PointCSP: Cross-Sample Semantic Propagation and Stability Preservation in Self-Supervised Point Cloud Learning

Supplementary Material

To facilitate a comprehensive understanding of PointCSP, this supplementary materials provides supplementary technical details. These details include theoretical analysis details, extended comparative experiments, and visualization.

1. Theoretical Analysis of Cross-Sample State Propagation

In this section, we provide a detailed probabilistic analysis of *cross-sample state propagation* in SSM-like modules. Our goal is to formalize the intuitive statement that mixing information across samples reduces the dispersion of same-class representations and thus improves their consistency. Since nonlinear and data-dependent modules are difficult to analyze exactly, we focus on the following question:

Under standard simplifying assumptions, how does cross-sample mixing affect the intra-class covariance of learned features?

We will show that, under mild assumptions, cross-sample propagation *tightens the worst-case upper bound* on the intra-class covariance compared to independent per-sample processing. This is precisely the sense in which cross-sample propagation acts as a variance-shrinking mechanism on class-conditional embeddings.

1.1. Problem Setup and Notation

We fix a semantic class c and consider N samples from this class, represented by feature vectors

$$u_i \in \mathbb{R}^d, \qquad i = 1, \dots, N. \tag{1}$$

We adopt the standard additive decomposition

$$u_i = \mu_c + \xi_i, \tag{2}$$

where $\mu_c \in \mathbb{R}^d$ is the class-wise semantic mean and ξ_i represents the intra-class deviation of sample i from the mean.

Assumption 1 (i.i.d. class-conditional embeddings)**.** *For a fixed class c, the perturbations $\{\xi_i\}_{i=1}^N$ are independent and identically distributed (i.i.d.) with*

$$\mathbb{E}[\xi_i] = 0, \qquad \mathrm{Cov}(\xi_i) = \Sigma_u \in \mathbb{R}^{d\times d}. \tag{3}$$

Equivalently, u_i are i.i.d. draws from the class-conditional distribution $p(u \mid c)$ with

$$\mathbb{E}[u_i \mid c] = \mu_c, \qquad \mathrm{Cov}(u_i \mid c) = \Sigma_u. \tag{4}$$

Measure of representation consistency. For any random vector $X \in \mathbb{R}^d$ with finite covariance, we define its (total) variance as

$$\mathcal{V}(X) := \mathrm{tr}\big(\mathrm{Cov}(X)\big). \tag{5}$$

In particular, the intra-class variance of u_i is

$$\mathcal{V}(u_i) = \mathrm{tr}(\Sigma_u). \tag{6}$$

A smaller $\mathcal{V}$ means that samples from the same class are more concentrated around the mean μ_c, i.e., the representation is more *consistent* within that class.

1.2. Two Architectures: Independent vs. Cross-Sample Mapping

We compare two architectural variants that share the same backbone but differ in whether they mix information across samples before applying a nonlinearity.

Case A (independent per-sample mapping). Each sample u_i is processed independently by a nonlinear mapping $g : \mathbb{R}^d \to \mathbb{R}^d$:

$$r_i = g(u_i), \qquad i = 1, \dots, N. \tag{7}$$

Here g can be a composition of linear layers, activation functions, and normalization layers, as long as it does not couple different samples.

Case B (cross-sample propagation). In the second variant, we first mix features across the sample dimension and then apply the same nonlinearity g:

$$\tilde{h}_i = \sum_{k=1}^{N} w_{ik}(U)\, u_k, \qquad h_i = g(\tilde{h}_i), \tag{8}$$

where $U := (u_1, \dots, u_N)$ and $W(U) := [w_{ik}(U)] \in \mathbb{R}^{N\times N}$ is an implicit *mixing matrix* induced by the SSM along the sample dimension. While the SSM architecture does not explicitly construct a mixing matrix over the sample dimension, its gated recurrent mechanism implicitly aggregates information across samples through a sequence of convex updates. Each state update is a normalized interpolation of the previous state and the current input, and unrolling these updates yields an effective convex mixing of past inputs. The matrix $W(U)$ used in our analysis is therefore an idealized abstraction of this implicit propagation mechanism. It captures the dominant statistical effect—namely, that the selective SSM performs mean-preserving convex averaging across samples—which is precisely the property required for the variance-shrinkage results established in this section.

Row-stochastic mixing abstraction. In many selective SSMs, the mixing weights arise from normalized gates (e.g., softmax-normalized or convex combinations of the current state and previous state). Motivated by this, we idealize the mixing operator as row-stochastic:

Assumption 2 (Row-stochastic mixing)**.** *For any fixed input batch U, the mixing matrix $W(U)$ satisfies*

$$w_{ik}(U) \geq 0, \qquad \sum_{k=1}^{N} w_{ik}(U) = 1, \qquad \text{for all } i = 1, \dots, N. \tag{9}$$

In Case A, Eq. (7) is recovered from Eq. (8) by taking $w_{ik}(U) = \delta_{ik}$ (Kronecker delta), i.e., $W(U)$ reduces to the identity and no cross-sample mixing occurs.

Conditional analysis of $W(U)$. In practice, $W(U)$ depends on U in a complex, nonlinear way. To obtain a tractable analysis, we consider the covariance *conditional on a fixed realization of W*. Intuitively, we "freeze" the mixing coefficients and study how they act on the random perturbations $\{\xi_i\}$.

Formally, all expectations and covariances below should be interpreted as conditional on W unless explicitly stated otherwise. Unconditional statements then follow by averaging over the randomness of W.

Nonlinearity assumption. We assume that the post-mixing mapping g is Lipschitz continuous.

Assumption 3 (Lipschitz nonlinearity)**.** *There exists a constant $L_g > 0$ such that*

$$\|g(a) - g(b)\|_2 \leq L_g \|a - b\|_2, \qquad \forall a, b \in \mathbb{R}^d. \tag{10}$$

Many practical neural network blocks (e.g., MLPs with smooth activations and bounded weight norms) admit such a global Lipschitz constant; we do not require its exact value, only its existence.

1.3. Linear Cross-Sample Mixing: Mean and Covariance

We now analyze the *linear* mixing stage in Eq. (8), fixing a particular realization of W and focusing on the random variation induced by the perturbations $\{\xi_i\}$.

Using Eq. (2), we can write

$$\tilde{h}_i = \sum_{k=1}^{N} w_{ik} u_k = \sum_{k=1}^{N} w_{ik}(\mu_c + \xi_k) \tag{11}$$

$$= \mu_c \sum_{k=1}^{N} w_{ik} + \sum_{k=1}^{N} w_{ik} \xi_k. \tag{12}$$

By the row-stochastic property Eq. (9), we have $\sum_{k=1}^{N} w_{ik} = 1$, and hence

$$\mathbb{E}[\tilde{h}_i \mid W] = \mu_c + \sum_{k=1}^{N} w_{ik} \mathbb{E}[\xi_k] = \mu_c. \tag{13}$$

Interpretation. Conditioned on the mixing weights W, cross-sample mixing does not move the class mean: it only averages the deviations $\{\xi_k\}$ across samples.

Next we compute the conditional covariance of $\tilde{h}_i$. Using Assumption 1 and the independence of $\{\xi_k\}$, we obtain

$$\mathrm{Cov}(\tilde{h}_i \mid W) = \mathrm{Cov}\Big(\sum_{k=1}^{N} w_{ik} \xi_k \,\Big|\, W \Big) \tag{14}$$

$$= \sum_{k=1}^{N} w_{ik}^2 \, \mathrm{Cov}(\xi_k) \tag{15}$$

$$= \Big(\sum_{k=1}^{N} w_{ik}^2 \Big) \Sigma_u. \tag{16}$$

Taking traces on both sides yields

$$\mathcal{V}(\tilde{h}_i \mid W) := \mathrm{tr}\big(\mathrm{Cov}(\tilde{h}_i \mid W)\big) = \Big(\sum_{k=1}^{N} w_{ik}^2 \Big) \mathrm{tr}(\Sigma_u). \tag{17}$$

Thus, the variance of the mixed feature $\tilde{h}_i$ depends only on the single scalar term $\sum_{k=1}^{N} w_{ik}^2$; all other components of the covariance structure remain preserved.

Lemma 1 (ℓ_2-norm of row-stochastic weights)**.** *Let $(w_{i1}, \dots, w_{iN})$ be a row-stochastic vector, i.e., $w_{ik} \geq 0$ and $\sum_{k=1}^{N} w_{ik} = 1$. Then*

$$\sum_{k=1}^{N} w_{ik}^2 \leq 1, \tag{18}$$

with equality if and only if $(w_{i1}, \dots, w_{iN})$ is a one-hot vector.

Proof. Under the constraints $w_{ik} \geq 0$ and $\sum_k w_{ik} = 1$, the quantity $\sum_k w_{ik}^2$ is maximized when all the mass is concentrated on a single coordinate, i.e., when one of the w_{ik} equals 1 and all others are 0. In that case, $\sum_k w_{ik}^2 = 1$. For any non-degenerate mixture with at least two strictly positive entries, the sum of squares is strictly smaller than 1. □

Combining Eq. (17) and Lemma 1, we obtain:

Corollary 1 (Conditional variance shrinkage of linear mixing)**.** *Under Assumption 1 and conditional on W satisfying Assumption 2,*

$$\mathcal{V}(\tilde{h}_i \mid W) = \Big(\sum_{k=1}^{N} w_{ik}^2 \Big) \mathrm{tr}(\Sigma_u) \leq \mathrm{tr}(\Sigma_u), \tag{19}$$

with strict inequality whenever the row $(w_{i1}, \dots, w_{iN})$ is non-degenerate (i.e., not one-hot).

Interpretation. The linear cross-sample mixing step preserves the class mean (see Eq. (13)) and strictly reduces the intra-class variance whenever each output sample aggregates information from at least two different inputs.

1.4. Effect of a Lipschitz Nonlinearity

We now incorporate the nonlinear mapping g and compare the two branches:

- Independent branch: $r_i = g(u_i)$;
- Cross-sample branch: $h_i = g(\tilde{h}_i)$.

The following lemma is standard and formalizes the idea that a Lipschitz mapping cannot increase variance arbitrarily.

Lemma 2 (Lipschitz mapping and variance)**.** *Let $X \in \mathbb{R}^d$ be a random vector with finite covariance, and let $g : \mathbb{R}^d \to \mathbb{R}^d$ be L_g-Lipschitz as in Assumption 3. Then*

$$\mathcal{V}(g(X)) = \operatorname{tr}\big(\operatorname{Cov}(g(X))\big) \le L_g^2 \operatorname{tr}\big(\operatorname{Cov}(X)\big) = L_g^2 \mathcal{V}(X). \tag{20}$$

Proof. Let X' be an independent copy of X. It is a standard identity that

$$\operatorname{tr}\big(\operatorname{Cov}(X)\big) = \frac{1}{2}\mathbb{E}\big[\|X - X'\|_2^2\big], \tag{21}$$

and similarly

$$\operatorname{tr}\big(\operatorname{Cov}(g(X))\big) = \frac{1}{2}\mathbb{E}\big[\|g(X) - g(X')\|_2^2\big]. \tag{22}$$

By the Lipschitz property of g,

$$\|g(X) - g(X')\|_2 \le L_g \|X - X'\|_2. \tag{23}$$

Squaring both sides and taking expectations gives

$$\mathbb{E}\big[\|g(X) - g(X')\|_2^2\big] \le L_g^2 \mathbb{E}\big[\|X - X'\|_2^2\big], \tag{24}$$

which implies the desired inequality for the traces of the covariances. □

Applying Lemma 2 to our two branches, we obtain

$$\mathcal{V}(h_i \mid W) \le L_g^2 \mathcal{V}(\tilde{h}_i \mid W), \tag{25}$$

and

$$\mathcal{V}(r_i) \le L_g^2 \mathcal{V}(u_i) = L_g^2 \operatorname{tr}(\Sigma_u). \tag{26}$$

Substituting the expression from Corollary 1 into Eq. (25), we get

$$\mathcal{V}(h_i \mid W) \le L_g^2 \mathcal{V}(\tilde{h}_i \mid W) \tag{27}$$

$$= L_g^2 \Big(\sum_{k=1}^{N} w_{ik}^2\Big) \operatorname{tr}(\Sigma_u). \tag{28}$$

Proposition 1 (Cross-sample propagation tightens the variance upper bound)**.** *Under Assumptions 1, 2 and 3, and conditioning on a fixed mixing matrix W, we have:*

$$\mathcal{V}(h_i \mid W) \le L_g^2 \Big(\sum_{k=1}^{N} w_{ik}^2\Big) \operatorname{tr}(\Sigma_u), \tag{29}$$

with strict inequality

$$\mathcal{V}(h_i \mid W) < L_g^2 \operatorname{tr}(\Sigma_u) \tag{30}$$

whenever the row $(w_{i1}, \ldots, w_{iN})$ is non-degenerate. By contrast, the independent branch only satisfies the looser bound

$$\mathcal{V}(r_i) \le L_g^2 \operatorname{tr}(\Sigma_u). \tag{31}$$

Thus, for any fixed realization of W, the cross-sample branch enjoys a strictly tighter worst-case upper bound *on intra-class variance than the independent per-sample branch.*

Although the derivation assumes a fixed realization of W, the variance-shrinking effect naturally generalizes to arbitrary convex or normalized mixing weights, and holds in expectation for stochastic or data-dependent W. Thus, the conclusion applies broadly to linear cross-sample propagation mechanisms.

Summary. The linear mixing step already shrinks variance while preserving the mean, and the subsequent Lipschitz nonlinearity cannot undo this shrinkage at the level of variance upper bounds. Therefore, cross-sample propagation provably leads to *more compact* class-conditional representations in this idealized setting.

1.5. Discussion and Connection to SSMs

The above analysis should be viewed as an idealized but informative model of cross-sample state propagation in SSM-like architectures.

- **Mean preservation.** By Eq. (13), the class-wise semantic mean μ_c is unchanged by cross-sample mixing: the operator only acts on the fluctuations $\{\xi_i\}$.
- **Variance shrinkage.** Corollary 1 shows that the linear mixing stage is a mean-preserving, variance-shrinking operator whenever each output aggregates multiple samples. Proposition 1 further shows that this shrinkage is preserved (in terms of upper bounds) after applying a Lipschitz nonlinearity.
- **Representation consistency.** A tighter upper bound on intra-class variance means that, in the worst case, same-class representations cannot spread out as much as in the independent branch. This provides a clear mathematical explanation for the improved representation consistency observed in our experiments with cross-sample SSM blocks.

In practical SSMs, the exact mixing operator may not be explicitly row-stochastic in every implementation. However, many variants employ normalized gates and convex-like propagation that approximate the row-stochastic behavior assumed here. Our analysis therefore captures the dominant statistical effect: *cross-sample state propagation acts as a mean-preserving variance-shrinking mechanism*

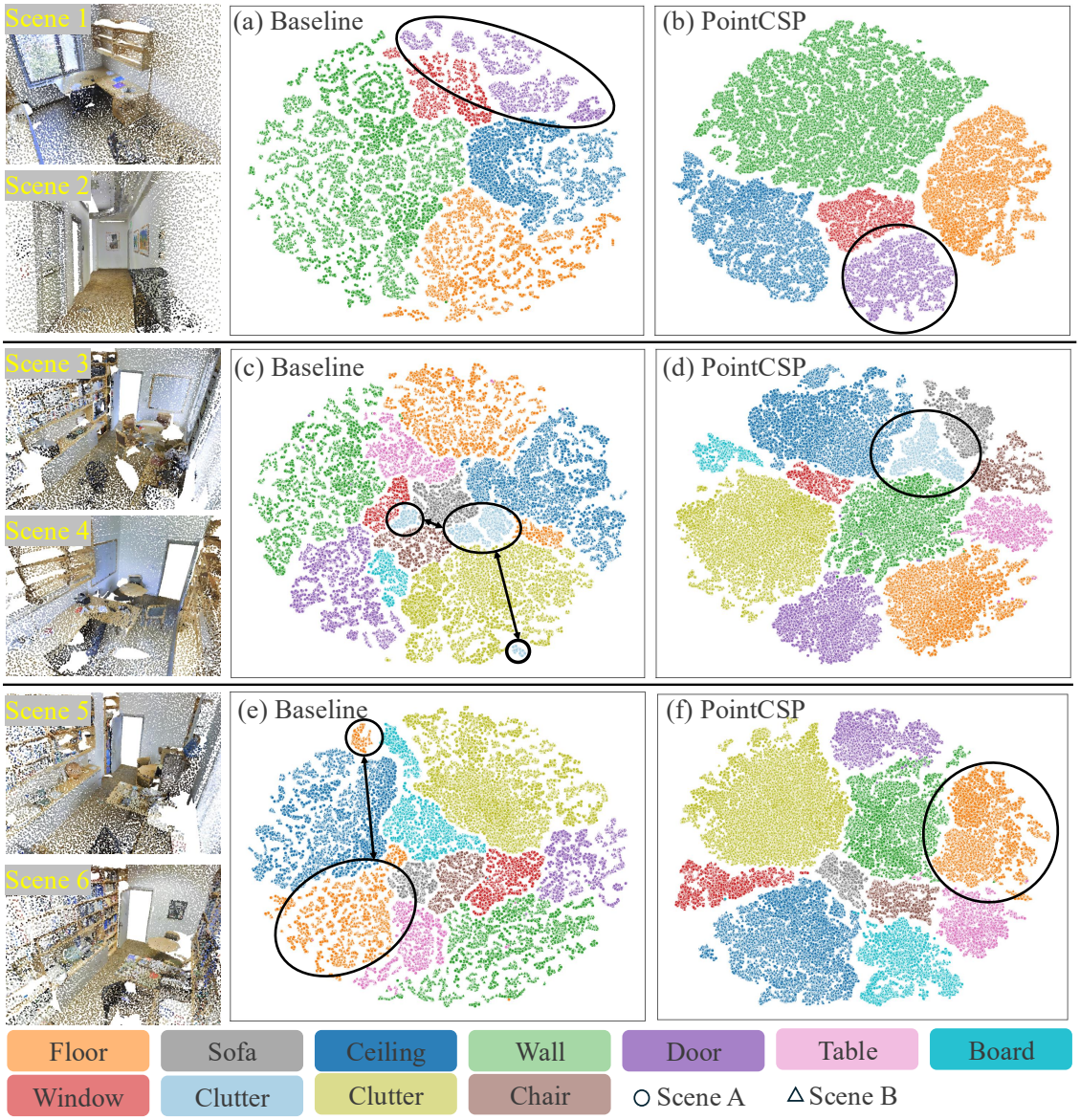

Figure 1. **t-SNE comparison of embeddings with and without cross-sample state propagation (reproduced for completeness).** This figure, previously shown in the main paper. The independent per-sample baseline produces dispersed and fragmented clusters, whereas our selective SSM with cross-sample propagation yields compact, well-aligned clusters with substantially reduced intra-class variance. These visualizations complement the theoretical analysis in theoretical analysis sections by demonstrating the empirically observed variance-shrinking behavior.

that stabilizes class-conditional embeddings. This theoretical perspective is consistent with the empirical performance gains observed when cross-sample propagation is enabled.

2. t-SNE Visualization of Representation Consistency

To further validate the theoretical analysis, we compare the geometric structure of the embedding spaces learned with and without cross-sample state propagation using t-SNE projections. Specifically, we visualize the distributions of sample embeddings under:

- **Without cross-sample propagation:** the independent per-sample mapping $r_i = g(u_i)$;
- **With cross-sample propagation:** the selective SSM mapping $h_i = g(\sum_k w_{ik}(U)\, u_k)$.

Fig. 1 illustrates the contrast between these two architectures on six scene instances (Sense 1–6), each pair corresponding to semantically similar environments.

Method	Ref.	Area-5			6-fold		
		mIoU(%)↑	mAcc(%)↑	OA(%)↑	mIoU(%)↑	mAcc(%)↑	OA(%)↑
PointNeXt	NeurIPS'22	71.1	77.2	91.0	74.9	83.0	90.3
PointMetaBase	CVPR'23	72.3	78.0	91.3	77.0	/	91.3
PointVector	CVPR'23	72.3	78.1	91.0	/	/	/
LinNet	NeurIPS'24	72.9	78.6	91.3	/	/	/
PTV3	CVPR'24	73.4	78.9	91.7	77.7	85.3	91.5
HPENet	AAAI'24	72.7	78.5	91.5	78.7	86.2	91.9
PDNet-XXL	AAAI'24	72.3	78.1	91.3	78.3	86.2	91.9
PCM	AAAI'25	74.1	81.6	92.9	/	/	/
CamPoint	CVPR'25	83.3	86.9	96.0	/	/	/
PointCSP	/	**88.2**	**90.6**	**98.1**	**93.1**	**96.3**	**98.7**

Table 1. Semantic segmentation results on the S3DIS dataset are evaluated on Area5 and 6-fold cross-validation.

(a) Sense 1 vs. (b) Sense 2 — Corridor and indoor-office scenes with similar semantic structure. Although the two scenes differ visually, a corridor versus an indoor office, their underlying spatial semantics are closely related. Under the independent per-sample mapping, embeddings from these two environments remain widely scattered, forming elongated and irregular clusters. With cross-sample propagation, however, the embeddings become significantly more compact and better aligned across the two scenes, indicating that the selective SSM suppresses scene-specific noise while preserving shared semantic structure.

(c) Sense 3 vs. (d) Sense 4 — Indoor office scenes. For office environments, the baseline again yields scattered clusters with samples drifting away from the semantic center. In contrast, the selective SSM compresses the embeddings into smoother, more coherent manifolds, demonstrating stronger stability under variations in layout and viewpoint.

(e) Sense 5 vs. (f) Sense 6 — Another pair of indoor scenes. A consistent pattern emerges in the final pair: the baseline embeddings occupy a broad, noisy region, whereas the cross-sample SSM forms tightly concentrated clusters. This shows that the benefits of cross-sample propagation generalize across different indoor geometries.

Summary. Across all three pairs of semantically equivalent scenes, the PointCSP consistently yields tighter, more compact, and more coherent clusters than the independent baseline. These qualitative observations closely align with our theoretical results: cross-sample propagation acts as a mean-preserving *variance-shrinking* operator, and the Lipschitz mapping g preserves this contraction. Importantly, the t-SNE results make the improvement in *semantic consistency* visually evident: samples drawn from different yet semantically related scenes collapse toward similar semantic centers, demonstrating that cross-sample propagation not only reduces variance but also promotes semantically aligned and scene-invariant representations.

3. More semantic segmentation results on S3DIS dataset

We also conducted additional experiments on the S3DIS dataset, comparing PointCSP with other state-of-the-art methods. The results are presented in Tab. 1. PointCSP achieves state-of-the-art performance with 88.2% mIoU, 90.6% mAcc, and 98.1% OA on Area-5, outperforming recent SOTA methods. The results demonstrate the effectiveness of PointCSP in semantic segmentation tasks, particularly in complex indoor environments.

In addition, we provide visualizations of the S3DIS Area 5 dataset in Fig. 2, Fig. 3 and Fig. 4, where the point clouds are colored according to their semantic labels. Figure. 3 is also presented in the main text and is enlarged here to allow more detailed inspection. The visualized results present a series of indoor scenes, including corridors, restrooms, meeting rooms, and office rooms. Semantic labels are represented by color codes within the point clouds, thus revealing the model's capacity for accurate segmentation across diverse semantic categories. The findings of this study provide unequivocal evidence of the efficacy of the model in navigating complex indoor environments.

Specifically, the results of the corridor segmentation in the second and fourth rows of Fig. 2, as well as the fifth row of Fig. 3, clearly show that PointCSP achieves a more accurate and clearer distinction of wall, bookcase and beam compared to CamPoint, with stronger semantic consistency. The third rows of Fig. 2 and the last rows of Fig. 3, which include windows, further demonstrate the ability of the model to accurately and robustly segment window regions. In other diverse scenes presented in Fig. 2, Fig. 3 and Fig. 4, PointCSP delivers precise segmentation in complex local regions. The visualization highlights the ability of the model to accurately segment different semantic categories in complex scenes.

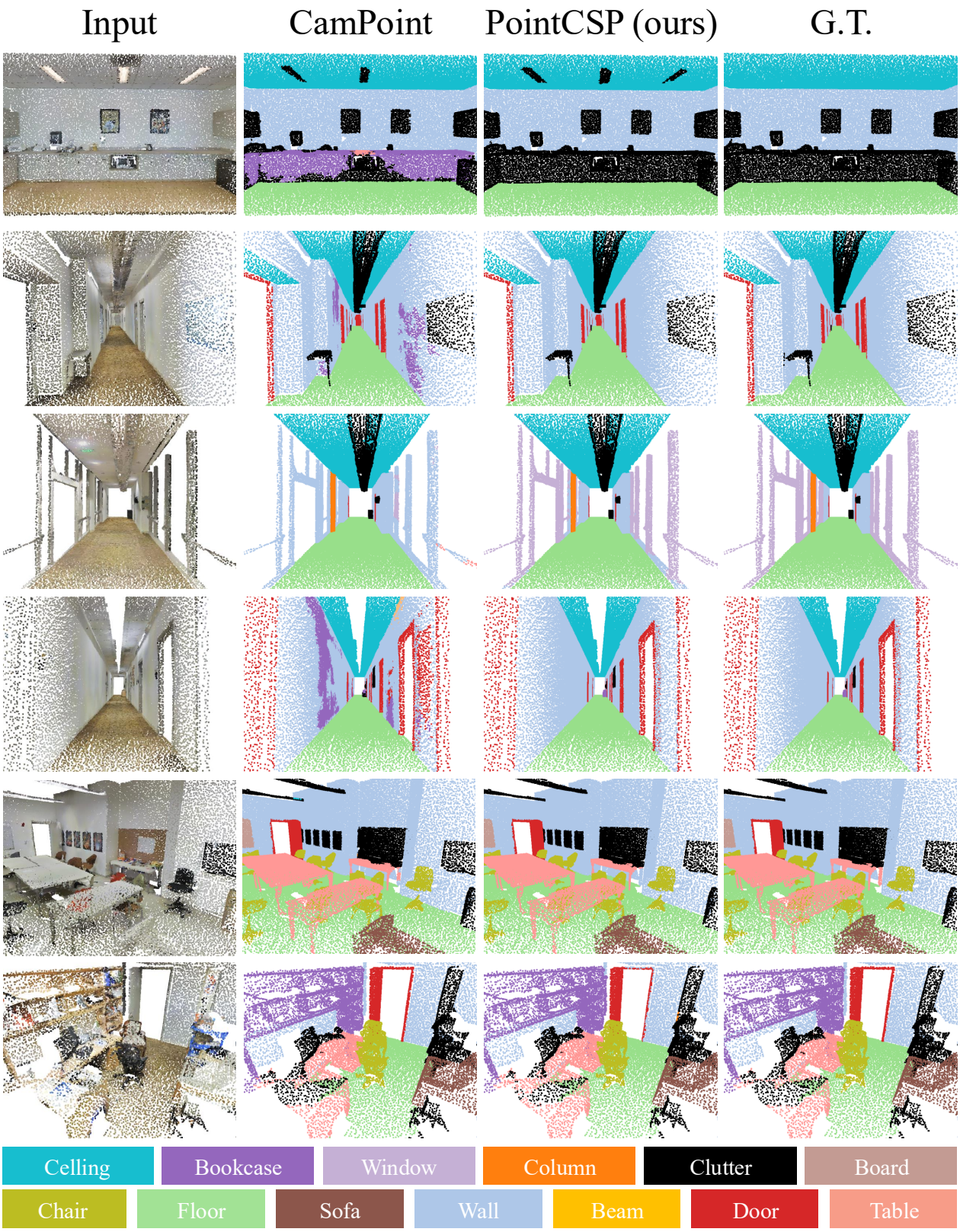

Figure 2. The visualization of the S3DIS dataset. The point clouds are colored by their semantic labels.

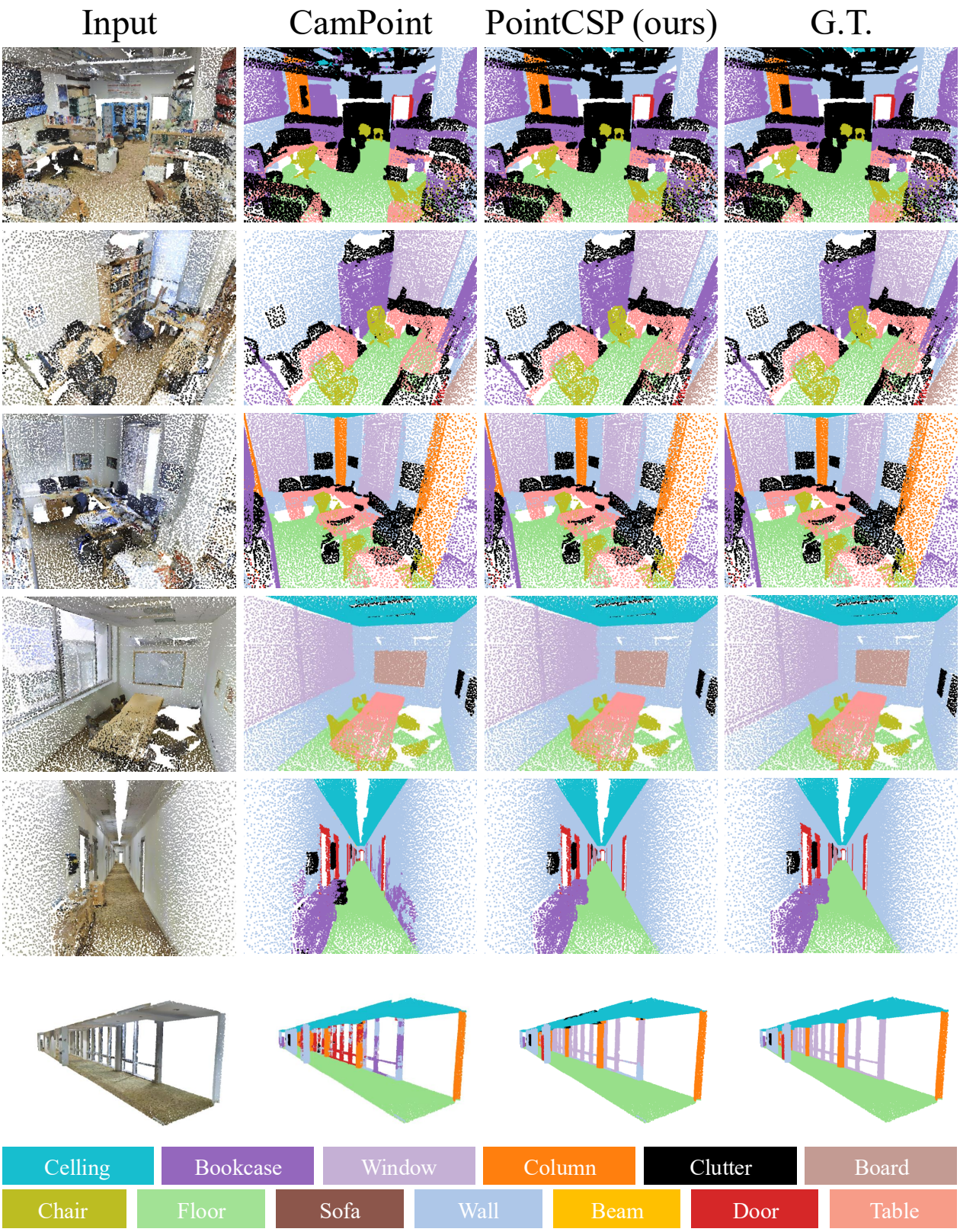

Figure 3. Visualization of the S3DIS dataset, where point clouds are color-coded based on semantic labels.

Figure 4. Visualization of the S3DIS dataset, where point clouds are color-coded based on semantic labels.