PaSCo: Urban 3D Panoptic Scene Completion with Uncertainty Awareness

Anh-Quan Cao[1] Angela Dai[2] Raoul de Charette[1]
[1]Inria [2]Technical University of Munich

https://astra-vision.github.io/PaSCo

Abstract

We propose the task of Panoptic Scene Completion (PSC) which extends the recently popular Semantic Scene Completion (SSC) task with instance-level information to produce a richer understanding of the 3D scene. Our PSC proposal utilizes a hybrid mask-based technique on the non-empty voxels from sparse multi-scale completions. Whereas the SSC literature overlooks uncertainty which is critical for robotics applications, we instead propose an efficient ensembling to estimate both voxel-wise and instance-wise uncertainties along PSC. This is achieved by building on a multi-input multi-output (MIMO) strategy, while improving performance and yielding better uncertainty for little additional compute. Additionally, we introduce a technique to aggregate permutation-invariant mask predictions. Our experiments demonstrate that our method surpasses all baselines in both Panoptic Scene Completion and uncertainty estimation on three large-scale autonomous driving datasets. ***Our code and data are available at https://astra-vision.github.io/PaSCo .***

1. Introduction

Understanding scenes holistically plays a vital role in various fields, including robotics, VR/AR, and autonomous driving. A fundamental challenge in this domain is the simultaneous estimation of complete scene geometry, semantics, and instances from incomplete 3D input data, which is often sparse, noisy, and ambiguous due to occlusions and the inherent complexity of the real scenes. Despite these challenges, achieving this level of understanding is crucial to enable machines to interact with their environment in a smart and safe manner.

Semantic Scene Completion (SSC) tackles 3D scene understanding by inferring the full scene geometry and semantics from a sparse observation. There have been significant advancements in SSC which has gained in popularity. Initial methods [7, 10, 16, 45, 66] focused on indoor scenes characterized by dense, regular, and small-scale input point clouds. The recent release of the Semantic KITTI dataset has ignited interest for SSC in outdoor driving scenarios [14, 63, 72, 73], which present unique challenges due to the sparsity, large scale, and varying densities of input point clouds [64].

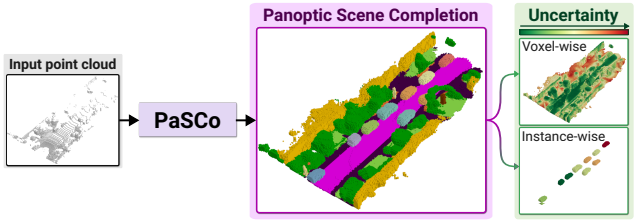

Figure 1. **PaSCo output.** Our method infers Panoptic Scene Completion (PSC) from a sparse input point cloud while concurrently assessing uncertainty at both the voxel and instance levels.

Despite its remarkable performance, current SSC techniques overlook instance-level information and uncertainty prediction. The absence of instance-level prediction hinders their utility in applications that require identification and tracking of individual objects while the lack of uncertainty estimation limits their deployment in real-world safety-critical applications.

To address these challenges, we propose the novel task of Panoptic Scene Completion (PSC), which aims to holistically predict the geometry, semantics, and instances of a scene from a sparse observation. We present the first method for this task, named PaSCo, which is a MIMO-inspired [30] ensemble approach boosting PSC performance and uncertainty estimation at minimal computational cost. It combines multi-scale generative sparse networks with a transformer decoder, implementing a mask-centric strategy for instance prediction [12, 13]. Consequently, we introduce a novel ensembling technique for combining unordered mask sets. Through extensive evaluations, our method demonstrates superior performance in PSC and provides valuable insights into the predictive uncertainty. Our contributions can be summarized as follows:

- We formulate the new task of Panoptic Scene Completion (PSC), extending beyond Semantic Scene Completion to reason about instances.

- Our proposed method, PaSCo, utilizes a sparse CNN-Transformer architecture with a multi-scale sparse generative decoder and transformer prediction, optimized for efficient PSC in extensive point cloud scenes.
- By adapting to the MIMO setting and introducing a novel ensembling strategy for unordered sets, our method boosts PSC performance and enhances uncertainty awareness, outperforming all baselines across three datasets.

2. Related works

Semantic Scene Completion (SSC). SSC was first proposed by SSCNet [66], and recently surveyed in [64]. Prior works mainly focus on indoor scenes [7, 10, 11, 15, 16, 18, 24, 33, 45, 46, 52, 68, 74, 75] with dense, uniform and small-scale point clouds. Semantic KITTI [3] sparked interest in SSC for urban scenes, which pose new challenges due to LiDAR sparsity, large scale, and varying density. To address this, a number of works rely on added modalities [14, 45, 52], while JS3CNet [73] improves by jointly training on semantic segmentation. Strategies for efficient SSC include 2D convolutions on BEV representation in LMSCNet [63] or group convolution in [74]. S3CNet [14] enhances SSC with spatial feature engineering, multi-view fusion, spatial propagation and geometric-aware loss. SCPNet [72] proposes a novel completion sub-network and distills knowledge from multi-frames model. Another line of work predicts SSC [8, 35, 49] and instances [69] from a 2D image. Despite impressive results, none of these works offer instance-level predictions and uncertainty estimation.

LiDAR Panoptic Segmentation Panoptic segmentation was initially introduced in [38] for images. Since then it was extended to 3D point clouds, first using range-based representations [36, 43, 59, 65] with 2D convolutions for efficiency which sacrifice spatial detail. Consequently, some leverage sparse convolutions [26, 32, 51, 62] for efficient 3D processing. Panoptic can be structured as a two-stage method, comprising a non-differentiable clustering followed by semantic segmentation [32, 43, 47, 59, 62], or as a proposal-based approach [36, 65], building on Mask R-CNN [31] with an added semantic head. CPSeg [44] and CenterLPS [58] were the first to propose proposal- and clustering-free end-to-end methods relying on pillarized point features [44] or center-based instance encoding and decoding [58]. MaskPLS [57] offers an end-to-end, mask-based architecture. While these methods exhibit strong performance, they only label the input points. Our work goes a step further by predicting a complete panoptic scene with incorporated uncertainty information, thereby facilitating a more comprehensive understanding of the scene.

Uncertainty Estimation with Efficient Ensemble. Early Bayesian Neural Networks (BNNs) [55] quantified uncertainty in shallow networks, but remain limited in scale [20], despite recent advances in variational inference techniques [5, 37]. Instead, Deep Ensemble [41] offers a practical approach to approximate BNNs' posterior weight distribution [71] and is acknowledged as the leading technique for uncertainty estimation and predictive performance [29, 42, 60]. Yet, its computational demand spurred alternatives like deep sparse networks [53] or BatchEnsemble [70] using partially shared weights. Multi-input multi-output (MIMO) [30] offers a lightweight alternative with diversified outputs, training independent subnetworks within a larger network. Techniques also involve selective dropping of neural weights [19, 23] or multiple model checkpoints of a training session [25, 34] but require multiple inferences. Alternatives also approximate the weight posterior during training to sample ensemble members [22, 56], or use grouped convolution [42]. Our work builds on the simplicity and single-inference MIMO [30], which we complement with a novel permutation-invariant mask ensembling.

3. Method

We introduce the task of Panoptic Scene Completion (PSC), taking an incomplete point cloud X as input and producing a denser output $Y{=}f(X)$ as K voxels masks each with semantic class, *i.e.* $\{(m_k, c_k)\}_{k=1}^{K}$. Inspired by Semantic Scene Completion [64] (SSC), we build a more holistic understanding by reasoning jointly about geometry, semantics and instances. Like panoptic segmentation [38] for semantic segmentation, PSC is a strict generalization of SSC.

To address PSC, we propose PaSCo, which leverages a multiscale sparse generative architecture and proxy completion in a mask-centric architecture [12, 13, 57]. As model calibration is critical for real-world applications like autonomous driving, we also seek to estimate uncertainty. This is crucial as generative tasks hallucinate part of the occluded scenery. Yet, to the best of our knowledge, uncertainty is overlooked in the SSC literature. To boost uncertainty awareness, we employ a multi-input multi-output strategy [30] with a *constant computational budget*, which outputs multiple PSC variations from augmentations of a single input point cloud. To then infer a unique PSC output, we introduce a custom permutation-invariant ensembling.

The schematic view of our method is in Fig. 2, highlighting how PaSCo enables panoptic scene completion with both semantic and instance-wise uncertainty. For simplicity, we first describe the architecture for panoptic scene completion in Sec. 3.1 and then extend to multi-input/output in Sec. 3.2 for uncertainty awareness. Finally, we detail the training strategy in Sec. 3.3.

3.1. Panoptic Scene Completion

Fig. 3 describes our PSC architecture which employs a mask-centric backbone [12, 13, 57]. We rely on multiscale geometric completion (Sec. 3.1.1) followed by a

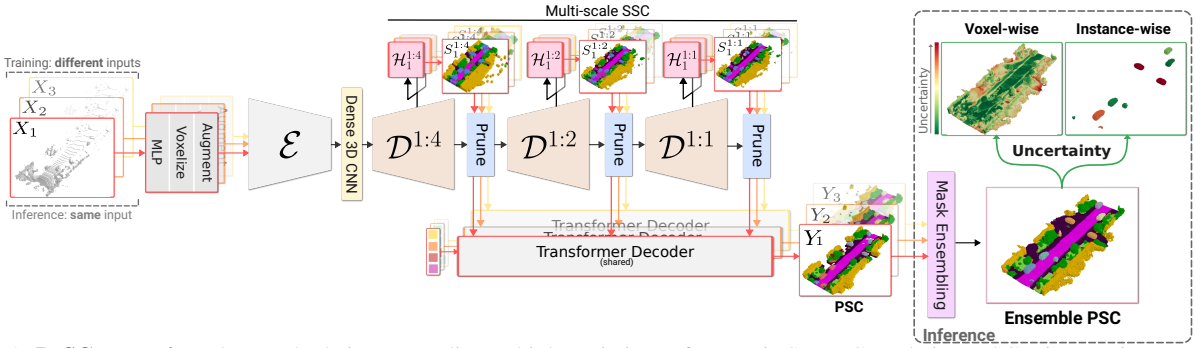

Figure 2. **PaSCo overview.** Our method aims to predict multiple variations of Panoptic Scene Completion (PSC) given an incomplete 3D point cloud, while allowing uncertainty estimation through mask ensembling. For PSC we employ a sparse 3D generative U-Net with a transformer decoder (Sec. 3.1). The uncertainty awareness is enabled using multiple subnets each operating on a different augmented version of an input data source (Sec. 3.2). PaSCo allows the first Panoptic Scene Completion while providing a robust method for uncertainty estimation. Instance-wise uncertainty shows only "things" classes for clarity.

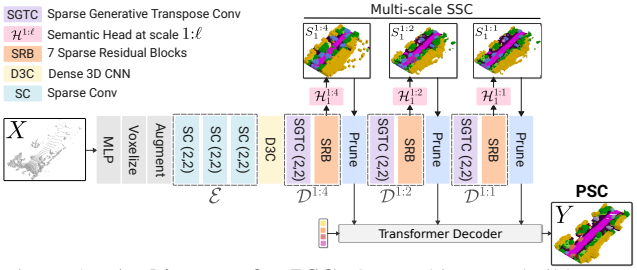

Figure 3. **Architecture for PSC.** Our architecture builds on a sparse generative U-Net coupled with a transformer decoder applied on pruned non-empty voxels to predict PSC.

transformer decoder for mask predictions of both stuff and things (Sec. 3.1.2) to produce panoptic scene completion.

3.1.1 Multiscale Geometric Guidance

We first extract multiscale semantic completion to serve as *geometric* guidance for PSC. For computational efficiency, we rely on sparse generative 3D U-Net as in [14, 17].

Technically, as detailed in Fig. 3, we process unstructured input point cloud X with an MLP and pass the voxelized features through a light-weight encoder $\mathcal{E}$ to produce 1:8 resolution features. To generate geometry beyond input manifold, we then employ a dense CNN, resulting in densified 1:8 features $\mathbf{f}^{1:8}$ which are decoded with sparse generative decoders $\{\mathcal{D}^{1:\ell}\}, \forall \ell \in \{4, 2, 1\}$ producing features, written as $\mathbf{f}^{1:\ell}$. At each scale, a lightweight segmentation head $\mathcal{H}^{1:\ell}$ extracts the proxy SSC, *i.e.*, $S^{1:\ell} = \mathcal{H}^{1:\ell}(\mathbf{f}^{1:\ell})$.

Importantly, we prune features after each decoder to preserve sparsity and thus computational efficiency:

$$\mathbf{f}^{1:\ell} = \mathcal{D}^{1:\ell}\big(\texttt{prune}(\mathbf{f}^{1:2\ell})\big), \forall \ell \in \{4, 2, 1\}. \quad (1)$$

Contrary to the literature [14, 17] using binary occupancy maps, we use semantic predictions for $\texttt{prune}(\cdot)$. We advocate that semantics better balance performance across small classes, while binary occupancy is dominated by large structural classes (*road*, *building*, *etc.*). As such, PSC can better inherit geometric guidance.

3.1.2 Semantic and Instance Prediction as Masks

We now estimate the panoptic completion $\{(m_k, c_k)\}_{k=1}^{K}$, with m_k being a voxel mask and c_k its corresponding semantic class, for both stuff and things. To do so, we follow the latest mask-centric transformer models [12, 13, 57] predicting PSC from the multiscale features of Sec. 3.1.1.

The transformer takes input queries as mask proposals and use the multi-scale features $\{\mathbf{f}^{1:4}, \mathbf{f}^{1:2}, \mathbf{f}^{1:1}\}$ to predict the final queries for mask prediction. We use the multi-scale decoder layer of Mask2Former [13, 57] with masked attention to foster spatial relationships, improving efficiency and training. Hence, each transformer decoder layer $\mathcal{T}^{1:\ell}, \forall \ell \in \{4, 2, 1\}$ is a tailored mix of masked cross attention, self-attention and feed forward network which ultimately produces a set of queries $Q^{1:\ell} \in \mathbb{R}^{K \times D}$ where K is the number of queries and D is feature dimension. Notably, unlike Mask2Former [13], our mask decoder queries and predicts only on pruned occupied voxels.

In practice, since empty voxels dominate 3D scenes [64] and contribute little to semantic understanding, we apply mask prediction only on non-empty voxels. In fine, transformer $\mathcal{T}^{1:\ell}$ takes in query embeddings of the lower scale $Q^{1:2\ell}$, and sparse query features $\hat{\mathbf{f}}^{1:\ell}$ from non-empty voxels of the same scale decoder $\mathcal{D}^{1:\ell}$, *i.e.*, $\hat{\mathbf{f}}^{1:\ell} = \texttt{nonempty}(\mathbf{f}^{1:\ell})$. Notably, at the lowest resolution (1:4) the input query embedding $Q^{1:8}$ is initialized and optimized during training.

For each query embedding $Q^{1:\ell}$, semantic probabilities $p \in \mathbb{R}^{K \times C}$ and mask scores $m \in \mathbb{R}^{K \times N}$ are extracted, where C is the number of classes and N the number of voxels. Probability p is derived by applying a linear layer to $Q^{1:\ell}$. The mask score m is computed from the dot prod-

uct of $Q^{1:\ell}$ and the full scale non-empty voxel features $\hat{\mathbf{f}}^{1:1}$: $m^{1:\ell} = \text{sigmoid}(\hat{\mathbf{f}}^{1:1} \cdot Q^{1:\ell\top})$. The resulting masks are obtained with argmax over p and m.

Similar to previous works, small masks occluded by others are filtered out to minimize false positives [12, 13, 57]. *In fine*, the PSC output is the 1:1 panoptic prediction, so $Y = \{(m, c)\}_{k=1}^{K} = \{(m^{1:1}, p^{1:1})\}_{k=1}^{K}$, while predictions from query embeddings at other scales (*i.e.*, $\ell \neq 1$) serve for additional guidance with mutiscale supervision.

3.2. Uncertainty awareness

We equip PaSCo with uncertainty awareness for both efficient and robust panoptic scene completion. Inspired by MIMO [30] doing image classification, we employ a subnetworks formulation to estimate uncertainty on the much more complex task of panoptic scene completion.

Hence, we adjust our PSC architecture (Sec. 3.1) to predict M variations of PSC outputs with different voxel sets and multi-scale contexts, using per-scale voxel pruning *in a single inference fashion*, as seen in Fig. 2, and at a *similar computation cost*. Intuitively, having several PSC outputs yield better predictive uncertainty estimation and robustness to out-of-distribution [30, 41, 70]. At inference, we use a permutation-invariant mask ensembling strategy to obtain a final unique PSC output.

3.2.1 MIMO Panoptic Scene Completion

In the general case of M subnets, PaSCo infers M outputs[1] given inputs $\{X_i\}_{i=1}^{M}$. Crucially, at training $\{X_i\}$ are distinct point clouds while, in inference, they are augmentations of the *same point cloud*. As in MIMO [30], only heads are duplicated and subnets are in fact trained concurrently in our architecture with minimal but effective adjustments, thus keeping the parameter number roughly constant irrespective of the subnets used. *E.g.*, a subnet of PaSCo(M=3) has 3 times less capacity than that of PaSCo(M=1).

Specifically, referring to Sec. 3.1.1 we share the MLP among subnets and then concat the voxelized representations along the features dimension before passing it to the encoder and decoders. A major difference, is that each subnet has its own semantic heads leading to $\{\mathcal{H}_i^{1:\ell}\}_{i=1}^{M}$ so that they infer distinct semantic outputs $\{S_i^{1:\ell}\}_{i=1}^{M}$. Notably also, the $\texttt{prune}(\cdot)$ operation of Eq. (1) prune only voxels predicted empty by *all* subnets.

To decode per-subnet panoptic output, we follow Sec. 3.1.2, using a dedicated set of query embeddings per subnet, with a shared transformer decoder to increase diversity of the masks predictions at little cost. Interestingly, we note that this also introduces more diversity into the mask predictions, as each query represents one mask.

[1] In this new light, our PSC architecture in Sec. 3.1 is equal to the special case of $M = 1$ subnetwork. *i.e.*, PaSCo(M=1).

Finally, PaSCo output is the combination of all subnets outputs, so $\{Y_i\}_{i=1}^{M}$ with $Y_i = \{(m_k, c_k)\}_{k=1}^{K}$.

3.2.2 Mask ensembling

Unlike classification in MIMO [30], ensembling several PSCs is complex since each subnet infers a set of masks that are permutation invariant. To ensemble these sets, we introduce a pair-wise alignment strategy.

Given two sets of K masks $Y = \{(m_k, p_k)\}_{k=1}^{K}$ and $\hat{Y} = \{(\hat{m}_k, \hat{p}_k)\}_{k=1}^{K}$. We densify the voxel grid, setting empty voxels to 0, such that both mask sets have the same dimension. As they are permutation invariant, we map the two sets using Hungarian matching [40] with the assignment cost matrix $\mathsf{C}(\cdot,\cdot) \in \mathbb{R}^{K\times K}$ where

$$\mathsf{C}(Y, \hat{Y})_{lk} = -\frac{m_l \hat{m}_k^\top}{|m_l| + |\hat{m}_k| - m_l \hat{m}_k^\top}, \tag{2}$$

l and k iterate over all mask indices. Rather than matching binary masks, we find that using "soft matching" with sigmoid probabilities improves results (see Sec. 4.3).
Once mapped together, the ensemble output is obtained by averaging the semantic probability p and binary mask probability m of these mapped queries.

With more than two sets, we arbitrarily use the first set of masks and iteratively align with the remaining sets.

3.3. Training

We train PaSCo end-to-end from scratch with pairs of input point cloud and semi-dense panoptic/semantic labeled voxels, applying losses only on voxels with ground truth labels as in [63, 72, 73].

Voxel-query semantic loss. For subnet i predicting binary mask $m_i \in \mathbb{R}^{N\times K}$ and mask softmax probability $p_i \in \mathbb{R}^{K\times C}$, we estimate a subsidiary per-voxel semantic prediction: $S_i' = m_i p_i$, $S_i' \in \mathbb{R}^{N\times C}$. As masks are predicted at full scale, S_i' is optimized with:

$$\mathcal{L}'_{\text{sem}} = \sum_{i=1}^{M} (\text{CE}(S_k'^{1:1}, \bar{S}^{1:1}) + \lambda_1 \text{lovasz}(S'^{1:1}, \bar{S}^{1:1})), \tag{3}$$

being $\bar{S}_i^{1:1}$ the labels, and $\lambda_1 = 0.3$ empirically fixed [4].

Semantic loss. For each scale $1{:}\ell$ and subnet i, we optimize the semantic output $S_i^{1:\ell}$ against the ground truth $\bar{S}_i^{1:\ell}$ (majority pooled to scale $1{:}\ell$), using a similar loss as Eq. (3), applied across all scales $\ell \in \{1, 2, 4\}$.

Masks matching loss. For each subnet i, we match the output masks $Y_i = \{(m_k, c_k)\}_{k=1}^{K}$ to the ground truth masks $\bar{Y}_i = \{(\bar{m}_{\bar{k}}, \bar{c}_{\bar{k}})\}_{\bar{k}=1}^{\bar{K}}$, using the Hungarian matching as in [12, 57] to learn an optimal mapping $\bar{\sigma}$ by minimizing the assignment map $\mathsf{C}(\cdot,\cdot) \in \mathbb{R}^{K\times\bar{K}}$. The latter is defined as $\mathsf{C}_{k,\bar{k}} = -p_k(\bar{c}_{\bar{k}}) + \mathcal{L}_{\text{mask}}$ with

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{dice}} \text{dice}(m_k, \bar{m}_{\bar{k}}) + \lambda_{\text{bce}} \text{BCE}(m_k, \bar{m}_{\bar{k}}). \tag{4}$$

Method	Semantic KITTI (val set) All PQ†	All PQ	All SQ	All RQ	Thing PQ	Thing SQ	Thing RQ	Stuff PQ	Stuff SQ	Stuff RQ	mIoU↑	Params↓	Time(s)↓	SSCBench-KITTI360 (test set) All PQ†	All PQ	All SQ	All RQ	Thing PQ	Thing SQ	Thing RQ	Stuff PQ	Stuff SQ	Stuff RQ	mIoU↑	Params↓	Time(s)↓
LMSCNet [63] +MaskPLS	13.81	4.17	36.13	6.82	1.62	29.87	2.68	6.02	40.69	9.82	17.02	**31.9M**	0.72	12.76	4.14	26.52	6.45	0.88	20.41	1.58	5.78	29.58	8.88	15.10	**31.9M**	0.87
JS3CNet [73] +MaskPLS	18.41	6.85	41.90	11.34	4.18	43.10	7.22	8.79	41.03	14.34	22.70	34.7M	1.46	16.42	6.79	51.16	10.71	3.36	48.41	5.83	8.51	52.54	13.15	21.31	34.7M	1.13
SCPNet [72] +MaskPLS	19.39	8.59	49.49	13.69	4.88	46.41	7.70	11.30	51.73	18.04	22.44	89.9M	0.91	16.54	6.14	51.18	10.15	4.23	48.46	7.05	7.09	52.55	11.70	21.47	89.9M	1.10
SCPNet* [72] +MaskPLS	23.21	10.89	48.29	17.80	7.35	42.98	12.75	13.46	52.15	21.46	27.89	91.9M	1.36	18.20	7.47	50.67	11.92	3.98	48.13	6.80	9.21	51.94	14.48	**22.66**	91.9M	1.31
PaSCo(M=1)	26.49	15.36	54.15	23.65	12.33	47.42	18.78	17.55	**59.05**	27.19	28.22	111.0M	**0.67**	19.53	9.91	**58.81**	15.40	3.46	**57.72**	6.10	13.14	**59.35**	20.05	21.17	111.0M	**0.39**
PaSCo (Ours)	**31.42**	**16.51**	**54.25**	**25.13**	**13.71**	**48.07**	**20.68**	**18.54**	58.74	**28.38**	**30.11**	120.0M	1.32	**26.29**	**10.92**	56.10	**17.09**	**4.88**	57.53	**8.48**	**13.94**	55.39	**21.39**	22.39	115.0M	0.65

Table 1. **Panoptic Scene Completion.** On both Semantic KITTI [3] (val) and SSCBench-KITTI360 [48] (test), our method PaSCo outperforms all baselines across almost all metrics, in particular, *All PQ*†. * denotes our own re-implementation of SCPNet.

We set K always greater than $\bar{K}$ the number of ground truth masks. Predicted masks without ground truth are mapped to a generic $\varnothing$ class. For the k-th mask of Y_i matched to the $\bar{\sigma}(k)$-th ground truth mask, the loss is

$$\mathcal{L}_{\text{matched}} = \sum_{k=1}^{\bar{K}} \lambda_{\text{CE}} \text{CE}(c_k, \bar{c}_{\bar{\sigma}(k)}) + \mathcal{L}_{\text{mask}}. \quad (5)$$

The $K-\bar{K}$ unmatched predicted masks are optimized to predict $\varnothing$ class $\mathcal{L}_{\text{unmatched}} = \sum_{k=\bar{K}+1}^{K} \lambda_{\varnothing} \text{CE}(c_k, \varnothing)$. where $\lambda_{\varnothing}$ is set to 0.1 as in [12]. λ_{dice} and λ_{bce} are emperically set to 1 and 40. We further apply auxiliary mask matching losses and $\mathcal{L}'_{\text{sem}}$ on the PSC outputs of intermediate scales, *i.e.*, $\{(m^{1:\ell}, p^{1:\ell})\}, \ell \neq 1$.

4. Experiments

We evaluate PaSCo on both panoptic scene completion and uncertainty estimation, while also reporting the subsidiary SSC metrics. As there are *no urban PSC datasets and baselines*, we produce our best effort to extend existing SSC datasets and baselines for fair evaluation.

Datasets. To evaluate PSC, we extend three large-scale urban LiDAR SSC datasets: Semantic KITTI, SSCBench-KITTI360 and Robo3D. **Semantic KITTI** [3] has 64-layer LiDAR scans voxelized into 256x256x32 grids of 0.2m voxels. We follow the standard train/val split [63, 72], leading to 3834/815 grids. **SSCBench-KITTI360** [48] is a very recent SSC benchmark derived from KITTI-360 [50] with urban scans encoded as in Semantic KITTI. We follow the standard train/val/test splits of 8487/1812/2566 grids. **Robo3D** [39] is a new robustness benchmark, extending popular urban datasets [3, 6, 27, 67] by modifying point cloud inputs with various type and intensity of corruptions (*e.g.*, fog, motion blur, *etc.*). We use corrupted input point clouds from the SemanticKITTI-C set of Robo3D to evaluate robustness to Out Of Distribution (OOD) effects.

To extract pseudo panoptic labels from semantic grids, we cluster *things* instances from ad-hoc classes using DBSCAN [9, 21] with a distance of $\epsilon = 1$ and groups with $\texttt{MinPts} = 8$. Following the original panoptic segmentation formulation [38], *stuff* masks are made of voxels with ad-hoc classes. For Semantic KITTI, labels cannot be generated on the hidden test set, so we evaluate on val. set only. We assess our pseudo labels quality in the supplementary.

PSC/SSC metrics. We evaluate panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ) following [38] on the complete scene. Due to the difficulty of the PSC task, most masks have low IoU w.r.t. ground truth. Hence, we note that the over-penalization effect of stuff classes described in [61] is amplified for PSC. Thus, we also evaluate PQ†, as in [61], removing the $>0.5-$IoU rule for stuff classes. We also complement our PSC study with subsidiary SSC metrics, *i.e.*, mean IoU (mIoU).

Uncertainty metrics. Following [42], we employ the maximum softmax probability as a measure of model confidence. Consistent with the established practices [30, 41, 42], we assess the model predictive uncertainty by evaluating its calibration [28] using the Expected Calibration Error (ECE) and Negative Log Likelihood (NLL). Notably, we distinguish between two forms of uncertainty: *voxel uncertainty* and *instance uncertainty*. The former is derived voxel-wise from semantic completion outputs, and the latter mask-wise from class probability predictions. To account for the dominance of empty voxels within 3D scenes, we calculate voxel uncertainties by averaging the uncertainties for empty and non-empty voxels. In line with [38], the label of predicted masks are assigned by finding the matched ground truth masks with $>0.5-$IoU rule. Unmatched masks are classified under a 'dustbin' category.

Training details. We train PaSCo for 30 epochs on Semantic KITTI and 20 epochs on SSCBench-KITTI360, both using AdamW [54] optimizer and batch size of 2. The learning rate is 1e-4, unchanged for Semantic KITTI but divided by 10 at epoch 15 on SSCBench-KITTI360. We apply random rotations in $[-30°, 30°]$ on Semantic KITTI and in $[-10°, 10°]$ on SSCBench-KITTI360, random crop to reduce the scene size to 80% along both the x and y axes, and random translations of ± 0.6m on x/y axes and ± 0.4m on z axis. Unless otherwise mentioned, PaSCo refers to the optimal number of subnets, which is $M = 3$ for Semantic KITTI and Robo3D, and $M = 2$ for SSCBench-KITTI360. This choice of subnets is justified in Tab. 4.

4.1. Panoptic Scene Completion

To evaluate PSC, we first establish baselines for this new task, and then report results on the aforementioned datasets.

Baselines. We combine existing SSC methods with 3D panoptic segmentation. We select three SSC open-source

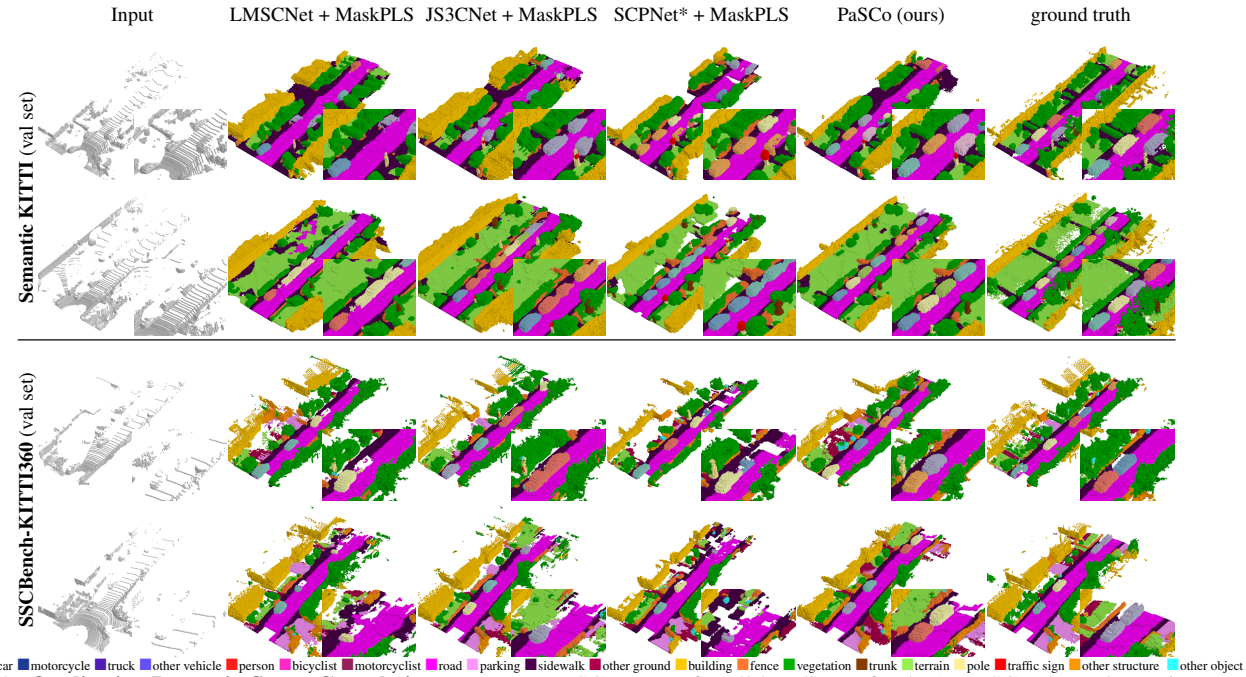

Figure 4. **Qualitative Panoptic Scene Completion**. We report PSC outputs for all baselines of Tab. 1. PaSCo shows better instance separation, with stronger instance shapes and scene structure, with fewer holes.

methods: LMSCNet [63], JS3CNet [73], SCPNet [72], and also add SCPNet* — our own reimplementation with much stronger performance. For 3D panoptic segmentation, we use MaskPLS [57], well-suited for dense voxelized scene and the best open-source 3D panoptic segmentation to date. All baselines are retrained with their reported parameters. We train the four PSC baselines using the SSC method to predict the complete semantic scene followed by the 3D panoptic segmentation method.

Performance. Tab. 1 compares PaSCo with the 4 baselines on Semantic KITTI and SSCBench-KITTI360. Our method is superior across all panoptic metrics (All, Things, Stuff) on both datasets. We see a major boost in All-PQ†/PQ of $+8.21$/$+5.62$ on Semantic KITTI and $+8.09$/$+3.45$ on SSCBench-KITTI360, due to our effective ensembling approach for PSC. Regarding inference time, PaSCo is only slower than LMSCNet+MaskPLS and SCPNet+MaskPLS but performs significantly better. Additionally, PaSCo outperforms baselines in individual metrics for both 'things'/'stuff' categories, showing significant improvements in PQ with $+6.36$/$+5.08$ and $+0.9$/$+4.73$ on each dataset. On the subsidiary mIoU metric we perform on-par, being first on Semantic KITTI ($+2.22$) and 2nd in SSCBench-KITTI360 (-0.27). Incidentally, we note that PSC and SSC metrics are not directly correlated since we improve the former drastically.

Fig. 4 shows that our qualitative PSC results similarly show visual superiority. Overall, we observe that instances are much better separated by PaSCo compared to SCPNet* (our best competitor), with less holes in the geometry.

4.2. Uncertainty estimation

We further evaluate uncertainty as it correlates with model calibration [1], and is crucial for many applications.

Baselines. Using our architecture, we design three uncertainty estimation baselines based on state-of-the-art uncertainty literature. Each baseline provides multiple outputs, enabling similar computation of uncertainty to ours from the maximum softmax probability across inferences. Test-Time Augmentation (TTA) is a classical strategy [2] to improve robustness using multiple inferences of a unique network with input augmentations. MC Dropout [23] provides a bayesian approximation of the model uncertainty by randomly dropping activations (*i.e.*, setting to 0) of neurons, applied with multiple inferences. Finally, we report Deep Ensemble [41], where duplicate networks solving the same task are trained independently and ensembled at test time for better predictive uncertainty than a single network.

For fair comparison, all baselines use our architecture. However, we note that in contrast to our approach, these baselines require more than one pass, either using multiple inferences for TTA and MC Dropout or multiple networks for Deep Ensemble which translate in more parameters.

Uncertainty estimation. Tab. 2 reports uncertainties for all baselines using our architecture, as well as for PaSCo and PaSCo(M=1) which uses a single subnet. To ensure comparable performance, we set the number of inferences (for TTA and MC Dropout) and number networks (for Deep

Semantic KITTI (val set)

method	ins ece↓	ins nll↓	voxel ece↓	voxel nll↓	All PQ†↑	All PQ↑	mIoU↑	Params↓	Passes ↓	Time(s) ↓
TTA	-	-	0.0456	0.7224	-	-	28.84	**111M**	3	1.78
MC Dropout [23]	-	-	0.0472	0.7437	-	-	28.82	**111M**	3	1.70
Deep Ensemble [41]	-	-	0.0428	0.6993	-	-	30.10	333M	3	1.69
PaSCo(M=1)	0.6181	4.6559	0.0610	0.8250	26.49	15.36	28.22	**111M**	**1**	**0.67**
PaSCo (ours)	**0.4922**	**3.9155**	**0.0426**	**0.5835**	**31.42**	**16.51**	**30.11**	120M	**1**	1.32

SSCBench-KITTI360 (test set)

method	ins ece↓	ins nll↓	voxel ece↓	voxel nll↓	All PQ†↑	All PQ↑	mIoU↑	Params↓	Passes ↓	Time(s) ↓
TTA	-	-	0.1580	2.1282	-	-	21.78	**111M**	2	0.85
MC Dropout [23]	-	-	0.1548	2.0737	-	-	21.73	**111M**	2	0.79
Deep Ensemble [41]	-	-	0.1540	2.0653	-	-	**22.51**	222M	2	0.90
PaSCo(M=1)	0.7899	5.4405	0.1749	2.3556	19.53	9.91	21.17	**111M**	**1**	**0.38**
PaSCo (ours)	**0.6015**	**4.1454**	**0.1348**	**1.6112**	**26.29**	**10.92**	22.39	115M	**1**	0.65

Table 2. **Uncertainty evaluation.** We evaluate uncertainty on Semantic KITTI (top) and SSCBench-KITTI360 (bottom). Baselines only produce voxel uncertainty ('voxel ece,' 'voxel nll') which we outperform while also estimating PSC uncertainty (All PQ/PQ†).

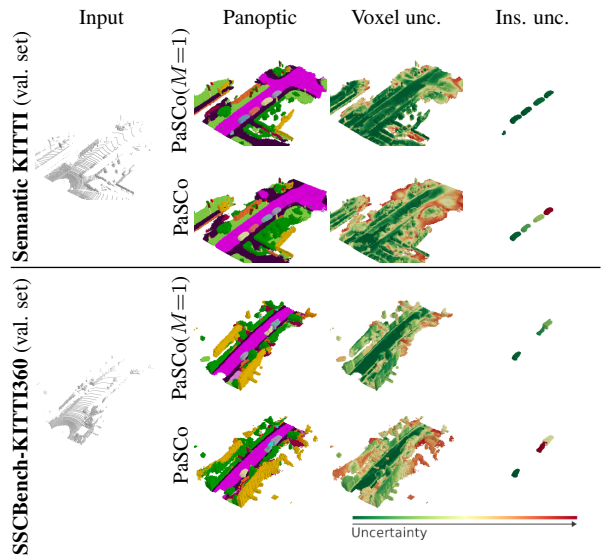

Figure 5. **Qualitative uncertainty comparison on SSCBench-KITTI360 and Semantic KITTI.** Note that "ins. unc." only shows examples from the "thing" class for clearer visualization. PaSCo(M=1) tends towards overconfidence in both voxel and ins. unc. In contrast, PaSCo gives more intuitive uncertainty estimates, e.g., at segment boundaries, in areas with hallucinated scenery, and in regions with low input point density.

Ensemble) equal to the number of subnets in PaSCo — i.e., 3 on Semantic KITTI and 2 on SSCBench-KITTI360. Notably, baselines can only estimate voxel-wise uncertainties, for which we outperform by a large margin. Only the voxel ece of Deep Ensemble for Semantic KITTI is a close second (0.0428 vs. 0.0426), though at the cost of ≈ 3 times our parameters count, 3 passes, and is ≈30% slower. Comparing PaSCo(M=1) and PaSCo highlights that our ensemble approach brings a clear boost on all metrics at a minor increase of number of parameters (111M vs 115M).

Fig. 5 visualizes uncertainty estimation from PaSCo(M=1) and PaSCo on Semantic KITTI and SSCBench-KITTI360. For clarity, instance-wise uncertainty shows only "thing" categories. PaSCo(M=1) often shows high confidence, likely due to deep networks' tendency for overconfidence [28]. For voxel-wise uncertainty, PaSCo exhibits increased uncertainty at segment boundaries (e.g., roads, sidewalks), low point density areas, and large missing regions. Instance-wise, PaSCo indicates more uncertainty in regions with ambiguous predictions, like sparse input points or close object proximity.

method	**Semantic KITTI** (val set) All PQ†↑	All PQ↑	ins ece↓	ins nll↓	Passes ↓	**SSCBench-KITTI360** (test set) All PQ†↑	All PQ↑	ins ece↓	ins nll↓	Passes ↓
TTA	28.16	15.95	0.5295	4.3804	3	23.31	9.76	0.6953	4.8958	2
MC Dropout	29.62	16.11	0.5684	4.8617	3	23.73	9.95	0.6804	4.6174	2
Deep Ensemble	30.71	16.41	0.5008	3.9181	3	23.85	9.88	0.6673	4.7809	2
PaSCo (Ours)	**31.42**	**16.51**	**0.4922**	**3.9155**	**1**	**26.29**	**10.92**	**0.6015**	**4.1454**	**1**

Table 3. **Effect of our ensembling.** We apply our permutation-invariant ensembling strategy (Sec. 3.2.2) to all baselines to enable PSC uncertainty estimation. Even when using our technique, we note PaSCo remains the best performing.

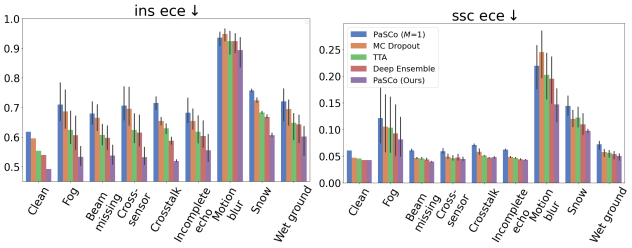

Figure 6. **Effects of Out Of Distribution.** We evaluate uncertainties on corruptions of the Robo3D [39], shown in the x axis. Each bar reports the metric average per corruption while its error bar indicates the per-intensity minimum and maximum metric. PaSCo outperforms all methods by a large margin on all corruptions for instance-wise uncertainty (left) and better on 7 of 8 conditions (xcept 'cross-talk') on voxel-wise uncertainty (right).

Mask ensembling. As the uncertainty-aware baselines do not estimate instance uncertainties, we apply our permutation-invariant ensembling (Sec. 3.2.2) to all baselines, in order to enable instance-wise uncertainty estimation for all. Tab. 3 shows that our MIMO-strategy performs better than the baselines on all metrics, using a single pass.

Effects of Out Of Distribution. In the literature, uncertainty is classically used as a proxy of robustness to Out Of Distribution (OOD). To complement our study, we evaluate on the Robo3D [39], which provides point cloud under eight types of corruptions (*e.g.*, fog, beam missing, cross-sensor, wet ground, *etc.*), each with three level of intensities (light, moderate, heavy). We evaluate on the complete set of 24 corruptions and plot instance and voxel uncertainties in Fig. 6, showing that PaSCo demonstrates consistent improvement over baselines. Each bar shows the mean uncertainty of a method on a given corruption, while the error bar shows the per-level minimum and maximum uncertainties. Interestingly, we note that instance (Fig. 6, left) and voxel (Fig. 6, right) uncertainties are not strongly correlated, although methods' rankings remain rather stable across conditions. For instance-wise uncertainty ('ins ece'), PaSCo is significantly better than *all* baselines on *all* corruptions, im-

# subnets	Semantic KITTI (val. set) All PQ†↑	All PQ↑	mIoU↑	ins ece↓	ins nll↓	voxel ece↓	voxel nll↓	SSCBench-KITTI360 (val. set) All PQ†↑	All PQ↑	mIoU↑	ins ece↓	ins nll↓	voxel ece↓	voxel nll↓	# Params↓
1	26.49	15.36	28.22	0.6181	4.6559	0.0610	0.8250	18.87	7.77	20.59	0.8355	6.2581	0.1744	2.5785	**111M**
2	30.34	**17.23**	30.04	0.5535	4.1474	0.0530	0.6449	**27.20**	**8.36**	**21.63**	0.6022	4.4120	0.1285	1.8063	115M
3	**31.42**	16.51	**30.11**	**0.4922**	**3.9155**	0.0426	0.5835	22.31	6.88	20.60	0.5293	3.4189	0.1233	1.6188	120M
4	31.20	16.33	29.41	0.5304	4.2681	**0.0349**	**0.5572**	23.23	6.49	20.34	**0.4098**	**2.4370**	**0.1011**	**1.4814**	125M

Table 4. **Performance when varying number of subnets on Semantic KITTI [3] and SSCBench-KITTI360 [48]** validation sets. PSC performance improves as the number of M increases, peaking at M=3 for Semantic KITTI and at M=2 for SSCBench-KITTI360. Further increasing the subnets can also help with uncertainty estimates. We choose M=3 for Semantic KITTI and $M = 2$ for SSCBench-KITTI360 to balance high PSC performance and uncertainty estimation.

	All PQ†↑	All PQ↑	mIoU↑	ins ece↓	ins nll↓	voxel ece↓	voxel nll↓
w/o augmentation	27.89	14.07	28.30	0.5031	4.4245	0.0442	0.6713
w/o rotation augmentation	28.84	14.95	28.95	0.5074	4.2987	0.0432	0.6309
w/o voxel-query sem. loss	28.82	15.55	29.87	0.5205	4.2909	0.0437	0.5878
w/o sem. pruning	30.12	15.04	29.04	0.5380	4.1814	0.0440	0.5980
PaSCo (Ours)	**31.42**	**16.51**	**30.11**	**0.4922**	**3.9155**	**0.0426**	**0.5835**

Table 5. **Method ablation.** We ablate inference (top) and training (bottom) components of our method, showing that each contributes to the best performance.

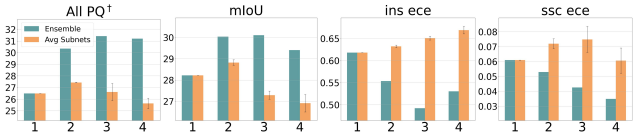

Figure 7. **Ensemble vs subnets averaging.** We compare our ensemble method with averaging individual subnets, across varying # subnets (x-axis). Error bars show standard deviation across subnets. Peak performance is at M=3, where our ensembling compensates for reduced per-subnet capacity with more subnets.

proving in 7 out of 8 on voxel-wise uncertainty ('ssc ece').

4.3. Ablation Studies

Method ablation. We ablate our method on SemanticKITTI [3] in Tab. 5 and report SSCBench-KITTI360 in the supp. The upper table ablates our inference augmentations (*i.e.*, rotation+translation), which benefit overall performance, especially All-PQ†/PQ. We attribute this to the increased variance profitable to the subnetworks as in MIMO [30]. In the lower table, we retrain PaSCo while removing some components. We show that removing our voxel-query semantic loss $\mathcal{L}'_{\text{sem}}$ (Eq. (3)) harms training; such proxy supervision boosts performance at no additional cost. Finally, 'w/o sem. pruning' replaces our semantic pruning with binary occupancy pruning [14, 17], resulting in degraded performance due to loss of smaller classes.

Subnets ablation. Tab. 4 ablates different numbers of subnets $M \in \{1, 2, 3, 4\}$ on the validation sets of our main datasets. Our main PQ metrics increase significantly with more subnets, though plateauing at $M = 3$ for SemanticKITTI and $M = 2$ for SSCBench-KITTI360. This is due to the preserved constant computational cost implying that more subnets mean less per-subnet capacity, leading to more noise in the ensembling. Our finding confirms that of MIMO [30] in the classification setting, though we argue our plateau is reached before since PSC being is a much more complex task than classification.

Mask matching. We ablate our mask matching, substituting our 'soft matching' (sigmoid probabilities) with 'hard matching' with binary mask IoU for assignment cost matrix calculation (Sec. 3.2.2). This results in a large drop in All-PQ†/PQ of -3.75/-1.12. Entirely removing mask matching severely impacts mask quality, dropping to 0.02/0.02.

Ensemble vs. Subnets averaging. To further shed light on subnets performance, Fig. 7 displays metrics of SemanticKITTI as a function of number of subnets M for our ensembling (Sec. 3.2.2) or the averaging of the individual subnet performance. When averaging subnets, optimal performance is reached at $M = 2$, with larger M unable to solve PSC efficiently. However, ensembling reaches improved performance at $M = 3$, showing that our ensembling effectively leverages weaker subnets.

Limitations. Like MIMO [30], our method can accommodate a limited number of subnets, depending on the task nature and the network's capacity. Our approach may overlook objects or mix up nearby objects, particularly when they are small and exhibit semantic resemblance. Our method does not distinguish between types of uncertainty, such as epistemic or aleatoric. Exploring this aspect could be a valuable direction for future research.

5. Conclusion

We first address Panoptic Scene Completion (PSC) which aims to complete scene geometry, semantics, and instances from a sparse observation. We introduce an efficient ensembling method complemented by a novel technique that combines predictions of unordered sets, enhancing the overall prediction accuracy and reliability in terms of uncertainty.

Acknowledgment. The research was funded by the ANR project SIGHT (ANR-20-CE23-0016), the ERC Starting Grant SpatialSem (101076253), and SAMBA collaborative project co-funded by BpiFrance in the Investissement d'Avenir Program. Computation was performed using HPC resources from GENCI–IDRIS (2023-AD011014102, AD011012808R2). We thank all Astra-Vision members for their valuable feedbacks, including Andrei Bursuc and Gilles Puy for excellent suggestions and Tetiana Martyniuk for her kind proofreading.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021. 6

[2] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *MIDL*, 2018. 6

[3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *ICCV*, 2019. 2, 5, 8

[4] 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 *CVPR*, 2018. 4

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015. 2

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[7] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, 2021. 1, 2

[8] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022. 2

[9] Hui Chen, Man Liang, Wanquan Liu, Weina Wang, and Peter Xiaoping Liu. An approach to boundary detection for 3d point clouds based on dbscan clustering. *Pattern Recognition*, 2022. 5

[10] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, 2020. 1, 2

[11] Yueh-Tung Chen, Martin Garbade, and Juergen Gall. 3d semantic scene completion from a single depth image using adversarial training. In *ICIP*, 2019. 2

[12] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 1, 2, 3, 4, 5

[13] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 2, 3, 4

[14] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Bingbing Liu. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *CoRL*, 2020. 1, 2, 3, 8

[15] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34: 8282–8293, 2021. 2

[16] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 1, 2

[17] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *CVPR*, 2020. 3, 8

[18] Aloisio Dourado, Teofilo E. de Campos, Hansung Kim, and Adrian Hilton. EdgeNet: Semantic scene completion from a single RGB-D image. In *ICPR*, 2020. 2

[19] Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *CVPR*, 2021. 2

[20] Michael W. Dusenberry, Ghassen Jerfel, Yeming Wen, Yi-An Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *ICML*, 2020. 2

[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 5

[22] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and Isabelle Bloch. Tradi: Tracking deep neural network weight distributions for uncertainty estimation. In *ECCV*, 2020. 2

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 6, 7

[24] Martin Garbade, Johann Sawatzky, Alexander Richard, and Juergen Gall. Two stream 3d semantic scene completion. In *CVPRW*, 2019. 2

[25] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *NeurIPS*, 2018. 2

[26] Stefano Gasperini, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, and Federico Tombari. Panoster: End-to-end panoptic segmentation of LiDAR point clouds. *RA-L*, 2021. 2

[27] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013. 5

[28] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*. PMLR, 2017. 5, 7

[29] Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. Evaluating scalable bayesian deep learning methods for robust computer vision. In *CVPRW*, 2020. 2

[30] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M. Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *ICLR*, 2021. 1, 2, 4, 5, 8

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[32] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *CVPR*, 2021. 2

[33] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 2

[34] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. In *ICLR*, 2017. 2

[35] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 2023. 2

[36] Juana Valeria Hurtado, Rohit Mohan, Wolfram Burgard, and Abhinav Valada. Mopt: Multi-object panoptic tracking. In *CVPR*, 2020. 2

[37] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An Introduction to Variational Methods for Graphical Models*, page 105–161. MIT Press, 1999. 2

[38] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2, 5

[39] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *ICCV*, 2023. 5, 7

[40] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955. 4

[41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 2, 4, 5, 6, 7

[42] Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed-ensembles for efficient uncertainty estimation. In *ICLR*, 2023. 2, 5

[43] Enxu Li, Ryan Razani, Yixuan Xu, and Bingbing Liu. Smac-seg: Lidar panoptic segmentation via sparse multi-directional attention clustering. In *ICRA*, 2022. 2

[44] Enxu Li, Ryan Razani, Yixuan Xu, and Bingbing Liu. Cpseg: Cluster-free panoptic segmentation of 3d lidar point clouds. In *ICRA*, 2023. 2

[45] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, 2020. 1, 2

[46] Jie Li, Laiyan Ding, and Rui Huang. Imenet: Joint 3d semantic scene completion and 2d semantic segmentation through iterative mutual enhancement. In *IJCAI*, 2021. 2

[47] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *CVPR*, 2022. 2

[48] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Sscbench: Monocular 3d semantic scene completion benchmark in street views. *arXiv*, 2023. 5, 8

[49] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, 2023. 2

[50] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *TPAMI*, 2022. 5

[51] Minzhe Liu, Qiang Zhou, Hengshuang Zhao, Jianing Li, Yuan Du, Kurt Keutzer, Li Du, and Shanghang Zhang. Prototype-voxel contrastive learning for lidar point cloud panoptic segmentation. In *ICRA*, 2022. 2

[52] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and Think: Disentangling Semantic Scene Completion. In *NeurIPS*, 2018. 2

[53] Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity. In *2022*, 2022. 2

[54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[55] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 1992. 2

[56] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *NeurIPS*, 2019. 2

[57] R. Marcuzzi, L. Nunes, L. Wiesmann, J. Behley, and C. Stachniss. Mask-Based Panoptic LiDAR Segmentation for Autonomous Driving. *RA-L*, 2023. 2, 3, 4, 6

[58] Jianbiao Mei, Yu Yang, Mengmeng Wang, Zizhang Li, Xiaojun Hou, Jongwon Ra, Laijian Li, and Yong Liu. Centerlps: Segment instances by centers for lidar panoptic segmentation. In *ACM MM*, 2023. 2

[59] Andres Milioto, Jens Behley, Chris McCool, and Cyrill Stachniss. Lidar panoptic segmentation for autonomous driving. In *IROS*, 2020. 2

[60] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019. 2

[61] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *CVPR*, 2019. 5

[62] Ryan Razani, Ran Cheng, Enxu Li, Ehsan Taghavi, Yuan Ren, and Liu Bingbing. Gp-s3net: Graph-based panoptic sparse semantic segmentation network. In *ICCV*, 2021. 2

[63] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 2020. 1, 2, 4, 5, 6

[64] Luis Roldão, Raoul De Charette, and Anne Verroust-Blondet. 3D Semantic Scene Completion: a Survey. *IJCV*, 2021. 1, 2, 3

[65] Kshitij Sirohi, Rohit Mohan, Daniel Büscher, Wolfram Burgard, and Abhinav Valada. Efficientlps: Efficient lidar panoptic segmentation. *T-RO*, 2021. 2

[66] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 1, 2

[67] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 5

[68] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. ForkNet: Multi-branch volumetric semantic completion from a single depth image. In *ICCV*, 2019. 2

[69] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *CVPR*, 2024. 2

[70] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. In *ICLR*, 2020. 2, 4

[71] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *NeurIPS*, 2020. 2

[72] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *CVPR*, 2023. 1, 2, 4, 5, 6

[73] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, 2021. 1, 2, 4, 5, 6

[74] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, 2018. 2

[75] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *ICCV*, 2019. 2