GGPT: Geometry-Grounded Point Transformer

Yutong Chen[1], Yiming Wang[1], Xucong Zhang[1,2], Sergey Prokudin[1], Siyu Tang[1]
ETH Zurich[1]; Delft University of Technology[2]

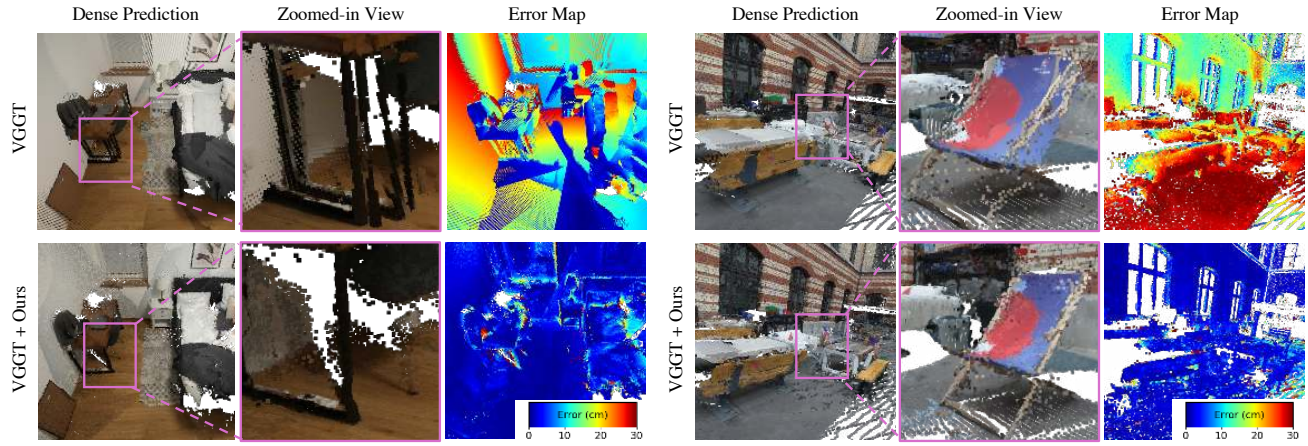

Figure 1. Top row: the dense point maps predicted by feed-forward methods (*e.g.* VGGT [42]) struggle with multi-view geometric consistency resulting in large error. Bottom row: with geometric guidance, our GGPT refines the dense point maps to enhance global alignment and 3D consistency, substantially reducing the reconstruction error.

Abstract

*Recent feed-forward networks have achieved remarkable progress in sparse-view 3D reconstruction by predicting dense point maps directly from RGB images. However, they often suffer from geometric inconsistencies and limited fine-grained accuracy due to the absence of explicit multi-view constraints. We introduce the **G**eometry-**G**rounded **P**oint **T**ransformer (GGPT), a framework that augments feed-forward reconstruction with reliable sparse geometric guidance. We first propose an improved Structure-from-Motion pipeline based on dense feature matching and lightweight geometric optimisation to efficiently estimate accurate camera poses and partial 3D point clouds from sparse input views. we propose a geometry-guided 3D point transformer that refines dense point maps under explicit partial-geometry supervision using an optimised guidance encoding. Extensive experiments demonstrate that our method provides a principled mechanism for integrating geometric priors with dense feed-forward predictions, producing reconstructions that are both geometrically consistent and spatially complete, recovering fine structures and filling gaps in textureless areas. Trained solely on ScanNet++ with VGGT predictions, GGPT generalises across architectures and datasets, substantially outperforming state-of-the-art feed-forward 3D reconstruction models in both in-domain and out-of-domain settings. Project website: https://chenyutongthu.github.io/research/ggpt*

1. Introduction

In recent years, feed-forward 3D reconstruction networks have rapidly emerged, aiming to recover complete scene geometry directly from sparse RGB inputs. Starting with DUSt3R [44], which first demonstrated that a single transformer can jointly predict camera poses and dense 3D point maps from uncalibrated image pairs, subsequent works such as MASt3R [21] and VGGT [42] have further advanced this paradigm. Trained on large-scale multi-view datasets, these vision transformers can now produce dense, visually coherent reconstructions in a single forward pass. As a result, feed-forward 3D reconstruction has become one of the most promising directions in 3D vision, offering fast, scalable, and unified reconstruction pipelines.

Yet, despite their impressive visual results, closer inspection reveals that these models still struggle with multi-view geometric consistency (see Fig. 1). Their reconstruc-

tions often exhibit multi-layer artifacts and deviate from the ground truth, particularly when applied outside their training distribution. In practice, although these networks can predict plausible 3D structures from limited views, they often fail to recover geometry that is sufficiently accurate and consistent across viewpoints. This limitation becomes evident in out-of-domain scenarios, such as medical or surgical scenes, or human data, where feed-forward predictions can deviate substantially from the true underlying geometry.

In contrast, Structure-from-Motion (SfM) [36] remains firmly grounded in geometric principles [14], producing geometrically consistent reconstructions. However, SfM pipelines remain fragile under wide baselines, low overlap, or limited viewpoints, and typically recover only sparse geometry structure. This contrast highlights a clear opportunity: to combine the completeness and efficiency of feed-forward 3D reconstruction with the geometric accuracy and generalisation of SfM.

Building on this motivation, several recent studies have explored similar ideas [12, 16, 17, 58], showing that incorporating sparse geometric guidance can improve dense feed-forward reconstructions. However, existing approaches remain limited in two key aspects. *First*, they often rely on unrealistic SfM guidance, such as pseudo SfM points sampled from ground truth [16, 17, 58] or SfM results obtained from densely captured video sequences [12], which are rarely available in real-world sparse-view scenarios. This dependency largely reflects the lack of robustness and efficiency of conventional SfM pipelines under limited input views. *Second*, prior methods refine predictions in 2D image space via depth-map diffusion [12] or image transformers [16, 17], which inherently constrain the model to view-dependent reasoning. As a result, they fail to exploit the explicit 3D structure of the scene and cannot enforce true cross-view geometric consistency.

In this work, we introduce GGPT (**G**eometry-**G**rounded **P**oint **T**ransformer), a geometry-guided framework that refines dense feed-forward reconstructions using accurate geometric guidance obtained from an improved SfM pipeline, and does so directly in 3D space.

First, we revisit SfM under limited input views and introduce an improved pipeline that integrates dense matchers [9, 55] with a lightweight optimisation procedure. Concretely, we estimate camera poses from a compact set of high-confidence correspondences and then triangulate all valid matches using a direct linear transform. Compared with state-of-the-art SfM methods designed for unconstrained captures, including VGGT+BA [41, 42] and MASt3R-SfM [8], our SfM pipeline achieves higher accuracy thanks to the rich geometric constraints provided by the dense correspondences, and also improved efficiency thanks to the separation of non-linear BA optimisation and linear triangulation. Its robustness and efficiency also enable the construction of realistic SfM supervision without requiring densely captured multi-view sequences.

Second, we introduce a variant of a lightweight 3D Point Transformer [48] that jointly processes dense point maps from feed-forward models and geometrically grounded partial point cloud from our SfM pipeline, predicting residual corrections for every point. By reasoning directly in a global 3D coordinate space, GGPT propagates the geometric accuracy of triangulated SfM points to dense but noisy feed-forward predictions. In contrast to prior depth completion methods that operate on 2D image tokens and channel-wise fuse geometry and image features [12, 16, 17, 58], GGPT performs attention on the two point clouds directly in 3D, where spatial proximity, not pixel coordinates, defines receptive fields. This design explicitly enforces multi-view geometric consistency and produces globally aligned and metrically coherent dense reconstructions.

We conduct extensive experiments to analyse both our improved SfM pipeline and Geometry-Grounded Point Transformer. Our SfM framework achieves better performance and efficiency than state-of-the-art SfM alternatives [8, 41], making it a practical and robust solution for applications beyond the scope of this work. When conditioned on the same SfM guidance, our proposed point transformer predicts significantly more accurate dense points than prior depth completion networks [12, 16, 17, 58].

Overall, thanks to its intuitive and modular design, our method can be seamlessly integrated with various feed-forward 3D reconstruction models at inference time, without requiring any fine-tuning. Remarkably, despite being trained only on ScanNet++ [49] using VGGT [42] dense predictions, GGPT generalises effectively across architectures and datasets, substantially improving the performance of state-of-the-art feed-forward models in both in-domain and out-of-domain settings. The strong generalisation results further suggest that it can serve as a broadly useful tool for a wide range of 3D reconstruction applications, extending well beyond conventional benchmark scenarios.

2. Related Work

Feed-forward 3D reconstruction models. Recent multi-view image models [10, 17, 26, 40, 42–44, 46, 54] predict camera poses and depths directly from RGBs. Trained on large-scale 3D datasets, these models offer greater robustness and efficiency than traditional SfM [36] in sparse-view settings. However, lacking explicit geometric constraints, their predictions often show multi-view inconsistency and global spatial drift, especially in out-of-domain scenarios. Our method alleviates these issues by leveraging geometry estimation to ground the dense predictions.

Structure-from-Motion. COLMAP [36], the traditional incremental SfM, is the standard approach to 3D reconstruction from sufficient overlapping views. MP-SfM [28]

enhances COLMAP with monocular priors, focusing on improving the camera pose in low-overlap cases. VGG-SfM [41] replaces the incremental pipeline with a simplified global optimisation initialised by feed-forward predictions. More recently, VGGT+BA [42] improves the results by using VGGT as initialisation for a global bundle adjustment, but it only estimates very sparse points due to the expensiveness of dense BA. MASt3R-SfM [8] adopts MASt3R [21] predictions for sparse matching and initialisation, and jointly optimize cameras and depths to produce dense prediction. However, its points estimation has limited accuracy due to the sparse matches. Our SfM adopts a global optimisation instead of incremental reconstruction [28, 36], which greatly simplifies the algorithm. In contrast to previous global SfM methods [8, 41, 42], our SfM pipeline combines efficient dense matching regressors [9, 55] with a lightweight sparse BA and dense linear triangulation, achieving improved accuracy and efficiency.

Geometry-conditioned 3D reconstruction. Previous work typically frames geometry-conditioned 3D reconstruction as a 2D depth completion task. Monocular depth completion models [1, 12, 22, 47, 58] inject sparse depth maps into pretrained depth networks, but their predictions suffer from multi-view inconsistency when applied to multi-view reconstruction. Recent multi-view transformers [16, 17] accept additional depth maps, yet depend on pseudo-SfM points sampled from ground truth. In contrast, we address a more practical and challenging setting, enhancing dense predictions using SfM points derived *solely* from the input RGB views. Unlike prior approaches that operate in 2D image space, our model employs a 3D point transformer to jointly process dense and sparse point clouds in 3D, yielding improved multi-view consistency and generalisation.

3D point cloud processing. Unlike 2D image processing, 3D point clouds exhibit irregular and non-uniform spatial distributions and require specialised backbone designs, such as MLP-based PointNets [29, 30], sparse convolutional networks [5], and attention-based Point Transformers [48, 56]. Early 3D networks mainly targeted scene understanding tasks, while recent work extends them to geometry-centric problems, including point cloud denoising [6, 25, 27, 32, 39], completion [52, 53], registration [24, 50], sampling [13, 33, 51], Gaussian splat processing [4, 20, 57], and point tracking [31]. We use point transformers to refine dense predictions under sparse geometric guidance, introducing dedicated encodings that capture positional relations and correspondence offsets between the two inputs. This bridges feed-forward image-based reconstruction and geometry-aware 3D reasoning into a unified, spatially grounded refinement framework.

3. Method

Our core idea is to refine dense point maps $\mathbf{X}_d$ predicted by feed-forward multi-view transformers [17, 42, 46] using geometrically accurate yet incomplete 3D point clouds $\mathbf{X}_s$. As illustrated in Fig. 2, our method consists of two stages. (1) **Efficient and Robust SfM**: a lightweight and robust SfM pipeline with dense matchers and sparse BA to produce an incomplete but geometrically consistent point map $\mathbf{X}_s$; and (2) **Geometry-Grounded Point Transformer** (GGPT): a 3D point transformer to refine the feed-forward prediction $\mathbf{X}_d$ with the geometric guidance $\mathbf{X}_s$.

3.1. Efficient and Robust SfM

Key insight. Our SfM pipeline aims to efficiently estimate accurate camera poses and a sufficient number of points from limited observations. We initialise global optimisation using camera parameters and points predicted by feed-forward models, improving robustness and efficiency compared with traditional incremental SfM pipelines [28, 36]. We leverage recent dense matching regressors, including RoMa [9] and UFM [55], to extract multi-view correspondences from pairwise matchings. Directly using all dense matchings for global non-linear optimisation is computationally expensive. Instead, we first perform sparse bundle adjustment (BA) on a compact set of high-confidence matches to estimate camera poses, followed by an efficient direct linear triangulation (DLT) to reconstruct dense points.

Initialisation from feed-forward models. Given N unposed RGB images $\{\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^N$, a multi-view transformers f_θ can predict camera parameters and dense 3D point maps:

$$\{\mathbf{g}_i, \mathbf{P}_i\}_{i=1}^N = f_\theta\big(\{\mathbf{I}_i\}_{i=1}^N\big)\,, \tag{1}$$

where each camera parameter vector $\mathbf{g}_i = [\mathbf{q}_i, \mathbf{t}_i, \mathbf{f}_i]$ contains rotation (unit quaternion $\mathbf{q}_i$), translation $\mathbf{t}_i$, and intrinsics $\mathbf{f}_i = [f_x, f_y, c_x, c_y]$. Each dense point map $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}$ assigns to every pixel (u, v) a predicted 3D coordinate $\mathbf{p}_{i,uv} \in \mathbb{R}^3$ in the global scene frame. The union of all $\mathbf{P}_i$ forms the initial dense prediction $\mathbf{X}_d = \{\mathbf{P}_i\}$, later refined by our transformer-based stage.

Dense feature matching. We adopt RoMa [9] and UFM [55] to obtain a global correspondence tensor across all image pairs

$$\mathbf{T} \in \mathbb{R}^{N\times N\times W\times H\times 2}, \quad \mathbf{C} \in [0,1]^{N\times N\times W\times H}, \tag{2}$$

where $\mathbf{T}[i,k,u,v]$ denotes the predicted location in view k corresponding to pixel (u, v) in view i, and $\mathbf{C}[i,k,u,v]$ its confidence. Invalid or inconsistent correspondences are removed by enforcing cycle consistency:

$$\big\|\mathbf{T}[k,i,\mathbf{T}[i,k,\mathbf{u}]] - \mathbf{u}\big\|_2 < \epsilon, \tag{3}$$

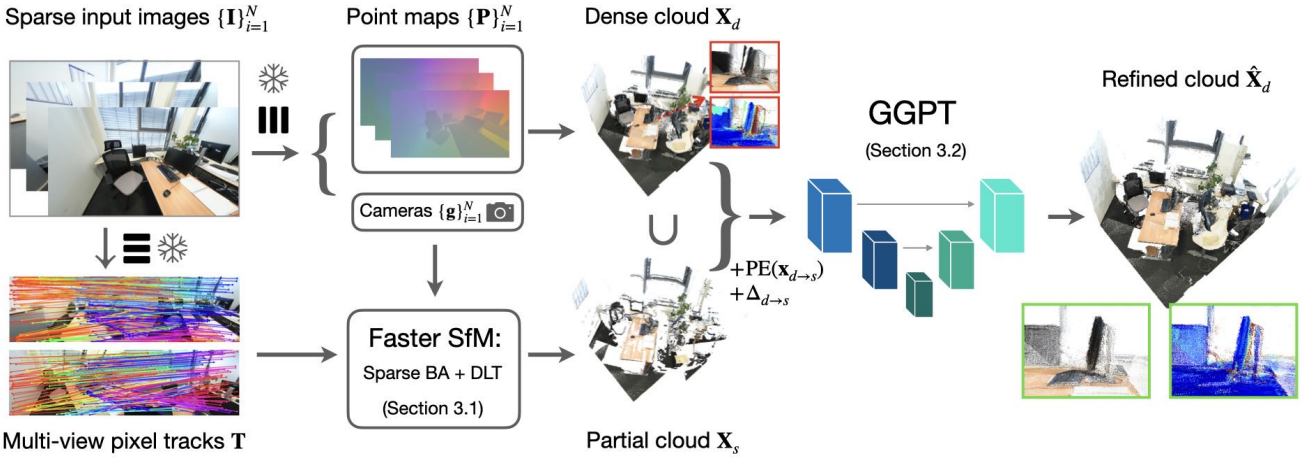

Figure 2. *Overview of our method.* We utilise off-the-shelf feed-forward multi-view transformers [17, 42] and dense matchers [9, 55] to predict dense point maps $\{\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{N}$, camera parameters $\{\mathbf{g}_i\}_{i=1}^{N}$, and multi-view correspondences $\mathbf{T} \in \mathbb{R}^{N\times N\times H\times W\times 2}$. These correspondences are used for sparse bundle adjustment and multi-view DLT triangulation, producing a geometrically consistent yet incomplete sparse cloud $\mathbf{X}_s$ via a lightweight alternative to standard SfM pipelines (Section 3.1). The dense yet multi-view inconsistent point cloud $\mathbf{X}_d$, obtained by combining the predicted per-view point maps, exhibits local misalignments (red boxes) that are then refined under the geometric guidance of $\mathbf{X}_s$ by our *Geometry-Grounded Point Transformer (GGPT)* (Section 3.2), a modified point transformer [48] equipped with specialised geometric embeddings $\mathrm{PE}(\mathbf{x}_{d\to s})$ and $\Delta_{d\to s}$, which encode spatial relations between dense and sparse points to provide stable geometric guidance. The final output is a refined, globally aligned, and geometry-consistent dense point cloud $\hat{\mathbf{X}}_d$.

producing a binary mask $\mathbf{M} \in \{0,1\}^{N\times N\times W\times H}$ that defines the set of geometrically valid matches.

Sparse bundle adjustment. From $\mathbf{T}$, $\mathbf{C}$, and $\mathbf{M}$ we derive two confidence-masked subsets for bundle adjustment and triangulation:

$$\mathbf{M}_{\text{BA}} = \big[\mathbf{C} > \epsilon_{\text{BA}}\big] \odot \mathbf{M},\ \mathbf{M}_{\text{DLT}} = \big[\mathbf{C} > \epsilon_{\text{DLT}}\big] \odot \mathbf{M}, \tag{4}$$

with $\epsilon_{\text{BA}} > \epsilon_{\text{DLT}}$, ensuring that BA uses only high-confidence tracks while DLT benefits from denser correspondences. Applying $\mathbf{M}_{\text{BA}}$ to $\mathbf{T}$ yields the masked track set $\mathbf{T}_{\text{BA}} = \mathbf{T} \odot \mathbf{M}_{\text{BA}}$. Each track $\mathbf{T}[i,:,u,v]$ matches an anchor pixel (u,v) in view i to pixels in other views. We use j to index the track anchored at (i,u,v) and denote its visible views as

$$\mathcal{V}^j = \{\, k \mid \mathbf{M}_{\text{BA}}[i,k,u,v] = 1 \,\},$$

As $\mathbf{T}_{\text{BA}}$ can still contain large numbers of tracks, we further select a maximal number of anchor pixels from each view and use their associated tracks for our BA. We choose n_{BA} anchor pixels per image with highest SuperPoint saliency [7] and visible in at least two views. Using the maximal $N \times n_{\text{BA}}$ tracks, we jointly refine cameras and the anchor points by minimising the 2D reprojection loss:

$$\{\mathbf{G}^*, \mathbf{X}^*\} = \arg\min_{\{\mathbf{g}_i\},\{\mathbf{x}^j\}} \sum_j \sum_{i\in\mathcal{V}^j} \mathcal{L}_{\text{r}}\big(\pi(\mathbf{g}_i, \mathbf{x}^j) - \mathbf{u}_i^j\big). \tag{5}$$

where $\pi(\mathbf{g}_i, \mathbf{x}^j)$ projects 3D point $\mathbf{x}^j$ into view i and $\mathbf{u}_i^j$ is the observation of track j in the view i, $\mathcal{L}_{\text{r}}$ is the robust Cauchy loss [2]. The optimisation is initialised from $\{\mathbf{g}_i, \mathbf{P}_i\}$ and converges quickly thanks to the feed-forward prior. While both cameras and sparse points are refined, only the updated camera estimates $\mathbf{G}^* = \{\mathbf{g}_i^*\}$ are propagated to the next stage.

Direct linear transform. Finally, we apply $\mathbf{M}_{\text{DLT}}$ to the correspondences to obtain $\mathbf{T}_{\text{DLT}} = \mathbf{T} \odot \mathbf{M}_{\text{DLT}}$ and reconstruct 3D points using all available valid correspondences jointly via a multi-view variant of the Direct Linear Transform [15]. Each track $\mathbf{T}^j$ across views $\mathcal{V}^j$ is triangulated into a 3D point $\mathbf{x}^j$ using the refined cameras $\mathbf{G}^*$. Points with large reprojection error or small triangulation angle are discarded. The resulting set

$$\mathbf{X}_s = \big\{\, \mathbf{x}^j \;\big|\; \mathbf{T}^j \subseteq \mathbf{T}_{\text{DLT}} \,\big\} \tag{6}$$

constitutes our final geometrically consistent scene representation. Note that $\mathbf{X}_s$ is incomplete for regions which are only observed in single views or do not have valid matchings due to lack of saliency. Additionally, as each $\mathbf{x}^j$ associates with pixels in $\mathbf{T}^j$, it has multiple correspondences (at least two) in the feed-forward multi-view point map $\mathbf{X}_d$. Next, we will use $\mathbf{X}_s$ as the geometric guidance to refine the inaccurate dense predictions $\mathbf{X}_d$.

3.2. Geometry-Grounded Point Transformer

Key insight. The geometry reconstruction $\mathbf{X}_s$ is geometrically accurate but incomplete, particularly lacking coverage

in textureless or occluded regions, whereas the dense prediction $\mathbf{X}_d$ from feed-forward models is complete yet can be multi-view inconsistent. To combine their complementary strengths, we introduce the *Geometry-Grounded Point Transformer*, which refines $\mathbf{X}_d$ under the guidance of $\mathbf{X}_s$. By jointly reasoning over both in a shared 3D space, GGPT transfers the geometric reliability of triangulated points to dense but inaccurate predictions. Unlike prior refinement methods [16, 17] that attend to 2D image tokens, GGPT performs attention directly in 3D, where spatial proximity rather than pixel location defines receptive fields, yielding 3D consistent dense reconstructions.

Input embeddings. Given the two point clouds $\mathbf{X}_d \in \mathbb{R}^{NHW\times 3}$ and $\mathbf{X}_s \in \mathbb{R}^{N_s\times 3}$, we first align them using the Kabsch–Umeyama transform [37] and embed each point $\mathbf{x}_i$ into an initial feature vector $\mathbf{z}_i^{(0)}$. For $\mathbf{x}_s \in \mathbf{X}_s$,

$$\mathbf{z}_s^{(0)} = [\mathrm{PE}(\mathbf{x}_s),\, \mathbf{e}_{\mathrm{type}(s)}], \tag{7}$$

where $\mathrm{PE}(\mathbf{x}_s)$ is a sinusoidal positional encoding [38, 56] with a frequency of 4 and $\mathbf{e}_{\mathrm{type}(s)} \in \mathbb{R}^{16}$ a learnable token marking guidance points. Some points $\mathbf{x}_d \in \mathbf{X}_d$ have corresponding points in the guidance, denoted as $\mathbf{x}_{d\to s} \in \mathbf{X}_s$, since they originate from the same image pixels. To make the network aware of this relation, we define

$$\mathbf{z}_d^{(0)} = [\mathrm{PE}(\mathbf{x}_d),\, \mathbf{e}_{\mathrm{type}(d)},\, \mathrm{PE}(\mathbf{x}_{d\to s}),\, \Delta_{d\to s}], \tag{8}$$

where $\mathbf{e}_{\mathrm{type}(d)}$ is a learnable token for dense points and $\Delta_{d\to s} = \mathbf{x}_{d\to s} - \mathbf{x}_s$ encodes positional offsets. For $\mathbf{x}_d$ without correspondences, we use only the first two terms. All embeddings are zero-padded to a uniform dimension, yielding the combined input $\mathbf{Z}^{(0)} = \{\mathbf{z}_s^{(0)}\} \cup \{\mathbf{z}_d^{(0)}\}$.

Point transformer backbone. The embeddings $\mathbf{Z}^{(0)}$ are processed by a 3D point transformer $g_\gamma(\cdot)$ with L layers:

$$\mathbf{Z}^{(L)} = g_\gamma(\mathbf{Z}^{(0)}), \tag{9}$$

We use an L=8-layer Point Transformer V3 [48] (PTv3), which applies patch-wise self-attention over spatial neighbourhood to capture fine-scale structure and long-range dependencies. This models point interaction by 3D proximity rather than exhaustive 2D grid attention, fusing $\mathbf{X}_d$ and $\mathbf{X}_s$ into 3D-aware features $\mathbf{Z}^{(L)}$. Our PTv3 backbone has 53M parameters in total, considerably lighter than 2D vision transformers (around 300M) used in prior geometry-conditioned methods [12, 17].

Prediction head. The geometry-aware features $\mathbf{Z}^{(L)}$ are decoded by a shared MLP $h_\psi(\cdot)$ with ReLU activations:

$$[\boldsymbol{\delta},\, \tilde{c}] = h_\psi(\mathbf{Z}^{(L)}), \tag{10}$$

where $\boldsymbol{\delta} \in \mathbb{R}^3$ is the predicted residual displacement and $\tilde{c} \in \mathbb{R}$ the raw confidence. The refined coordinates are computed as $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$, and the final confidence as $c = \exp(\tilde{c}) + 1$ following [18, 42, 44]. The refined dense and sparse sets are $\hat{\mathbf{X}}_d = \{\hat{\mathbf{x}}_d\}$ and $\hat{\mathbf{X}}_s = \{\hat{\mathbf{x}}_s\}$, with $\hat{\mathbf{X}}_d$ used as the final output.

Patch-based processing. To handle large-scale point clouds efficiently, GGPT operates on spatially local patches instead of the full scene. We represent the dense point cloud as overlapping subsets $\mathbf{X}_d = \{\mathbf{X}_i\}_{i=1}^{N_p}$, where each patch $\mathbf{X}_i$ corresponds to a cubic region centered at a 3D location and normalised to the unit cube $[0, 1]^3$. During *training*, we randomly sample anchor points from $\mathbf{X}_s$ and extract local cubes around them. During *inference*, we apply a sliding cube with some overlap to cover the whole point cloud. Each patch is processed independently by the transformer and prediction head. For points included in multiple overlapping patches, their final 3D coordinates are computed by averaging the predictions in multiple patches. Such patch-based decomposition, also used in prior point cloud denoising [39], balances computational efficiency with fine-grained geometric fidelity across large 3D scenes.

Training objectives. GGPT is trained with the loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{conf}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}. \tag{11}$$

Confidence-weighted regression. We use a heteroscedastic formulation [18]:

$$\mathcal{L}_{\mathrm{conf}} = \sum_{\mathbf{x}\in\mathbf{X}_d\cup\mathbf{X}_s} c\,\|\hat{\mathbf{x}} - \mathbf{x}_{\mathrm{GT}}\| - \alpha \log c. \tag{12}$$

The predicted confidence c modulates each residual, reducing the effect of uncertain regions while the $-\alpha \log c$ term penalises trivial solution with overly low confidence. Ground-truth points $\mathbf{X}_{\mathrm{GT}}$ are pre-aligned to $\mathbf{X}_s$ via Umeyama alignment, ensuring supervision focuses on local corrections rather than global shifts.

Identity consistency. If a point $\hat{\mathbf{x}} \in \mathbf{X}_d$ has a valid correspondence in the guidance, *i.e.* $\mathbf{x}_{d\to s} \in \mathbf{X}_s$, we apply an anchoring term to its prediction $\hat{\mathbf{x}}$:

$$\mathcal{L}_{\mathrm{id}} = \sum_{\mathbf{x}\in\mathbf{X}_d^{d\to s}} \|\hat{\mathbf{x}} - \mathbf{x}_{d\to s}\|, \tag{13}$$

where $\mathbf{X}_d^{d\to s}$ denotes the set of points with geometry guidance. This term encourages the model to predict points aligned with the geometry guidance.

Key architectural insights. We provide ablation studies in the supplementary to highlight two factors improving the performance of our point transformer design: (i) incorporating partial geometry $\mathbf{X}_s$ as guidance, both as an auxiliary input and via encodings $\mathrm{PE}(\mathbf{x}_{d\to s})$ and $\Delta_{d\to s}$, and (ii) adopting patch-based processing, which enhances efficiency while retaining fine geometric detail.

Table 1. **Multi-view 3D Reconstruction on Standard Test Sets.** We report AUC@5/10 cm (% ↑). Our method (our SfM and GGPT) can improve dense predictions of various models. Despite GGPT being trained solely on ScanNet++ and VGGT's dense prediction, as we highlight the top-left cells as **within-domain**, it generalises well to **cross-domain datasets** and other methods' predictions (rows 3–10).

	ScanNet++ [49]			**ETH3D** [35]			**T&T** [19]		
	4 imgs	8 imgs	16 imgs	4 imgs	8 imgs	16 imgs	4 imgs	8 imgs	16 imgs
VGGT [42]	23/37	19/32	16/29	27/41	23/36	19/32	26/40	25/39	24/38
[42] + Ours	**38/53**	**45/60**	**50/66**	**41/55**	**47/61**	**49/63**	**34/47**	**42/57**	**43/57**
Pi3 [46]	**54/69**	**56**/71	58/**74**	31/47	25/41	23/38	25/39	26/42	25/40
[46] + Ours	**54**/68	**56/72**	**59/74**	**36/53**	**36/53**	**37/54**	**27/43**	**32/50**	**33/50**
MapAnything [17]	40/57	38/57	38/58	10/20	7/15	5/12	10/21	9/20	8/17
[17] + Ours	**44/61**	**48/64**	**52/68**	**32/43**	**33/45**	**34/47**	**29/43**	**40/55**	**42/56**
MAtCha [11]	12/19	15/26	18/32	40/52	41/53	42/56	34/47	36/50	33/47
[11] + Ours	**27/37**	**40/52**	**48/63**	**42/55**	**47/60**	**50/65**	**35/48**	**43/57**	**43/57**
MASt3R-SfM [8]	12/22	14/35	16/31	37/50	39/51	40/54	**34**/46	37/50	36/49
[8] + Ours	**30/41**	**39/52**	**48/64**	**41/55**	**46/59**	**48/63**	33/**47**	**41/56**	**42/56**

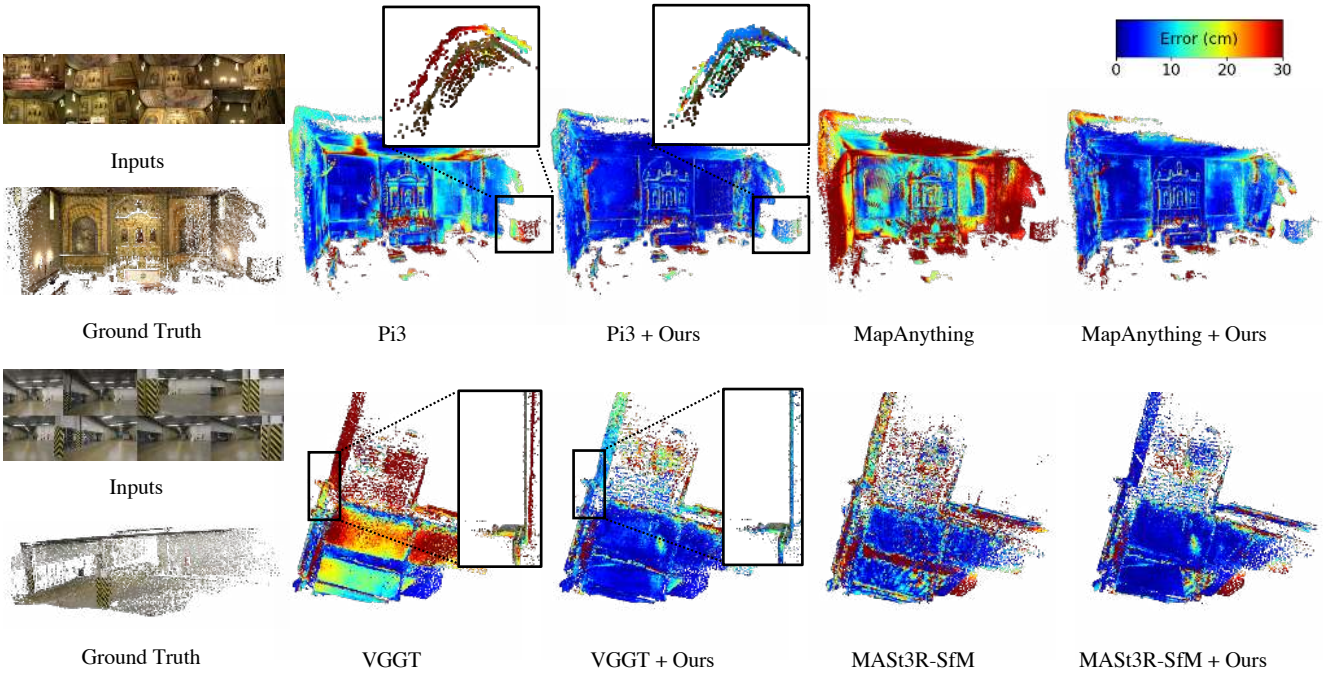

Figure 3. **3D reconstruction results on T&T [19] and ETH3D [35].** Points with confidence above the 90% quantile are visualised. We compare the error maps of the reconstruction before and after our refinement. In the zoomed-in regions, the ground truth (colored by input RGBs) is overlaid with the predictions (colored by error) to highlight how our method corrects the misalignment of input points.

4. Experiments

4.1. Implementation Details

SfM configurations. To filter matches for BA and DLT, we set $\epsilon = 4$, $\epsilon_{\text{BA}} = 0.6$, $\epsilon_{\text{DLT}} = 0.1, n_{\text{BA}} = 2048$. After DLT, we filter out points with reprojection error above 4 pixels and maximal triangulation angle below 3 degrees.

GGPT training. The training set has 20k multi-view sequences sampled from 856 training scenes in ScanNet++ [49]. We use VGGT [42] for $\mathbf{X}_d$ prediction and SfM initialisation. Each input patch has a half width of $0.4\times$ the scene radius. Each forward pass processes up to 400k points. We set $\lambda_{id} = 1, \alpha = 0.2$. Training is performed on 8 NVIDIA GH200 GPUs for one day.

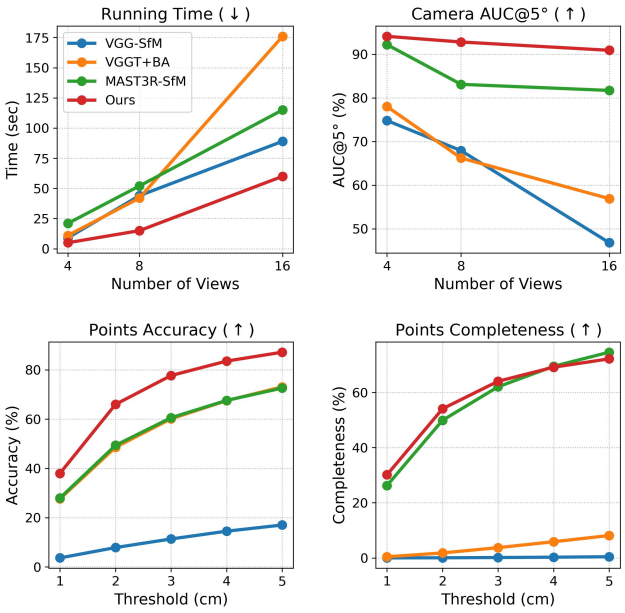

Figure 4. **Comparison between SfMs on ETH3D [35].** Across 4/8/16-view setups, our SfM pipeline achieves consistently better camera pose accuracy, points accuracy, and good points completeness, while retaining the shortest running time.

Inference runtime. We provide a detailed runtime breakdown for different input views in the Sec. C of the supplementary, which shows that dense matching [9, 55] dominates the runtime while our proposed BA, DLT, and GGPT refinement add minor overhead.

Evaluation protocols. We align the prediction to the metric ground truth with robust Umeyama implemented by pycolmap [34], compute Euclidean error between predicted point and its corresponding ground truth, and calculate the area under the curve at varying error thresholds.

4.2. Within-domain and Cross-domain Evaluation

We evaluate methods on three benchmarks: ScanNet++ [49], ETH3D [35], and T&T [19].mEach dataset contains varying baseline coverages with 4, 8, and 16 input views and each view split has 32 sequences. For ScanNet++ and ETH3D, we use a subset of test sets in [17]. For T&T, we adopt the same view-sampling protocol [17] to construct a test set. Since GGPT is exclusively trained on ScanNet++ using VGGT predictions, evaluations on unseen ScanNet++ scenes are referred to as **within-domain**, while results on ETH3D and T&T are referred to as **cross-domain**.

Despite GGPT is trained solely on VGGT's dense prediction, it can be applied to improve the dense prediction of other approaches, including feed-forward methods [17, 46] and forward-optimization hybrid methods [8, 11]. Concretely, we replace VGGT's BA initialisation and $\mathbf{X}_d$ with the alternative prediction while using the same SfM pipeline and GGPT model weights.

Table 2. **Multi-view 3D reconstruction on out-of-domain datasets.** AUC@1/5 cm (% ↑) for human body reconstruction (4D-DRESS [45]) and AUC@1/5 mm (% ↑) for surgical scene reconstruction (MV-dVRK [3]).

	4D-DRESS [45]	MV-dVRK [3]
VGGT [42]	10/45	8/33
[42] + Ours	**66/77**	**45/61**
Pi3 [46]	8/50	18/51
[46] + Ours	**63/80**	**40/67**
MapAnything [17]	2/12	3/13
[17] + Ours	**42/52**	**35/47**
MAtCha [11]	48/68	37/62
[11] + Ours	**62/75**	**50/67**
MASt3R-SfM [8]	54/71	39/62
[8] + Ours	**64/75**	**49/66**

Tab. 1 shows that our method can seamlessly enhance a range of advanced 3D reconstruction approaches, on both within-domain and cross-domain datasets, across different view setups. Pi3 [46], a concurrent work, is trained on ScanNet++ alongside a vast number of similar indoor scenes, achieving highly optimised performance on the ScanNet++ dataset. Nevertheless, GGPT can still largely improve Pi3 on cross-domain datasets. Fig. 3 shows that our refinement reduces geometric misalignments and mitigates multi-layer artifacts frequently observed in feed-forward predictions.

4.3. Out-of-domain Evaluation

To further test the generalisation of our method, we also evaluate it on **out-of-domain** datasets where the input images differ substantially in appearance or geometry from the training data of both our model and existing 3D reconstruction methods. We consider two challenging out-of-distribution domains: human bodies and robotic abdominal surgery. We render high-resolution clothed body scans from 4D-DRESS [45], and use photorealistic Blender renderings from the MV-dVRK [3] dataset. Each dataset contains 24 sequences, with 4–12 input views per sequence.

Tab. 2 shows our method achieves superior performance, particularly in AUC@1 cm/mm, which reflects fine-grained geometric accuracy. Fig. 5 illustrates that our approach effectively corrects distortions and misalignments produced by feed-forward models [17, 42, 46], yielding more accurate and consistent reconstructions than hybrid methods [8, 11].

4.4. Evaluation of the SfM Pipeline

Direct comparison. Following standard SfM evaluation protocols [23, 41], we measure the camera pose with AUC@5° and the partial point map with Accuracy/Completeness based on its Chamfer distance to the ground truth. Fig. 4 shows that our SfM outperforms state-

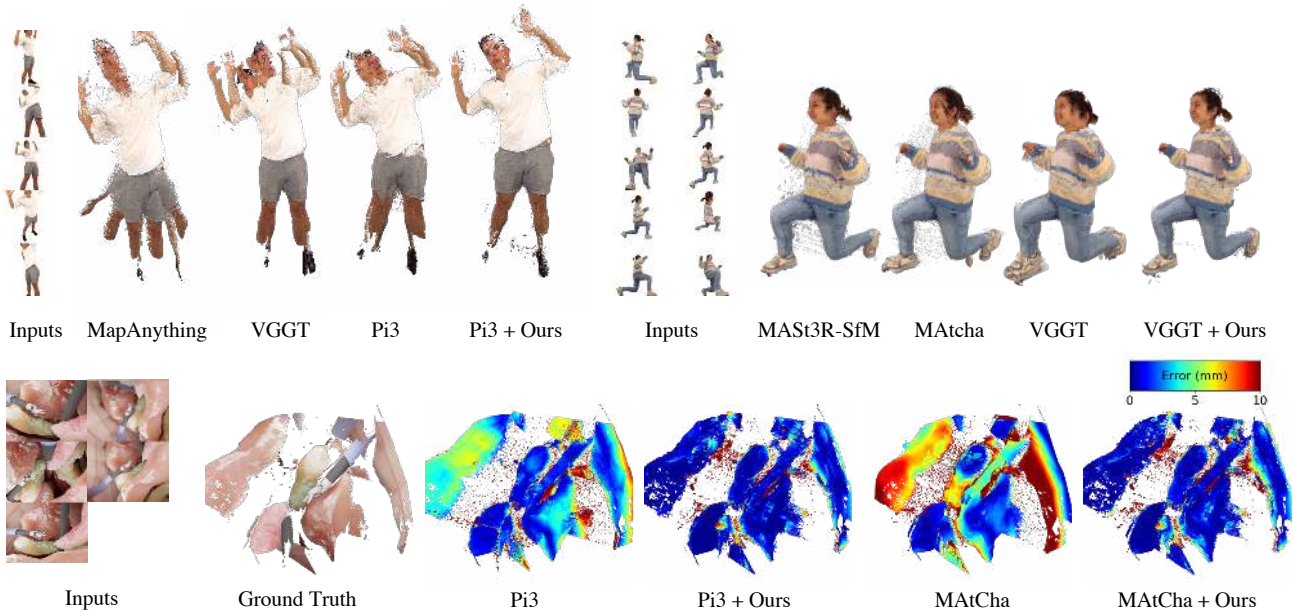

Figure 5. **Examples on out-of-domain 4D-DRESS [45]** with visualisation (top row), and **MV-dVRK** [3] with both visualisation and reconstruction error (bottom row), predicted by feed-forward networks and enhanced by our method. Our GGPT can significantly improve the multi-view consistency in the reconstruction result.

of-the-art global SfM methods [8, 41, 42] in camera pose and points accuracy. Compared with MASt3R-SfM [8], which optimises a dense grid for near-complete scene coverage, our method attains competitive completeness without sacrificing accuracy. Our pipeline consistently achieves the lowest runtime across all input view counts.

Impacts on geometry-conditioned models. To further assess the utility of our SfM results, we feed $\mathbf{X}_s$, from either our method or [8], into state-of-the-art geometry-guided models [12, 16, 17, 58], and compare their dense predictions using each geometric condition. As shown in Tab. 3, replacing the SfM input from [8] with ours yields substantial performance gains across all models. This verifies that our SfM pipeline provides a stronger geometric signal for the dense reconstruction task.

4.5. Geometry-conditioned Models

We compare our geometry-guided 3D point transformer with 2D depth completion methods [12, 16, 17, 58]). All models receive the same partial points $\mathbf{X}_s$ produced either by our SfM or by prior state-of-the-art SfM [8]. For monocular methods, we process each view independently and unproject per-view depths into a global point map using the SfM camera poses. Tab. 3 shows that given the same partial geometry guidance, GGPT consistently produces more accurate dense point maps than all baselines, regardless of the underlying SfM algorithm. This verifies the effectiveness of the 3D point transformer architecture design.

Table 3. **Comparison of geometry conditions and geometry-grounded models.** Points AUC@5/10 cm (% ↑) on the ETH3D [35] 8-view test set. **Two conclusions can be drawn here.** (1) Column-wise, our SfM $\mathbf{X}_s$ provides a stronger geometry signal than prior SfM [8] (We highlight the gain from using our SfM). (2) Row-wise, given the same SfM condition, our 3D GGPT outperforms prior 2D-based methods.

Method	$\mathbf{X}_s$ from [8]	$\mathbf{X}_s$ from our SfM
Murre [12]	9/23	26/40 (+17/+17)
OMNI-DC [58]	25/44	31/44 (+6/+0)
POW3R [16]	13/29	32/45 (+19/+16)
MapAnything [17]	13/32	27/40 (+14/+8)
VGGT [42] + GGPT	**36/50**	**47/61** (+11/+11)

5. Conclusion

We introduced GGPT, a framework that enhances feed-forward 3D reconstruction by employing reliable geometric guidance from SfM. Our efficient SfM uses dense feature matching and linear triangulation to recover accurate sparse geometry. We further introduce a lightweight 3D point transformer to refine the feed-forward dense predictions directly in 3D space. This design provides an effective way to fuse sparse geometric priors with dense feed-forward predictions. Our GGPT demonstrates strong generalisation to various feed-forward methods and diverse datasets, highlighting its practicality as a broadly applicable refinement module for sparse-view 3D reconstruction.

Acknowledgements

This study was conducted within the national "Proficiency" research project (No. PFFS-21-19) funded by the Swiss Innovation Agency Innosuisse in 2021 as one of 15 flagship initiatives. This work was supported by the Swiss AI Initiative under project IDs a136 and a144, funded through a grant from the ETH Domain. We gratefully acknowledge the computational resources provided by the Swiss National Supercomputing Centre (CSCS) under the Alps infrastructure. We sincerely thank Guido Caccianiga and others at the Max Planck Institute for Intelligent Systems for providing the MV-dVRK dataset. We thank Philipp Lindenberger, Shaohui Liu, Zador Pataki, Xudong Jiang, Frano Rajic, Linfei Pan, Nischal Maharjan, Johannes Weidenfeller, Zinuo You, and Malte Prinzler for valuable discussions and support.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. In *BMVC*, 2022. 3

[2] Jonathan T. Barron. A general and adaptive robust loss function. In *CVPR*, 2019. 4

[3] Guido Caccianiga, Sergey Prokudin, Bernard Javot, Yutong Chen, Omer Burak Aladag, Rachael L'Orsa, Yarden Sharon, Jens Rolinger, Ivan Capobianco, Siyu Tang, Anton Deguet, and Katherine J. Kuchenbecker. MV-dVRK: Integrated methods for multi-viewpoint surgical telerobotics. *In submission*, 2026. Available at: https://mv-dvrk.is.mpg.de/. 7, 8

[4] Yutong Chen, Marko Mihajlovic, Xiyi Chen, Yiming Wang, Sergey Prokudin, and Siyu Tang. Splatformer: Point transformer for robust 3d gaussian splatting. In *ICLR*, 2025. 3

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 3

[6] Dasith de Silva Edirimuni, Xuequan Lu, Zhiwen Shao, Gang Li, Antonio Robles-Kelly, and Ying He. Iterativepfn: True iterative point cloud filtering. In *CVPR*, 2023. 3

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *CVPRW*, 2018. 4

[8] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 2, 3, 6, 7, 8

[9] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. In *CVPR*, 2024. 2, 3, 4, 7

[10] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, and Chengfei Lyu. Dens3r: A foundation model for 3d geometry prediction. In *ICLR*, 2026. 2

[11] Antoine Guédon, Tomoki Ichikawa, Kohei Yamashita, and Ko Nishino. Matcha gaussians: Atlas of charts for high-quality geometry and photorealism from sparse views. *CVPR*, 2025. 6, 7

[12] Haoyu Guo, He Zhu, Sida Peng, Haotong Lin, Yunzhi Yan, Tao Xie, Wenguan Wang, Xiaowei Zhou, and Hujun Bao. Multi-view reconstruction via sfm-guided monocular depth estimation. In *CVPR*, 2025. 2, 3, 5, 8

[13] Juhyung Ha, Vibhas Kumar Vats, Soon-heung Jung, Alimoor Reza, and David J. Crandall. Hvpunet: Hybrid-voxel point-cloud upsampling network. In *ICCV*, 2025. 3

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[15] Richard I. Hartley and Peter Sturm. Triangulation. *Comput. Vis. Image Underst.*, 1997. 4

[16] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *CVPR*, 2025. 2, 3, 5, 8

[17] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction, 2025. 2, 3, 4, 5, 6, 7, 8

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 2017. 5

[19] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM ToG*, 2017. 6, 7

[20] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. In *ICLR*, 2025. 3

[21] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 1, 3

[22] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. 2025. 3

[23] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 7

[24] Jiuming Liu, Guangming Wang, Zhe Liu, Chaokang Jiang, Marc Pollefeys, and Hesheng Wang. Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In *ICCV*, 2023. 3

[25] Wang Liu and Wei Gao. Omni-scene perception-oriented point cloud geometry enhancement for coordinate quantization. In *ICCV*, 2025. 3

[26] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *CVPR*, 2025. 2

[27] Aihua Mao, Zihui Du, Yu-Hui Wen, Jun Xuan, and Yong-Jin Liu. Pd-flow: A point cloud denoising framework with normalizing flows. In *ECCV*, 2022. 3

[28] Zador Pataki, Paul-Edouard Sarlin, Johannes L. Schönberger, and Marc Pollefeys. MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion. In *CVPR*, 2025. 2, 3

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 2017. 3

[30] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 3

[31] Frano Rajič, Haofei Xu, Marko Mihajlovic, Siyuan Li, Irem Demir, Emircan Gündoğdu, Lei Ke, Sergey Prokudin, Marc Pollefeys, and Siyu Tang. Multi-view 3d point tracking. In *ICCV*, 2025. 3

[32] Marie-Julie Rakotosaona, Vittorio La Barbera, Paul Guerrero, Niloy J Mitra, and Maks Ovsjanikov. Pointcleannet: Learning to denoise and remove outliers from dense point clouds. In *Computer Graphics Forum*, 2020. 3

[33] Yi Rong, Haoran Zhou, Kang Xia, Cheng Mei, Jiahao Wang, and Tong Lu. Repkpu: Point cloud upsampling with kernel point representation and deformation. In *CVPR*, 2024. 3

[34] Johannes Lutz Schönberger. Colmap – general-purpose structure-from-motion and multi-view stereo pipeline. https://colmap.github.io/, 2025. Version 3.13.0.dev0. 7

[35] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 6, 7, 8

[36] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 3

[37] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 2002. 5

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 5

[39] Mathias Vogel, Keisuke Tateno, Marc Pollefeys, Federico Tombari, Marie-Julie Rakotosaona, and Francis Engelmann. P2p-bridge: Diffusion bridges for 3d point cloud denoising. In *ECCV*, 2024. 3, 5

[40] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *3DV*, 2025. 2

[41] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, 2024. 2, 3, 7, 8

[42] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 3, 4, 5, 6, 7, 8

[43] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025.

[44] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 2, 5

[45] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *CVPR*, 2024. 7, 8

[46] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 2, 3, 6, 7

[47] Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025. 3

[48] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 2, 3, 4, 5

[49] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *CVPR*, 2023. 2, 6, 7

[50] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *CVPR*, 2022. 3

[51] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *CVPR*, 2018. 3

[52] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, 2021. 3

[53] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *3DV*, 2018. 3

[54] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025. 2

[55] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. Ufm: A simple path towards unified dense correspondence with flow, 2025. 2, 3, 4, 7

[56] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 3, 5

[57] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. Diffgs: Functional gaussian splatting diffusion. In *NeurIPS*, 2024. 3

[58] Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omnidc: Highly robust depth completion with multiresolution depth integration. In *ICCV*, 2025. 2, 3, 8