FootRecon: Quadrupedal Terrain Reconstruction from Sparse Foot Contacts with Geometric Prior

Yujin Park[1], Seungoh Han[1], Taebok Lee[2], Chanwoo Lee[1], Hyungyu Park[1], Kyungdon Joo[1†]

***Abstract*— Reliable terrain geometry is essential for safe legged locomotion in unstructured environments. However, existing methods rely on dense exteroceptive sensing, whose reliability degrades under adverse conditions. In contrast, proprioceptive foot–terrain contacts remain available but are inherently sparse, rendering terrain reconstruction fundamentally ill-posed. We propose `FootRecon`, a proprioceptive-only terrain reconstruction framework that formulates geometry estimation as a contact-conditioned generative inference problem. A conditional variational autoencoder (cVAE) learns structural terrain priors to resolve ambiguity in underconstrained regions, while a geometry-aware optimization enforces contact consistency and preserves discontinuities. The refined local patches are incrementally fused into a globally consistent 2.5D height map during locomotion. Real-world experiments across diverse outdoor terrains demonstrate improved geometric fidelity over contact-only baselines while maintaining online performance.**

I. Introduction

For legged robots that operate in challenging outdoor environments, estimating reliable terrain geometry is essential for safe and efficient exploration [1]. Accurate reconstruction of the ground surface directly affects downstream tasks. Most existing approaches construct terrain geometry using exteroceptive sensing, such as LiDAR, cameras and vision sensors [2]–[5]. However, exteroceptive sensing can be degraded under adverse illumination, motion blur, or sensor failures in real-world environments, introducing significant noise into geometric measurements [6], [7]. When such degradation occurs, observations become unavailable, leading to incomplete or noisy terrain estimates. This motivates the problem of reconstructing terrain geometry without relying on exteroceptive sensing.

To alleviate the reliance on exteroceptive sensing, proprioceptive sensing can serve as a natural modality for physical interaction with the terrain. Foot–terrain contact provides direct physical measurements and remains available even when visual sensing fails. However, contact observations are available only at discrete foothold locations during locomotion, leaving large regions of the terrain unobserved between contacts. As a result, the underlying surface cannot be reliably inferred from sparse contacts alone. To address this issue, classical interpolation or smoothness-based regularizations can fill such gaps, but they inherently impose a bias toward locally smooth surfaces. Real-world terrains, however, often contain discontinuities, sharp elevation changes, and complex clutter. Therefore, reconstruction from sparse contacts cannot be addressed by estimating a single deterministic surface. Instead, it requires reasoning about multiple plausible terrain geometries consistent with the observed contacts.

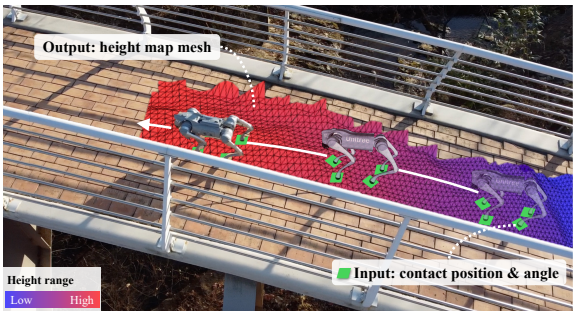

Fig. 1. `FootRecon` performs global terrain reconstruction given only sparse foot contacts during quadrupedal locomotion.

This ambiguity can be addressed by introducing prior knowledge about terrain structure beyond sparse contact constraints. This perspective is consistent with biological locomotion, where humans integrate prior experience with sparse tactile and inclination cues to infer unseen surface structure. Inspired by this, we incorporate a generative prior that captures terrain geometry patterns during training. The learned prior enables plausible completion of local terrain patches conditioned solely on sparse contact measurements.

Building upon this generative geometric prior, we propose `FootRecon`, a framework that models terrain reconstruction as a contact-conditioned generation of local terrain given contact observations. Sparse contacts serve as geometric constraints, while terrain structure is inferred through a contact-conditioned latent generative model. To learn the terrain structure, we employ a cVAE [8] architecture. By leveraging this prior, geometrically underconstrained regions can be efficiently generated conditioned on sparse contact cues. To ensure global consistency, we introduce a geometry-aware refinement. As a result, `FootRecon` incrementally accumulates contact observations to construct a globally consistent terrain structure (see Fig. 1).

We demonstrate improved geometric fidelity over contact-only baselines through extensive real-world experiments across diverse outdoor terrains. Both qualitative and quantitative results show improved fidelity over baselines. The contributions of this paper are summarized as follows:

- We formulate proprioceptive-only terrain reconstruction as contact-conditioned inference, enabling geometry es-

[1]Yujin Park, [1]Seungoh Han, [1]Chanwoo Lee, [1]Hyungyu Park, and [1†]Kyungdon Joo are with the Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. {yujinpark, sohan, dlcksdn1201, hyungyu, kyungdon}@unist.ac.kr
[2]Taebok Lee is with the School of Electrical Engineering, Kookmin University, Seoul, South Korea. {plkj3078}@kookmin.ac.kr

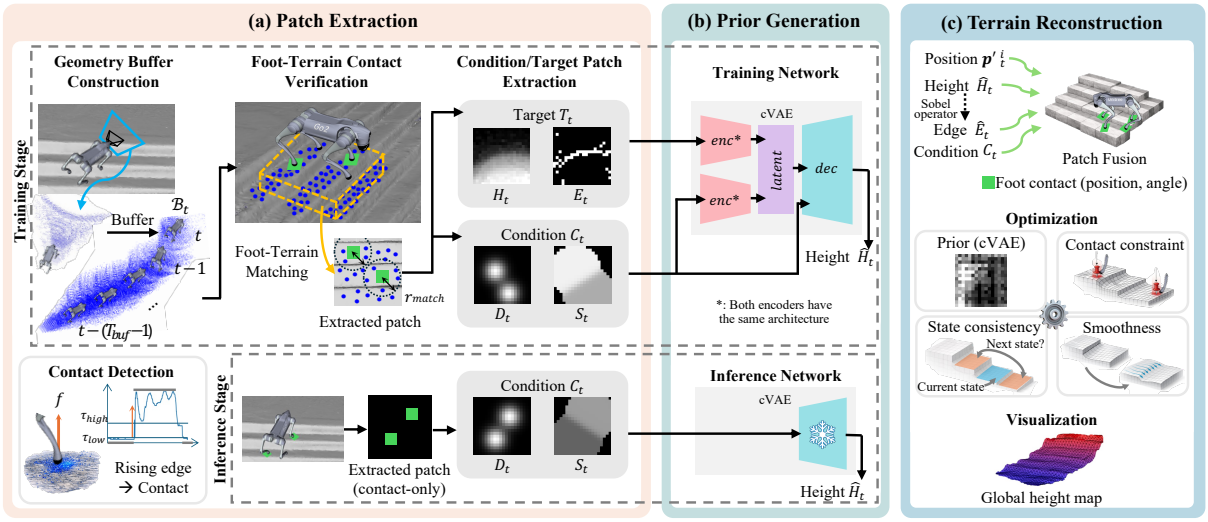

Fig. 2. **Overview of the proposed `FootRecon` framework.**

timation without relying on exteroceptive sensing.

- We introduce a contact-conditioned generative prior based on a cVAE to resolve the geometric ambiguity arising from sparse foot–terrain contacts.
- We propose a geometry-aware refinement framework that enforces physical contact consistency while preserving terrain discontinuities in a fused height map.

II. Method

We propose `FootRecon`, a framework for proprioceptive terrain reconstruction during quadrupedal locomotion. The proposed method constructs globally consistent terrain geometry from sparse foot–terrain contacts by integrating learned local terrain priors with geometry-aware optimization. In the following, we introduce the contact-conditioned terrain reconstruction problem in quadrupedal robots. We then present our pipeline composed of three stages, as illustrated in Fig. 2.

A. Problem Definition

At time t, a set of contact observations is defined as

$$\mathcal{O}_t = \left\{ (\mathbf{p}_i^t, \boldsymbol{a}_i^t, f_i^t) \mid i \in \{1,2,3,4\} \right\}, \tag{1}$$

where $\mathbf{p}_i^t$ denotes the 3D contact position in the world frame, $\boldsymbol{a}_i^t$ is a proprioceptive angle extracted from the orientation of the foot–terrain contact, and f_i^t denotes the vertical contact force. Since contacts are observed only at sparse locations, multiple height maps can satisfy identical contact constraints. Thus, reconstructing the terrain surface directly from sparse contacts remains an ill-posed problem.

To address this underconstrained nature, we formulate terrain reconstruction as a contact-conditioned generation problem. Instead of estimating a single deterministic surface, we model a distribution over plausible local geometries consistent with the observed contacts and incrementally integrate them into a globally consistent terrain map.

B. Patch Extraction

We define two types of spatial patches for contact-conditioned terrain reconstruction: a target patch derived from terrain geometry and a condition patch constructed from contact observations. The target patch provides supervision for terrain geometry, while the condition patch encodes sparse proprioceptive cues obtained from foot contacts. Target patch extraction is used only during training, whereas condition patch extraction is applied during both training and inference, as shown in Fig. 2(a).

Geometry Buffer Construction. To maintain sufficient terrain geometry for patch extraction, we construct a geometry buffer $\mathcal{B}_t$ that accumulates 3D terrain observations within a fixed temporal window T_{buf}. The 3D points are aggregated into the geometry buffer $\mathcal{B}_t$. Specifically, only the most recent T_{buf} frames are retained in the buffer $\mathcal{B}_t$ to improve computational efficiency. To obtain reliable contact observations, we detect foot–terrain contact events based on proprioceptive force measurements f_i^t during locomotion.

Foot–Terrain Contact Verification. Since detected contacts are subject to noise, false positives may occur. For each detected contact location $\mathbf{p}_i^t$, we check whether 3D points of the buffer $\mathcal{B}_t$ exist within a radius r_{match}. We use this finite matching radius to tolerate positional uncertainty.

Condition and Target Patch Extraction. Based on the verified contacts, we extract a pair of 2D patches that serve as the target height map and the corresponding contact condition for prior generation. For the target patch, we define the target patch set $T_t = \{H_t, E_t\}$, where H_t is a local height patch representing the terrain geometry and E_t is the corresponding edge map. The height patch H_t is constructed from the geometry buffer $\mathcal{B}_t$.

Next, we define the condition patch set $C_t = \{D_t, S_t\}$ that encodes cues for proprioceptive conditioning. The distribution patch D_t models the spatially distributed influence of each verified contact using a Gaussian kernel. Contact location alone is insufficient to characterize local geometry,

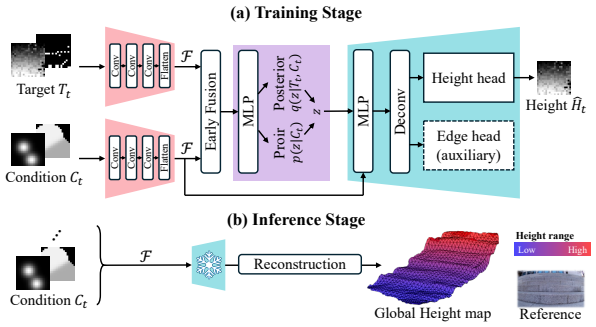

Fig. 3. **cVAE-based training and inference architecture of `FootRecon`.**

Fig. 4. **Training environments.**

we therefore construct a slope patch S_t that encodes the contact angle $\boldsymbol{a}_i^t$ as an additional conditioning variable.

T_t provides supervision for the terrain reconstruction network, while C_t serves as proprioceptive input.

C. Contact-Conditioned Prior Generation

We formulate terrain reconstruction as a contact-conditioned generative problem given the spatially aligned target patch set T_t condition patch set C_t as depicted in Fig. 2(b). We adopt a cVAE-based framework that models the distribution of the target patch set T_t conditioned on the contact patch set C_t. During inference, the trained model directly generates the predicted height patch $\hat{H}_t$ using only C_t, without requiring exteroceptive sensing.

Contact-Conditioned cVAE. Due to the underconstrained nature of sparse inputs $\mathcal{O}_t$, multiple plausible terrain geometries may correspond to the same contact condition. To capture this ambiguity, we adopt a cVAE that learns the distribution of H_t given C_t. The architecture follows the basic cVAE network (see Fig. 3).

Objective Function. The overall training objective combines height reconstruction, edge supervision, and latent regularization as:

$$\mathcal{L} = \lambda_{height}\,\mathcal{L}_{height} + \lambda_{edge}\,\mathcal{L}_{edge} + \lambda_{\mathrm{KL}}\,\mathcal{L}_{\mathrm{KL}}. \quad (2)$$

We supervise the predicted height map using a Huber loss. The predicted edge map is trained using a binary cross-entropy loss. We regularize the latent space by minimizing the KL divergence between the posterior conditioned on T_t and C_t and the contact-conditioned prior derived from C_t.

D. Terrain Reconstruction

To combine learned geometric priors with observed contacts, we refine a local height patch $\tilde{H}_t$ using energy-based optimization and incrementally fuse it into a persistent global height map, as illustrated in Fig. 2(c). The contact-conditioned estimate from the learned network provides a plausible terrain completion, but it does not explicitly enforce physical consistency with contact observations.

We define a set of energy terms for optimization. The contact energy $\mathcal{E}_{contact}$ enforces consistency between the reconstructed height and measured contact heights by penalizing their squared differences. To maintain temporal consistency and reduce drift, the state energy $\mathcal{E}_{state}$ penalizes deviations from the initialized height patch and enforces agreement between the estimated gradients and the stored slope state. The prior energy $\mathcal{E}_{prior}$ constrains the solution to remain close to the predicted prior, ensuring plausible geometry in underconstrained regions. The edge-aware regularization term $\mathcal{E}_{edge}$ suppresses noise while preserving structural discontinuities by reducing smoothing near predicted edges, allowing sharp features to be maintained.

With these components, the energy function is defined as:

$$\begin{aligned}\mathcal{E}(\tilde{H}_t) = \beta_{contact}\mathcal{E}_{contact} + \beta_{state}\mathcal{E}_{state} \\ + \beta_{prior}\mathcal{E}_{prior} + \beta_{edge}\mathcal{E}_{edge}.\end{aligned} \quad (3)$$

The refined patch is obtained by minimizing the energy. After optimization, the refined patch are fused into the global map via bilinear interpolation.

III. Experiment

A. Experimental Setup

Implementation Details. All experiments are conducted on a workstation equipped with an Intel i7-13700K CPU and a single NVIDIA RTX 4090 GPU. Datasets are collected using a Unitree Go2 platform. For training, the robot collects RGB images captured from its onboard monocular camera, and we estimate metric depth maps [9] for geometric supervision. Initial camera poses are obtained from the robot's state estimator and refined using RGB-D ORB-SLAM3 [10].

Terrain Dataset. The training environments (Fig. 4) include *Monticule*, *Mild Slope Bridge*, and *Tight Staircase*. The test environments (Fig. 5) comprise *Wide Staircase*, *Coir Mat*, *Bamboo Forest*, *Steep Slope*, and *Cave*. All test environments are excluded from training to evaluate generalization to unseen structures.

Contact-only Reconstruction Baselines. We compare against three contact-only baselines that reconstruct a 2.5D terrain height map from sparse contact points.

(1) Gaussian Kernel Regression. Patch heights are reconstructed via Gaussian kernel regression [11], [12], where neighboring contact heights are averaged within a truncation radius with bandwidth-controlled smoothing.

(2) Moving Least Squares (MLS) Plane Fitting. MLS plane fitting computes a locally weighted first-order surface to adjacent contact points within a fixed radius [13].

(3) TSDF-based Implicit Fusion. We maintain a sparse TSDF volume in the world frame and integrate each contact by updating voxels within a truncation band using weighted averages of normalized signed distances.

Evaluation. Reconstruction performance is evaluated at both global and local scales using complementary height map and patch-level metrics. For the global map evaluation, reconstructed meshes are compared against a reference height

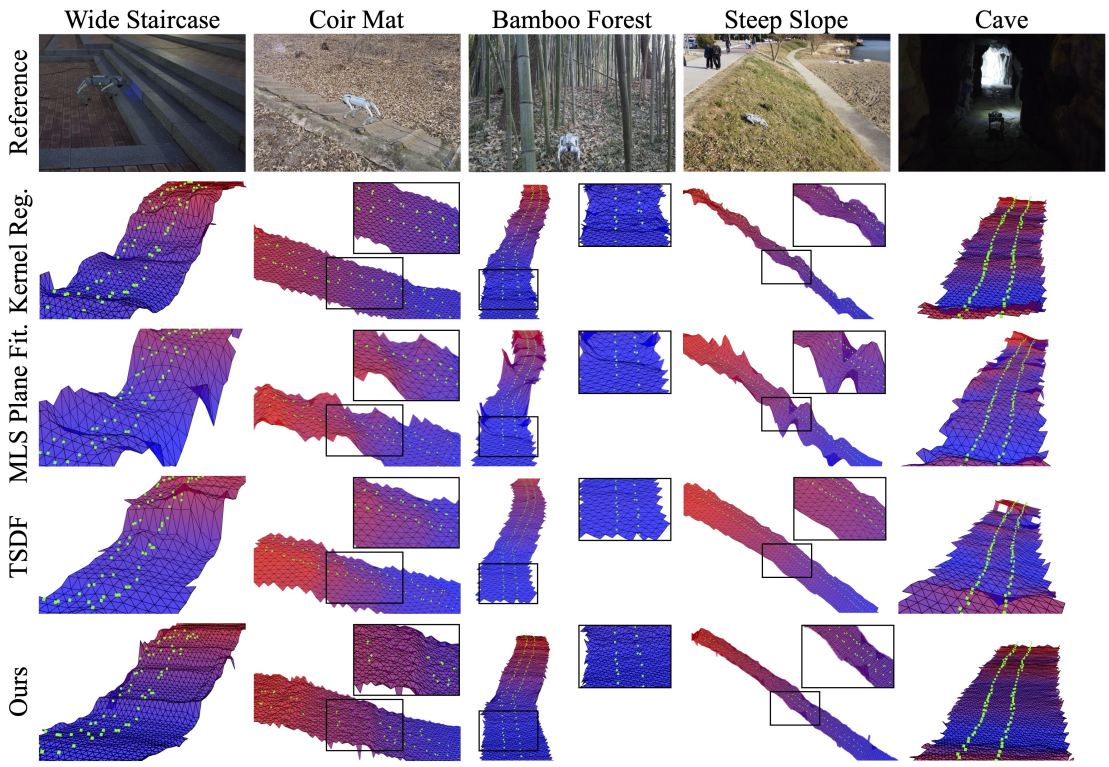

Fig. 5. **Qualitative results across unseen terrains.** All meshes are reconstructed using identical contact-conditioned inputs.

TABLE I

QUANTITATIVE EVALUATION ACROSS DIVERSE TERRAINS.

Terrain	Method	$RMSE_{global}\downarrow$	$MAE_{patch}\downarrow$	$RMSE_{patch}\downarrow$
Wide Staircase	Kernel Reg.	0.1428	0.5503	0.5579
	MLS Plane Fit.	0.2200	0.5903	0.6096
	TSDF	0.2696	0.4642	0.4819
	Ours	**0.0917**	**0.1228**	**0.1384**
Coir Mat	Kernel Reg.	0.1009	0.7669	0.7690
	MLS Plane Fit.	0.1155	0.7510	0.7619
	TSDF	0.2018	0.4764	0.4857
	Ours	**0.0772**	**0.1080**	**0.1309**
Bamboo Forest	Kernel Reg.	0.1362	1.1365	1.1397
	MLS Plane Fit.	0.1797	1.1177	1.1234
	TSDF	0.2637	1.0213	1.0264
	Ours	**0.0560**	**0.1078**	**0.1269**
Steep Slope	Kernel Reg.	0.1504	2.3107	2.3214
	MLS Plane Fit.	0.2118	2.1983	2.2126
	TSDF	0.3373	2.0107	2.0164
	Ours	**0.0892**	**0.2640**	**0.3030**
Cave	Kernel Reg.	0.2677	0.0883	0.0946
	MLS Plane Fit.	**0.2555**	0.0853	0.1008
	TSDF	0.3039	**0.0541**	**0.0641**
	Ours	0.2657	0.0748	0.0819

Bold and underline denote the best and second-best results, respectively.

map derived from the onboard LiDAR of the robot. Global accuracy is measured using the root-mean-square height error, $RMSE_{global}$. At the local scale, predicted patches are compared with corresponding reference patches. MAE_{patch} and $RMSE_{patch}$ are computed over valid cells and averaged across all patches in each sequence.

IV. RESULT

Fig. 5 shows the qualitative comparisons on multiple baselines given proprioceptive sensing. Classical approaches exhibit strong biases toward smoothness and planarity; regression-based methods attenuate discontinuities and reduces gradient magnitude, while TSDF introduces voxel-induced rounding near boundaries. On the other hand, our method preserves sharp discontinuous regions and maintains consistent slope-like geometry. Across irregular and rough terrains, the proposed pipeline retains fine structure details.

Table I summarizes the quantitative evaluation across collected environments. The proposed method achieves the highest local patch-level accuracy while maintaining competitive or superior global accuracy in most environments. In *Cave*, the terrain is predominantly flat with weak slope variations. Therefore, naively interpolating nearby area can produce relatively less height error, resulting in a reduced performance gap in the evaluated metrics. As a result, these quantitative results demonstrate that the proposed method handles various real-world terrain attributes.

V. CONCLUSION

`FootRecon` formulates proprioceptive terrain reconstruction a contact-conditioned generative inference problem to address the ill-posed nature of sparse observations. By coupling a learned prior with geometry-aware optimization, it resolves structural ambiguity while preserving discontinuities and global consistency. While our evaluation relies on height-map comparisons against exteroceptive references, the physically contact-consistent surface may not always coincide with the visually observed geometry. Developing evaluation criteria that better reflect functional locomotion performance remains an important direction for future work.

REFERENCES

[1] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, and M. Hutter, "Elevation mapping for locomotion and navigation using gpu," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022.

[2] R. Yu, Q. Wang, H. Li, Z. Jun, Z. Wang, J. Wu, and Q. Zhu, "Start: Traversing sparse footholds with terrain reconstruction," *IEEE Robotics and Automation Letters*, vol. 11, no. 2, pp. 2194–2201, 2025.

[3] B. Yang, Q. Zhang, R. Geng, L. Wang, and M. Liu, "Real-time neural dense elevation mapping for urban terrain with uncertainty estimations," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 696–703, 2022.

[4] G. Chen and L. Hong, "Research on environment perception system of quadruped robots based on lidar and vision," *Drones*, vol. 7, no. 5, p. 329, 2023.

[5] P. Fankhauser, M. Bloesch, and M. Hutter, "Probabilistic terrain mapping for mobile robots with uncertain localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3019–3026, 2018.

[6] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science robotics*, vol. 7, no. 62, 2022.

[7] Y. Cheng, H. Liu, G. Pan, H. Liu, and L. Ye, "Quadruped robot traversing 3d complex environments with limited perception," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.

[8] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[9] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[10] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, p. 1874–1890, Dec. 2021.

[11] T. Homberger, L. Wellhausen, P. Fankhauser, and M. Hutter, "Support surface estimation for legged robots," in *2019 International Conference on Robotics and Automation*, 2019, pp. 8470–8476.

[12] A. Li, C. Yang, J. Frey, J. Lee, C. Cadena, and M. Hutter, "Seeing through the grass: Semantic pointcloud filter for support surface learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7687–7694, 2023.

[13] S.-L. Liu, H.-X. Guo, H. Pan, P.-S. Wang, X. Tong, and Y. Liu, "Deep implicit moving least-squares functions for 3d reconstruction," 2021.