Failure Modes for Deep Learning–Based Online Mapping: How to Measure and Address Them

Supplementary Material

7. Metric Definitions and Details

In the following, we provide the mathematical definitions for the Chamfer distance and the corresponding AP metric in Sec. 3.2 as well as the discrete Fréchet distance [6] used in $\mathrm{sim}(\cdot,\cdot)$ in Sec. 3.1 and M in Sec. 3.2. Finally, we provide the correlation between M and mAP as a verification for the eligibility of M as a performance measure.

7.1. Chamfer Distance

Let $P, Q \subset \mathbb{R}^d$ be two finite point sets, each representing one map element (polygon or polyline) after uniformly resampling it to N_{pts} vertices. The Chamfer distance d_{ch} between P and Q is defined as

$$d_{\text{ch}}(P,Q) = \frac{1}{2}\left(\frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2\right). \tag{10}$$

7.2. Average Precision

For each class $c_i \in C_{\text{map}}$, we compute the AP across all samples in the set by declaring a prediction a true positive if its Chamfer distance D_{Chamfer} to the ground truth is below a threshold τ. We evaluate the AP separately for each threshold $\tau \in T$, where $T = \{0.5, 1.0, 1.5\}$, and then take the average across all thresholds T and all classes C_{map} to obtain the final mean average precision (mAP) score used for comparison:

$$\text{mAP} = \frac{1}{|C_{\text{map}}||T|}\sum_{c_i\in C_{\text{map}}}\sum_{\tau\in T}\text{AP}_{c_i,\tau}. \tag{11}$$

7.3. Discrete Fréchet Distance

Let P and Q be two map elements represented as polygons or polylines in $\mathbb{R}^d$, and let

$$\sigma(P) = (p_1,\dots,p_n), \qquad \sigma(Q) = (q_1,\dots,q_m) \tag{12}$$

be the corresponding sequences of uniformly resampled vertices, with $p_i, q_j \in \mathbb{R}^d$.

A *coupling* L between P and Q is a sequence of distinct pairs between $\sigma(P)$ and $\sigma(Q)$,

$$L = \big((p_{a_1}, q_{b_1}),\dots,(p_{a_K}, q_{b_K})\big), \tag{13}$$

δ	0	5	10	20	30	40	50	60
$r(s(v), M(v))$	0.345	0.535	0.555	0.567	0.572	0.574	0.572	0.568

Table 3. Pearson correlation r between $s(v)$ and $M(v)$ on the nuscenes original split for different δ for unmatched elements.

where $(a_k)_{k=1}^K$ and $(b_k)_{k=1}^K$ are nondecreasing surjective index sequences, i.e.

$$a_1 = 1,\ a_K = n,\ b_1 = 1,\ b_K = m, \tag{14}$$

$$\{a_1,\dots,a_K\} = \{1,\dots,n\}, \quad \{b_1,\dots,b_K\} = \{1,\dots,m\}, \tag{15}$$

and for all $r < s$,

$$a_r \le a_s, \qquad b_r \le b_s. \tag{16}$$

The norm $\|L\|$ of a coupling L is the length of its longest pair,

$$\|L\| = \max_{k=1,\dots,K}\|p_{a_k} - q_{b_k}\|_2. \tag{17}$$

The discrete Fréchet distance d_{fr} between P and Q is then defined as

$$d_{\text{fr}}(P,Q) = \min\big\{\|L\| \;\big|\; L \text{ is a coupling between } P \text{ and } Q\big\}. \tag{18}$$

7.4. Effect of δ on Geometric Similarity $s(v)$

To examine the sensitivity of δ, we computed the Pearson correlation r between $s(v)$ and $M(v)$ on the nuScenes original split for different penalties δ for unmatched elements (cf. Sec. 7.3). As expected, the correlation stays almost constant from $\delta = 20$ onwards, since resulting large costs for sample pairs with lots of unmatched elements rarely match the minimum condition from $s(v)$.

7.5. Details on Performance Measure M

In the following, we provide additional details on M which is defined in Sec. 3.2 regarding map element cardinality, handling of multiple predictions per ground truth element and practical confidence threshold tuning. Finally, we correlate M and mAP to prove the eligibility as a performance measure.

For all examined online mapping architectures in the experiments [17, 18, 21, 23], the predicted map elements have the same number of points compared to the ground

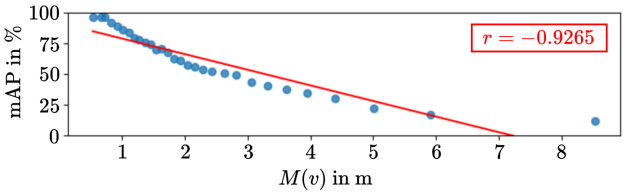

Figure 7. Correlation between $M(v)$ and mAP for bins with 200 samples from the nuScenes original validation split

truth to match cardinality. If the number of vertices per map element do not match, we advise to subsample or simplify to predicted map elements w.r.t. to the number of vertices in the ground truth, since the discrete Fréchet distance comparison between prediction and ground truth relies on discrete points and large distances between points could lead to inaccurate results. Differences in map element cardinality are handled analogously to $\mathrm{sim}(\cdot,\cdot)$ by considering all permutations: For polylines, original and reversed; for polygons, all cyclic permutations in both orientations (original and reversed).

Note that similarly to the Chamfer-based AP metric, multiple predictions for the same ground truth element are not penalized in order to not introduce overwhelming complexity to the metric. In practice, we examine that the predicted map elements that are matched to a ground truth map element yield significantly higher confidence scores in comparison to the unmatched predictions, so the proposed measure could also be used to determine a suitable confidence threshold for deployment of a model.

Correlation between M and mAP. We validate M as a performance measure by correlating it with mAP on sufficiently large sample sets. Specifically, we partition the nuScenes validation set into bins of 200 samples sorted by M, compute mAP per bin, and report the resulting correlation in Fig. 7. The two measures exhibit a strong negative correlation ($r = -0.9265$), confirming that $M(v)$ is a suitable metric for evaluation.

8. Mathematical Details on Evaluation Set and Failure Mode Measure Derivation

In this section, we provide mathematical background regarding the derivation of evaluation sets for quantifying localization and geometric overfitting in Sec. 3.1 and geometric overfitting Sec. 3.3.

8.1. Similarity Distribution Alignment between Geographically Separated Validation Sets

We derived the evaluation sets V_{close} and V_{far} and want to align the distributions w.r.t. $s(v)$ to mitigate the effect caused by correlation to $d(v)$ and correctly draw conclusions regarding localization overfitting.

The empirical cumulative distribution $F_{s,V_{\text{sub}}}$ of $s(v_{\text{sub}})$ for a subset $V_{\text{sub}} \subset V$ is described by

$$F_{s,V_{\text{sub}}}(t) \;:=\; \frac{1}{|V_{\text{sub}}|}\,\big|\{\, v_{\text{sub}} \in V_{\text{sub}} \mid s(v_{\text{sub}}) \le t\}\big|. \quad (19)$$

To match the distributions w.r.t. $s(v)$ between V_{close} and V_{far}, we sample the subsets $V_{\text{close}*} \subset V_{\text{close}}$ and $V_{\text{far}*} \subset V_{\text{far}}$ by approximating

$$F_{s,V_{\text{close}*}}(t) \approx F_{s,V_{\text{far}*}}(t)\ \forall\, t. \quad (20)$$

To obtain $V_{\text{close}*}$ and $V_{\text{far}*}$, we perform bipartite matching of $s(v)$ across both sets, followed by filtering based on a predefined threshold for the absolute value of the differences between $s(v)$ for the matched pairs. To ensure that the geometric similarity distributions of $s(v)$ for V_{close} and $V_{\text{far}*}$ are closely aligned, we choose the threshold such that their Kullback–Leibler divergence remains below 0.01.

8.2. Bin Separation for Geographically Distant Validation Set

We aim to quantify geometric overfitting by stratifying the geographically distant subset V_{far}.

Let $\tau_0 \le \tau_1 \le \ldots \le \tau_b$ partition $[\min_{v\in V_{\text{far}}} s(v),\, \max_{v\in V_{\text{far}}} s(v)]$ into b equal-width intervals. For $i = 1, \ldots, b$, we define

$$B_i := \{\, v \in V_{\text{far}} \mid \tau_{i-1} < s(v) \le \tau_i \,\}. \quad (21)$$

8.3. Definition of Geometric Overfitting Score

We use the mean value of $s(v)$ within each interval $[\tau_{i-1}, \tau_i]$, denoted as $\mu_{\text{s,far},i}$, together with the corresponding $M_{\text{far},i}$ value for the linear regression. To account for differing bin sizes, we weight the regression data points by their sample counts w_i. Formally, we propose the following geometry overfitting score $\mathcal{O}_{\text{geom}}$:

$$\mathcal{O}_{\text{geom}} = \frac{\sum_{i=1}^{b} p_i \big(\mu_{\text{s,far},i} - \mu_x\big)\big(M_{\text{far},i} - \mu_y\big)}{\sum_{i=1}^{b} p_i \big(\mu_{\text{s,far},i} - \mu_x\big)^2} \quad \text{where}$$

$$p_i = \frac{w_i}{\sum_{j=1}^{b} w_j},\quad \mu_x = \sum_{i=1}^{b} p_i\, \mu_{\text{s,far},i},\quad \mu_y = \sum_{i=1}^{b} p_i\, M_{\text{far},i}. \quad (22)$$

9. Topology-Based Similarity Measure

Motivation. Besides examining the effect of geometrical similarity, we investigate whether online mapping models are capable of learning translation- and rotation-invariant features. Recent NeRF-based simulation studies have shown that such models suffer severe performance degradation under ego-centric translation and rotation perturbations, even

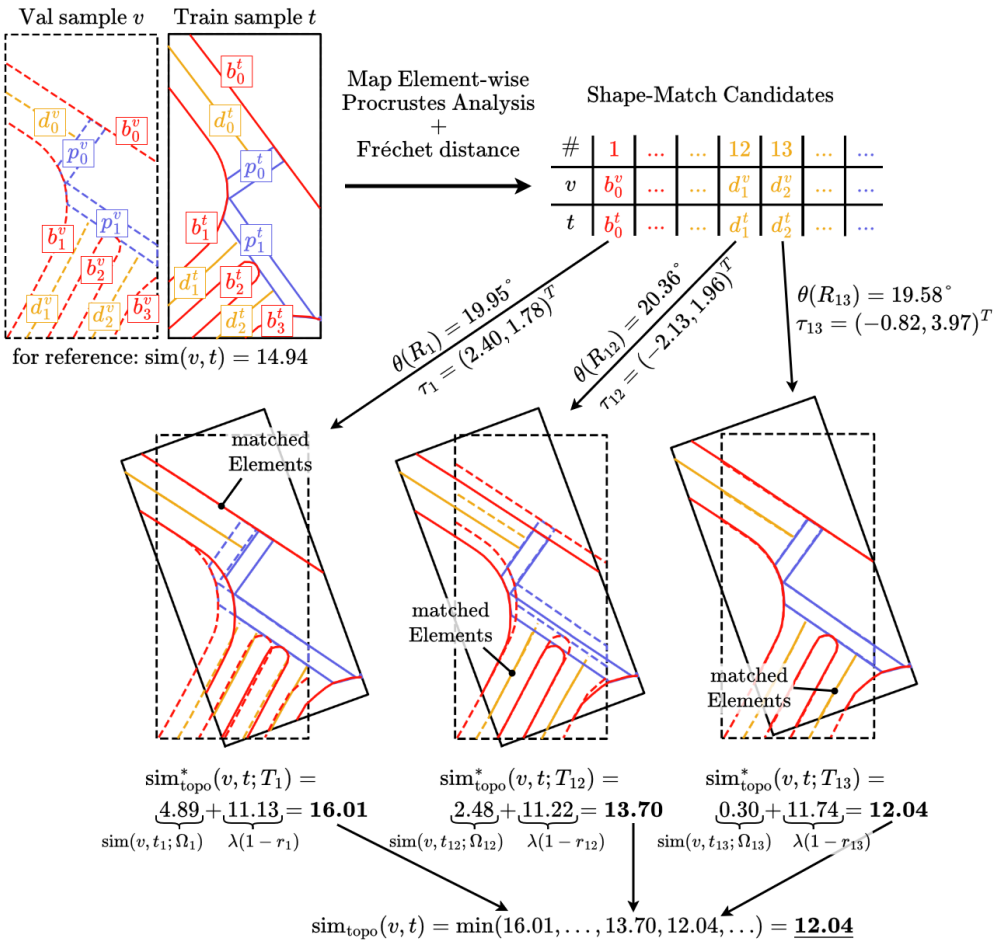

Figure 8. Visualization of the topology-based similarity measure $\mathrm{sim}_{\mathrm{topo}}(v,t)$ and its computation steps. Candidate rigid transformations are derived via class and map element-wise Procrustes analysis and discrete Fréchet distance matching. For each transformed training sample, similarity is evaluated within the overlapping field-of-view and penalized for low overlap. The final similarity is obtained by selecting the minimum across all candidates.

when the underlying road topology remains unchanged [20]. This raises the question of whether current models truly exploit pose-invariant topological structure or instead rely on pose-dependent geometric cues. To this end, we introduce the similarity measure $\mathrm{simtopo}(v,t)$, which captures translation- and rotation-invariant similarity between samples by comparing their topological structures rather than the geometrical patterns in the field of view as in $\mathrm{sim}(v,t)$. We base $\mathrm{simtopo}(v,t)$ on $\mathrm{sim}(v,t)$ with a preceding alignment step. A visualization of the process used to derive the topology-based similarity measure for an exemplary validation sample is shown in Fig. 8, we advise following the steps in the figure alongside the mathematical definition below.

Definition. Let $\Omega_v \subset \mathbb{R}^2$ and $\Omega_t \subset \mathbb{R}^2$ denote the FOV of the validation sample v and the training sample t, respectively. We first obtain a finite set of candidate rigid transformations

$$\mathcal{T}(v,t) = \{T_k(x) = R_k x + \tau_k \mid k = 1, \ldots, K\}, \quad (23)$$

by comparing the shapes of all map elements within the same class using Procrustes analysis (without uniform scaling) followed by measuring similarity using discrete Fréchet distance (cf. Sec. 7.3) and selecting the k top matches. Each transformation T_k is defined by a rotation matrix $R_k \in \mathrm{SO}(2)$ and a translation vector $\tau_k \in \mathbb{R}^2$.

For a given candidate T_k, we transform the training sample

$$t_k := T_k(t), \quad (24)$$

and consider only the part of the scene that lies in the overlapping FOV

$$\Omega_k := \Omega_v \cap T_k(\Omega_t). \quad (25)$$

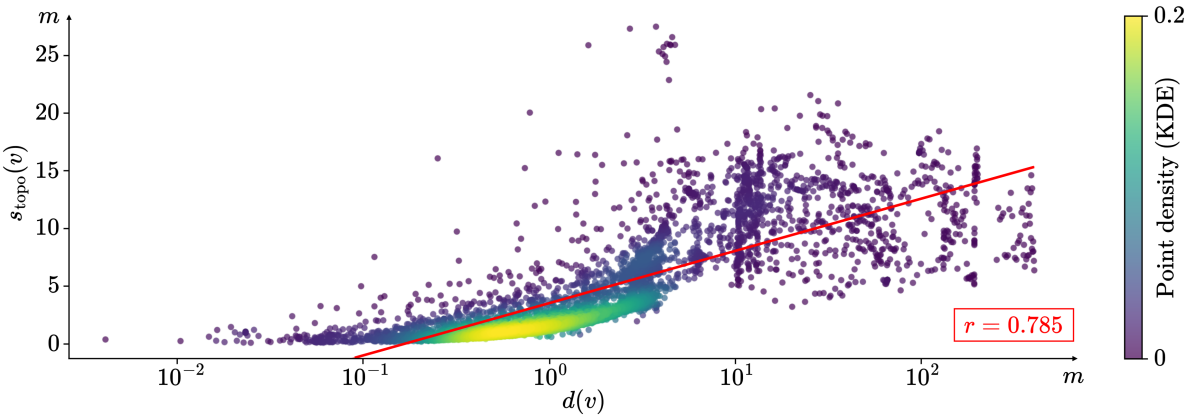

Figure 9. Correlation between the topology-based similarity $s_{\text{topo}}(v)$ and the geographical distance $d(v)$ for the nuScenes original split. Compared to the geometric similarity $s(v)$, the topology-based score exhibits a stronger correlation and a less cluttered distribution (Pearson correlation coefficient $r = 0.785 > 0.724$, cf. Fig. 3).

We denote by $\text{sim}(v, t_k; \Omega_k)$ the original similarity sim evaluated only on those map elements of v and t_k whose geometry lies inside or intersects Ω_k. All elements outside Ω_k are discarded and intersecting elements are clipped, since elements outside of Ω_k cannot occur in the opposing sample.

To avoid degenerate alignments through trivial solutions where the overlap between FOVs becomes very small and no map elements lie inside or intersect with Ω_k, we penalize low-overlap candidates. We define the overlap ratio

$$r_k \;:=\; \frac{|\Omega_k|}{|\Omega_v|} \in [0, 1] \tag{26}$$

and add a penalty term that increases linearly as the overlap decreases. With a weight parameter $\lambda \geq 0$, we define the topology-based similarity for a candidate transform T_k as

$$\text{sim}^*_{\text{topo}}(v, t; T_k) \;:=\; \text{sim}(v, t_k; \Omega_k) \;+\; \lambda\big(1 - r_k\big). \tag{27}$$

Finally, the topology-based similarity between v and t is obtained by minimizing over all candidate transforms:

$$\text{sim}_{\text{topo}}(v, t) \;:=\; \min_{T_k \in \mathcal{T}(v,t)} \text{sim}^*_{\text{topo}}(v, t; T_k). \tag{28}$$

Analogous to $s(v)$, we define $s_{\text{topo}}(v)$ as the lowest similarity cost between v and any training sample $t \in T$ with the new topological similarity measure across all candidate transforms

$$s_{\text{topo}}(v) := \min_{t \in T} \text{sim}_{\text{topo}}(v, t). \tag{29}$$

Results. For this ablation study, we aim to validate our translation- and rotation-invariant similarity measure s_{topo} and compare it against s in terms of their correlation with performance. This allows us to determine whether overfitting is more strongly driven by geometric patterns or by topological structure.

We begin by examining the correlation between $s_{\text{topo}}(v)$ and $d(v)$ in the nuScenes original split in Fig. 9. As expected, the correlation is stronger compared to $s(v)$, and the plot is visibly less cluttered (Pearson correlation coefficient $r = 0.785 > 0.724$, cf. Fig. 3). This is because samples that lie close to each other tend to share similar topological structure, whereas their geometric structure, which is sensitive to translation and rotation, differs more significantly. This suggests that the metric accurately captures topological structure rather than pure geometric alignment. To further substantiate this claim, we compare $s(v)$ and $s_{\text{topo}}(v)$ in Fig. 10. Several samples deviate from the bisector of the two axes, most often with $s(v) > s_{\text{topo}}(v)$, indicating that a closer topological match has been identified (cf. examples in Fig. 10).

We also reexamine the correlation between $s_{\text{topo}}(v)$ and the per-sample performance $M(v)$ on the original nuScenes split in Fig. 11. The Pearson correlation is slightly lower for $s_{\text{topo}}(v)$ than for $s(v)$ ($r = 0.552 < 0.568$, cf. Fig. 5), suggesting that the online mapping model relies more on rotation- and translation-dependent geometric features than on invariant topological features.

To support this claim, we examine the correlation between $s(v)$ and $M(v)$ against the correlation between $s_{\text{topo}}(v)$ and $M(v)$ for all splits that are examined in Sec. 5. The results are shown in Tab. 4. While the Pearson correlation is in a similar range per split, we see slightly stronger correlation for the geometrical similarity measure $s(v)$ across all original and geographical splits. However, in the geometrical splits we introduced, the correlation for s_{topo} is higher, even though both values indicate minimal correlation to M. This indicates, that for these validation sets where geometries are

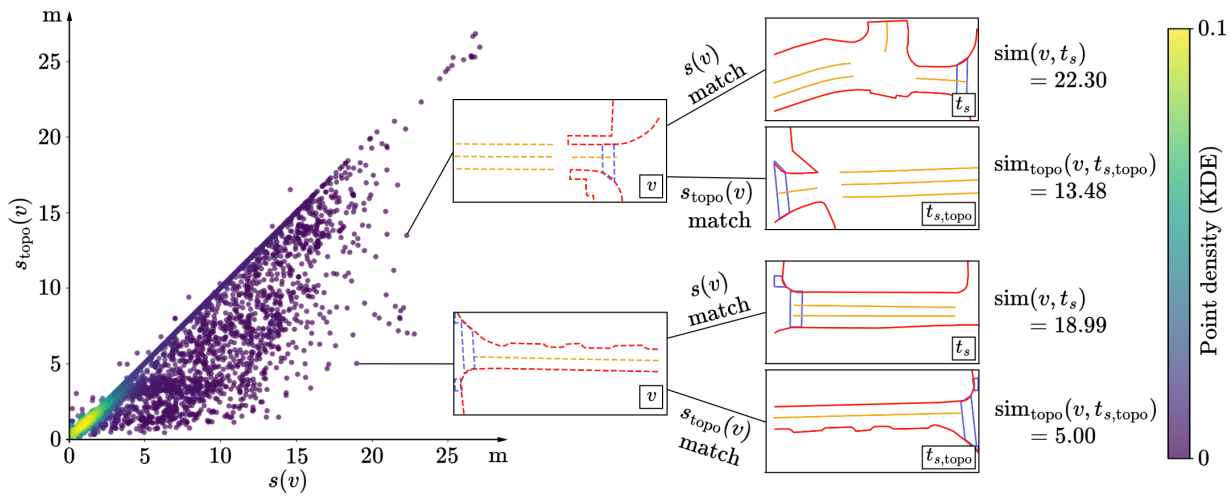

Figure 10. Comparison of geometric similarity $s(v)$ and topology-based similarity $s_{\text{topo}}(v)$ for the nuScenes original split. Samples deviating from the diagonal most often satisfy $s(v) > s_{\text{topo}}(v)$, demonstrating that $s_{\text{topo}}(v)$ identifies topologically similar scenes that differ geometrically due to translation or rotation. Two exemplary validation samples are displayed along their best matches from the training set for $s(v)$ and $s_{\text{topo}}(v)$.

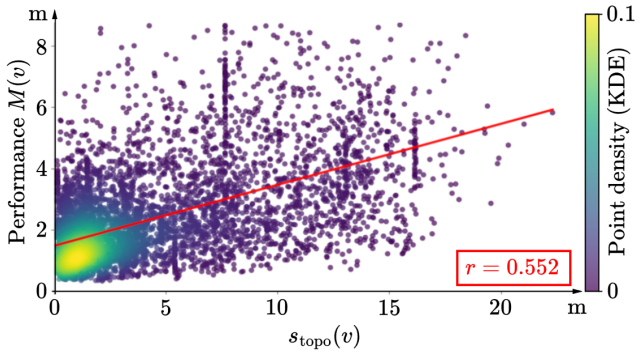

Figure 11. Correlation between the topology-based similarity $s_{\text{topo}}(v)$ and the per-sample performance $M(v)$ on the nuScenes original split. The slightly weaker correlation compared to $s(v)$ (Pearson correlation coefficient $r = 0.552 < 0.568$, cf. Fig. 5) indicates that current models rely more strongly on geometric patterns than on topological structure.

Dataset and split		$r(s(v), M(v))$	$r(s_{\text{topo}}(v), M(v))$
nuScenes	original	**0.568**	0.552
	geo. [19]	**0.226**	0.211
	geo. [37]	**0.275**	0.270
	geometric	0.137	**0.171**
Argoverse 2	original	**0.392**	0.298
	geo. [19]	**0.433**	0.366
	geo. [37]	**0.330**	0.202
	geometric	-0.071	**0.209**

Table 4. Pearson correlation between $M(v)$ and geometric similarity $s(v)$ versus topology-based similarity $s_{\text{topo}}(v)$ across all examined dataset splits. While $s(v)$ correlates more strongly with performance for original and geographical splits, $s_{\text{topo}}(v)$ shows mildly higher correlation for the geometric splits, where geometric similarity is intentionally minimized.

contained that are minimize similarity to the training split, topological alignment is still beneficial for model performance.

Our results indicate that, despite the existence of topologically consistent alignments, geometric similarity correlates more strongly with mapping performance across original and geographically disjoint splits, suggesting that current models encode strong ego-centric geometric priors rather than true pose-invariant representations. This observation is consistent with findings from NeRF-based perturbation analyses, which demonstrate that deviations from canonical ego poses cause predictions to collapse toward memorized geometric templates, highlighting geometry-dependent overfitting as a dominant failure mode [20].

Due to the lack of significant differences in the correlation, we expect similar results for any measures derived from $s_{\text{topo}}(v)$ in place of $s(v)$. Since the correlation with performance for all original and geographical splits is lower for $s_{\text{topo}}(v)$ compared to $s(v)$ and the effect of benefiting from topological alignment only seems to show for sets with very dissimilar geometries, we refrain from additional experiments.

10. Geometrical Dataset Split Derivation

Besides the original and geographical training splits, we want to examine model performance on a geometric dataset split, focusing on maximum geometric dissimilarity between training and evaluation sets. To be coherent with the other splits, the geometric split should also partition the data into 70/15/15% for the training, validation, and test sets, respectively.

We base our geometric split on the geometric similarity MST for the whole dataset. At first, we identify the edges in the MST with highest similarity costs, suggesting highest geometric dissimilarity. We then consider these edges as first-cut candidates and retain those that yield a subset of size near the desired training set size (70% of the dataset size). For each retained first cut, we search for a second cut in the remaining opposite subset (30% of the dataset size) to separate it into the validation and test set (each 15% of the dataset size). We then evaluate the resulting three subsets against size tolerances, and score candidates by a balance–separation criterion (small size deviation, large cut edge weights). Finally, we remove the chosen two edges and assign the three connected subtrees to train/val/test sets.

11. Exact Numbers for Dataset Sparsification

In the following, we list the exact numbers for Fig. 6 for the reader to reference. For each split, we report the sparsification threshold, remaining samples, MST length, and mAP values on the validation set corresponding to the plotted results.

Sparsification Threshold	Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
–	28130 (100.00 %)	96828.00 m (100.00 %)	60.95 (±0.00 %)
0.1	23084 (82.06 %)	96815.33 m (99.99 %)	61.47 (+0.85 %)
0.2	22752 (80.88 %)	96805.15 m (99.98 %)	61.56 (+1.00 %)
0.5	21417 (76.14 %)	96613.71 m (99.78 %)	61.48 (+0.87 %)
1	18397 (65.40 %)	95585.00 m (98.72 %)	59.71 (-2.03 %)
2	13969 (49.66 %)	92051.27 m (95.07 %)	56.71 (-6.96 %)
5	6512 (23.15 %)	72116.01 m (74.48 %)	48.11 (-21.07 %)
10	2184 (7.76 %)	37363.45 m (38.59 %)	28.50 (-53.24 %)

Table 5. Original nuScenes split

Sparsification Threshold	Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
–	27840 (100.00 %)	80572.21 m (100.00 %)	24.96 (±0.00 %)
0.1	22815 (81.95 %)	80559.35 m (99.98 %)	25.73 (+3.08 %)
0.2	22427 (80.56 %)	80543.07 m (99.96 %)	25.48 (+2.08 %)
0.5	20381 (73.21 %)	80197.95 m (99.54 %)	25.64 (+2.72 %)
1	16515 (59.32 %)	78843.58 m (97.85 %)	26.46 (+6.01 %)
2	11970 (43.00 %)	74909.77 m (92.97 %)	25.07 (+0.44 %)
5	5575 (20.03 %)	57577.21 m (71.46 %)	23.62 (-5.37 %)
10	1697 (6.10 %)	26597.42 m (33.01 %)	15.59 (-37.54 %)

Table 6. Geographically disjoint nuScenes split from [19] (Near Extrapolation Split)

Sparsification Threshold	Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
–	28008 (100.00 %)	90172.57 m (100.00 %)	28.53 (±0.00 %)
0.1	22801 (81.41 %)	90159.43 m (99.99 %)	27.56 (-3.40 %)
0.2	22451 (80.16 %)	90148.35 m (99.97 %)	28.44 (-0.32 %)
0.5	20840 (74.41 %)	89897.52 m (99.69 %)	27.19 (-4.70 %)
1	17414 (62.18 %)	88718.11 m (98.39 %)	27.95 (-2.03 %)
2	12825 (45.79 %)	85090.37 m (94.36 %)	28.19 (-1.19 %)
5	6051 (21.60 %)	67000.17 m (74.30 %)	25.57 (-10.38 %)
10	1964 (7.01 %)	34123.62 m (37.84 %)	18.10 (-36.56 %)

Table 7. Geographically disjoint nuScenes split from [37]

Sparsification Threshold	Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
–	22279 (100.00 %)	91003.65 m (100.00 %)	63.97 (±0.00 %)
0.1	19892 (89.29 %)	90995.80 m (99.99 %)	64.32 (+0.55 %)
0.2	19580 (87.89 %)	90991.00 m (99.99 %)	64.97 (+1.56 %)
0.5	18633 (83.63 %)	90909.49 m (99.90 %)	63.71 (-0.41 %)
1	17159 (77.02 %)	90555.71 m (99.51 %)	64.22 (+0.39 %)
2	14440 (64.81 %)	88289.72 m (97.02 %)	63.40 (-0.89 %)
5	6882 (30.89 %)	66745.42 m (73.34 %)	59.42 (-7.11 %)
10	2116 (9.50 %)	31739.04 m (34.88 %)	45.72 (-28.53 %)

Table 8. Original Argoverse 2 split, remaining samples in T and $\mathrm{geomdiv}(T)$ are computed for 2 Hz to be comparable to nuScenes. Performance results are for 10 Hz.

Sparsification Threshold	Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
–	22223 (100.00 %)	87302.62 m (100.00 %)	49.53 (±0.00 %)
0.1	19777 (88.99 %)	87294.12 m (99.99 %)	50.47 (+1.90 %)
0.2	19454 (87.54 %)	87289.24 m (99.98 %)	50.13 (+1.21 %)
0.5	18445 (83.00 %)	87188.20 m (99.87 %)	49.50 (-0.06 %)
1	16731 (75.29 %)	86747.95 m (99.36 %)	50.18 (+1.31 %)
2	13827 (62.22 %)	84389.69 m (96.66 %)	49.78 (+0.50 %)
5	6636 (29.86 %)	63978.80 m (73.28 %)	46.85 (-5.41 %)
10	1919 (8.64 %)	28972.13 m (33.19 %)	36.35 (-26.61 %)

Table 9. Geographically disjoint Argoverse 2 split from [19] (Near Extrapolation Split), remaining samples in T and $\mathrm{geomdiv}(T)$ are computed for 2 Hz to be comparable to nuScenes. Performance results are for 10 Hz.

Sparsification Threshold	Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
–	23954 (100.00 %)	97153.16 m (100.00 %)	57.61 (±0.00 %)
0.1	21391 (89.30 %)	97144.47 m (99.99 %)	57.55 (-0.10 %)
0.2	21077 (87.99 %)	97138.62 m (99.99 %)	57.74 (+0.23 %)
0.5	20106 (83.94 %)	97050.71 m (99.89 %)	58.36 (+1.30 %)
1	18403 (76.83 %)	96625.41 m (99.46 %)	57.88 (+0.47 %)
2	15348 (64.07 %)	93994.73 m (96.75 %)	56.75 (-1.49 %)
5	7353 (30.70 %)	71206.04 m (73.29 %)	54.08 (-6.13 %)
10	2251 (9.40 %)	33621.14 m (34.61 %)	45.45 (-21.11 %)

Table 10. Geographically disjoint Argoverse 2 split from [37], remaining samples in T and $\mathrm{geomdiv}(T)$ are computed for 2 Hz to be comparable to nuScenes. Performance results are for 10 Hz.

Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
28130 (100.00 %)	96828.00 m (100.00 %)	60.95 (±0.00 %)
23084 (82.06 %)	87906.66 m (90.79 %)	59.79 (-1.90 %)
22752 (80.88 %)	87715.78 m (90.59 %)	59.50 (-2.38 %)
21417 (76.14 %)	85104.96 m (87.89 %)	59.36 (-2.61 %)
18397 (65.40 %)	78757.05 m (81.34 %)	58.48 (-4.05 %)
13969 (49.66 %)	68450.78 m (70.69 %)	56.79 (-6.83 %)
6512 (23.15 %)	44951.67 m (46.42 %)	48.79 (-19.95 %)
2184 (7.76 %)	22131.76 m (22.86 %)	31.70 (-47.99 %)

Table 11. Random sampling for remaining sample amounts from Tab. 5 from the original nuScenes split

Remaining Samples in T	$\mathrm{geomdiv}(T)$	Performance on V (mAP)
22279 (100.00 %)	91003.65 m (100.00 %)	63.97 (±0.00 %)
19892 (89.29 %)	86989.97 m (95.59 %)	63.79 (-0.28 %)
19580 (87.89 %)	86184.47 m (94.70 %)	63.67 (-0.47 %)
18633 (83.63 %)	84021.93 m (92.33 %)	63.06 (-1.42 %)
17159 (77.02 %)	80187.60 m (88.11 %)	63.95 (-0.03 %)
14440 (64.81 %)	73394.62 m (80.65 %)	62.34 (-2.55 %)
6882 (30.89 %)	46629.68 m (51.24 %)	62.04 (-3.02 %)
2116 (9.50 %)	21827.41 m (23.98 %)	57.18 (-10.61 %)

Table 12. Random sampling for remaining sample amounts from Tab. 8 from the original Argoverse 2 split, remaining samples in T and $\mathrm{geomdiv}(T)$ are computed for 2 Hz to be comparable to nuScenes. Performance results are for 10 Hz.