This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MESC-3D:Mining Effective Semantic Cues for 3D Reconstruction from a Single Image

Shaoming Li[1]*, Qing Cai[1]*†, Songqi Kong[1], Runqing Tan[1], Heng Tong[1],
Shiji Qiu[1], Yongguo Jiang[1], Zhi Liu[2]
Faculty of Computer Science and Technology, Ocean University of China[1]
School of Information Science and Engineering, Shandong University[2]
{lishaoming,kongsongqi,tanrunqing,tongheng,qiushiji}@stu.ouc.edu.cn
{cq,jiangyongguo}@ouc.edu.cn liuzhi@sdu.edu.cn

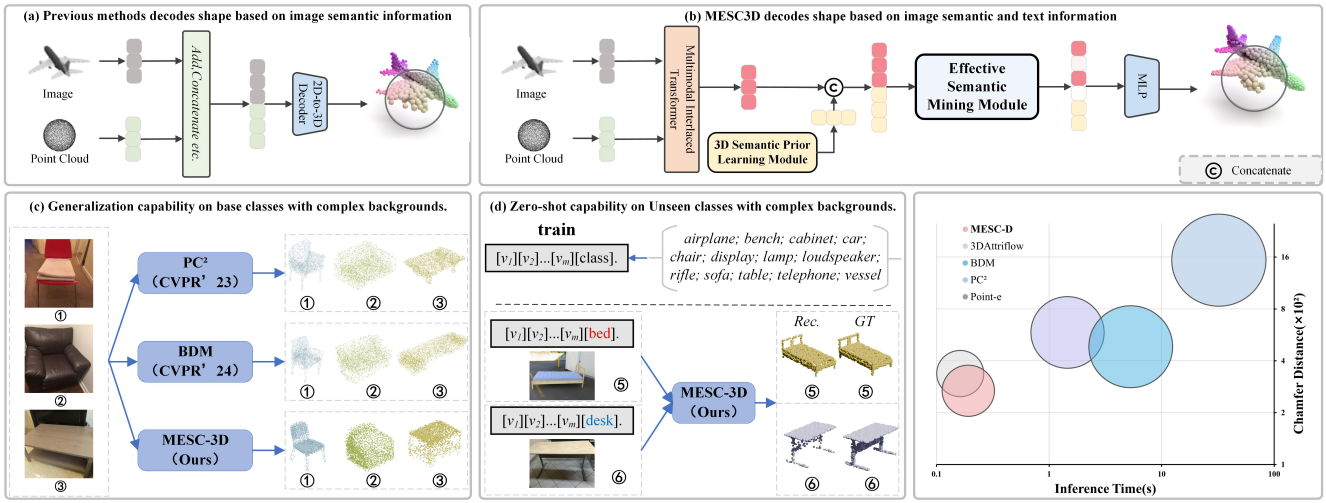


Figure 1. (a) Previous methods simply performed basic operations on the extracted 2D image information and 3D point cloud without establishing a connection between them. (b) Compared to that, we introduced two key designs: First, the Effective Semantic Mining Module, which effectively mines semantic information from the entangled features and enables point cloud to select the information. Second, the 3D Semantic Prior Learning Module, which aims to enable the model to interpret 3D structures as humans do in 3D reconstruction from a single image. (c) The generalization comparsion between the proposed MESC-3D and SOTA methods on base classes with complex backgrounds. (d) MESC-3D's zero-shot on unseen classes. (e) Comparison with state-of-the-art methods on ShapeNet [1] Dataset on Chamfer Distance (y-axis), parameter count (size of the area), and inference time (x-aixs) which show that MESC-3D achieve the best performance with comparable computational cost.

Abstract

Reconstructing 3D shapes from a single image plays an important role in computer vision. Many methods have been proposed and achieve impressive performance. However, existing methods mainly focus on extracting semantic information from images and then simply concatenating it with 3D point clouds without further exploring the concatenated semantics. As a result, these entangled semantic features significantly hinder the reconstruction performance. In this paper, we propose a novel single-image 3D reconstruction method called Mining Effective Semantic Cues for 3D Reconstruction from a Single Image (MESC-3D), which can actively mine effective semantic cues from entangled features. Specifically, we design an Effective Semantic Mining Module, which incorporates a point-selection map establish connections between point clouds and semantic attributes, enabling the point clouds to autonomously select the necessary information. Furthermore, to address the potential insufficiencies in semantic information from a single image, such as occlusions, inspired by the human ability to represent 3D objects using prior knowledge drawn

*Equal contribution.
†The corresponding author is Qing Cai.

from daily experiences, we introduce a 3D Semantic Prior Learning Module. This module employs contrastive learning paradigm along with a learnable text prompt to incorporates semantic understanding of spatial structures, enabling the model to interpret and reconstruct 3D objects with greater accuracy and realism, closely mirroring human perception of complex 3D environments. Extensive evaluations show that our method achieves significant improvements in reconstruction quality and robustness compared to prior works. Additionally, further experiments validate the strong generalization capabilities and excels in zero-shot preformance on unseen classes. Code is available at https://github.com/QINGQINGLE/MESC-3D.

1. Introduction

Image-based 3D object reconstruction plays an important role in computer vision [13, 15, 16, 21, 31], with widespread applications in areas such as virtual reality [14, 19, 24], autonomous driving [10, 11, 18], video games, robotics [39] and 3D content creation [36]. So far, many image-based 3D object reconstruction method have been proposed, which can be broadly categorized into single-view 3D shape reconstruction [29, 34] and multi-view 3D shape reconstruction [26, 40]. This paper focuses on single-view point cloud reconstruction and conducts a series of studies on this topic. The typical paradigm for single-view 3D point cloud reconstruction involves extracting semantic attributes from a single image, obtaining implicit or explicit features through these attributes, parsing the image's class and geometric information, and finally decoding it into a 3D shape via a 3D decoder. The typical paradigm followed by most existing methods [27–29, 32, 33] indicates that the key to 2D-to-3D reconstruction lies in accurately translating the semantic attributes of the image into 3D space.

However, most existing methods [27–29, 32, 33] simply transfer image semantic information to the 3D decoder by performing element-wise addition or channel concatenation between semantic features and the 3D point cloud as seen in Fig. 1(a). Since these semantic features are entangled, directly using them for reconstruction would severely affect the reconstruction quality. Moreover, due to the limited visual information in a single view, 3D reconstruction from a single image often produce rough geometric shapes in occluded regions. Despite the difficulty of this task, humans are adept at using a range of monocular cues and prior knowledge to infer the 3D structure of common objects from single view. Inspired by daily life, humans can easily and naturally infer the shape of the back of an object in a photograph based on life experiences.

Based on the above discussion, this paper aims to design a novel network that, like humans, can sketch the shape of an object in theirs mind based on prior understanding when observing the object in an image. To achieve this goal, this paper has made two key designs, as shown in Fig. 1 (b). Specifically, to simulate the process of humans obtaining prior knowledge of the object when seeing it in a single image, we design a 3D Semantic Prior Learning Module (3DSPL), which integrates object global semantic prior into the network by utilizing a learnable text prompt. Furthermore, to further untangle semantics, we have designed an Effective Semantic Mining Module (ESM), which can enable point cloud to autonomously select the necessary semantic information and establish a connection between point cloud and semantic attributes. Experimental results also demonstrate the effectiveness of this design.

In summary, our contributions are three-fold:

- We design an Effective Semantic Mining Module (ESM). Unlike traditional 3D reconstruction[27–29, 32, 33], our approach uses a point-selection map to directly guide the point cloud in autonomously selecting key semantic information, achieving a more precise mapping from features to points. This mechanism establishes a connection between point clouds and semantic attributes.
- We design a 3D Semantic Prior Learning Module (3DSPL). Through a learnable text prompt and contrastive learning between text and point cloud modalities, this module effectively compensates for the lack of semantic information in single view, enabling it to infer object shapes from a single image, much like how humans can intuitively deduce the complete structure of an object from a single image.
- We conduct extensive experiments to demonstrate the versatile capabilities of MESC-3D. It exhibits superior performance on synthetic dataset [1] and real-world dataset [25], generalization capability and zero-shot capability on 3D reconstruction.

2. Related Work

Single-View Point Clouds Reconstruction. 3D reconstruction is a fundamental task of translating the 2D world into a 3D representation. Unlike meshes and voxels, point clouds offer a memory-efficient way to represent the shape of arbitrary objects. Point cloud reconstruction methods are currently divided into traditional deep learning-based approaches [4, 27, 29, 32] and diffusion-based methods [16, 20, 21]. Traditional deep learning methods include approaches like 3DR2N2 [4], PSGN [7], Pix2Mesh [27], AtlasNet [8], OccNet [17] and 3DAttriflow [29]. These methods typically involve extracting semantic attribute from single or multiple views of images and then decoding these 2D features into 3D space. They have demonstrated effective performance in single-view and multi-view reconstruction [26]. In recent years, researchers [5, 16, 20, 21] have explored various pipelines based on diffusion probabilistic models to achieve 3D shape generation. For in-

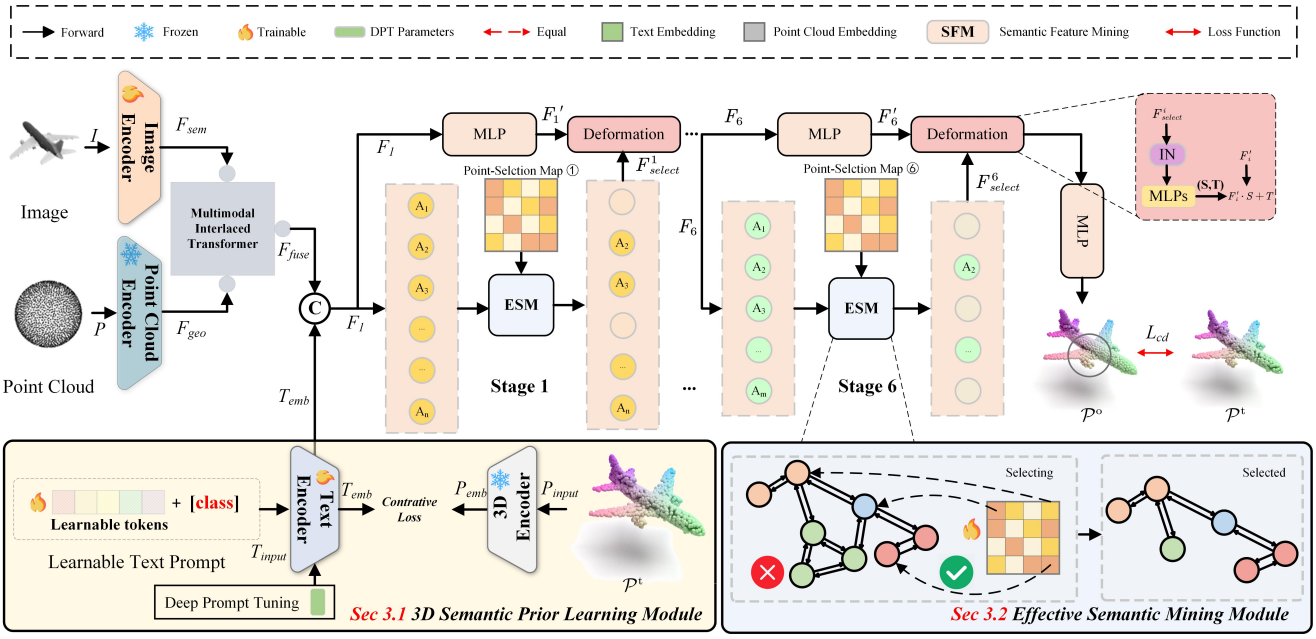


Figure 2. **The overall architecture of MESC-3D.** Our network is composed of two main components. (a) The 3DSPL align point cloud modality features with text features, aiming to capture the unique 3D geometric characteristics of each category. (b) The ESM establishes a connection between the semantic feature F_i and the 3D point cloud at i^{th} stage, allowing each point to autonomously select the most valuable semantic information.

stance, methods such as Point-E [21], PC² [16] and BDM [34] have successfully generated sparse point clouds from single images through a denoising diffusion process. They use a diffusion model which conditions on image. Additionally, PC² takes its camera pose and gradually denoises a set of 3D points, aiming to project local image features onto the partially-denoised point cloud at each step of the diffusion process. Unfortunately, these methods [27–29, 32, 33] either focus on enhancing the richness of semantic information or improving communication between 2D and 3D encoders, without adequately addressing the connection between these modalities. As a result, these methods [27–29, 32, 33] fail to fully exploit the effective semantic information in the image. Furthermore, a single view image often suffer from insufficient semantic information primarily due to occlusion.

Prompt Learning. This topic [2, 6, 30, 37, 38] has emerged as an effective fine-tuning strategy to adapt Vision models. It adds a small number of learnable embeddings along with model inputs which are optimized during training while the rest of the model is kept frozen. This approach has already shown outstanding performance in the 2D domain. It is worth noting that the CoOp [38] early presents the first comprehensive study about adapting large vision models with prompt learning. Specifically, CoOp [38] models a prompt's context words with learnable vectors while the entire pre-trained parameters are kept fixed. It demonstrates that prompt learning could bring much stronger robustness. Following this, they introduced CoCoOp [37]. Inspired by their work, we aim to extend prompt learning to the 3D domain, addressing prompt engineering for 3D and learning prior of 3D structure like humans, making use of the 3D rich knowledge encoded in the parameters to optimize the context. We attempt to incorporate 3D geometric prior into the Text Encoder to enhance guidance for subsequent 3D reconstruction tasks.

3. Method

The overall architecture of MESC-3D is visualized in Fig. 2. Given a single view Image $I \in [224, 224]$ and a random sphere points $P \in (2048, 3)$. We first adopt ResNet18 [9] and pre-trained PointMAE [22] to obtain the image feature F_{sem} and the point cloud feature F_{geo} respectively. Then, using the Multimodal Interlaced Transformer MIT [35], we enrich these features through mutual interaction, resulting in the fused feature $F_{fuse} = MIT(F_{geo}, F_{sem})$. Next, a prior feature T_{emb} obtained by one of the key component 3DSPL, which is then concatenated with the original feature to form the stage 1 feature as $F_1 = Concat[F_{fuse}, T_{emb}]$. Finally, after six stages of effective feature selection by using another main component ESM, the 3D coordinates of each point are predicted by a tiny MLP.

3.1. 3D Semantic Prior Learning Module

In order to endow the model with the ability to infer like humans when seeing an object, we utilize contrastive learning to imitate the human acquisition of prior knowledge in daily life. This enables the model to recognize an object's category and associate it with learned prior knowledge for accurate 3D reconstruction. We use the ShapeNet [1] dataset as our 3D modality, which is one of the most extensive public 3D CAD datasets. However, instead of using these specific textual descriptions, we replace them with learnable text prompt.

Deep Prompt Tuning(DPT) By using Deep Prompt Tuning (DPT) [41], we can retain the original CLIP parameters intact while effectively learning 3D geometric prior. Formally, we denote the original text embeddings in CLIP as $K^l = \{k_1^l, k_2^l, \ldots, k_N^l\}$. Deep prompt tuning appends learnable tokens $P^l = \{p_1^l, p_2^l, \ldots, p_M^l\}$ to this token sequence in each Vision Transformer (ViT) layer of the CLIP text encoder. Then, the l-th Multi-Head Attention (MHA) module processes the input token as:

$$[_, K^l] = \text{Layer}^l[P^{l-1}, K^{l-1}], \tag{1}$$

where the output embeddings of $\{p_1^l, p_2^l, \ldots, p_M^l\}$ are discarded (denoted as $_$) and are not fed into the next layer. Therefore, $\{p_1^l, p_2^l, \ldots, p_M^l\}$ merely acts as a set of learnable parameters to adapt the MHA model.

Learnable Text Prompt Inspired by CoOp [38], we designed a learnable text prompt specifically for the 3D reconstruction task. To extract useful information for 3D reconstruction, we allow the model to autonomously summarize the corresponding 3D textual descriptions for point cloud. The learnable text prompt shares the same context vectors with all classes. The design of continuous representations also allows full exploration in the word embedding space, which facilitates the learning of 3D-relevant context. In this process, we only need to provide an appropriate prompt, such as the 3D shape category. Concretely, the original 3D categories are first transformed into class tokens through the CLIP tokenizer, i.e., $t_{\text{init}} = \text{Tokenizer}(\text{Category})$, where Category is the discrete 3D text category, e.g., airplane, vessel, etc. We then concatenate t_{init} with the learnable text prompt $\{c_1, \ldots, c_l\}$, which contains l context tokens, to form a complete sentence token. Thus, the input to the text encoder is presented as follows:

$$t_p = \{c_1, ..., t_{init}, ..., c_l\}. \tag{2}$$

Here, we place the class token at the middle of a sequence. Then this sequence token is added to the positional embedding to obtain positional information, and finally, the text encoder of CLIP takes as input T_{input} and generates prompt embedding T_{emb}.

3.2. Effective Semantic Mining Module

The model acquires 3D prior knowledge T_{emb} through the 3DSPL. This prior knowledge is then integrated with enriched semantic features F_{fuse}, resulting in a more comprehensive 3D attribute representation F_1. Through multi-stage selection, the model progressively refines attribute information. Notably, each stage focuses on different attribute information. To achieve this, we designed an ESM that is applied to the $F_i \in \mathbb{R}^{B\times C\times N}$ features, allowing each point in the point cloud to autonomously select the features it requires, where i denotes i^{th} stage. For convenience, we denote the process of selecting features from the point cloud as:

$$F_{select}^i = ESM(F_i), \tag{3}$$

Specifically, ESM considers a point-selection map with the same shape as F_i. First, the map is initialized, and during training, the values in the point-selection map are optimized. The initial values follow a normal distribution map $N\left(\mu, \sigma^2\right)$ with mean $\mu = 0$ and standard deviation $\sigma = 0.01$, which provides the network with a moderate amount of perturbation to gradually optimize the parameters in the point-selection map within the [0, 1] range. To enable the point cloud to select information based on the point-selection map, it selects attributes according to the varying values within the map. Then, map is used as a mask, and multiplied element-wise with the input features F_i to yield the final output F_{select}^i:

$$F_{select}^i = F_i \cdot map \tag{4}$$

Meantime, we obtain the feature representation of each sample in a common embedding space by projecting the feature F_i to a common dimension represented by:

$$F_i' = MLP(F_i) \tag{5}$$

Next, each point selects effective semantic features to compute its deformation of scaling and translation (S, T), give by:

$$(S, T) = \phi(F_{select}^i), \tag{6}$$

where ϕ represents MLPs, S denoted the scaling factor and T denoted the translation.

Finally, during the decoding process into 3D space, this information is used to guide the corresponding features.

$$F_{i+1} = F_i' \cdot S + T \tag{7}$$

The benefits of this approach are threefold: first, it reduces feature entanglement by allowing each point to compute its positive and negative features from the multiple channels of the fused features, thereby effectively extracting the semantic cues each point needs; second, it reduces

Methods	CD					F-score@1%				
	Point-E[21] (Arxiv 22)	3DAttriFlow [29] (CVPR 22)	PC²[16] (CVPR 23)	BDM-B [34] (CVPR 24)	**Ours**	Point-E [21] (Arxiv 22)	3DAttriFlow [29] (CVPR 22)	PC² [16] (CVPR 23)	BDM-B [34] (CVPR 24)	**Ours**
Airplane	17.44	2.11	4.12	4.01	**1.91**	0.486	0.983	0.920	0.928	**0.987**
Bench	32.00	2.71	4.18	3.74	**2.41**	0.188	0.978	0.921	0.930	**0.985**
Cabinet	22.59	2.66	6.08	6.03	**2.49**	0.280	0.984	0.839	0.842	**0.986**
Car	23.55	2.50	4.74	4.82	**2.36**	0.292	0.990	0.929	0.926	**0.993**
Chair	16.89	3.33	5.25	5.15	**3.02**	0.374	0.966	0.883	0.888	**0.976**
Display	24.27	3.60	6.24	6.03	**3.15**	0.300	0.952	0.832	0.838	**0.971**
Lamp	27.06	4.55	8.12	8.12	**4.26**	0.290	0.898	0.705	0.711	**0.907**
Loudspeaker	20.33	4.16	7.55	7.35	**3.72**	0.340	0.928	0.744	0.758	**0.943**
Rifle	15.90	1.94	2.70	2.71	**1.77**	0.498	0.986	0.964	0.966	**0.990**
Sofa	26.26	3.24	6.57	6.48	**3.00**	0.249	0.970	0.840	0.844	**0.975**
Table	25.38	2.85	5.91	5.83	**2.67**	0.239	0.966	0.843	0.843	**0.985**
Phone	28.46	2.66	4.30	4.37	**2.38**	0.239	0.976	0.943	0.940	**0.987**
Vessel	18.02	2.96	4.36	4.28	**2.70**	0.439	0.973	0.920	0.924	**0.982**
Average	22.93	3.02	5.39	5.30	**2.76**	0.324	0.965	0.868	0.872	**0.974**

Table 1. 2D-to-3D reconstruction on ShapeNet dataset in terms of per-point L1 CD $\times 10^2$ and F-score@1%.

the computational load by excluding marginalized features that are less relevant from the computations; and third, it reduces inference time, as demonstrated by our comparison of the inference time of diffusion-based methods with our approach. This method not only streamlines the computational process but also enhances the precision and efficiency of the 3D reconstruction task.

3.3. Training loss

For 3D text encoder pre-training, as shown in Fig. 2, for an object i, features T^i_{emb} and P^i_{emb} are extracted from learnable text prompt and 3D point cloud encoders. Then contrastive loss among each pair of modalities is computed as follows,

$$L_{(T,P)} = \sum_{(i,j)} -\frac{1}{2}\log\frac{\exp\left(\frac{T^i_{emb}P^j_{emb}}{\tau}\right)}{\sum_k \exp\left(\frac{T^i_{emb}P^k_{emb}}{\tau}\right)} - \frac{1}{2}\log\frac{\exp\left(\frac{T^i_{emb}P^j_{emb}}{\tau}\right)}{\sum_k \exp\left(\frac{T^k_{emb}P^j_{emb}}{\tau}\right)}. \tag{8}$$

where T and P represent two modalities and (i, j) indicates a positive pair in each training batch. We use a learnable temperature parameter τ as well, similar to CLIP [23].

We train our MESC-3D model fully supervised using the Chamfer Distance(CD) loss. The CD measures the distance of each point to the other set:

$$\mathcal{L}_{\mathrm{CD}}(\mathcal{P}^{\mathrm{o}}, \mathcal{P}^{\mathrm{t}}) = \frac{1}{2N}\sum_{p^{\mathrm{o}}\in\mathcal{P}^{\mathrm{o}}}\min_{p^{\mathrm{t}}\in\mathcal{P}^{\mathrm{t}}}\left\|p^{\mathrm{o}} - p^{\mathrm{t}}\right\|_2 + \frac{1}{2N}\sum_{p^{\mathrm{t}}\in\mathcal{P}^{\mathrm{t}}}\min_{p^{\mathrm{o}}\in\mathcal{P}^{\mathrm{o}}}\left\|p^{\mathrm{t}} - p^{\mathrm{o}}\right\|_2, \tag{9}$$

where $\mathcal{P}^{\mathrm{o}}$ and $\mathcal{P}^{\mathrm{t}}$ represent the predicted point cloud and the corresponding ground truth point cloud respectively.

4. Experiments

In this section, we experimentally evaluate the effectiveness of MESC-3D in 3D reconstruction task, and analyze the quantitative and qualitative results. The ablation studies will focus on effectiveness of each part of MESC-3D, visually analyze the generalization ability on base classes. To demonstrate the benefits of 3DSPL, we conduct experiments on the zero-shot 3D reconstruction task, where the model is evaluated on previously unseen classes.

4.1. Datasets And Evalution Metrics

Following [16, 29, 34] and [3, 12], we evaluate our method on the synthetic dataset ShapeNet [1] and the real-wolrd dataset Pix3D [25].

ShapeNet. The ShapeNet Dataset is a collection of 3D CAD models corresponding to categories in the WordNet lexical database. We use a subset of the ShapeNet dataset consisting of 44K models and 13 major categories following the experimental setup used in 3DAttriflow [29]. More specifically, renderings of each 3D model contains 24 random views of size 137 × 137.

Pix3D. The Pix3D dataset is a large-scale single-image 3D shape modeling dataset that features pixel-level 2D-3D alignments. The dataset contains 395 3D models of nine object classes such as chairs, tables, and sofas.

Evalution Metrics. To evaluate the reconstruction quality of the proposed methods, we calculate Chamfer Distance (CD) between points clouds uniformly sampled from the ground truth and our prediction to measure the surface accuracy. We use Chamfer disatance described by Eq. 9 to evaluate the performance of 3D reconstruction results. In addition, following PC² [16], we also take F-score@1% as an extra metric.

For CD, the smaller is better. For F-score@1%, the larger is better.

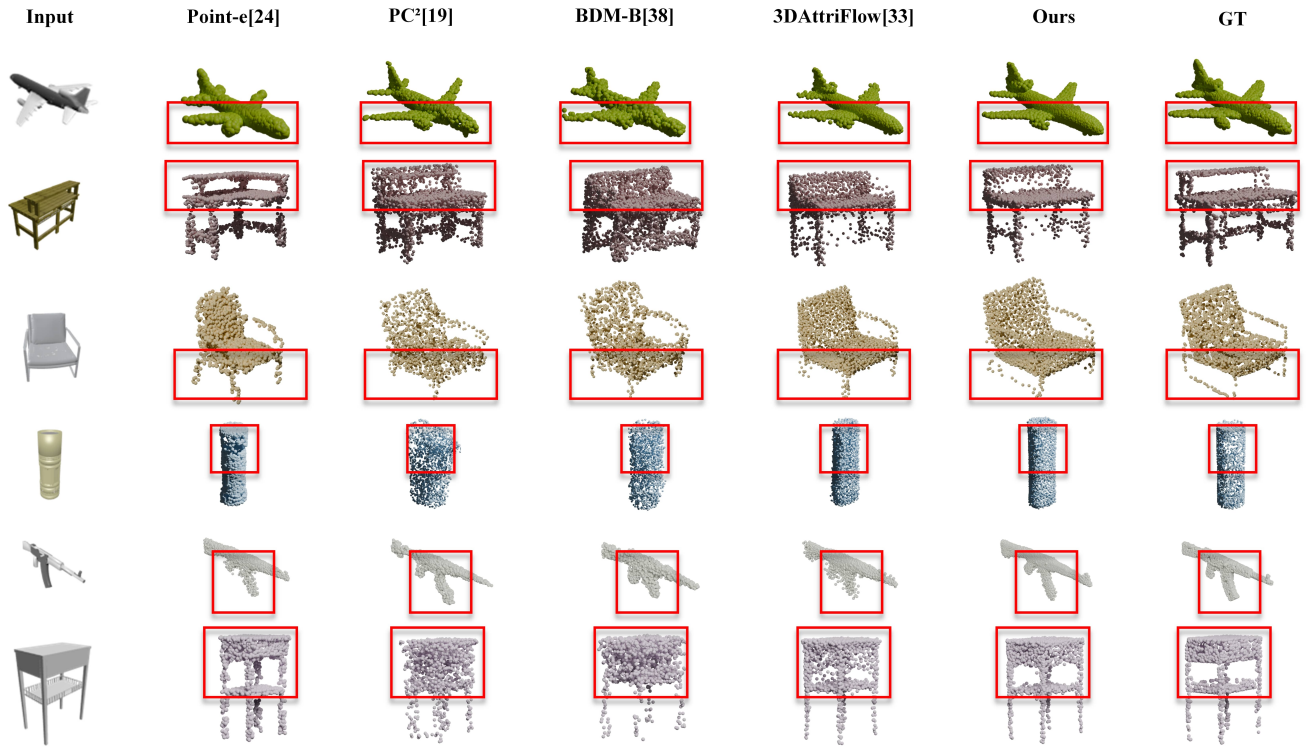


Figure 3. Visual comparison of 2D-to-3D reconstruction results with different methods under ShapeNet dataset. Additional qualitative results are provided in the supplementary material.

4.2. Implementation Details

Pre-training. For 3DSPL, we use the point cloud data from the ground truth of our training and validation sets, without utilizing the ground truth of the test set for training. We freeze the point cloud encoder and only update the text encoder's parameters during pre-training. MESC-3D is trained for 200 epochs. We use 64 as the batch size, 10^{-4} as the learning rate, and Adam as the optimizer.

3D Reconstruction. On ShapeNet, we use the learning rate of 10^{-3} and Adam as the optimizer.The initial learning rate is set to 10^{-3} and gradually decays to 10^{-5}. We train for 400 epochs with batch size 24 on an RTX 3090 GPU.

4.3. Comparison with SOTAs

We benchmark our method against SOTA approaches in point cloud reconstruction, including those based on traditional deep learning [4, 7, 8, 17, 27, 29] and diffusion methods [12, 16, 21, 34].

Quantitative Comparison. We present results for 13 categories in the widely-used ShapeNet benchmark. Tab. 1 shows a quantitative comparison between our method and state-of-the-art point cloud reconstruction methods. Our method consistently outperforms the SOTA method across all 13 classes and on average, demonstrating its capability to achieve superior performance comparable to prior methods. Through the improvement in metrics across various categories, we see that our method performs better on both simple and complex objects, such as "display" and "airplane". Overall, our approach achieves an average Chamfer Distance of 2.76 and an F-score@1% of 0.974, outperforming the best competing method by a notable margin.

We also evaluate on the Pix3D[25] dataset in Tab. 2. It can be seen that our method remarkably improves the performance and achieves state-of-the-art. On average, we achieve 1.37 CD, while other diffusion methods have around 7, showing that our approach attains a much better overall object reconstruction quality.

Table 2. 2D-to-3D reconstruction on Pix3D dataset in terms of per-point L1 CD $\times 10^2$ and F-score@1%.

Methods	CD	F-score@1%
SDFusion(CVPR'23) [3]	7.22	0.772
3DQD(CVPR'23) [12]	8.65	0.693
Ours	**1.37**	**0.990**

Qualitative Comparison. To provide a more comprehensive comparison with SOTAs, We also presents the qualitative results as seen in the Fig. 3. It is evident that 3DAttriFlow struggles to model regions with high uncertainty from

Table 3. Impact of different modules in our MESC-3D in terms of L2 CD$\times 10^3$. We show the results of integrating different modules into the baseline. ESM denoted as Effective Semantic Mining Module and 3DSPL denoted as 3D Semantic Prior Learning Module.

Model	Avg-CD	Airplane	Bench	Cabinet	Car	Chair	Display	Lamp	Loud.	Rifle	Sofa	Table	Tele.	Vessel
Baseline	5.68	2.22	4.14	3.59	2.21	4.83	8.66	14.61	12.66	1.87	5.74	4.48	4.89	3.88
Baseline + ESM	3.69	1.79	2.68	2.62	1.87	3.36	4.85	10.20	7.17	1.62	3.45	3.29	2.29	2.74
Baseline + ESM + 3DSPL	**3.20**	**1.65**	**2.52**	**2.36**	**1.83**	**3.24**	**3.58**	**8.45**	**5.32**	**1.53**	**3.32**	**3.06**	**2.18**	**2.63**

a single view, such as the two ram air turbines on the airplane and the backrest of the bench, which are all occluded in this perspective and lack sufficient information. In contrast, by incorporating 3D prior, our method effectively reconstructs these occluded areas. Additionally, BDM-B produces very blurred reconstructions for some complex objects, such as the shape of the chair and the magazine of the rifle. Relative to prior methods, our approach generates shapes with significantly finer levels of detail.

Fig. 4 presents qualitative results on Pix3D dataset. Our method trained on real-world data recovers not only realistic geometry, but also fine-grained, such as the thickness of the bookcase and the components of the bed with a ladder. This further demonstrates that its better mining effective semantic cues to guide 3D reconstruction.

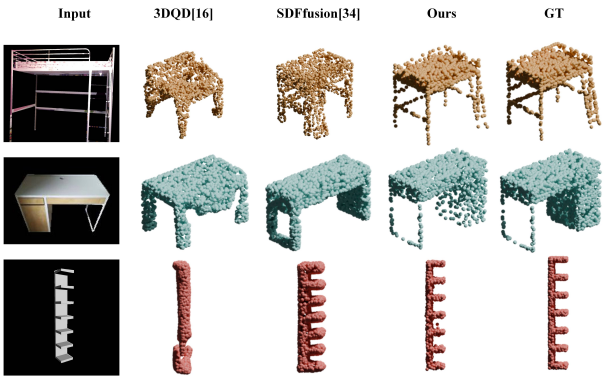


Figure 4. Visual comparison of 2D-to-3D reconstruction results with different methods under Pix3D dataset.

4.4. Ablation Study

Abaltion Study on Model Architecture. Tab. 3 shows the impact of integrating different components into the baseline. The baseline adopts original Multimodal Interlaced Transformer (MIT) and 3D Decoder to predict the point cloud. When the ESM was introduced, the score improved from 5.68 to 3.69, outperforming the baseline by 35%. Furthermore, when the prior was introduced through the 3DSPL, the occluded areas in the image have been reconstructed with enhanced details, as seen in Fig. 5. The results validate that each component contributes to the performance of our method.

Abaltion Study on Image Encoder. We conduct ablation study on the backbone without the introduction of text prompt, and the evaluation metrics were measured only on the vessel. Tab. 4 presents the ablation study on the pre-training and fine-tuning of the image encoder. In this ablation, our model incorporates only the ESM, and the experimental results are measured exclusively on the vessel.

Table 4. Abaltion Study on Image Encoder. We present results in terms of L2 CD$\times 10^3$, using a frozen CLIP model, a fine-tuned CLIP model, and a trainable ResNet18 network architecture on vessel.

Image Encoder	CD	F-score@1%
CLIP(Pre-trained)	4.71	0.954
CLIP(Fine-tuned)	4.24	0.966
Resnet18(Ours)	**2.74**	**0.980**

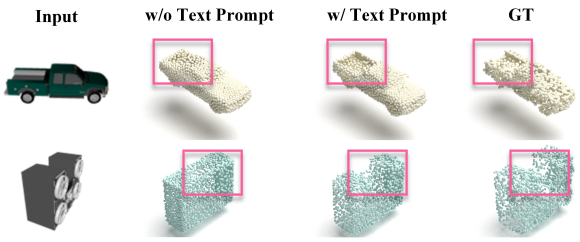


Figure 5. Ablation Study on learnable text prompt. Visual results on ShapeNet.

Table 5. Ablation Study on Learnable Text Prompt Training Strategy. we evaluated them by L2 CD$\times 10^3$ metrics on vessel.

Text Encoder	PC Encoder	CD	F-score@1%
CLIP (Pre-trained)	ULIP	2.80	0.978
CLIP (Pre-trained)	PointMAE	2.74	0.980
CLIP (DPT)(Ours)	**ULIP**	**2.63**	**0.982**

Ablation Study on Learnable Text Prompt Training Strategy. Tab. 5 shows different training strategies for the first-stage learnable text prompt. We employed various Text Encoders, PC Encoders, and Image Encoders. In the first two strategies, we used pre-trained encoders and applied learnable text prompt with point cloud and image triad alignment strategies. The L2 CD scores for these strategies are similar. In the final strategy, we introduced DPT to fine-

tune the Text Encoder, combined with learnable text prompt and point cloud pair alignment strategies. The experiments demonstrate that incorporating DPT leads to a better learning of 3D geometric prior.

Ablation Study on Introduce Manner of Learnable Text Prompt. Tab. 6 shows different methods for incorporating 3D prior into the point cloud reconstruction stage, including addition, fusion, and concatenation of point cloud and image fusion features. The experimental results indicate that the three-modal fusion strategy performs the worst. Directly adding the 3D prior to the fused features also did not significantly improve performance. However, concatenating the 3D prior with the fused features led to an improvement in the L2 CD score to 2.80. This method's advantage lies in preserving the original 3D prior, thereby providing more intuitive feature selection for point cloud reconstruction.

Table 6. Introduce of Learnable Promp. Here, Ⓜ represents the MultiModal Interlace Transformer, Ⓒ stands for the concatenation operation between two vectors, and ⊕ denotes element-wise addition between two vectors. The following operations are performed sequentially from left to right. We evaluated them on vessel in terms of L2 CD$\times 10^3$.

Methods	CD	F-score@1%
F_{sem} Ⓜ T_{emb} Ⓜ F_{geo}	3.26	0.974
F_{sem} Ⓜ F_{geo} ⊕ T_{emb}	2.98	0.976
F_{sem} Ⓒ T_{emb} Ⓜ F_{geo}	2.92	0.977
F_{sem} Ⓜ F_{geo} Ⓒ T_{emb}	**2.80**	**0.978**

Table 7. Ablation study on category token placement.

Position	Start	Middle	End
Avg-CD L2	3.48	3.20	3.51

Hyperparameter Analysis. We experimented on different embedding positions of categories as illustrated in Tab. 7. The results indicate that the best position is in the middle.

4.5. Further Analysis

Generalization on Base Classes. We use ShapeNet as the training set and Pix3D as the test set, keeping other settings consistent and without introducing prompt learning. By visually comparing the model predictions with the Ground Truth, as seen in Fig. 6, we observe that our model successfully reconstructs the square-shaped tabletop and the four table legs. This result demonstrates the model's effective capability in mining semantic features and its strong generalization ability to unseen basic classes.

Zero-shot on Unseen Classes. We conducted a quantitative analysis of zero-shot performance through an ablation study on learnable text prompt, where the text prompt were trained only on the 13 categories (excluding bed, bookcase, desk, and tool). As shown in Fig. 7, introducing the text prompt led to a significant improvement in the metrics. By incorporating 3D prior, the reconstruction of unseen classes was effectively aided, thereby demonstrating the zero-shot capability enabled by learnable text prompt.

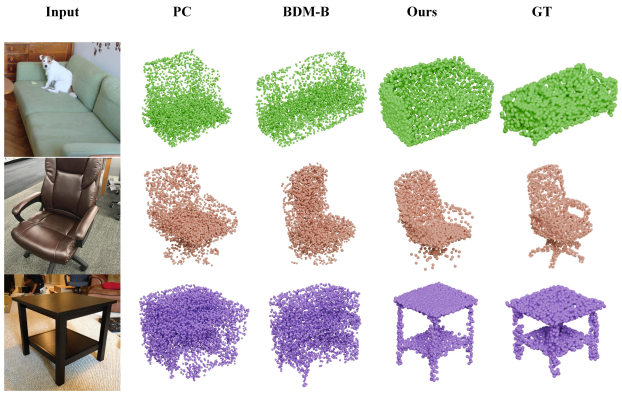


Figure 6. Generalization on base classes with various methods.

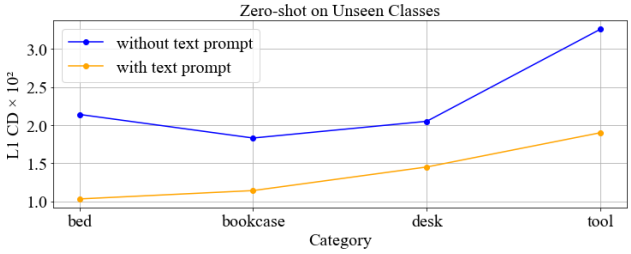


Figure 7. Zero-shot on Unseen Classes in Pix3D dataset.

5. Conclusion

In this work, we propose a novel 3D point cloud reconstruction method from a single-view image by mining effective semantic cues. To ensure precision and efficiency in reconstruction, we designed two key components to achieve competitive results on the reconstruction task. The first is the ESM, which enables the point cloud to autonomously select relevant semantic information during the decoding process, overcoming the issue of deeply entangled with each other. The second component is the 3DSPL, which introduces learnable text prompts and contrastive learning to emulate the human-like ability to learn 3D prior knowledge. To demonstrate the effectiveness of our method, we conduct extensive evaluations on both synthetic and real-world datasets, achieving performance that surpasses previous state-of-the-art methods. Furthermore, our work explores generalization to base classes and zero-shot capability on unseen classes, which may provide inspiration for future research.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 4, 5

[2] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. 2022. 3

[3] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 5, 6

[4] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision*, pages 628–644. Springer, 2016. 2, 6

[5] Ruikai Cui, Xibin Song, Weixuan Sun, Senbo Wang, Weizhe Liu, Shenzhou Chen, Taizhang Shang, Yang Li, Nick Barnes, Hongdong Li, et al. Lam3d: Large image-point-cloud alignment model for 3d reconstruction from single image. *arXiv preprint arXiv:2405.15622*, 2024. 2

[6] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Bayesian prompt learning for image-language model generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15237–15246, 2023. 3

[7] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 605–613, 2017. 2, 6

[8] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 216–224, 2018. 2, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[10] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022. 2

[11] Varun Ravi Kumar, Ciarán Eising, Christian Witt, and Senthil Kumar Yogamani. Surround-view fisheye camera perception for automated driving: Overview, survey & challenges. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):3638–3659, 2023. 2

[12] Yuhan Li, Yishun Dou, Xuanhong Chen, Bingbing Ni, Yilin Sun, Yutian Liu, and Fuzhen Wang. Generalized deep 3d shape prior via part-discretized diffusion process. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16784–16794, 2023. 5, 6

[13] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2

[14] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 2

[15] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 2

[16] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 2, 3, 5, 6

[17] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2, 6

[18] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4213–4220. IEEE, 2019. 2

[19] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 2

[20] Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *Advances in Neural Information Processing Systems*, 36:67960–67971, 2023. 2

[21] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2, 3, 5, 6

[22] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*, pages 604–621. Springer, 2022. 3

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[24] Manli Shu, Le Xue, Ning Yu, Roberto Martín-Martín, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. Model-agnostic hierarchical attention for 3d object detection. *arXiv preprint arXiv:2301.02650*, 2023. 2

[25] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum,

and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 2, 5, 6

[26] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5731, 2021. 2

[27] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision*, pages 52–67, 2018. 2, 3, 6

[28] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1042–1051, 2019.

[29] Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, and Zhizhong Han. 3d shape reconstruction from 2d images with disentangled attribute flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3803–3813, 2022. 2, 3, 5, 6

[30] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6074–6082, 2024. 3

[31] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8929–8939, 2023. 2

[32] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2690–2698, 2019. 2, 3

[33] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12): 2919–2935, 2020. 2, 3

[34] Haiyang Xu, Yu Lei, Zeyuan Chen, Xiang Zhang, Yue Zhao, Yilin Wang, and Zhuowen Tu. Bayesian diffusion models for 3d shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10628–10638, 2024. 2, 3, 5, 6

[35] Cheng-Kun Yang, Min-Hung Chen, Yung-Yu Chuang, and Yen-Yu Lin. 2d-3d interlaced transformer for point cloud segmentation with scene-level supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 977–987, 2023. 3

[36] Kai Yang, Hong Shang, Tianyang Shi, Xinghan Chen, Jingkai Zhou, Zhongqian Sun, and Wei Yang. Asm: Adaptive skinning model for high-quality 3d face modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20708–20717, 2023. 2

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 4

[39] Peng Zhou, Rui Peng, Maggie Xu, Victor Wu, and David Navarro-Alarcon. Path planning with automatic seam extraction over point cloud models for robotic arc welding. *IEEE Robotics and Automation Letters*, 6(3):5002–5009, 2021. 2

[40] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 2

[41] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 4