. Linking points with labels in 3d: A review of point cloud semantic segmentation. *IEEE Geoscience and Remote Sensing Magazine*, 8(4):38–59, 2020. 2, 3

[95] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5588–5597, 2020. 4

[96] Yuwei Yang, Munawar Hayat, Zhao Jin, Chao Ren, and Yinjie Lei. Geometry and uncertainty-aware 3d point cloud class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21759–21768, 2023. 5

[97] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2, 4, 6

A. Additional Visualizations

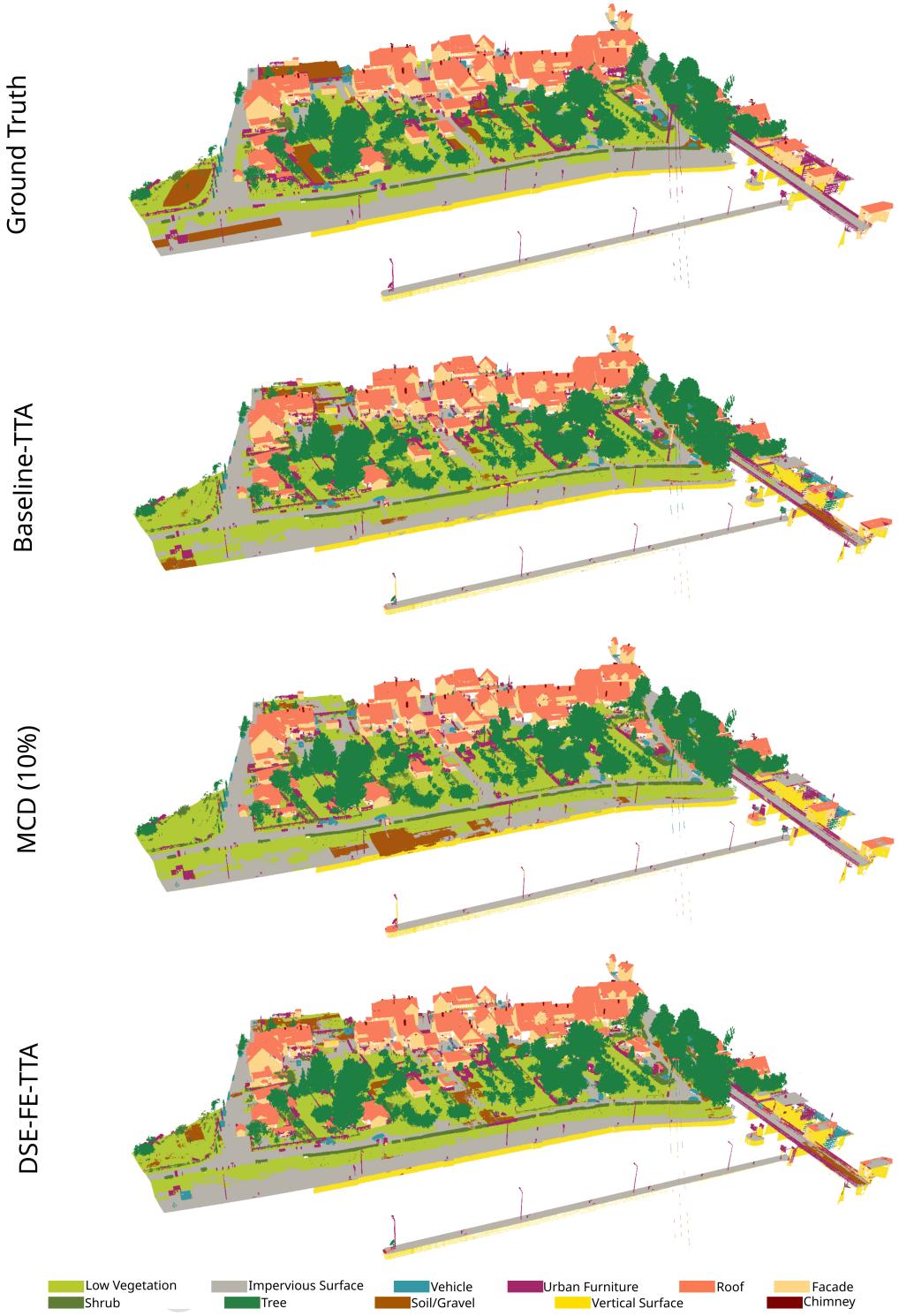


Figure 8. Labels for the H3D dataset in side view.

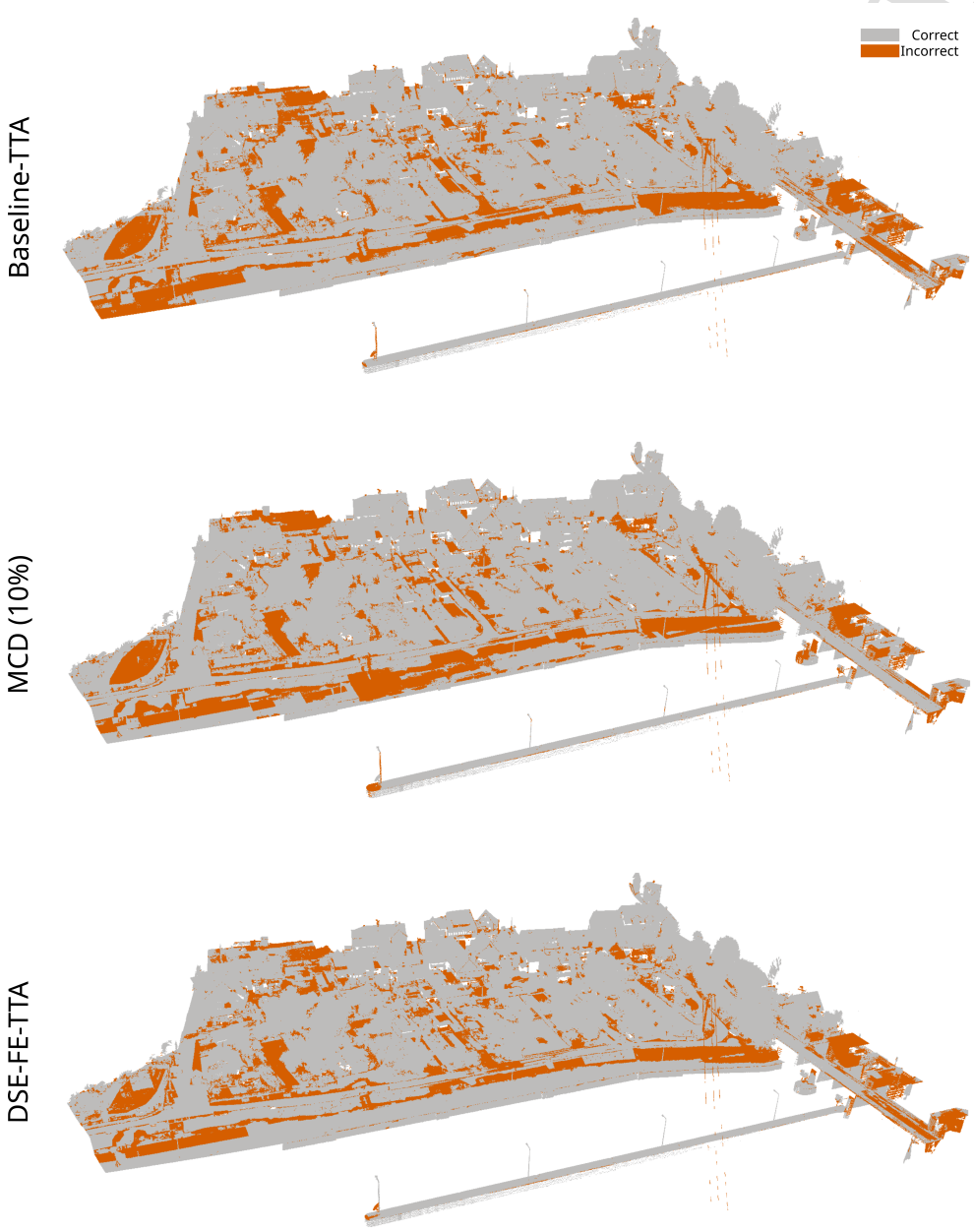


Figure 9. Accuracy map for the H3D dataset in side view.

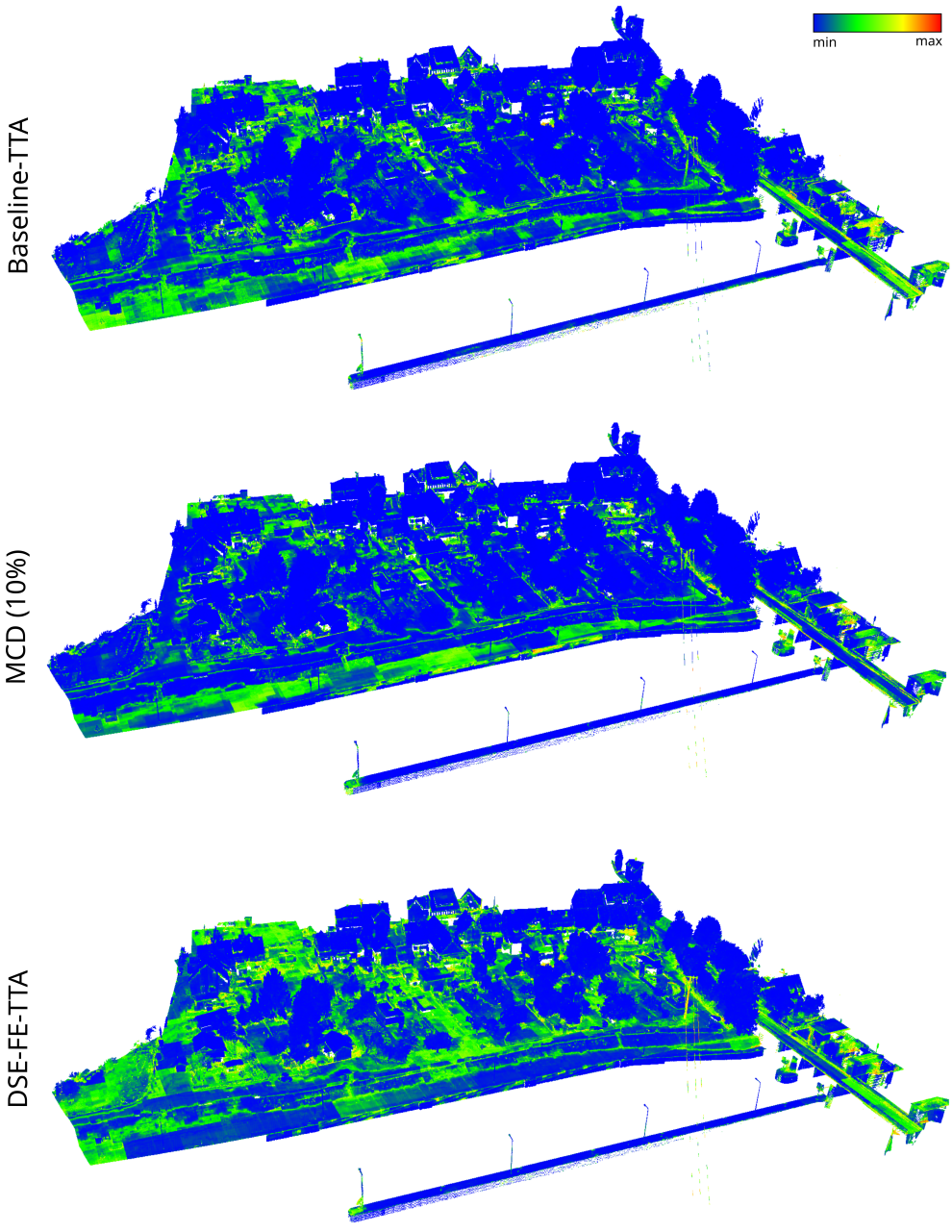


Figure 10. Uncertainty for the H3D dataset in side view.

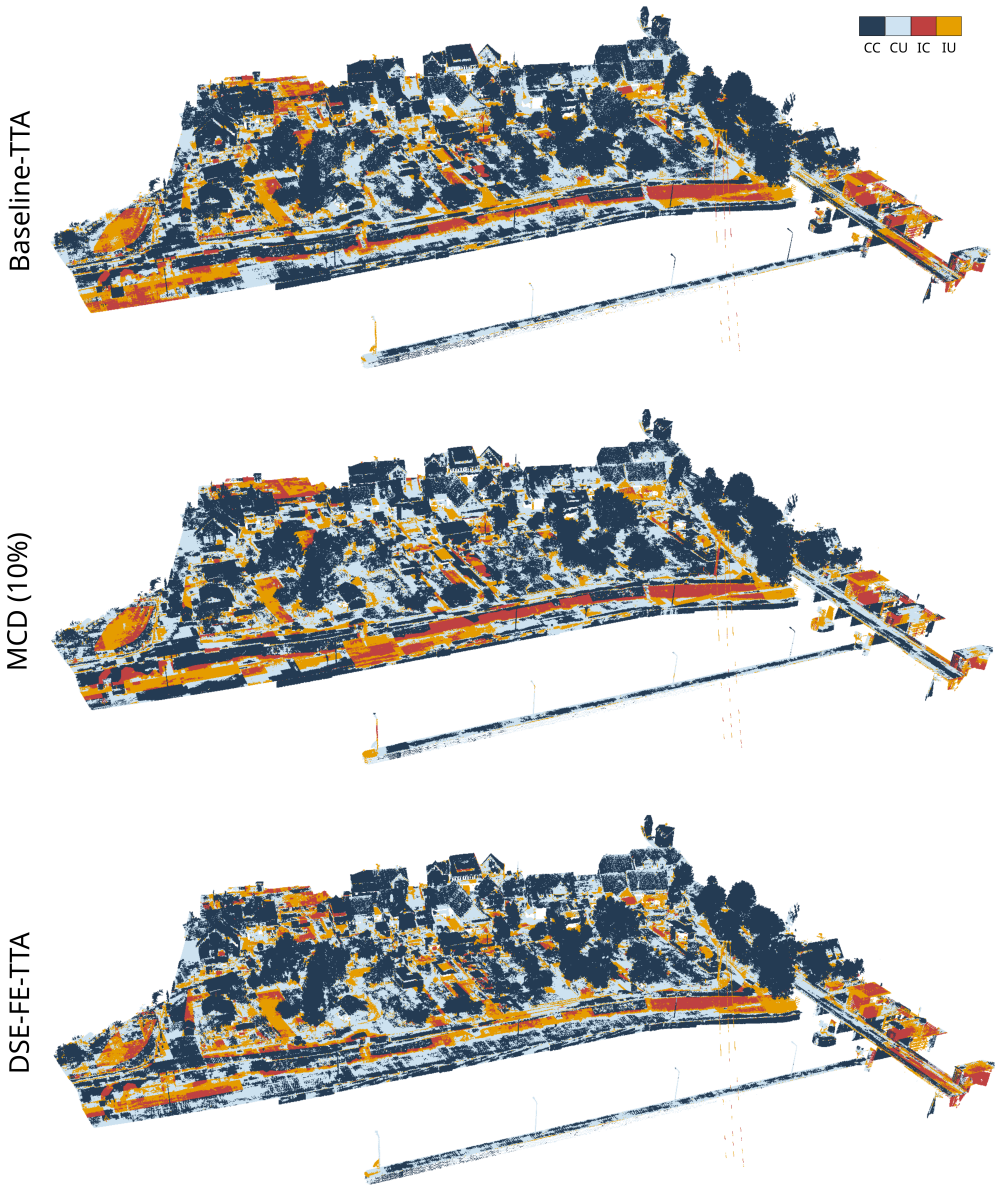


Figure 11. CC/CU/IC/IU representation for the H3D dataset in side view.

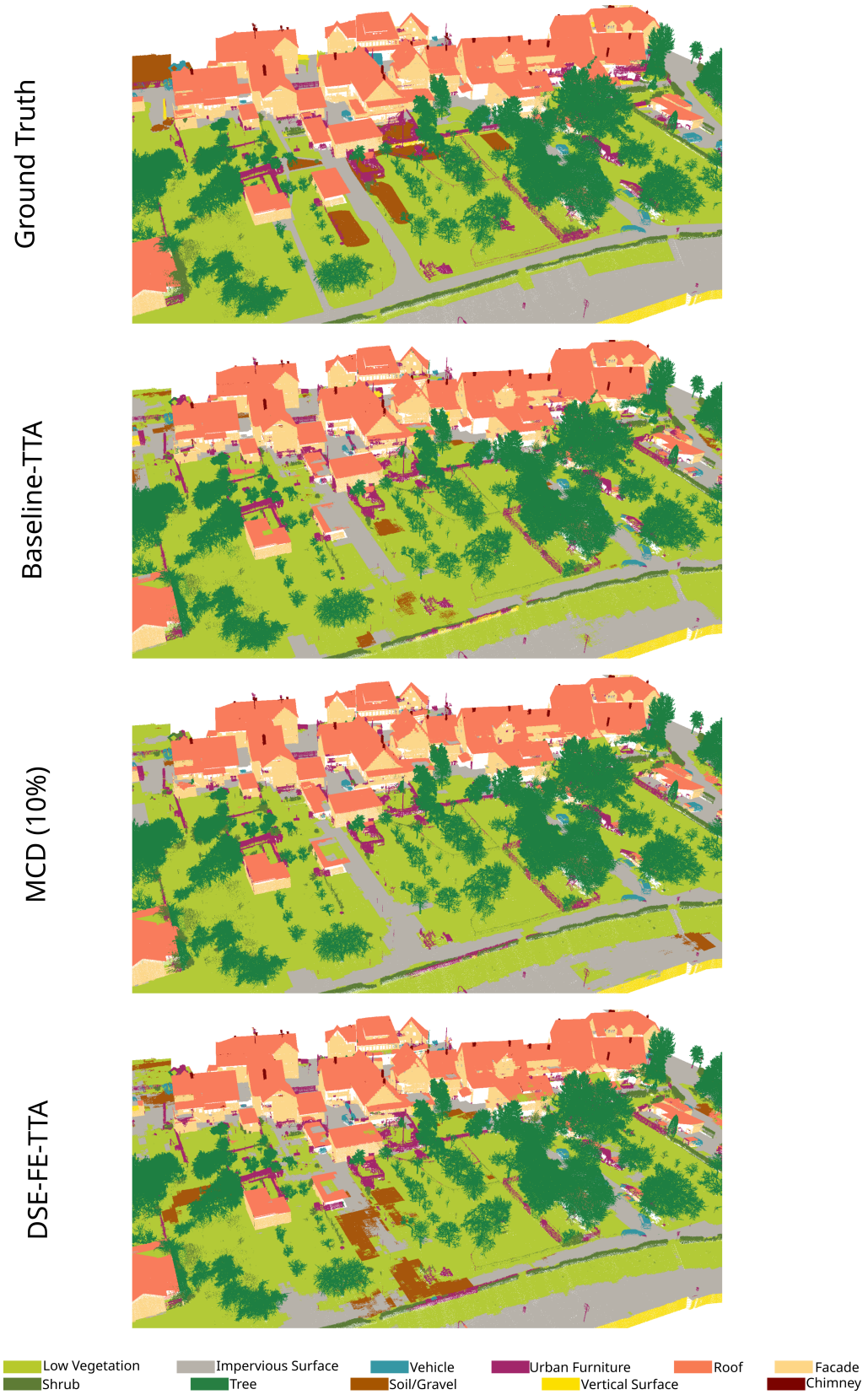


Figure 12. Labels for the H3D dataset in closer view.

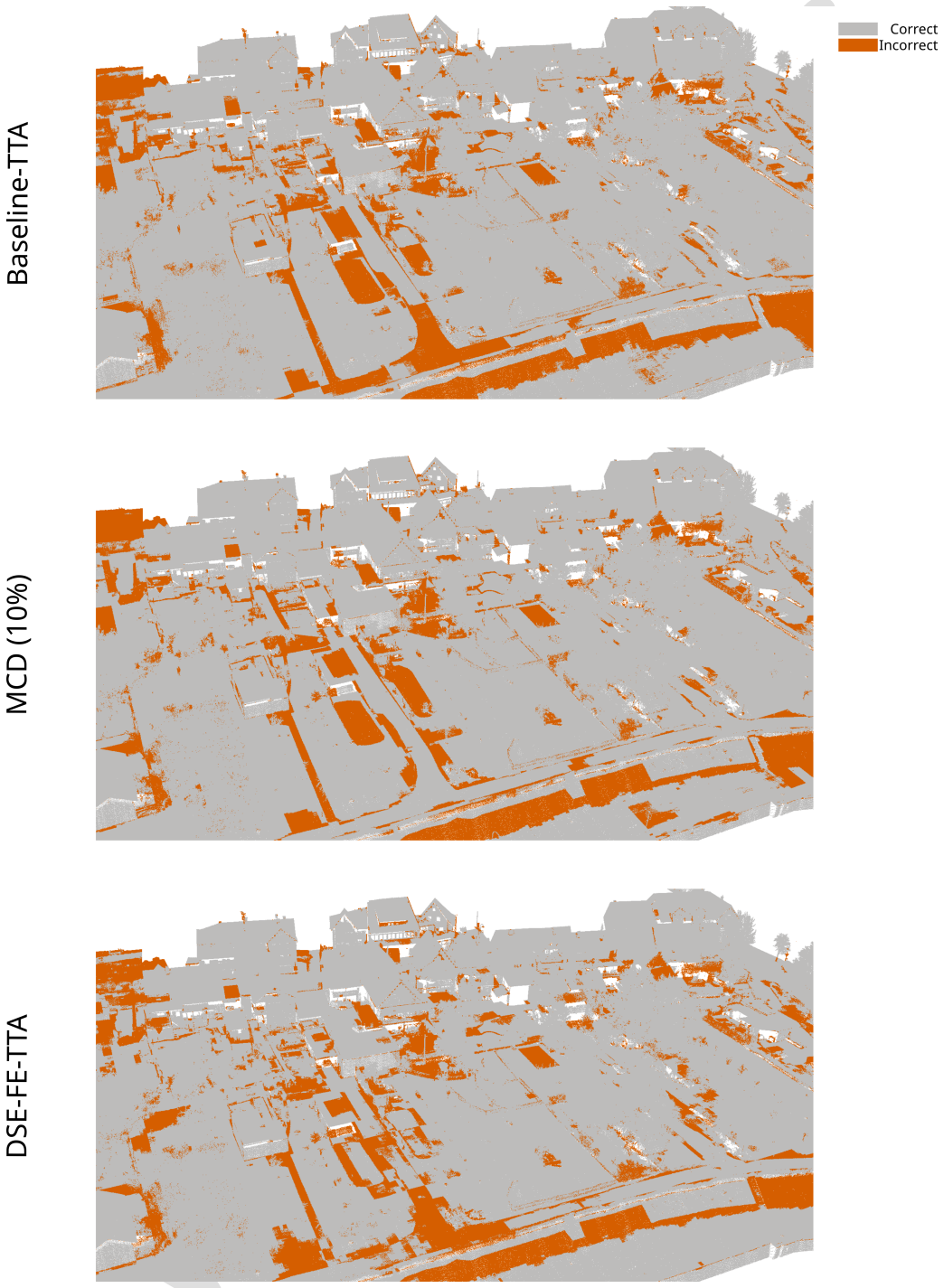


Figure 13. Accuracy map for the H3D dataset in closer view.

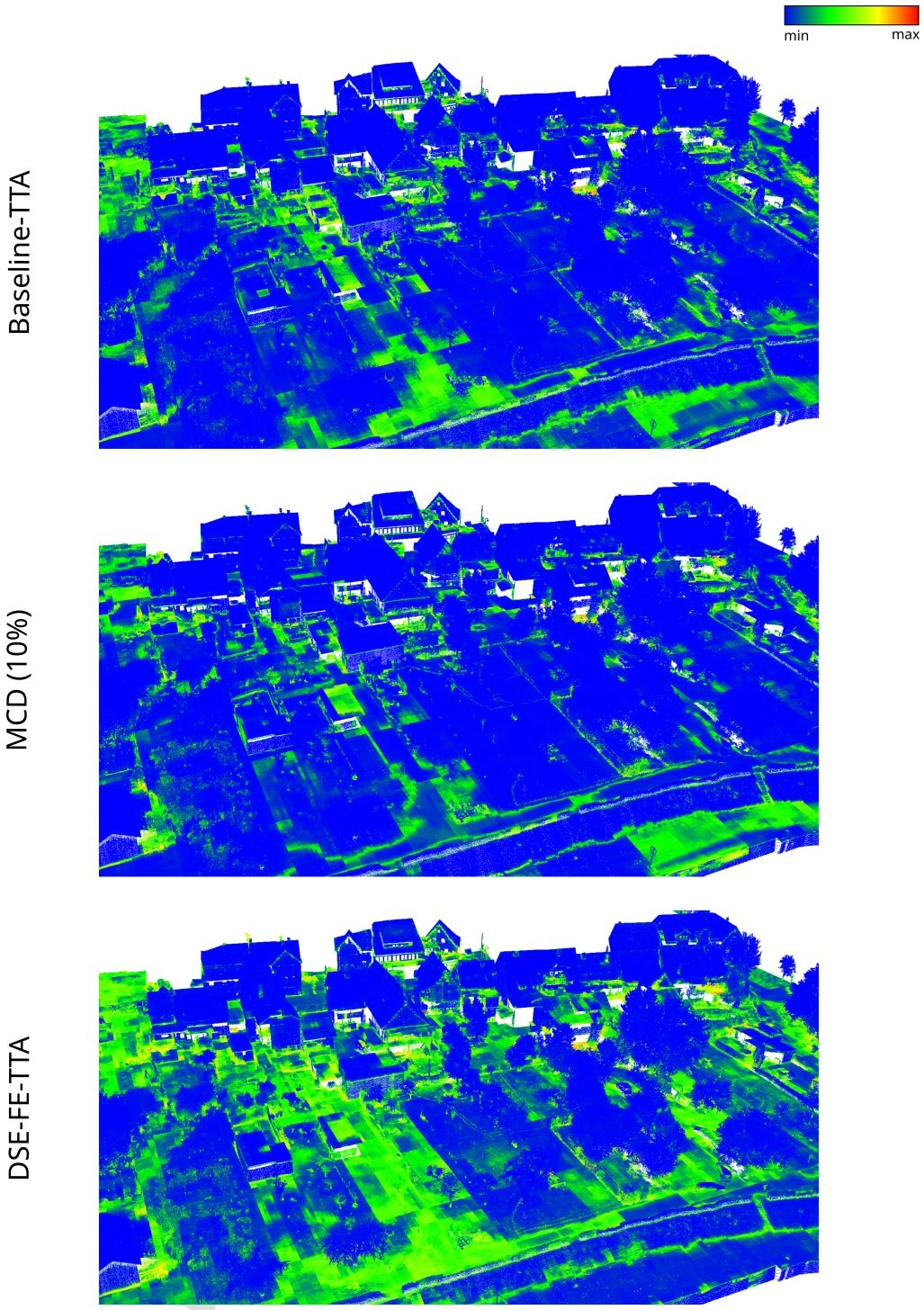


Figure 14. Uncertainty for the H3D dataset in closer view.

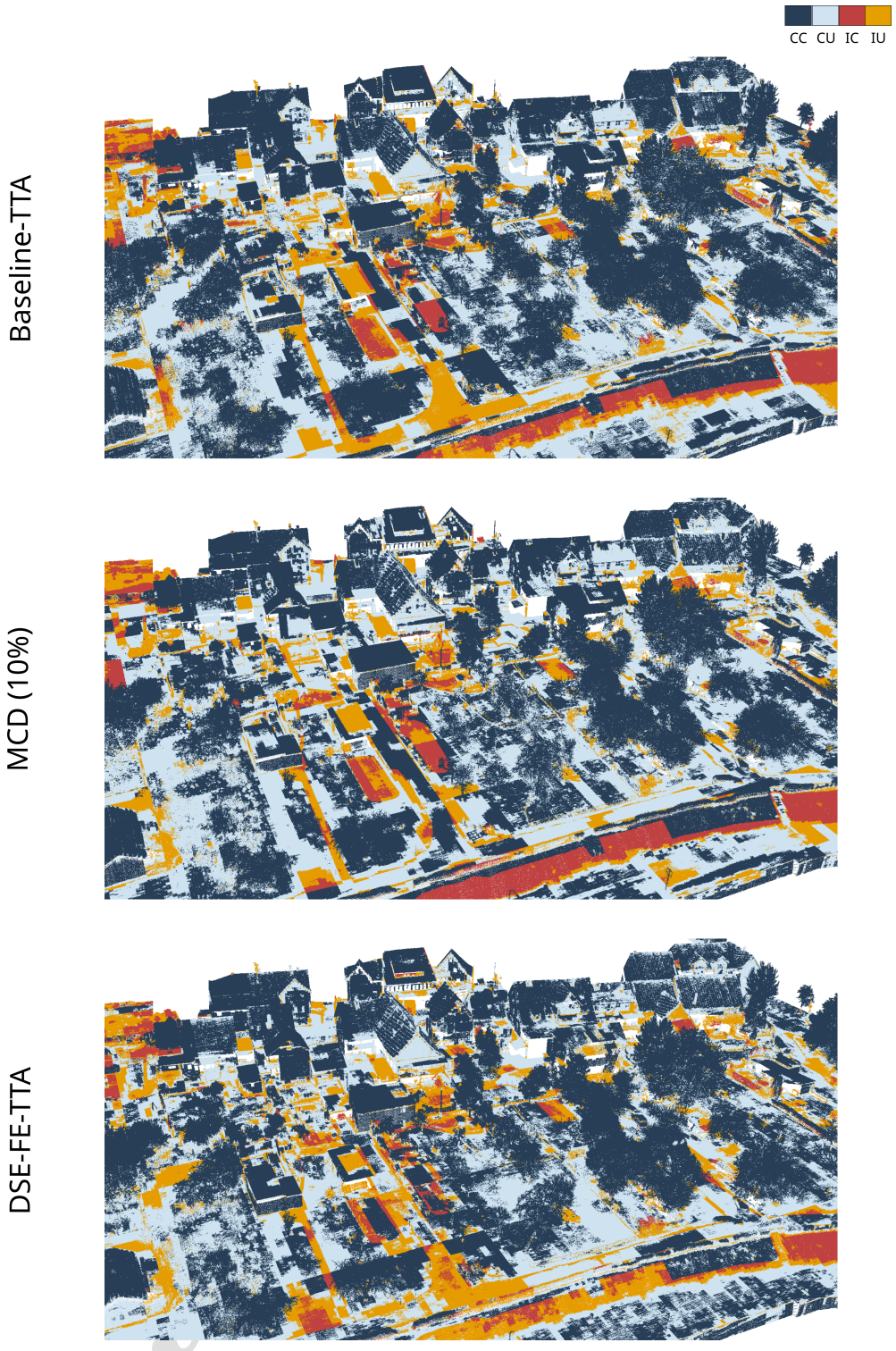


Figure 15. CC/CU/IC/IU representation for the H3D dataset in closer view.

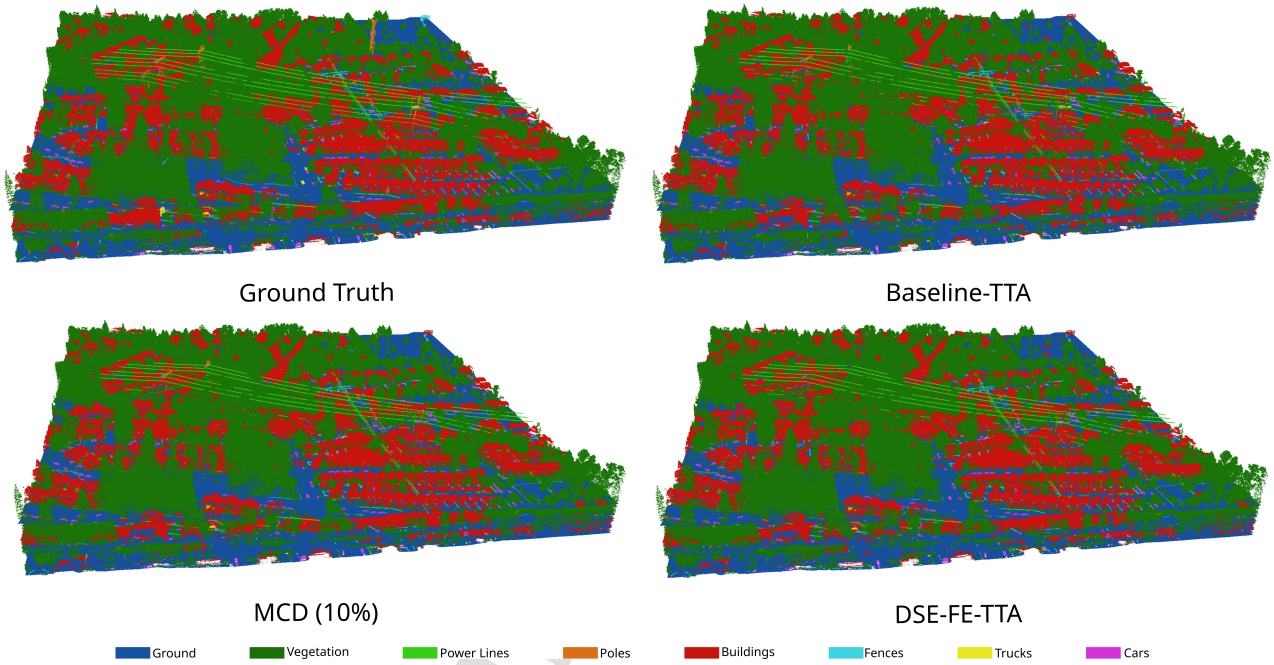


Figure 16. Labels for the DALES dataset in side view.

Journal Pre-proof

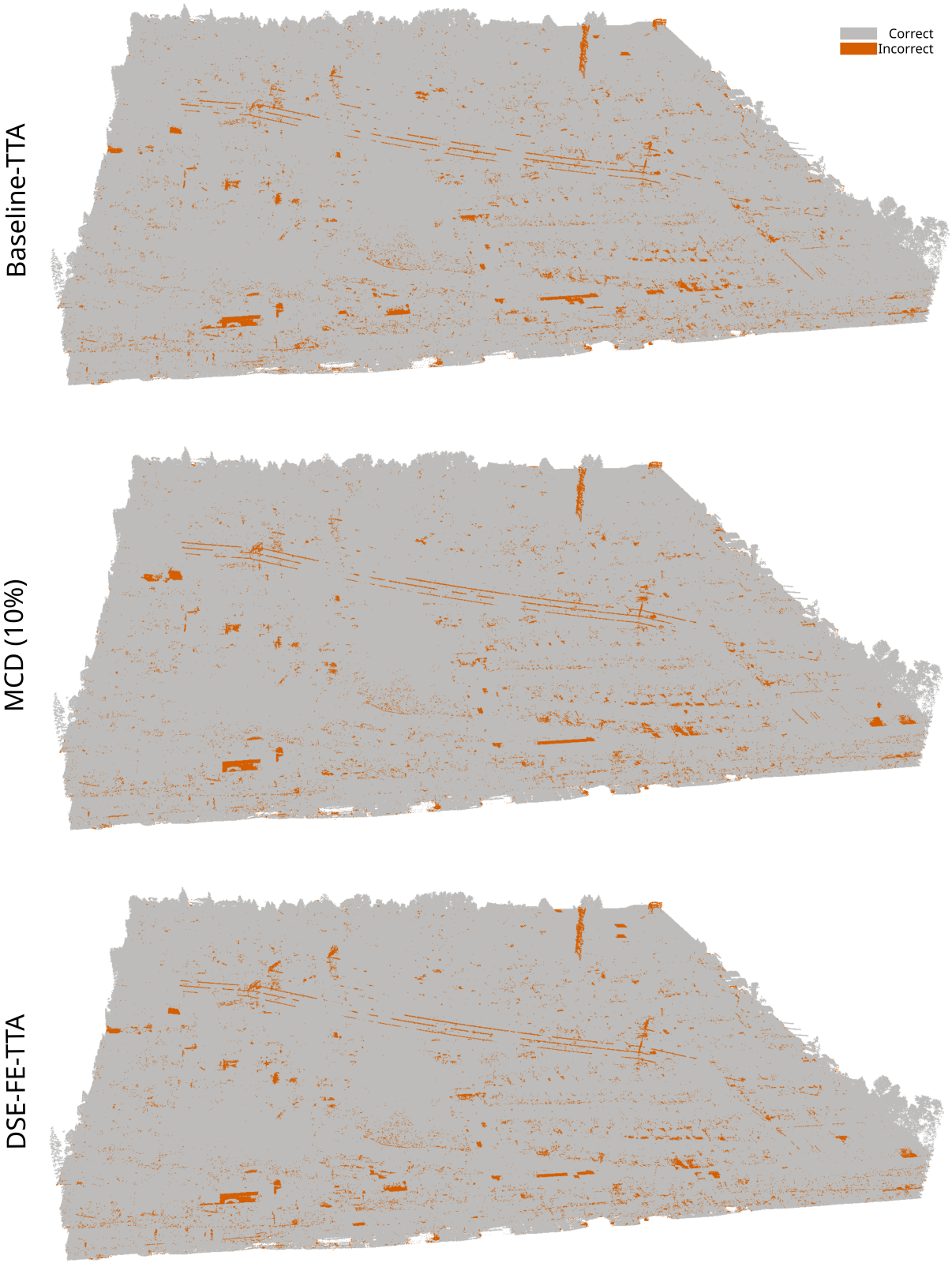


Figure 17. Accuracy map for the DALES dataset in side view.

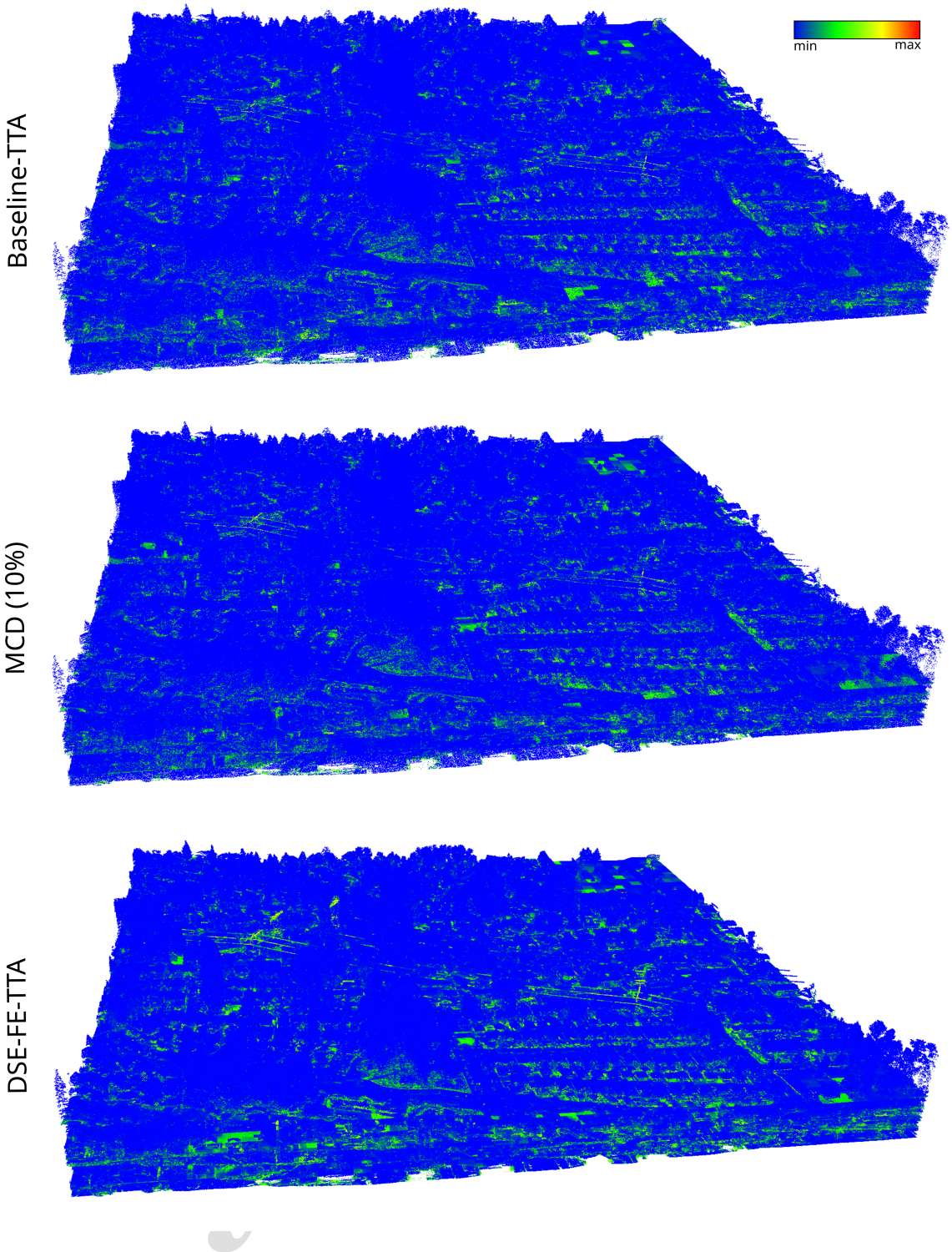


Figure 18. Uncertainty for the DALES dataset in side view.

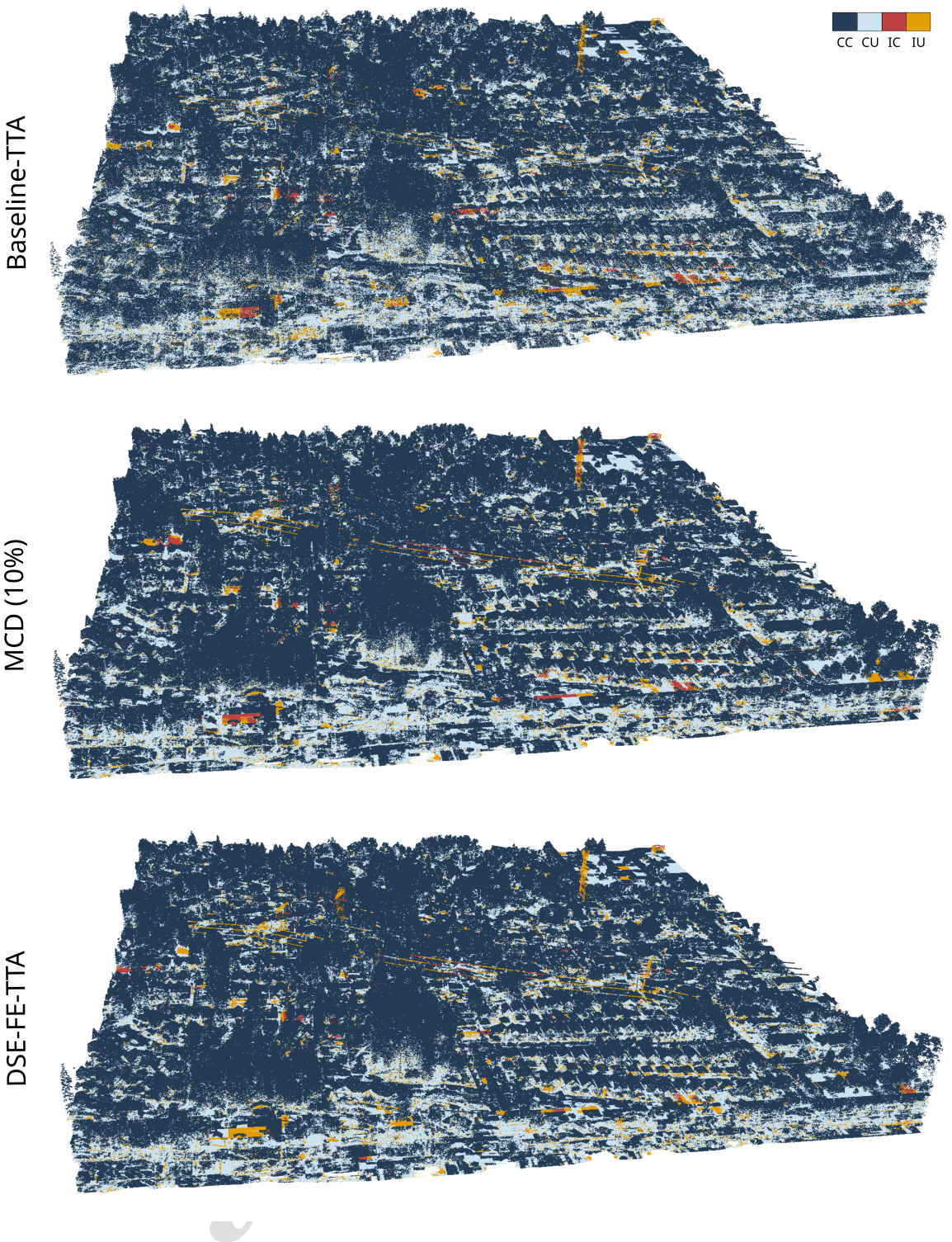


Figure 19. CC/CU/IC/IU for the DALES dataset in side view.

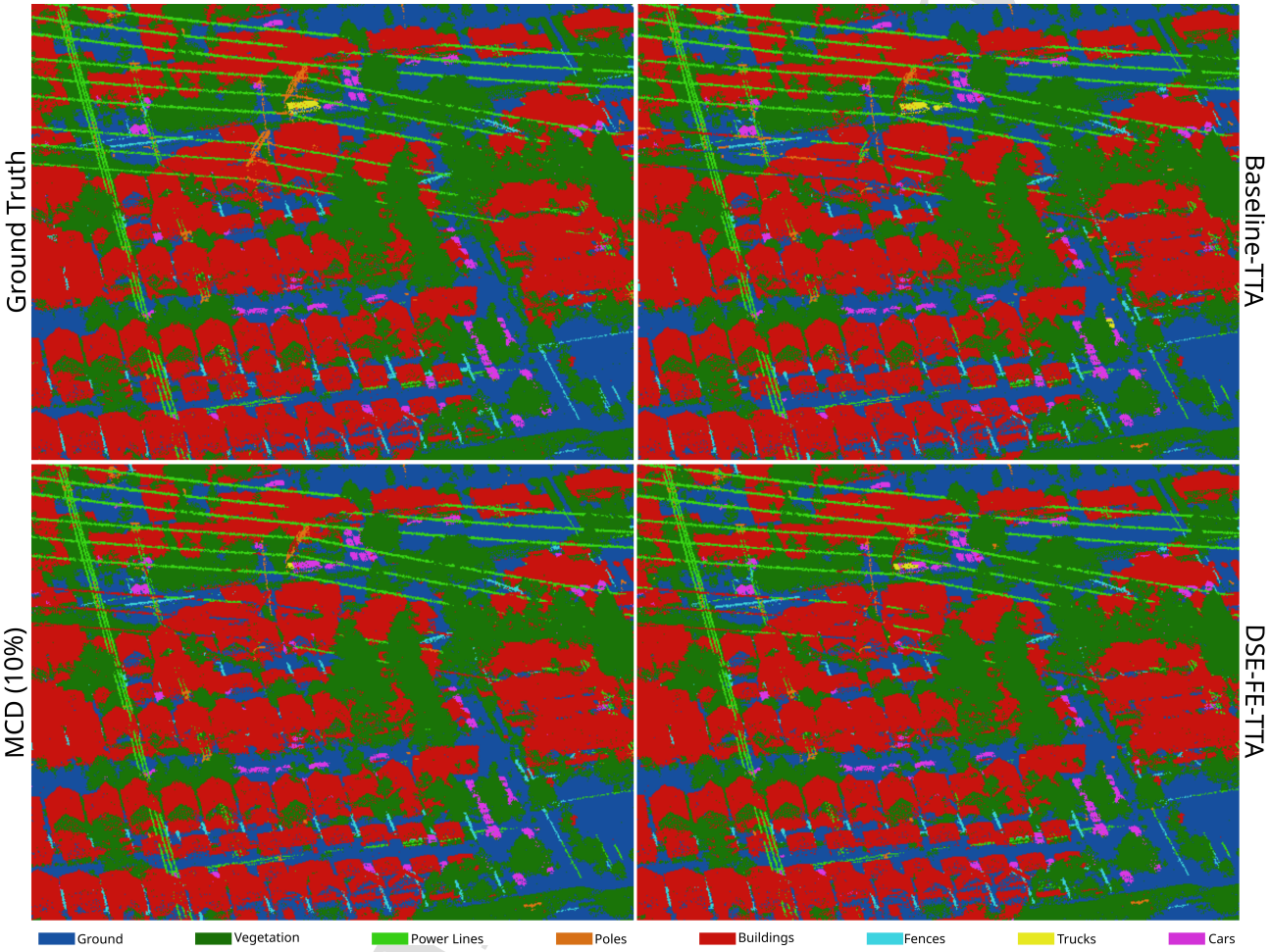


Figure 20. Labels for the DALES dataset in closer view.

Journal Pre-proof

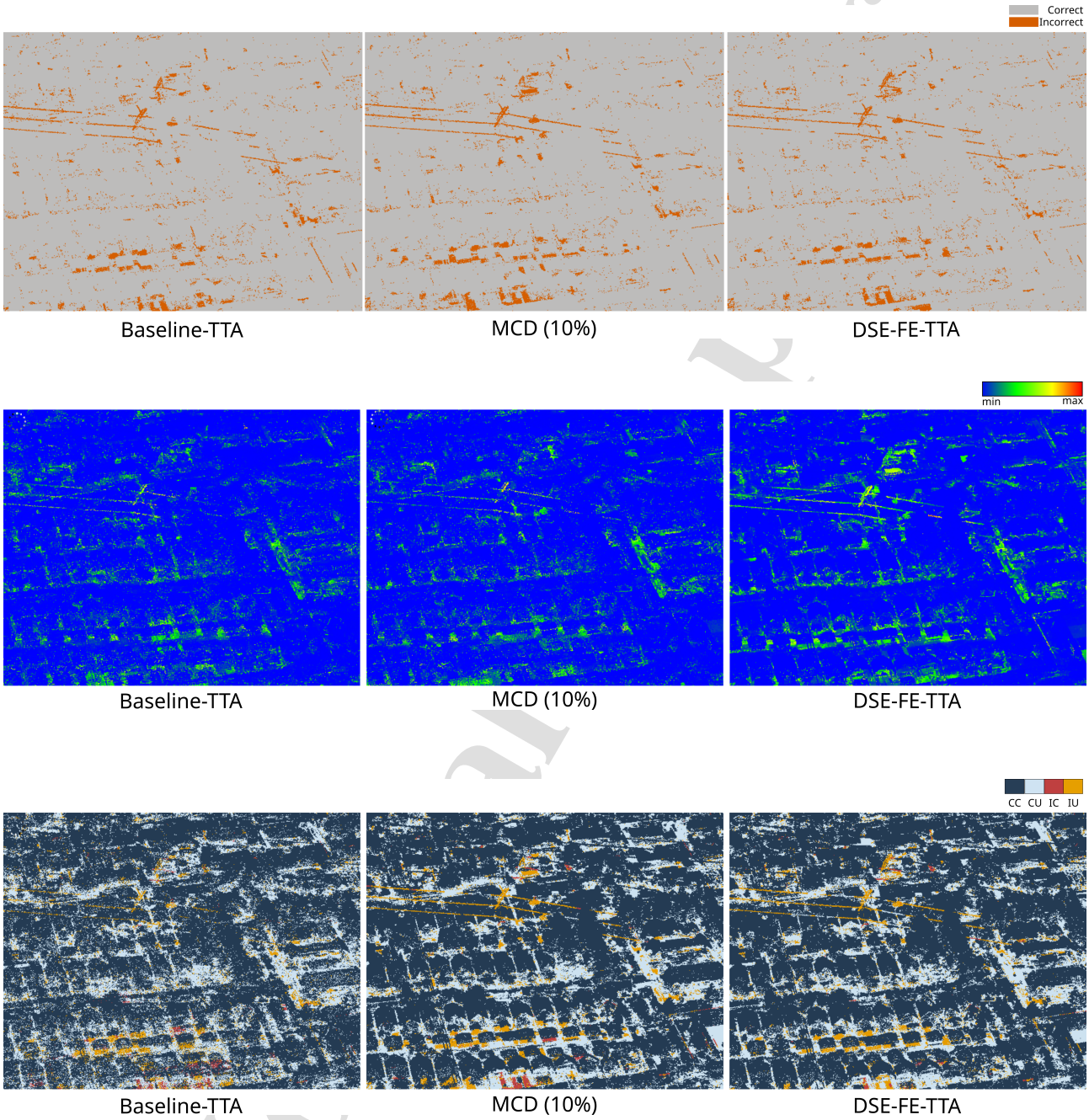
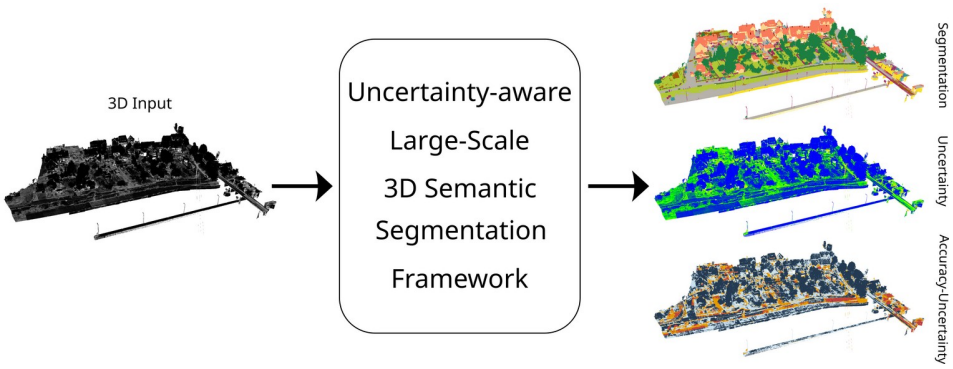


Figure 21. Accuracy map (first row), uncertainty map (second row), and CC/CU/IC/IU representation (third row) in closer view for the DALES dataset.


3D Input
Uncertainty-aware
Large-Scale
3D Semantic
Segmentation
Framework
Segmentation
Uncertainty
Accuracy-Uncertainty

Declaration of Interest Statement

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Please note that Martin Weinmann is co-author of this paper but also Editor in Chief of this Journal.

Journal Pre-proof