C-GenReg: Training-Free 3D Point Cloud Registration by Multi-View-Consistent Geometry-to-Image Generation with Probabilistic Modalities Fusion

Supplementary Material

Yuval Haitman Amit Efraim Joseph M. Francos
Ben-Gurion University, Beer-Sheva, Israel

1. Probabilistic Fusion Derivation

1.1. Noisy-AND Derivation (Eq. 6)

Proposition 1.1 (Joint Posterior Fusion *(Noisy-AND)*) *Under the conditional independence assumption of the image and geometric evidences given the match event, the Joint Posterior Fusion probability satisfies:*

$$p_{ij}^{fuse} = \frac{p_{ij}^{img}\, p_{ij}^{geo}\,(1-\pi_{ij})}{p_{ij}^{img}\, p_{ij}^{geo}\,(1-\pi_{ij}) + \left(1-p_{ij}^{img}\right)\left(1-p_{ij}^{geo}\right)\pi_{ij}}. \tag{1}$$

Proof (Proposition 1.1) Define the odds $o(x) = \frac{x}{1-x}$.

By Bayes rule in odds form and conditional independence of $S_{ij}^{img} \perp\!\!\!\perp S_{ij}^{geo} \mid M_{ij}$,

$$\begin{aligned} O_{ij}^{\text{fuse}} &= \frac{\Pr(M_{ij}{=}1 \mid S_{ij}^{img}, S_{ij}^{geo})}{\Pr(M_{ij}{=}0 \mid S_{ij}^{img}, S_{ij}^{geo})} \\ &= O_{ij}^{\pi} \cdot \mathrm{LR}_{ij}^{\text{img}} \cdot \mathrm{LR}_{ij}^{\text{geo}}, \end{aligned} \tag{2}$$

where $O_{ij}^{\pi} = o(\pi_{ij})$ and $\mathrm{LR}_{ij}^{m} = \frac{p(S_{ij}^{m}|M_{ij}=1)}{p(S_{ij}^{m}|M_{ij}=0)}$. Applying Bayes rule to each modality gives:

$$O_{ij}^{\text{img}} = o(p_{ij}^{\text{img}}) = O_{ij}^{\pi}\,\mathrm{LR}_{ij}^{\text{img}}, \tag{3}$$

$$O_{ij}^{\text{geo}} = o(p_{ij}^{\text{geo}}) = O_{ij}^{\pi}\,\mathrm{LR}_{ij}^{\text{geo}}. \tag{4}$$

Hence:

$$\mathrm{LR}_{ij}^{\text{img}} = \frac{o(p_{ij}^{\text{img}})}{o(\pi_{ij})}, \tag{5}$$

$$\mathrm{LR}_{ij}^{\text{geo}} = \frac{o(p_{ij}^{\text{geo}})}{o(\pi_{ij})}. \tag{6}$$

Substitute Eq. (5) and Eq. (6) into Eq. (2) to obtain

$$O_{ij}^{\text{fuse}} = O_{ij}^{\pi} \cdot \frac{o(p_{ij}^{\text{img}})}{o(\pi_{ij})} \cdot \frac{o(p_{ij}^{\text{geo}})}{o(\pi_{ij})} = \frac{o(p_{ij}^{\text{img}})\, o(p_{ij}^{\text{geo}})}{o(\pi_{ij})}. \tag{7}$$

Finally, writing odds in terms of probabilities and simplifying gives the closed form

$$p_{ij}^{\text{fuse}} = \frac{p_{ij}^{\text{img}}\, p_{ij}^{\text{geo}}\,(1-\pi_{ij})}{p_{ij}^{\text{img}}\, p_{ij}^{\text{geo}}\,(1-\pi_{ij}) + \left(1-p_{ij}^{\text{img}}\right)\left(1-p_{ij}^{\text{geo}}\right)\pi_{ij}}.$$

□

1.2. Noisy-OR Derivation (Eq. 7)

Proposition 1.2 (Disjunctive Posterior Fusion *(Noisy-OR)*) *Let A_{ij}^{m} be a modality activation random variable indicating whether modality m supports the correspondence (i,j). We model A_{ij}^{m} as a Bernoulli random variable depending only on its own similarity signal,*

$$A_{ij}^{m} \mid S_{ij}^{m} \sim \text{Bernoulli}\big(p_{ij}^{m}\big), \quad \Pr(A_{ij}^{m}=1 \mid S_{ij}^{m}) = p_{ij}^{m}. \tag{8}$$

*Then, under the conditional independence of activations given their respective modality signals, the Disjunctive Posterior Fusion (*Noisy-OR*) is given by:*

$$p_{ij}^{Noisy\text{-}OR} = 1 - (1-p_{ij}^{img})(1-p_{ij}^{geo}). \tag{9}$$

Proof (Proposition 1.2) By construction, each activation depends only on its own signal and is therefore independent of the other modality's evidence once that signal is known,

$$A_{ij}^{\text{img}} \perp\!\!\!\perp S_{ij}^{\text{geo}} \mid S_{ij}^{\text{img}}, \qquad A_{ij}^{\text{geo}} \perp\!\!\!\perp S_{ij}^{\text{img}} \mid S_{ij}^{\text{geo}}. \tag{10}$$

As a result, the two activations are independent given both similarity signals,

$$A_{ij}^{\text{img}} \perp\!\!\!\perp A_{ij}^{\text{geo}} \mid (S_{ij}^{\text{img}}, S_{ij}^{\text{geo}}). \tag{11}$$

The Disjunctive Posterior Fusion is defined as the probability that at least one modality supports the correspondence:

$$p_{ij}^{\text{Noisy-OR}} = \Pr(A_{ij}^{\text{img}} = 1 \vee A_{ij}^{\text{geo}} = 1 \mid S_{ij}^{\text{img}}, S_{ij}^{\text{geo}}). \tag{12}$$

Using the complementary event and applying Eqs. (10) and (11) gives:

$$\begin{aligned} p_{ij}^{\text{Noisy-OR}} &= 1 - \Pr(A_{ij}^{\text{img}} = 0, A_{ij}^{\text{geo}} = 0 \mid S_{ij}^{\text{img}}, S_{ij}^{\text{geo}}) \\ &= 1 - \Pr(A_{ij}^{\text{img}} = 0 \mid S_{ij}^{\text{img}})\,\Pr(A_{ij}^{\text{geo}} = 0 \mid S_{ij}^{\text{geo}}). \end{aligned} \tag{13}$$

Substituting the Bernoulli definition from (8), $\Pr(A_{ij}^{m}=0 \mid S_{ij}^{m}) = 1 - p_{ij}^{m}$, into (13) yields (9). □

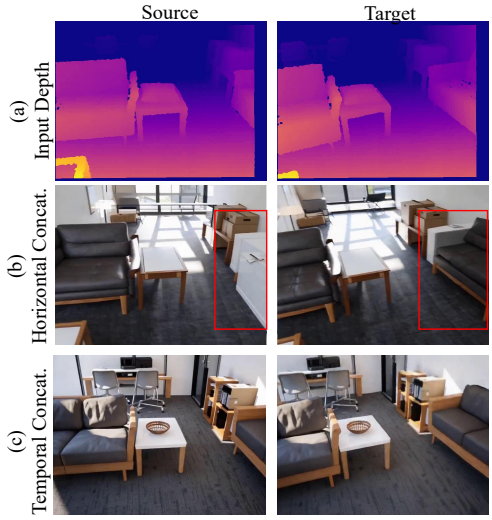


Figure 1. **WFM Input Formatting.** (a) Input depth maps of the source and target views. (b) Feeding the pretrained WFM with *horizontally concatenated* depth inputs causes cross-view inconsistencies, e.g., the sofa is mistakenly replaced in the generated source image. (c) Using *temporal concatenation* produces RGB outputs that are geometrically coherent and appearance-consistent between the two views.

2. Methodological Details

2.1. Consistent Multi-View Generation: Horizontal vs. Temporal concatenation

Our method relies on Cosmos-Transfer, a World Foundation Model (WFM) capable of generating multi-view consistent RGB videos from depth inputs. Since Cosmos-Transfer is trained to operate on depth videos, it expects a temporally ordered sequence of depth maps as input.

To adapt this interface to the point cloud registration setting, we construct the WFM input by concatenating the source and target depth sequences along the temporal axis:

$$D^{\text{in}} = \left[D_1^{\text{src}}, \ldots, D_L^{\text{src}},\ D_1^{\text{tgt}}, \ldots, D_L^{\text{tgt}} \right]. \tag{14}$$

This temporal concatenation forms a single depth video containing the two fragments sequentially. While this configuration does not perfectly match the model's original training distribution, it remains physically plausible and statistically closer to natural camera motion than alternative layouts such as spatial horizontal concatenation.

The key advantage of temporal concatenation is that it preserves the pretrained multi-view consistency priors of the WFM. Because the model was trained on temporally coherent video sequences, it naturally propagates geometric and semantic information across adjacent frames, including across the transition between the source and target segments.

In contrast, horizontally concatenating the two fragments would introduce an artificial spatial discontinuity that is not present in the model's training distribution, often resulting in weaker geometric coherence between the generated views. A qualitative comparison between the two strategies is illustrated in Fig. 1.

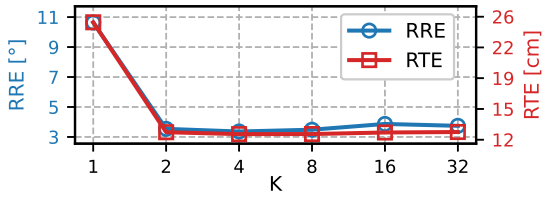


Figure 2. **Effect of View Selection (K).** Registration performance measured by Relative Rotation Error (RRE) and Relative Translation Error (RTE) as a function of the number of selected views K. Performance saturates for $K \geq 4$, indicating that only a few representative views are sufficient for stable registration.

Due to the autoregressive nature of video generation, minor visual transients may appear near the transition between the source and target segments. To mitigate this effect, we discard several "safeguard" frames around the midpoint of the generated video before extracting features for correspondence estimation.

2.2. View Selection Strategy

MASt3R extracts features using a cross-attention decoder that jointly processes pairs of images, meaning that the representation of a given image depends on the image with which it is paired. Consequently, evaluating a source image against different target views produces distinct conditioned feature maps. To exploit this property, we sample K views uniformly from the L frames of the generated source and target sequences and evaluate all pairwise combinations. This produces K^2 conditioned feature maps per domain, denoted as $F_{\text{src}}^{\text{img}}, F_{\text{tgt}}^{\text{img}} \in \mathbb{R}^{K^2 \times H \times W \times d_{\text{img}}}$.

While increasing K improves viewpoint coverage, it also leads to quadratic growth in the number of evaluated image pairs, motivating a careful view-selection strategy. As shown in Fig. 2, both RRE and RTE quickly saturate as K increases, reflecting the high correlation between frames in the generated sequences. This suggests that selecting $K \ll L$ is sufficient to maintain view diversity while keeping the computational cost manageable.

2.3. C-GenReg for LiDAR Data

Since Cosmos-Transfer expects depth images as input, we introduce a preprocessing stage that converts raw 3D LiDAR scans into a depth-image representation. Following [3], we project each LiDAR point cloud onto a *virtual camera*. Choosing an appropriate virtual camera model is essential, as LiDAR sensors cover an extremely wide field of view (FOV).

In line with Cosmos-Transfer, we use an f-θ virtual camera instead of a standard pinhole model. Wide-angle

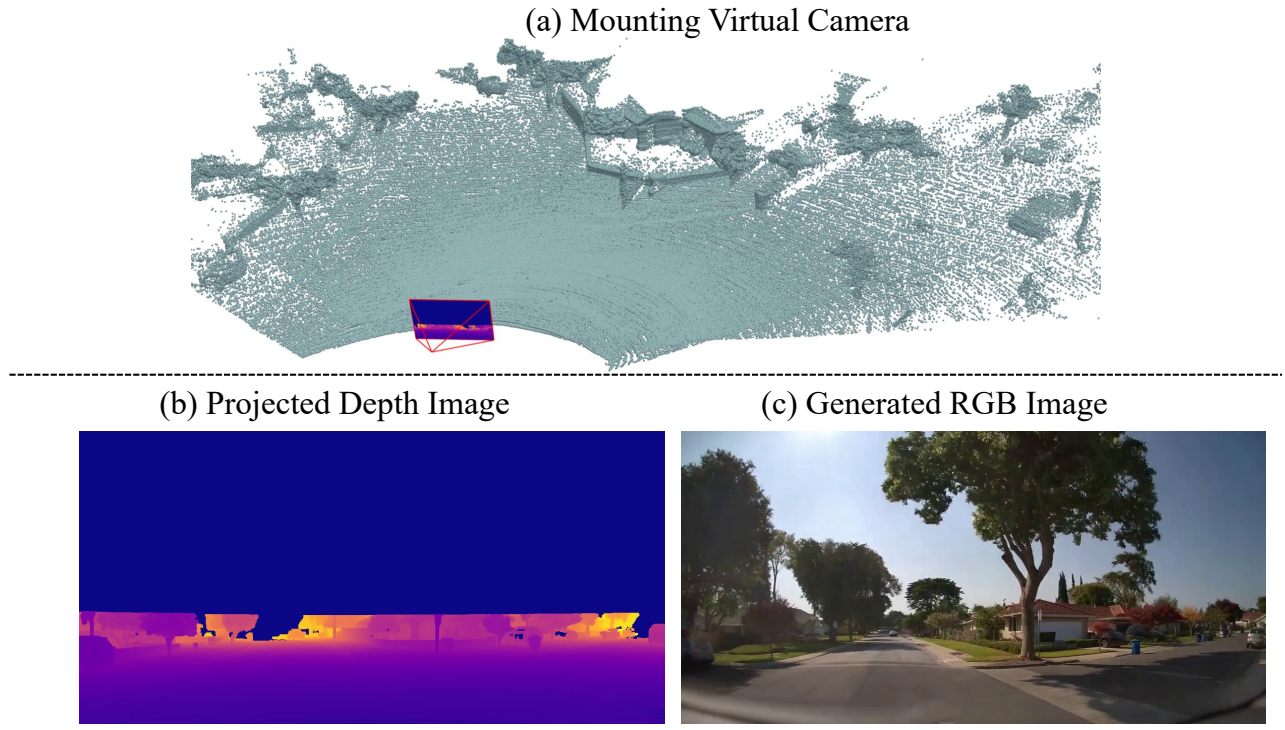


Figure 3. **C-GenReg LiDAR Input Pipeline:** (a) A virtual camera is configured into the LiDAR scan. (b) The LiDAR points are projected into a depth image. (c) The resulting depth map is fed into the generative model to produce an aligned RGB image.

perception systems, typical in autonomous driving, require ray–angle–based projection models that accurately handle $180°+$ FOV and avoid the extreme nonlinear distortions produced by perspective projection at large incidence angles [1, 4]. The f-θ formulation also aligns well with the optics of wide-FOV imaging systems commonly used in robotics and AV platforms.

Figure 3 shows the full conversion pipeline: attaching the virtual camera, rendering a depth map from the LiDAR scan, and passing the resulting depth image to the World Foundation Model to generate the corresponding RGB frame. For $360°$ LiDAR, our approach can be naturally extended by deploying multiple virtual cameras with overlapping FOVs, leveraging the WFM's multi-view consistency to produce a coherent full-view RGB generation.

2.4. Fusion Method - Point Matching Performance

Accurate registration relies heavily on producing reliable point-to-point correspondences. To evaluate the effect of our probabilistic fusion operators on this stage, we compare Noisy-AND and Noisy-OR on the point-matching task. Fig. 4 shows the resulting precision–recall curves evluated on the 3DMatch dataset. A correspondence is counted as correct if, under the ground-truth transformation, the matched points lie within $5cm$ of each other. Across the entire recall range, Noisy-AND consistently attains higher precision than Noisy-OR. This behavior is expected: Noisy-AND emphasizes matches that are simultaneously confident in both modalities, whereas Noisy-OR tends to admit a larger set of correspondences, including lower-quality ones, which reduces precision.

3. Additional Implementation Details

For RGB generation, we employ *Cosmos-Transfer1-7B (Depth)* for indoor datasets and Cosmos-*Transfer1-7B-Sample-AV (LiDAR)* for outdoor datasets. Input depth maps are resized to 960×704 for indoor data and 1280×640 for outdoor data to match the expected Cosmos input resolutions. Cosmos is run with the following parameters: CFG=7, $\sigma_{\max} = 80$, a spatiotemporal control weight of 1, 35 denoising steps, and a target frame rate of 30 fps. The output resolution matches the input depth resolution, and all inference is performed on an NVIDIA RTX A6000 GPU.

For the VFM pathway, we use MASt3R with an Encoder ViT-L and Decoder ViT-B configuration. Input RGB images are resized to 512×384 prior to feature extraction, and we use only the descriptor head ($d_{\text{img}} = 24$).

For the geometric branch, indoor point clouds are voxelized at 2.5 cm and outdoor point clouds at 5 cm. We extract geometric features using GeoTransformer, employing the official 3DMatch checkpoint for indoor scenes and the

Method	WFM (s)	VFM (s)	Geom. (s)	Pose (s)	Total (s)
GeoTransformer	-	-	0.075	1.558	1.633
Ours	507.0	0.973	0.075	0.066	508.1

Table 1. **Runtime Analysis.** Runtime per registration problem measured on a single NVIDIA RTX A6000 GPU.

Method	Rotation [deg]					Translation [cm]				
	Accuracy ↑			Error ↓		Accuracy ↑			Error ↓	
	5	10	45	Mean	Med.	5	10	25	Mean	Med.
FCGF	70.2	87.7	96.2	9.5	3.3	27.5	58.3	82.9	23.6	8.3
GeoTransformer	94.0	96.8	98.1	4.3	1.0	79.2	92.0	96.7	8.2	2.5
C-GenReg (Ours)	**99.4**	**99.8**	**99.9**	**1.1**	**0.9**	**87.5**	**97.1**	**99.3**	**3.0**	**1.9**

Table 2. **ScanNet Original Benchmark.** Rotation and translation accuracy (% of pairs within RRE/RTE thresholds in deg and cm respectively) and mean/median error on the ScanNet original Split benchmark. Best results are in **bold**.

KITTI checkpoint for outdoor scenes, producing the geometric descriptors ($d_{\text{geo}} = 256$).

All components, Cosmos, MASt3R, and GeoTransformer, are used with their publicly released pretrained weights and remain completely frozen in our pipeline.

4. Runtime Analysis

We report the runtime breakdown of C-GenReg and compare it with GeoTransformer in Tab. 1. Since GPCR code is not publicly available, a direct runtime comparison with that method cannot be provided. The runtime of C-GenReg is dominated by the World Foundation Model (WFM), which generates multi-view RGB videos from the input depth sequences. As shown in Tab. 1, this stage accounts for almost the entire runtime (507s), while the remaining components are lightweight: the Vision Foundation Model (VFM) used for correspondence extraction requires less than one second, and the geometric matching and pose estimation stages take only a fraction of a second.

This cost reflects the use of a powerful pretrained generative prior in a fully training-free pipeline. Importantly, recent work on distilling Cosmos Transfer models [2] reports up to a 72× inference speedup, which would reduce the runtime of our pipeline to approximately $\sim$ 7s. Additional speedups may also be achieved by lowering the video generation rate.

5. Additional Results

5.1. ScanNet Original Benchmark

In Tab. 2, we report registration performance on the original ScanNet benchmark, which consists of relatively easy pairs sampled 20 frames apart, resulting in modest motion between the source and target scans. C-GenReg achieves clear improvements across all metrics, with particularly strong gains in translation accuracy compared to FCGF and GeoTransformer, thus demonstrating that augmenting geometric features with our RGB-generated branch provides a consistent performance boost and serves as an effective enhancement to existing registration pipelines.

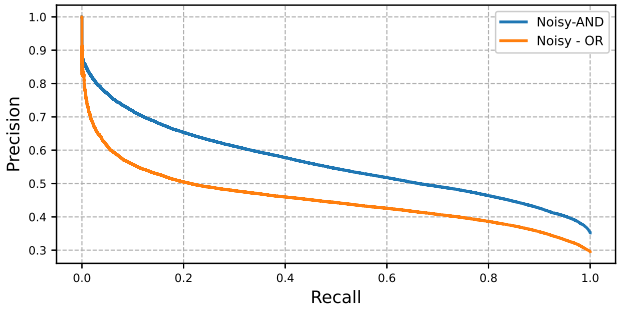


Figure 4. **Matching Performance Comparison of Noisy-AND vs. Noisy-OR.** Precision–recall curves comparing the two probabilistic fusion operators on the point-matching task (a match is correct if within $5cm$ under the ground-truth transformation). Noisy-AND consistently achieves higher precision at similar recall rates.

	Lo3DMatch		LoWaymo	
Method	**RRE**	**RTE**	**RRE**	**RTE**
GeoTransformer	21.10	53.46	19.72	9.04
Ours	**14.57**	**45.49**	**4.95**	**1.66**

Table 3. **Low-Overlap Results.** Mean RRE (degrees) and mean RTE (cm for Lo3DMatch and m for LoWaymo).

5.2. Low-Overlap Benchmarks

To further evaluate the robustness of C-GenReg in challenging scenarios, we conduct experiments on low-overlap registration benchmarks. Specifically, we evaluate on the Lo3DMatch benchmark and on a low-overlap split of the Waymo dataset, where the overlap between point clouds is limited to be less than 30% (Tab. 3). As expected, performance degrades compared to high-overlap cases due to the reduced geometric overlap between scans. Nevertheless, C-GenReg consistently outperforms the geometry-only baseline GeoTransformer across both datasets. On Lo3DMatch, C-GenReg reduces the rotation error from 21.10° to 14.57° and the translation error from 53.46 cm to 45.49 cm. A more substantial improvement is observed on the low-overlap Waymo benchmark, where C-GenReg reduces the rotation error from 19.72° to 4.95° and the translation error from 9.04 m to 1.66 m.

These results highlight the benefit of incorporating generative priors into the registration pipeline. Even when the geometric overlap between scans is limited, the WFM-based image generation remains consistent in the shared regions of the scene. Combined with the probabilistic match-then-fuse formulation, this enables C-GenReg to recover re-

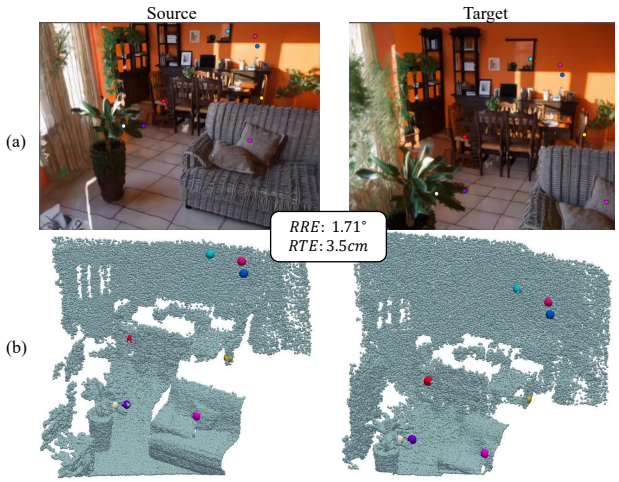


Figure 5. **Qualitative registration example from C-GenReg on the 3DMatch dataset.** Generated source and target images with a subset of matched keypoints (same color indicates correspondence), and the same correspondences visualized on the source and target 3D point clouds. The resulting rotation error (RRE, °) and translation error (RTE, cm) are reported as well.

liable correspondences despite the sparsity of overlapping geometry, leading to improved registration accuracy.

5.3. Qualitative Examples

We present additional qualitative results of C-GenReg across both indoor and outdoor benchmarks. Fig. 5 and Fig. 6 illustrate registration outcomes on the 3DMatch and Waymo datasets, respectively. Each example shows the generated RGB views with a subset of color-coded correspondence matches, together with the same matches visualized directly on the 3D point clouds.

Fig. 7, Fig. 8 and Fig. 9 highlight the generative capabilities of the employed World Foundation Model across the evaluated datasets, and especially the geometric coherence and multi-view consistency of the views generated by the C-GenReg pipeline, as these are the essential components towards the success of the C-GenReg registration. For each scene, we visualize the input depth maps and the generated RGB outputs, demonstrating strong multi-view consistency between source and target views as well as geometric coherence between the underlying depth structure and the synthesized images.

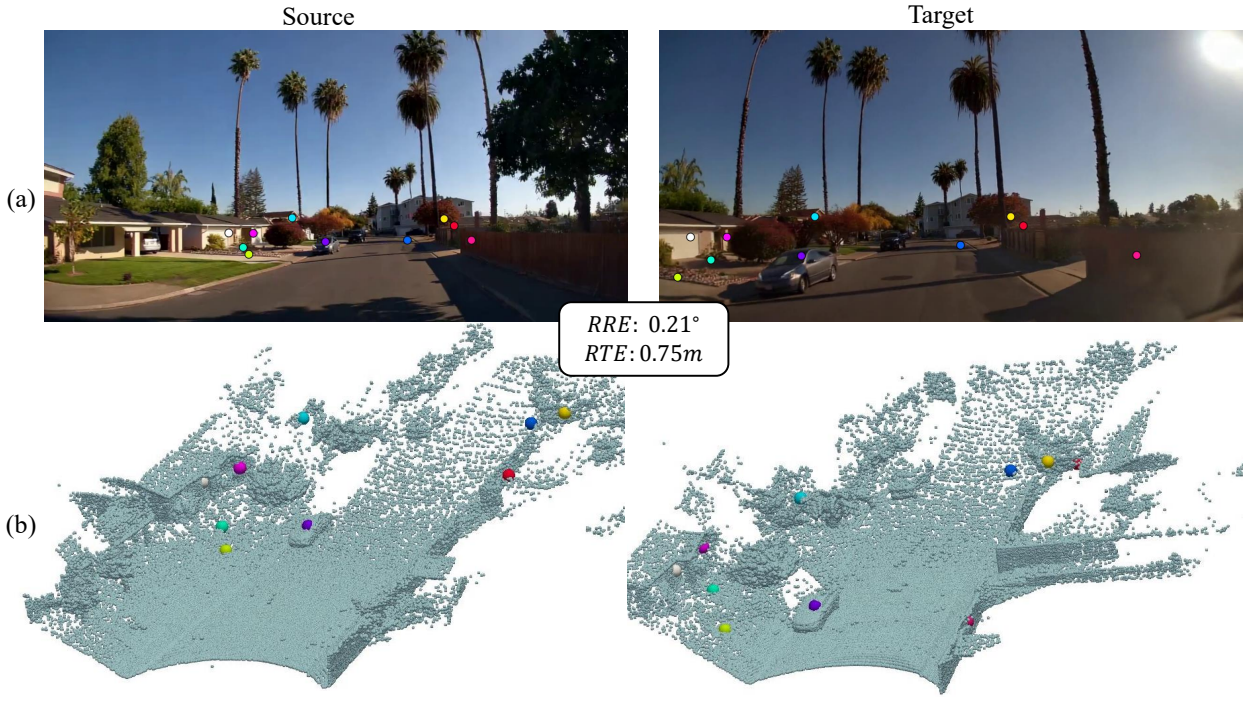
References

[1] Jonathan Courbon, Youcef Mezouar, Laurent Eckt, and Philippe Martinet. A generic fisheye camera model for robotic applications. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1683–1688. IEEE, 2007. 3

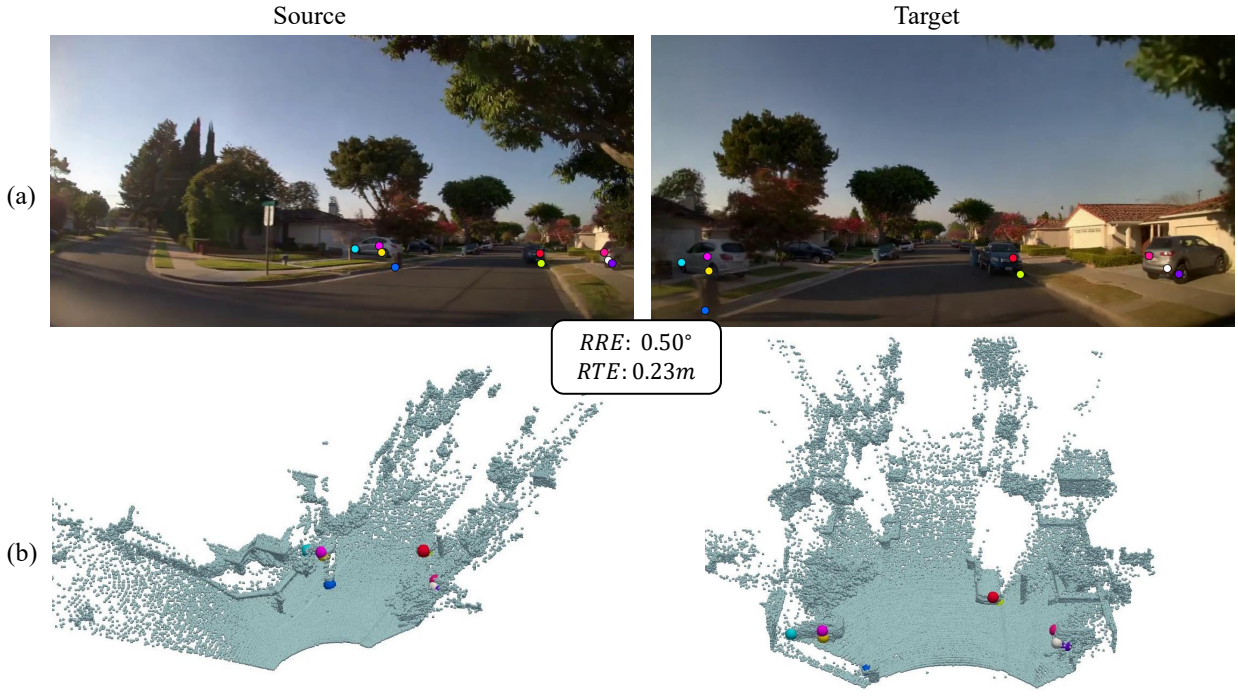
[2] Grace Lam. Distilling cosmos transfer 1 models. https://nvidia-cosmos.github.io/cosmos-cookbook/core_concepts/distillation/distilling_transfer1.html, 2025. NVIDIA Cosmos Cookbook. 4

[3] Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, Seung Wook Kim, Jun Gao, Laura Leal-Taixe, Mike Chen, Sanja Fidler, and Huan Ling. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models, 2025. 2

[4] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A flexible technique for accurate omnidirectional camera calibration and structure from motion. In *Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, pages 45–45. IEEE, 2006. 3

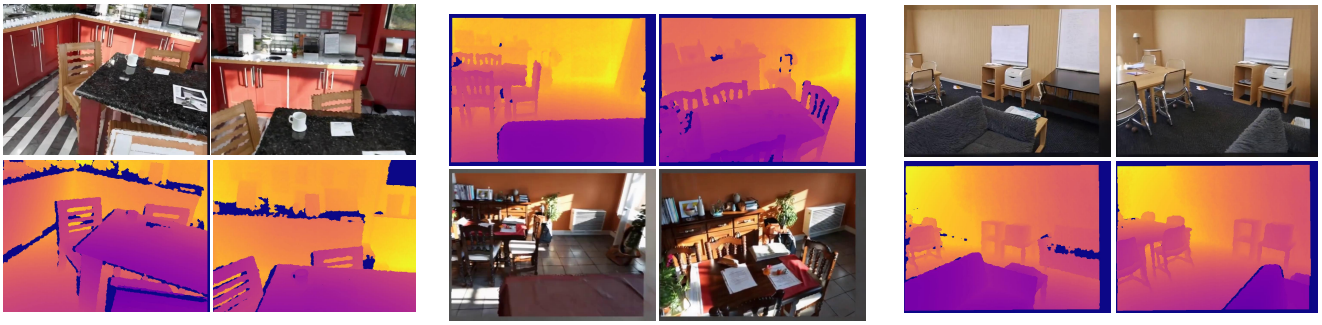


Example 1



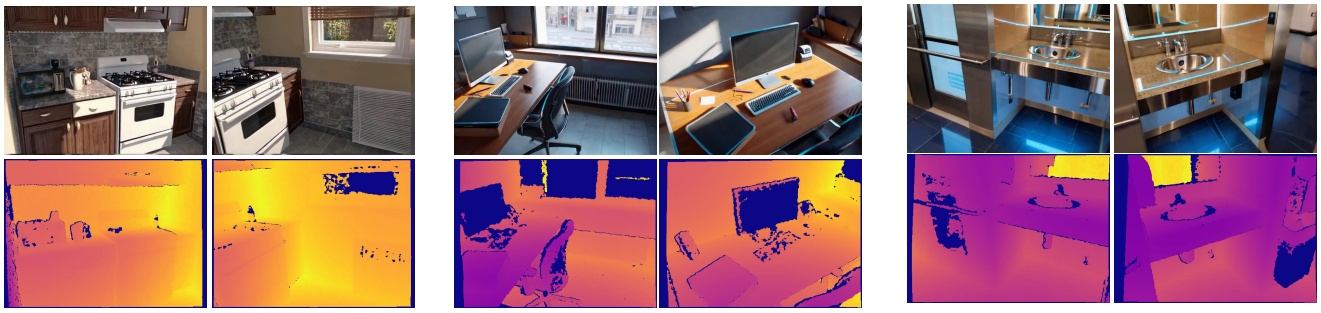
Example 2

Figure 6. **Qualitative registration examples of C-GenReg on the Waymo dataset.** Row (a) shows generated source and target images with a subset of matched keypoints (same color indicates correspondence). Row (b) shows the same correspondences visualized on the source and target 3D point clouds. The resulting rotation error (RRE, °) and translation error (RTE, m) are also reported.



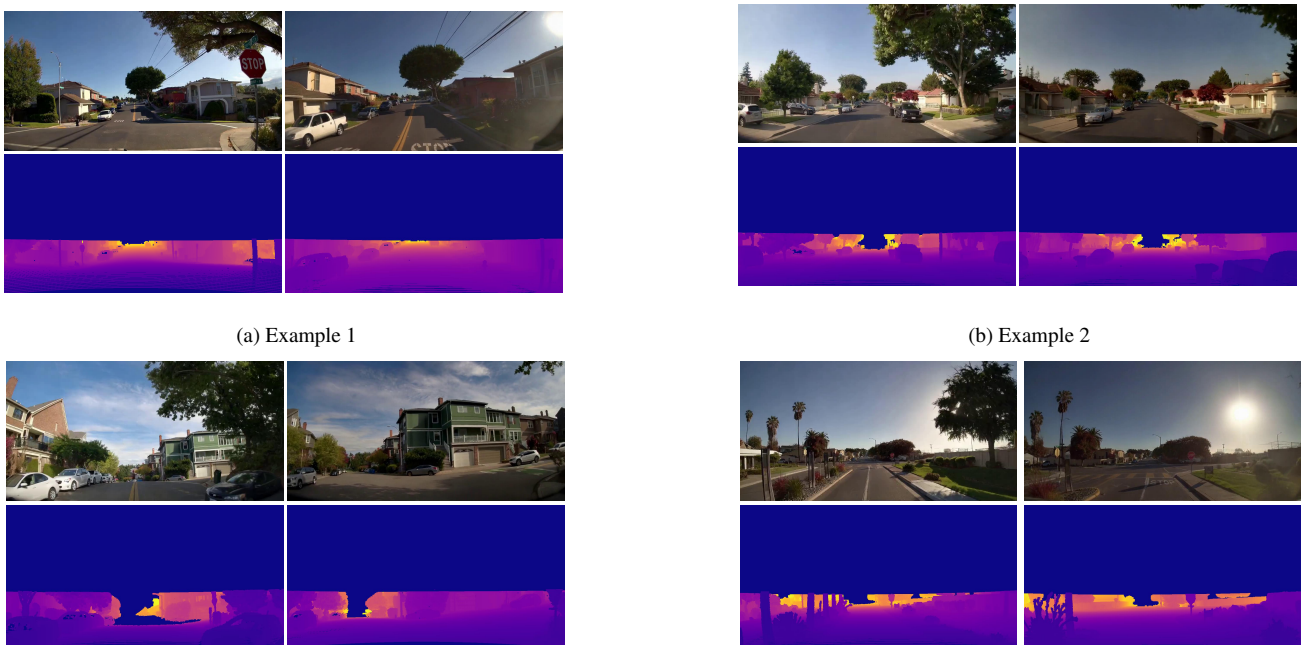
(a) Example 1 (b) Example 2 (c) Example 3

Figure 7. **Multi-view consistent RGB generation from depth on 3DMatch.** Three representative synthetic RGB examples generated from depth. The paired views remain geometrically and visually consistent.



(a) Example 1 (b) Example 2 (c) Example 3

Figure 8. **Multi-view consistent RGB generation from depth on ScanNet.** Three representative synthetic RGB examples from indoor depth scans. The synthesized frames preserve layout and structure across viewpoints.



(a) Example 1 (b) Example 2

(c) Example 3 (d) Example 4

Figure 9. **Multi-view consistent RGB generation from depth on Waymo.** Four representative synthetic RGB examples generated from LiDAR-projected depth. The synthesized frames preserve scene geometry across viewpoints.