FREEREG: IMAGE-TO-POINT CLOUD REGISTRATION LEVERAGING PRETRAINED DIFFUSION MODELS AND MONOCULAR DEPTH ESTIMATORS

Haiping Wang[1,*] **Yuan Liu**[2,*] **Bing Wang**[3] **Yujing Sun**[2] **Zhen Dong**[1,†]
Wenping Wang[4] **Bisheng Yang**[1,†]
[1]LISMARS & Hubei Luojia Laboratory, Wuhan University [2]The university of Hong Kong
[3]The Hong Kong Polytechnic University [4]Texas A&M University
{hpwang,dongzhenwhu,bshyang}@whu.edu.cn yuanly@connect.hku.hk
bingwang@polyu.edu.hk yjsun@cs.hku.hk wenping@tamu.edu

ABSTRACT

Matching cross-modality features between images and point clouds is a fundamental problem for image-to-point cloud registration. However, due to the modality difference between images and points, it is difficult to learn robust and discriminative cross-modality features by existing metric learning methods for feature matching. Instead of applying metric learning on cross-modality data, we propose to unify the modality between images and point clouds by pretrained large-scale models first, and then establish robust correspondence within the same modality. We show that the intermediate features, called diffusion features, extracted by depth-to-image diffusion models are semantically consistent between images and point clouds, which enables the building of coarse but robust cross-modality correspondences. We further extract geometric features on depth maps produced by the monocular depth estimator. By matching such geometric features, we significantly improve the accuracy of the coarse correspondences produced by diffusion features. Extensive experiments demonstrate that without any training on the I2P registration task, direct utilization of both features produces accurate image-to-point cloud registration. On three public indoor and outdoor benchmarks, the proposed method averagely achieves a 20.6% improvement in Inlier Ratio, a $3.0\times$ higher Inlier Number, and a 48.6% improvement in Registration Recall than existing state-of-the-arts. The code and additional results are available at `https://whu-usi3dv.github.io/FreeReg/`.

1 INTRODUCTION

Image-to-point cloud (I2P) registration requires estimating pixel-to-point correspondences between images and point clouds to estimate the SE(3) pose of the image relative to the point cloud. It is a prerequisite for many tasks such as Simultaneous Localization and Mapping (Zhu et al., 2022), 3D reconstruction (Dong et al., 2020), segmentation (Guo et al., 2020), and visual localization (Sarlin et al., 2023).

To establish pixel-to-point correspondences, we have to match features between images and point clouds. However, it is difficult to learn robust cross-modality features for images and point clouds. Most existing methods (Feng et al., 2019; Wang et al., 2021; Pham et al., 2020; Jiang & Saripalli, 2022; Li et al., 2023) resort to metric learning methods like contrastive loss, triplet loss or InfoCE loss to force the alignment between the 2D and 3D features of the same object. However, due to the inherent data disparities that images capture appearances while point clouds represent structures, directly aligning cross-modal data inevitably leads to poor convergence. Consequently, cross-modality metric learning suffers from poor feature robustness (Wang et al., 2021) and limited generalization ability (Li et al., 2023).

[*]Equal contribution.
[†]Corresponding Authors.

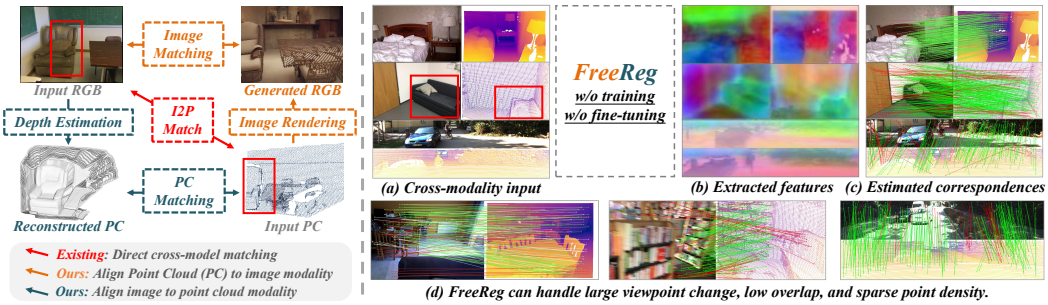

Figure 1: ***Left***: FreeReg unifies the modalities of images and point clouds, which enables mono-modality matching to build cross-modality correspondences. ***Right***: FreeReg does not require any training on the I2P task and is able to register RGB images to point clouds in both indoor and outdoor scenes, even for challenging cases with small overlaps, large viewpoint changes, and sparse point density.

In this paper, we propose a novel method, called *FreeReg*, to build robust cross-modality correspondences between images and point clouds with the help of recent large-scale diffusion models (Rombach et al., 2022; Zhang & Agrawala, 2023; Mou et al., 2023) and monocular depth estimators (Bhat et al., 2023; Yin et al., 2023). FreeReg avoids the difficult cross-modality metric learning and does even not require training on the I2P task. As shown in Fig. 1, the key idea is to unify the modality between images and point clouds by these large-scale pretrained models so FreeReg allows robust correspondence estimation within the same modality for cross-modality matching.

In order to convert point clouds to the image modality, a straightforward way is to project points onto an image plane to get a depth map and then convert the depth map to an image by a depth-to-image diffusion model ControlNet (Zhang & Agrawala, 2023). However, as shown in Fig. 2 (I), a depth map may correspond to multiple possible images so that the generated image from the point cloud would have a completely different appearance from the input image, which leads to incorrect matching results even with SoTA image matching methods (Sarlin et al., 2020; DeTone et al., 2018; Sun et al., 2021). To address this problem, we propose to match the semantic features between the generated images and the input image because the generated images show strong semantic consistency with the input image in spite of different appearances. Inspired by recent diffusion-based semantic correspondence estimation methods (Tang et al., 2023; Zhang et al., 2023), we utilize the intermediate feature maps in the depth-to-image ControlNet to match between depth maps and images. As shown in Fig. 2 (II), we visualize the diffusion features of the depth map and the RGB image. Then, we utilize the nearest neighbor (NN) matcher with mutual check (Wang et al., 2022a) to establish correspondences between them. We find that such semantic features show strong consistency even though they are extracted on depth maps and images separately, making it possible to build robust cross-modality correspondences. However, the semantic features are related to a large region of the image. Such a large receptive field leads to coarse-grained features and only sparse correspondences in feature matching.

We further improve the accuracy of our cross-modality correspondences with the help of the monocular depth estimators (Bhat et al., 2023). Recent progress in monocular depth estimators enables metric depth estimation on a single-view image. However, directly matching features between the point cloud and the estimated depth maps from the input image leads to poor performance as shown in Fig. 2 (III). The main reason is that the predicted depth maps are plausible but still contain large distortions in comparison with the input point cloud. The distortions prevent us from estimating robust correspondences. Though the global distortions result in noisy matches, the local geometry of the estimated depth maps still provides useful information to accurately localize keypoints and densely estimate fine-grained correspondences. Thus, we combine the local geometric features (Choy et al., 2019) extracted on the estimated depth maps with the semantic features extracted from diffusion models as the cross-modality features, which enable dense and accurate correspondence estimation between images and point clouds, as shown in Fig. 2 (IV).

In summary, FreeReg has the following characteristics. 1) FreeReg combines coarse-grained semantic features from diffusion models and fine-grained geometric features from depth maps for accurate cross-modality feature matching. 2) FreeReg does not require training on the I2P task,

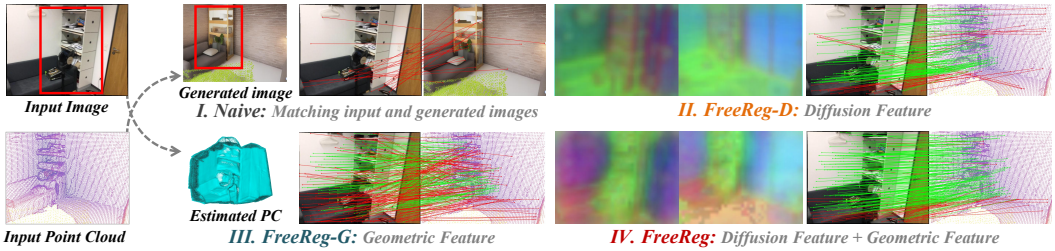

Figure 2: To unify the modalities of point clouds (PCs) and images, ***I***: a straightforward way is to generate RGB images from point clouds by depth-to-image diffusion models. However, the generated images usually have large appearance differences from the query images. ***II***: We find that the intermediate features of diffusion models show strong semantic consistency between RGB images and depth maps, resulting in *sparse but robust* correspondences. ***III***: We further convert RGB images to point clouds by a monocular depth estimator and extract geometric features to match between the input and the generated point clouds, yielding *dense but noisy* correspondences. ***IV***: We propose to fuse both types of features to build *dense and accurate* correspondences.

which avoids the unstable and notorious metric learning to align local features of point clouds and images. 3) FreeReg significantly outperforms existing fully-supervised cross-modality registration baselines (Pham et al., 2020; Li et al., 2023). Specifically, on the indoor 3DMatch and ScanNet datasets and the outdoor KITTI-DC dataset, FreeReg roughly achieves over 20% improvement in Inlier Ratio, a $3.0\times$ more Inlier Number, and a 48.6% improvement in Registration Recall.

2 RELATED WORK

Image-to-point cloud registration. In order to establish correspondences between images and point clouds for pose recovery, most existing methods (Li et al., 2015; Xing et al., 2018; Feng et al., 2019; Lai et al., 2021; Wang et al., 2021; Pham et al., 2020; Liu et al., 2020; Jiang & Saripalli, 2022; Li et al., 2023) rely on metric learning to align local features of images and point clouds (Feng et al., 2019; Pham et al., 2020; Lai et al., 2021; Jiang & Saripalli, 2022), or depth maps (Liu et al., 2020; Wang et al., 2021; Li et al., 2023). However, these methods often require cross-modal registration training data (Pham et al., 2020; Wang et al., 2021; Jiang & Saripalli, 2022; Li et al., 2023; Kim et al., 2023) and show limited generalization ability (Pham et al., 2020; Wang et al., 2021; Li et al., 2023; Ren et al., 2022; Yao et al., 2023) due to the difficulty in the cross-modality metric learning. In contrast, FreeReg does not require any training and fine-tuning on the I2P registration task and exhibits strong generalization ability to both indoor and outdoor scenes.

Some other methods directly solve image-to-point cloud registration as an optimization problem (David et al., 2004; Campbell et al., 2019; Arar et al., 2020; Wang et al., 2023a; Zhou et al., 2023), which regresses poses by progressively aligning keypoints (Li & Lee, 2021; Ren et al., 2022; Campbell et al., 2019), pole structures (Wang et al., 2022b), semantic boundaries (Liao et al., 2023), or cost volumes (Wang et al., 2023a) of RGB images and depth maps. However, these methods heavily rely on an accurate initial pose (Wang et al., 2021; Liao et al., 2023) to escape from local minima in optimizations. FreeReg does not require such a strictly accurate initialization because FreeReg matches features to build correspondences to handle large pose changes.

Diffusion feature extraction. Recently, a category of research (Ho et al., 2020; Song et al., 2020a;b; Karras et al., 2022; Song & Ermon, 2019; Dhariwal & Nichol, 2021; Liu et al., 2023), known as diffusion models, has demonstrated impressive generative capabilities. Based on that, with the advent of classifier-free guidence (Ho & Salimans, 2022) and billions of text-to-image training data (Schuhmann et al., 2022), a latent diffusion model, specifically stable diffusion (Rombach et al., 2022), has shown remarkable text-to-image generation capabilities. Building upon this, existing methods have demonstrated the exceptional performance of Stable Diffusion internal representations (Diffusion Feature) (Kwon et al., 2022; Tumanyan et al., 2023) in various domains such as segmentation (Amit et al., 2021; Baranchuk et al., 2021; Chen et al., 2022b; Jiang et al., 2018; Tan et al., 2022; Wolleb et al., 2022), detection (Chen et al., 2022a), depth estimation (Duan et al., 2023; Saxena et al., 2023b;a). These methods only extract diffusion features on RGB images utilizing Stable Diffsuion. Our method extracts diffusion features on RGB and depth maps based on recent finetuned diffusion

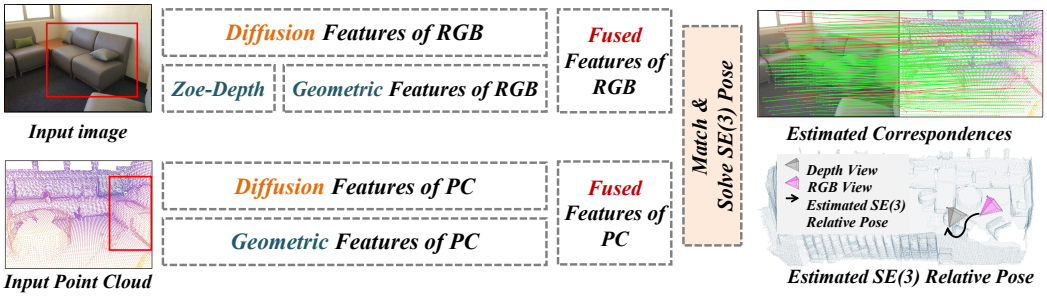

Figure 3: *FreeReg pipeline.* Given a point cloud (PC) and a partially overlapping RGB image, FreeReg extracts diffusion features and geometric features for the point cloud and the image. These two features are fused and matched to establish pixel-to-point correspondences, on which we compute the SE(3) relative pose between the image and the point cloud.

models ControlNet (Zhang & Agrawala, 2023) or T2IAdaptor (Mou et al., 2023), which efficiently leverage depth, semantic maps, and sketches to guide stable diffusion in image generation.

Diffusion feature for matching. Some recent works utilize diffusion features for representation learning (Kwon et al., 2022) and semantic matching (Luo et al., 2023; Tang et al., 2023; Hedlin et al., 2023; Zhang et al., 2023) among RGB images capturing objects across instances and categories. In comparison, our method shows the effectiveness of diffusion features in learning cross-modality features for image-to-point cloud registration.

Monocular depth estimator Monocular depth estimation inherently suffers from scale ambiguity (Chen et al., 2016; 2020; Xian et al., 2018; 2020). With more and more monocular depth training data (Guizilini et al., 2023; Antequera et al., 2020; Wilson et al., 2023), recent works (Bhat et al., 2021; 2022; Jun et al., 2022; Li et al., 2022; Yang et al., 2021; Yin et al., 2021; 2019; Yuan et al., 2022; Guizilini et al., 2023; Yin et al., 2023) learn scene priors to regress depth values in real metric space and show impressive results. We employ a SoTA metric depth estimator Zoe-Depth (Bhat et al., 2023) to recover point clouds in the same metrics corresponding to the RGB images.

3 METHOD

Let $I \in \mathbb{R}^{H \times W \times 3}$ be an RGB image and $P \in \mathbb{R}^{N \times 3}$ be a point cloud. We first project P to a depth map $D \in \mathbb{R}^{H' \times W'}$ on a camera pose, which is calculated from the depth or LiDAR sensor center and orientation. More details about this projection are given in the supplementary material. FreeReg aims to match the cross-modality features extracted on I and D to establish correspondences and solve the relative pose between them. The pipeline of FreeReg is illustrated in Fig. 3. Specifically, We extract diffusion features (Sec. 3.2) and geometric features (Sec. 3.3) for feature matching and then estimate the I2P transformation estimation from the matching results. We begin with a brief review of diffusion methods, which we utilize to extract cross-modality features.

3.1 PRELIMINARY: STABLE DIFFUSION AND CONTROLNET

The proposed cross-modality features are based on ControlNet (Zhang & Agrawala, 2023) (CN) so we briefly review the related details of ControlNet in this section. Diffusion models contain a forward process and a reverse process, both of which are Markov chains. The forward process gradually adds noise to the input image in many steps and finally results in pure structure-less noise. The corresponding reverse process gradually denoises the noise step-by-step to gradually recover the structure and generate the image. Stable Diffusion (Rombach et al., 2022) (SD) is a widely-used diffusion model mainly consisting of a UNet which takes noisy RGB images as input and predicts the noise. The original Diffusion model only allows text-to-image generation. Recent ControlNet (Zhang & Agrawala, 2023), as shown in Fig. 4 (b), adds an additional encoder to process depth maps and utilizes the extracted depth features to guide the reverse process of SD, enabling SD to generate images coherent to the input depth map from a pure Gaussian noise. In FreeReg, we utilize CN and SD to extract cross-modality features for feature matching.

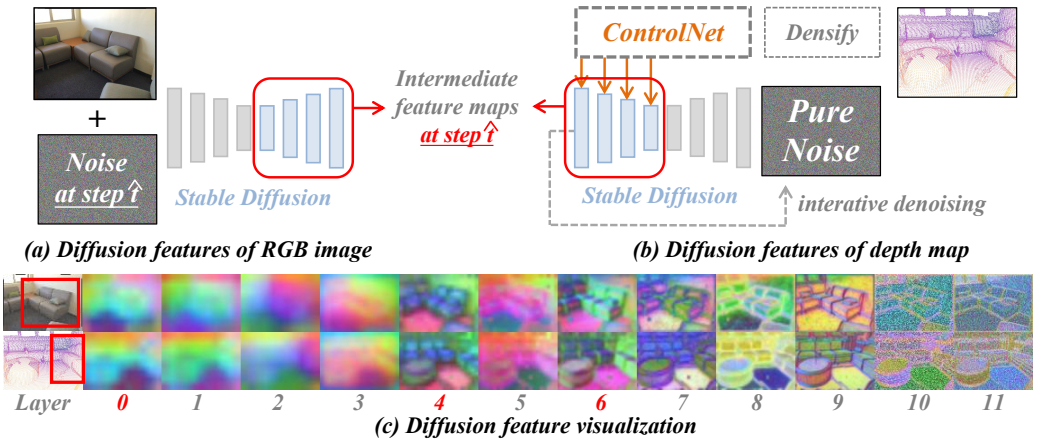

Figure 4: *Diffusion feature extraction* on (a) images and (b) depth maps. (c) Visualization of diffusion features.

3.2 Diffusion Features on Cross-Modality Data

Directly generating an image from the input depth map suffers from appearance inconsistency with the input image, which results in inaccurate feature matching. Instead of generating an explicit image, we resort to the intermediate feature maps of stable diffusion models for cross-modality feature matching. The overview is shown in Fig. 4.

RGB diffusion feature. As shown in Fig. 4(a), we perform the forward process of SD (Rombach et al., 2022) to add noise to the input RGB image, which results in a noisy image on a predefined step $\hat{t}$. The noisy image is fed to the UNet of the SD and the intermediate feature maps of the UNet decoder are used as the diffusion feature for the input RGB image.

Depth diffusion feature. Given the depth maps, we first densify them using traditional erosion and dilation operations (Ku et al., 2018). As shown in Fig. 4 (b), we propose to feed the depth map to a CN (Zhang & Agrawala, 2023) as a condition to guide the reverse process of SD. With such a condition, SD gradually denoise a pure Gaussian noise until the same predefined step $\hat{t}$ and then we use the feature maps in the SD UNet decoder as the depth diffusion features. An alternative way is to directly treat the depth map as an RGB image for diffusion feature extraction, which however leads to poor performance as shown in the supplementary material.

Layer selection. The remaining problem is about which layer to be used for feature extraction. Visualization of extracted diffusion features on RGB images and depth maps are given in Fig. 4(c). It can be observed that the features of early upsampling layers with layer index $l \leq 6$ show strong consistency between RGB and depth data. Features of later upsampling layers with an index larger than 6 show more fine-grained details like textures that no longer exhibit consistency. Therefore, we use features of early layers 0,4,6 as our diffusion features. To reduce the feature dimension on each layer, we apply a Principal Component Analysis (PCA) to reduce the feature dimension to 128. The resulting diffusion features of RGB image I and depth map D are F_d^I and F_d^D respectively, both of which are obtained by concatenating the features from different layers and L2 normalized.

3.3 Geometric Features on Cross-Modality Data

The above diffusion feature is extracted from a large region on the image, which struggles to capture fine-grained local details and estimates only sparse correspondences as shown in Fig. 5 (b/e). To improve the accuracy of these correspondences, we introduce a so-called geometric feature, leveraging the monocular depth estimator Zoe-Depth (Bhat et al., 2023).

Specifically, we utilize Zoe-Depth to generate per-pixel depth D^Z for the input RGB image I and recover a point cloud from the generated depth map. Then, we employ a pre-trained point cloud feature extractor FCGF (Choy et al., 2019) to extract per-point features, which serve as the geometric features of their corresponding pixels in the image I. We construct geometric features for pixels of the depth map D in the same way. As illustrated in Fig. 5 (c/f), solely matching geometric features produces many outlier correspondences due to large distortion in the single-view depth

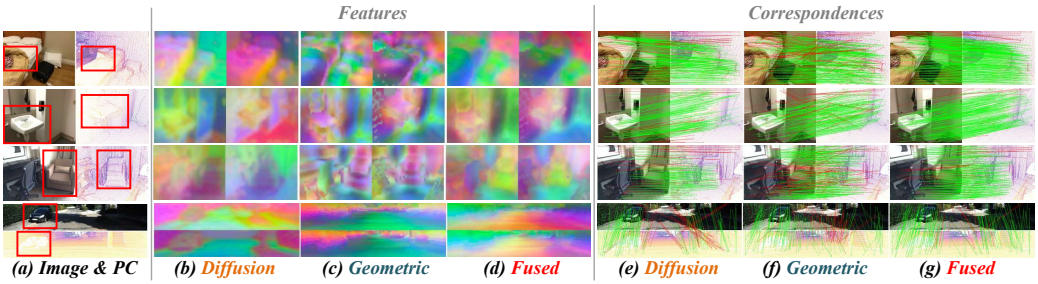

Figure 5: *Visualization of features and estimated correspondences.* (a) Input images and point clouds. (b), (c), and (d) show the visualization of diffusion, geometric, and fused feature maps respectively. (e), (f), and (g) show the pixel-to-point correspondences estimated by the nearest neighbor (NN) matcher using diffusion, geometric, and fused features respectively. Diffusion features estimate reliable but sparse correspondences. Geometric features yield dense matches but with more outliers. Fused features strike a balance between accuracy and preserving fine-grained details, resulting in accurate and dense matches.

estimation. However, these geometric features provide local descriptions of the geometry, which are more localized and enable more accurate correspondence in cooperation with the diffusion features.

3.4 Fuse both features for I2P registration

Fuse features. In this section, we propose to fuse the two types of features to enable accurate correspondence estimation, as shown in Fig. 5. Note that we uniformly sample a dense grid of keypoints on both the depth map and the image. Then, we extract the above diffusion features and geometric features on the keypoints. Both features are normalized by their L2 norm before the fusion. Specifically, we follow (Zhang et al., 2023) to fuse two kinds of features on each keypoint in I or D by

$$F = [wF_d, (1-w)F_g], \tag{1}$$

w is a fusion weight, $[\cdot,\cdot]$ means concatenating on feature dimension, and F is the resulting FreeReg feature.

Pixel-to-point correspondences. Given two sets of fused features F^I on RGB image I and F^D on depth map D, we conduct nearest neighborhood (NN) matching with a mutual nearest check (Wang et al., 2022a) to find a set of putative correspondences. Note that the pixel from the depth map D in each match corresponds to a 3D point in point cloud P.

Image-to-point cloud registration. To solve SE(3) poses of RGB image I relative to P. A typical approach is to conduct the Perspective-n-Point (PnP) algorithm (Lepetit et al., 2009) on the established pixel-to-point correspondences. However, we have estimated a depth map corresponding to RGB using Zoe-Depth (Bhat et al., 2023). Thus, we can convert the pixel-to-point correspondences to 3D point-to-point correspondences, and estimate the SE(3) relative pose using the Kabsch algorithm (Kabsch, 1976). In the supplementary material, we empirically show that using the PnP algorithm leads to a more accurate pose estimation but fails in many cases, while the Kabsch algorithm works in more cases but the estimated transformations exhibit larger errors.

4 Experiments

4.1 Experimental protocol

Datasets. We evaluate the proposed method on three widely used datasets: (1) The *3DMatch* (Zeng et al., 2017) testset comprises RGB images and point clouds (called *I2P pairs*) from 8 indoor scenes. The point clouds used here are collected by an Asus Xtion depth sensor. We manually exclude the I2P pairs with very small overlaps resulting in 1210 I2P pairs with over 30% overlaps. (2) The *ScanNet* (Dai et al., 2017) testset consists of 4,660 I2P pairs from 31 indoor scenes with more than 30% overlap. To further increase the difficulty, we downsampled the input point clouds using a voxel size of 3cm, which leads to highly sparse point clouds. (3) The *Kitti-DC* (Uhrig et al., 2017) testset has 342 I2P pairs from 4 selected outdoor scenes. The sparse point clouds come from a 64-line LiDAR scan. The distance between each I2P pair is less than 10 meters.

Table 1: *Cross-modality registration performance of different methods.* "InvCP." means Inverse Camera Projection (Li & Lee, 2021).

Method		LCD	SG	DeepI2P	CN+SG	I2P-Matr	FreeReg-D	FreeReg-G	FreeReg	FreeReg
SE(3) Solver		*PnP*	*PnP*	*InvCP.*	*PnP*	*PnP*	*PnP*	*Kabsch*	*PnP*	*Kabsch*
3DMatch	*FMR(%)*	40.1	50.3	/	64.7	90.6	91.9	90.7	**94.6**	**94.6**
	IR(%)	35.1	11.1	/	18.4	24.9	39.6	31.4	**47.0**	**47.0**
	IN(#)	4.3	3.1	/	10.9	49.0	60.8	49.4	**82.8**	**82.8**
	RR(%)	/	1.8	/	6.5	28.2	33.2	50.4	40.0	**63.8**
ScanNet	*FMR(%)*	55.1	53.2	/	64.1	87.0	95.3	96.4	**98.5**	**98.5**
	IR(%)	30.7	13.4	/	18.3	14.3	45.7	40.5	**56.8**	**56.8**
	IN(#)	5.0	4.7	/	9.1	24.8	61.5	84.5	**114.4**	**114.4**
	RR(%)	/	1.2	/	5.5	8.5	42.3	69.4	57.6	**78.0**
Kitti-DC	*FMR(%)*	/	73.4	/	94.2	/	**100.0**	94.4	99.7	99.7
	IR(%)	/	18.1	/	34.4	/	**59.4**	41.2	58.3	58.3
	IN(#)	/	12.6	/	51.1	/	103.6	93.6	**132.9**	**132.9**
	RR(%)	/	8.2	20.9	20.4	/	68.1	43.3	**70.5**	67.5

Metrics. Following (Choy et al., 2019; Wang et al., 2023c;b), we adopt four evaluation metrics: (1) *Feature Matching Recall (FMR)* is the fraction of I2P pairs with more than 5% correct estimated correspondences. A correspondence is regarded as correctly matched if its ground truth 3D distance is smaller than τ_c. τ_c is set to 0.3m for 3DMatch/ScanNet and 3m for Kitti-DC. (2) *Inlier Ratio (IR)* is the average correct correspondence proportions among all I2P pairs. (3) *Inlier Number (IN)* is the average number of correct correspondences on each I2P pair. and (4) *Registration Recall (RR)* is the percentage of correctly-aligned I2P pairs with rotation and translation errors less than τ_R and τ_t respectively. (τ_R, τ_t) is set to ($20°$, 0.5m) for 3DMatch/ScanNet and ($10°$, 3m) for Kitti-DC. We provide additional results under different threshold conditions in the supplementary material.

Baselines. We compare FreeReg with fully supervised registration baselines. The image registration method SuperGlue (SG) (Sarlin et al., 2020) is modified to match RGB images and point clouds. LCD (Pham et al., 2020) learns to construct I2P cross-modality descriptors utilizing metric learning. DeepI2P (Li & Lee, 2021) resolve I2P registration by optimizing an accurate initial pose. We implement a cross-modality feature extraction method I2P-Matr following a concurrent work 2D3D-Matr (Li et al., 2023), where the official codes are not released yet. Meanwhile, we compare FreeReg with P2-Net (Wang et al., 2021) and 2D3D-Matr (Li et al., 2023) under their experimental protocol (Li et al., 2023) in the supplementary material, where FreeReg also achieves the best registration performance. We also adopt a baseline as mentioned in Fig. 2 (I) that first utilizes ControlNet (Zhang & Agrawala, 2023) to generate an RGB image from the target point cloud and then conducts SuperGlue (Sarlin et al., 2020) to match the input and the generated image (CN+SG). For our method, we report the results using only the diffusion feature (FreeReg-D, i.e. $w = 1$), only the geometric feature (FreeReg-G, i.e. $w = 0$), and the fused feature (FreeReg, i.e. $w = 0.5$ by default) for matching. More implementation details and analysis are provided in the supplementary material.

4.2 RESULTS ON THREE BENCHMARKS

The quantitative results of FreeReg and baselines on the three cross-modality registration benchmarks are given in Table 1. Some quantitative results are shown in Fig. 6.

Correspondence quality is reflected by *FMR*, *IR*, and *IN*. For LCD and I2P-Matr, utilizing a metric learning method to directly align cross-modality features leads to poor performance. CN+SG suffers from the appearance difference between generated images and the input images and thus fails to build reliable correspondences. For FreeReg, using solely diffusion features (FreeReg-D) or geometric features (FreeReg-G) can already yield results superior to the baselines. Utilizing both features, FreeReg achieves the best correspondence quality and outperforms baselines by a large margin with 54.0% in *FMR*, 20.6% in *IR*, and a $3.0\times$ higher *IN*. Note that, unlike baseline methods, FreeReg does not even train on the I2P task.

Registration quality is indicated by *RR*. Benefited by the high-quality correspondences, FreeReg significantly outperforms the baseline methods by a 48.6% RR and FreeReg-D/G by a $22.9\%/16.4\%$ RR. Moreover, FreeReg utilizing Kabsch significantly surpasses PnP on indoor 3DMatch/ScanNet but is 3% lower than PnP on the outdoor Kitti-DC. The main reason is that Zoe-Depth performs better on these two indoor datasets with an average 0.27m error but worse on the KITTI with an

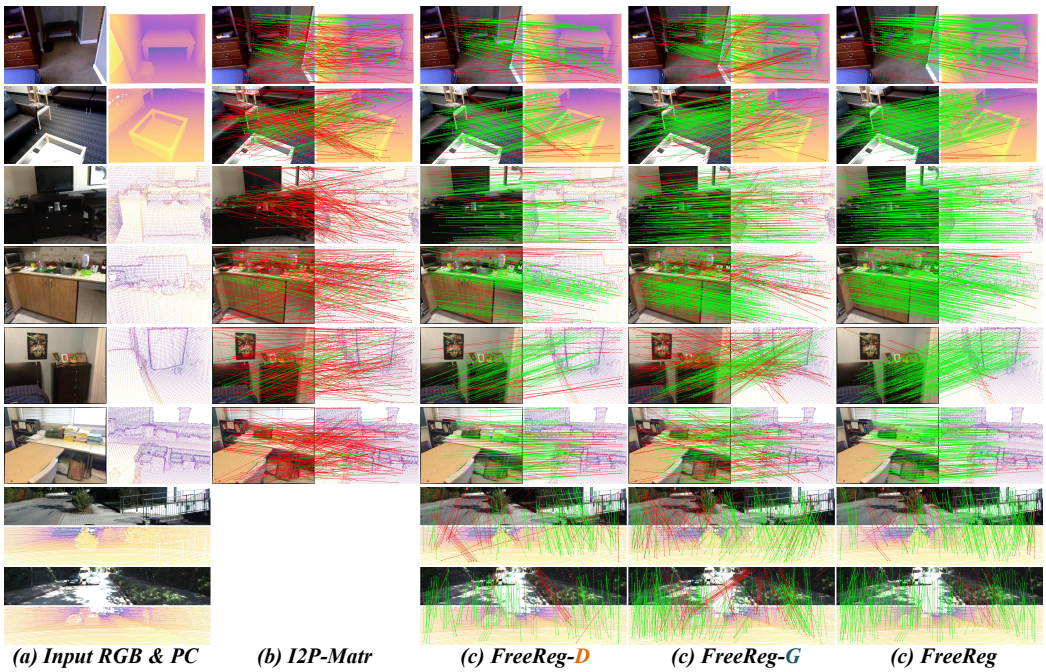

Figure 6: *Visualization of correspondences.* (a) Input RGB images and point clouds for registration. (b) Estimated correspondences from I2P-Matr. (c / d / e) Estimated correspondences from FreeReg-D / FreeReg-G / FreeReg.

average 3.4m error. In the supplementary material, we further provide more analysis and find that PnP achieves more accurate results while Kabsch provides plausible results in more cases.

4.3 More analysis

We conduct comprehensive ablation experiments on FreeReg in the following. More analysis about how to fine-tune FreeReg for performance gains, runtime analysis and acceleration strategies of FreeReg, FreeReg performance under different thresholds, more qualitative comparisons and ablation studies are provided in the supplementary material.

4.3.1 Ablating diffusion feature extraction

In this section, we evaluate on a validation scene "bundlefusion-office0" (BFO) which is not included in the testset to tune hyperparameters in diffusion feature layer selection and diffusion step $\hat{t}$ selection. Subsequently, we report their performances on the 3DMatch dataset.

Diffusion layer selection. In table 2 (a-i), we report the size of output feature maps of 8 layers in the UNet of Stable Diffusion. The feature map size is divided into three levels, i.e. small group (8×11, layers 0-2), medium group (16×22, layers 3-5), and large group (32×44, layers 6-8). We select the layers with the best registration performance and reasonable matching quality from each level on BFO, specifically layers 0, 4, and 6, to construct our Diffusion Features. Then, in table 2 (j-m), we ablate the layer selection in constructing diffusion features. It can be seen that concatenating features of 0,4,6 layers significantly improves the correspondence quality and registration performance. The results from 3DMatch further validate the effectiveness of our choice. More ablation studies on the diffusion layer selection are provided in the supplementary material.

Diffusion step selection. In Table 3, we aim to determine the diffusion step $\hat{t}$. The experimental results demonstrate that the Diffusion Features from $\hat{t} = 150$ achieve the best registration performance on BFO. Results on 3DMatch confirm its effectiveness.

4.3.2 Ablating feature fusion weight

Table 2: *Layer selection in diffusion feature extraction*. "Feature map" means the size of the feature map in the form of "channel×width×length".

ID	*Layer*	*Feature Map* (*channel*$\times h \times w$)	***BFO***				***3DMatch***			
			FMR(%)	*IR(%)*	*IN(#)*	*RR(%)*	*FMR(%)*	*IR(%)*	*IN(#)*	*RR(%)*
(a)	0	$1280 \times 8 \times 11$	88.9	42.7	18.9	14.4	89.5	39.7	17.6	16.7
(b)	1	$1280 \times 8 \times 11$	91.5	42.1	19.1	12.4	86.9	39.7	18.1	15.8
(c)	2	$1280 \times 8 \times 11$	86.3	42.9	21.2	14.4	84.2	39.7	20.2	16.9
(d)	3	$1280 \times 16 \times 22$	87.6	42.7	45.9	23.5	88.4	41.0	47.2	23.0
(e)	4	$1280 \times 16 \times 22$	91.5	36.0	31.7	24.2	92.1	35.3	32.9	26.0
(f)	5	$1280 \times 16 \times 22$	89.5	35.4	28.1	22.9	91.7	35.5	28.9	25.6
(g)	6	$1280 \times 32 \times 44$	92.8	31.3	45.5	30.1	89.4	31.4	51.7	28.7
(h)	7	$640 \times 32 \times 44$	90.8	19.9	34.3	14.4	85.2	19.6	35.1	22.1
(i)	8	$640 \times 32 \times 44$	88.9	17.2	28.4	9.8	82.9	16.8	27.5	17.3
(j)	[0,4]	$256 \times 32 \times 44$	93.5	**44.6**	25.9	25.5	**92.5**	**41.3**	34.7	26.5
(k)	[0,6]	$256 \times 32 \times 44$	92.8	40.2	53.9	34.0	91.4	38.5	**62.2**	32.9
(l)	[4,6]	$256 \times 32 \times 44$	91.5	36.4	45.8	32.0	91.4	35.6	56.2	30.7
(m)	[0,4,6]	$384 \times 32 \times 44$	**94.8**	42.3	**58.2**	**35.9**	91.9	39.6	60.8	**33.2**

Table 3: Determining $\hat{t}$ in diffusion feature extraction.

$\hat{t}$	***BFO***				***3DMatch***			
	FMR(%)	*IR(%)*	*IN(#)*	*RR(%)*	*FMR(%)*	*IR(%)*	*IN(#)*	*RR(%)*
300	94.1	40.0	55.8	33.3	91.7	39.4	60.4	31.4
200	92.8	41.0	58.1	35.3	91.8	39.8	61.4	31.2
150	94.8	42.3	58.2	**35.9**	91.9	39.6	60.8	**33.2**
100	92.8	41.3	57.3	35.3	91.8	38.8	59.3	31.6
50	92.8	40.0	54.6	32.7	92.0	38.1	57.3	30.6

We ablate the fusion weight w to fuse diffusion and geometric features in Table. 4 based on the baseline model FreeReg. It can be seen that FreeReg achieves the best registration performance when w is set to 0.5. Moreover, we find that relying more on diffusion features, i.e., $w = 0.6$ achieves a much similar result to the default FreeReg. While relying more on geometric features, i.e., $w = 0.4$ causes a sharp performance drop of a 8.7% lower IR and a 2.5% lower RR. This demonstrates the robustness of the proposed diffusion features.

Table 4: Determining the fusion weight to fuse diffusion and geometric features.

w	*FMR(%)*	*IR(%)*	*IN(#)*	*RR(%)*
1.0	91.9	39.6	60.8	52.6
0.7	94.8	45.1	74.1	58.5
0.6	**95.3**	**47.1**	81.7	62.3
0.5	94.6	47.0	**82.8**	**63.8**
0.4	93.8	42.9	73.5	60.3
0.3	91.8	37.5	61.9	56.5
0.0	90.7	31.4	49.4	50.4

4.4 LIMITATIONS

The main limitation is that FreeReg requires about 9.3s and 12.7G GPU memory to match a single I2P pair on a 4090 GPU, yielding much higher RR but longer time usage than baselines LCD (0.6s, 3.5G, I2P-Matr (1.7s, 2.7G), and CN+SG (6.4s, 11.6G). The reason is that we need to run multiple backward process steps of ControlNet to denoise the pure noises to reach a specific step $\hat{t}$ for feature extraction. In the supplementary material, we show how to accelerate FreeReg by $\sim 50\%$ with only a $\sim 1.4\%$ RR drop. Meanwhile, though we show the superior performance of using diffusion features for I2P registration, we manually select layers and denoising steps in the diffusion feature extraction, which could be improved by future works to automatically select good features.

5 CONCLUSION

We propose an I2P registration framework called FreeReg. The key idea of FreeReg is the utilization of diffusion models and monocular depth estimators for cross-modality feature extraction. Specifically, we leverage the intermediate representations of diffusion models to construct multi-modal diffusion features that show strong consistency across RGB images and depth maps. We further introduce so-called geometric features to capture distinct local geometric details on RGB images and depth maps. Extensive experiments demonstrate that FreeReg shows strong generalization and robustness in the I2P task. Without any training or fine-tuning on I2P registration task, FreeReg achieves a 20.6% improvement in Inlier Ratio, a $3.0\times$ higher Inlier Number, and a 48.6% improvement in Registration Recall on three public indoor and outdoor benchmarks.

6 ACKNOWLEDGEMENT

This research is jointly sponsored by the National Key Research and Development Program of China (No.2022YFB3904102), the Open Fund of Hubei Luojia Laboratory (No. 2201000054), the National Natural Science Foundation of China (No.42301520), and the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative and Ref. T45-205/21-N of Hong Kong RGC.

REFERENCES

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *ECCV*, 2020.

Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *CVPR*, 2020.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *ECCV*, 2022.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Dylan Campbell, Lars Petersson, Laurent Kneip, Hongdong Li, and Stephen Gould. The alignment of the spheres: Globally-optimal spherical mixture alignment for camera pose estimation. In *CVPR*, 2019.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022a.

Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*, 2022b.

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *NeurIPS*, 2016.

Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *CVPR*, 2020.

Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, 2019.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Philip David, Daniel Dementhon, Ramani Duraiswami, and Hanan Samet. Softposit: Simultaneous pose and correspondence determination. *IJCV*, 59:259–284, 2004.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Zhen Dong, Fuxun Liang, Bisheng Yang, Yusheng Xu, Yufu Zang, Jianping Li, Yuan Wang, Wenxia Dai, Hongchao Fan, Juha Hyyppä, et al. Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark. *ISPRS J.*, 163:327–342, 2020.

Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023.

Mengdan Feng, Sixing Hu, Marcelo H Ang, and Gim Hee Lee. 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In *ICRA*, 2019.

Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rares Ambrus, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. *arXiv preprint arXiv:2306.17253*, 2023.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE TPAMI*, 43(12):4338–4364, 2020.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *arXiv preprint arXiv:2305.15581*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Peng Jiang and Srikanth Saripalli. Contrastive learning of features between images and lidar. In *CASE*, 2022.

Peng Jiang, Fanglin Gu, Yunhai Wang, Changhe Tu, and Baoquan Chen. Difnet: Semantic segmentation by diffusion networks. In *NeurIPS*, 2018.

Jinyoung Jun, Jae-Han Lee, Chul Lee, and Chang-Su Kim. Depth map decomposition for monocular depth estimation. In *ECCY*, 2022.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Sec.A*, 32(5):922–923, 1976.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Minjung Kim, Junseo Koo, and Gunhee Kim. Ep2p-loc: End-to-end 3d point to 2d pixel localization for large-scale visual localization. In *ICCV*, pp. 21527–21537, 2023.

Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *CRV*, 2018.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2022.

Baiqi Lai, Weiquan Liu, Cheng Wang, Xuesheng Bian, Yanfei Su, Xiuhong Lin, Zhimin Yuan, Siqi Shen, and Ming Cheng. Learning cross-domain descriptors for 2d-3d matching with hard triplet loss and spatial transformer network. In *ICIG*, 2021.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the p n p problem. *IJCV*, 81:155–166, 2009.

Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *CVPR*, 2021.

Minhao Li, Zheng Qin, Zhirui Guo, Renjiao Yi, Chengyang Zhu, and Kai Xu. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *ICCV*, 2023.

Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM TOG*, 34(6):1–12, 2015.

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

Youqi Liao, Jianping Li, Shuhao Kang, Qiang Li, Guifang Zhu, Shenghai Yuan, Zhen Dong, and Bisheng Yang. Se-calib: Semantic edges based lidar-camera boresight online calibration in urban scenes. *IEEE TGRS*, 2023.

Liu Liu, Dylan Campbell, Hongdong Li, Dingfu Zhou, Xibin Song, and Ruigang Yang. Learning 2d-3d correspondences to solve the blind perspective-n-point problem. *arXiv preprint arXiv:2003.06752*, 2020.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv preprint arXiv:2305.14334*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. Lcd: Learned cross-domain descriptors for 2d-3d matching. In *AAAI*, 2020.

Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *IEEE T-CSVT*, 33(3):1198–1208, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *CVPR*, 2023.

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *arXiv preprint arXiv:2306.01923*, 2023a.

Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023b.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021.

Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. In *NeurIPS*, 2022.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *3DV*, 2017.

Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, et al. P2-net: Joint description and detection of local features for pixel and point matching. In *ICCV*, 2021.

Guangming Wang, Yu Zheng, Yanfeng Guo, Zhe Liu, Yixiang Zhu, Wolfram Burgard, and Hesheng Wang. End-to-end 2d-3d registration between image and lidar point cloud for vehicle localization. *arXiv preprint arXiv:2306.11346*, 2023a.

Haiping Wang, Yuan Liu, Zhen Dong, and Wenping Wang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *ACM MM*, 2022a.

Haiping Wang, Yuan Liu, Zhen Dong, Yulan Guo, Yu-Shen Liu, Wenping Wang, and Bisheng Yang. Robust multiview point cloud registration with reliable pose graph initialization and history reweighting. In *CVPR*, 2023b.

Haiping Wang, Yuan Liu, Qingyong Hu, Bing Wang, Jianguo Chen, Zhen Dong, Yulan Guo, Wenping Wang, and Bisheng Yang. Roreg: Pairwise point cloud registration with oriented descriptors and local rotations. *IEEE TPAMI*, 2023c.

Yuan Wang, Yuhao Li, Yiping Chen, Mingjun Peng, Haiting Li, Bisheng Yang, Chi Chen, and Zhen Dong. Automatic registration of point cloud and panoramic images in urban scenes based on pole matching. *JAG*, 115:103083, 2022b.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *MIDL*, 2022.

Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018.

Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 2020.

Xiaoxia Xing, Yinghao Cai, Tao Lu, Shaojun Cai, Yiping Yang, and Dayong Wen. 3dtnet: Learning local features using 2d and 3d cues. In *3DV*, 2018.

Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 2021.

Gongxin Yao, Yixin Xuan, Yiwei Chen, and Yu Pan. Cfi2p: Coarse-to-fine cross-modal correspondence learning for image-to-point cloud registration. *arXiv preprint arXiv:2307.07142*, 2023.

Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE TPAMI*, 44(10):7282–7295, 2021.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. *arXiv preprint arXiv:2307.10984*, 2023.

Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022.

Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. *arXiv preprint arXiv:2312.04060*, 2023.

Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *CVPR*, 2022.