Mask3D: Pre-training 2D Vision Transformers by Learning Masked 3D Priors

Ji Hou[1] Xiaoliang Dai[1] Zijian He[1] Angela Dai[2] Matthias Nießner[2]
[1]Meta Reality Labs [2]Technical University of Munich

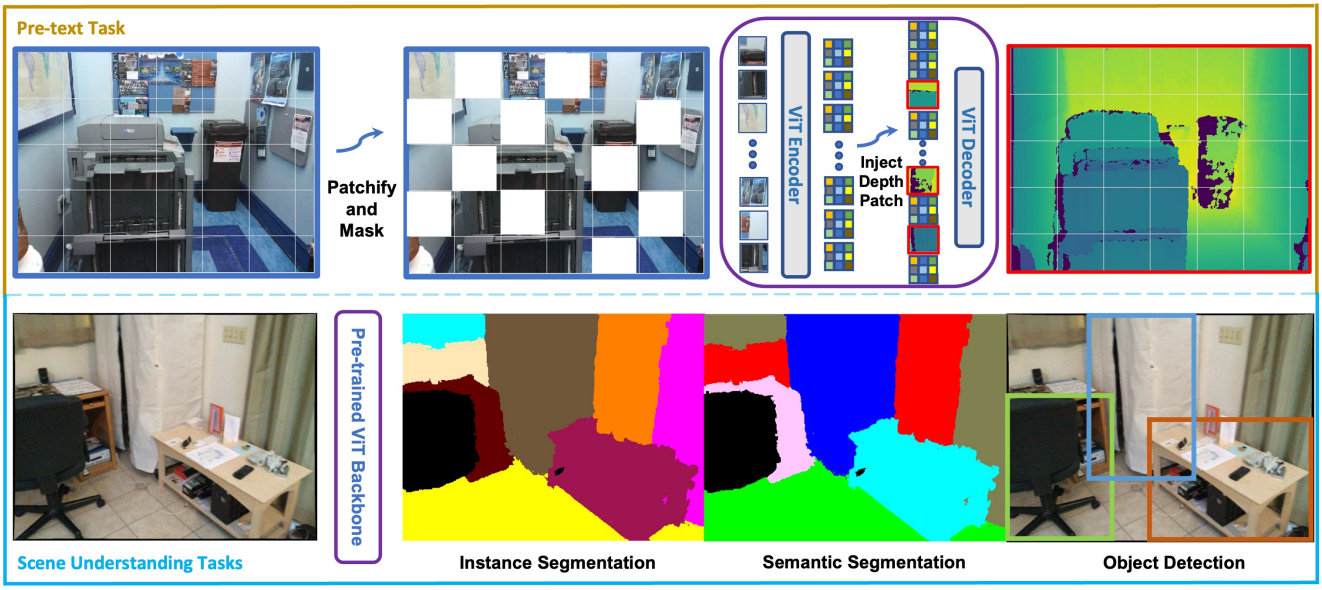

Figure 1. We present Mask3D, which learns to embed 3D priors to 2D representations for image understanding tasks, based on a self-supervised pre-training formulation from single RGB-D views, without requiring any camera pose or multi-view correspondence information. Our pre-training takes masked RGB and depth patches as input to reconstruct the dense depth map, and the pre-trained color backbone is used to fine-tune various downstream image understanding tasks. This results in effective ViT pre-training for a variety of downstream tasks and datasets.

Abstract

Current popular backbones in computer vision, such as Vision Transformers (ViT) and ResNets are trained to perceive the world from 2D images. However, to more effectively understand 3D structural priors in 2D backbones, we propose Mask3D to leverage existing large-scale RGB-D data in a self-supervised pre-training to embed these 3D priors into 2D learned feature representations. In contrast to traditional 3D contrastive learning paradigms requiring 3D reconstructions or multi-view correspondences, our approach is simple: we formulate a pre-text reconstruction task by masking RGB and depth patches in individual RGB-D frames. We demonstrate the Mask3D is particularly effective in embedding 3D priors into the powerful 2D ViT backbone, enabling improved representation learning for various scene understanding tasks, such as semantic segmentation, instance segmentation and object detection. Experiments show that Mask3D notably outperforms existing self-supervised 3D pre-training approaches on ScanNet, NYUv2, and Cityscapes image understanding tasks, with an improvement of +6.5% mIoU against the state-of-the-art Pri3D on ScanNet image semantic segmentation.

1. Introduction

Recent years have seen remarkable advances in 2D image understanding as well as 3D scene understanding, although their representation learning has generally been treated separately. Powerful 2D architectures such as ResNets [21] and Vision Transformers (ViT) [14] have achieved notable success in various 2D recognition and segmentation tasks, but focus on learning from 2D image data. Current large-scale RGB-D datasets [1,4,10,34,35] provide an opportunity to learn key geometric and structural priors to provide more informed reasoning about the scale and cir-

cumvent view-dependent effects, which can provide more efficient representation learning. In 3D, various successful methods have been leveraging the RGB-D datasets for constrastive point discrimination [6, 24, 39, 44] for downstream 3D tasks, including high-level scene understanding tasks as well as low-level point matching tasks [15, 42, 43]. However, the other direction from 3D to 2D is less explored.

We thus aim to embed such 3D priors into 2D backbones to effectively learn the structural and geometric priors underlying the 3D scenes captured in 2D image projections. Recently, Pri3D [25] adopted similar multi-view and reconstruction-based constraints to induce 3D priors in learned 2D representations. However, this relies on not only acquiring RGB-D frame data but also the robust registration of multiple views to obtain camera pose information for each frame. Instead, we consider how to effectively learn such geometric priors from only single-view RGB-D data in a more broadly applicable setting for 3D-based pre-training.

We thus propose Mask3D, which learns effective 3D priors for 2D backbones in a self-supervised fashion by pre-training with single-view RGB-D frame data. We propose a pre-text reconstruction task to reconstruct the depth map by masking different random RGB and depth patches of an input frame. These masked input RGB and depth are encoded simultaneously in separate encoding branches and decoded to reconstruct the dense depth map. This imbues 3D priors into the RGB backbone which can then be used for fine-tuning downstream image based scene understanding tasks.

In particular, our self-supervised approach to embedding 3D priors from single-view RGB-D data to 2D learned features is not only more generally applicable, but we also demonstrate that it is particularly effective for pre-training vision transformers. Our experiments demonstrate the effectiveness of Mask3D on a variety of datasets and image understanding tasks. We pre-train on ScanNet [10] with our masked 3D pre-training paradigm and fine-tune for 2D semantic segmentation, instance segmentation, and object detection. This enables notable improvements not only on ScanNet data but also generalizes to NYUv2 [34] and even Cityscapes [8] data. We believe that Mask3D makes an important step to shed light on the paradigm of incorporating 3D representation learning to powerful 2D backbones.

In summary, our contributions are:

- We introduce a self-supervised pre-training approach to learn masked 3D priors for 2D image understanding tasks based on learning from only single-view RGB-D data, without requiring any camera pose or 3D reconstruction information, and thus enabling more general applicability.
- We demonstrate that our masked depth reconstruction pre-training is particularly effective for the modern, powerful ViT architecture, across a variety of datasets and image understanding tasks.

2. Related Work

Pre-training in Visual Transformers. Recently, visual transformers have revolutionized computer vision and attracted wide attention. In contrast to popular CNNs that operate in a sliding window fashion, Vision Transformers (ViT) describe the image as patches of 16x16 pixels. The Swin Trasnformer [28] has set new records with its hierarchical transformer formulation on major vision benchmarks. The dominance of visual transformers in many vision tasks has inspired study into how to pre-training such backbones. MoCoV3 [5] first investigated the effects of several fundamental components for self-supervised ViT training. MAE [19] then proposed an approach inspired by BERT [13], which randomly masks words in sentences and leveraged masked image reconstruction for self-supervised pre-training that achieved state-of-the-art results in ViT. A similar self-supervision has also been proposed by MaskFeat [37] for self-supervised video pre-training. MaskFeat randomly masks out pixels of the input sequence and then predicts the Oriented Gradients (HOG) of the masked regions. However, such ViT pre-training methods focus on image or video data, without exploring how 3D priors can potentially be exploited. MultiMAE [2] on the other hand introduces depth priors. However, it requires depth as input not only in pre-training but also in downstream tasks. In addition to depth, human annotations (e.g., semantics) are also leveraged in the pre-training. To achieve a self-supervised pre-training, we do not use semantics in the pre-training and only use RGB images as input in downstream tasks.

RGB-D Scene Understanding. Research in 3D scene understanding have been spurred forward with the introduction of larger-scale, annotated real-world RGB-D datasets [1, 4, 10]. This has enabled data-driven semantic understanding of 3D reconstructed environments, where we have now seen notable progress, such as for 3D semantic segmentation [7, 11, 17, 31, 32, 36], object detection [29, 30, 45], instance segmentation [16, 18, 22, 23, 26, 27, 40, 41], and recently panoptic segmentation [9]. Such 3D scene understanding tasks have been analogously defined to 2D image understanding, which considers RGB-only input without depth information. However, learning from 3D enables geometric reasoning without requiring learning view-dependent effects or resolving depth/scale ambiguity that must be learned when considering solely 2D data. We thus take advantage of existing large-scale RGB-D data to explore how to effectively embed 3D priors for better representation learning for 2D scene understanding tasks.

Embedding 3D Priors in 2D Backbones. Learning cross-modality features has been seen in extensive studies of the ties between languages and images. In particular, CLIP [33] learns visual features from language supervision during pre-training, showing promising results in zero-

shot learning for image classification. Pri3D [25] explores 3D-based pre-training for image-based tasks by leveraging multi-view consistency and 2D-3D correspondence with contrastive learning to embed 3D priors into ResNet backbones. This results in enhanced features over ImageNet pre-training on 2D scene understanding tasks. However, Pri3D requires camera pose registration across RGB-D video sequences and is specifically designed for CNNs-based architectures. In contrast, we formulate a self-supervised pre-training that operates on only single-view RGB-D frames and leverages masked 3D priors that can effectively pre-train powerful ViT backbones.

3. Method

We introduce Mask3D to embed 3D priors into learned 2D representations by self-supervised pre-training from only single-view RGB-D frames. To effectively learn 3D structural priors without requiring any camera pose information or multi-view constraints, we formulate a pre-text depth reconstruction task to inform the RGB feature extraction to be geometrically aware. Randomly masked color and depth images are used as input to reconstruct the dense depth map, and the RGB backbone can then be used to fine-tune downstream image understanding tasks. In particular, we show in Sec. 4 that this single-frame self-supervision is particularly well-suited for powerful vision transformer (ViT) backbones, even without any multi-view information.

3.1. Learning Masked 3D Priors

We propose to learn masked 3D priors to embed to learned 2D backbones by pre-training to reconstruct dense depth from RGB images with the guidance of sparse depth. That is, for an RGB-D frame $F = (C, D)$ with RGB image C and depth map D, we train to reconstruct D from masked patches of C guided with sparse masked patches of D. An overview of our approach is shown in Fig. 2.

To create masked color and depth M_c and M_d from C and D as input for reconstruction, a 240x320 RGB image C is uniformly divided into 300 16x16 patches, from which we randomly keep a percentage p_c of patches, masking out the others, to obtain M_c. M_d is created similarly by keeping only a percentage p_d of patches, such that the resulting depth patches do not coincide with the RGB patches in M_c.

We then train color and depth encoders Ψ_c and Ψ_d to separately encode RGB and depth signals. RGB patches are fed into Ψ_c and concatenated with a positional embedding, following the ViT architecture, and similarly for depth. The positional embedding used encodes the patch location by a cosine function. Patches and their positional embeddings are then mapped into higher dimensional feature vectors via Ψ_c and Ψ_d. The encoders Ψ_c and Ψ_d are built by blocks composed of linear and norm layers. The features from Ψ_c and Ψ_d are then fused in the bottleneck; since depth patches were selected in regions where no RGB patches were selected, there are no duplicate patches representing the same patch location.

For those regions which do not have any associated RGB or depth patch, we use patches of constant values as mask tokens to create a placeholder in the bottleneck to enable reconstructing dense depth at the original image resolution. In the bottleneck, the RGB and depth patch feature vectors, along with the mask tokens, form the input to the decoder. This formulates a reconstruction task from sparse RGB and depth; the joint RGB-D pre-training enables reconstruction from very sparse input, as shown by our ablation on the masked input rations in Sec. 4.5. Note that the depth encoder is trained only during pre-training, and only the color ViT encoder (and decoder, if applicable) are used for downstream fine-tuning.

To demonstrate the effectiveness of the pre-training task, we demonstrate the depth completion results from the pre-training phase in Fig. 6. A detailed analysis of masking different ratios of color and depth signals is shown in Sec. 4.5.

Pre-training Loss In contrast to the widely used contrastive loss in 3D representation learning, we train for dense depth reconstruction with an ℓ_2 reconstruction loss. Similar to MAE [19], we normalize the output patches as well as the target patches prior to computing the loss, which we found to empirically improve the performance.

4. Results

We demonstrate the effectiveness of Mask3D pre-training for ViT [14] backbones on semantic segmentation, instance segmentation, and object detection tasks. We pre-train on ScanNet [10] data and demonstrate the effectiveness of learned masked 3D priors for not only ScanNet downstream tasks but also their transferability to NYUv2 [34] and even across the indoor/outdoor domain gap to Cityscapes [8] data.

4.1. Experimental Setup

In this section, we introduce the pre-training and fine-tuning procedures. Our method uses a two-stage pre-training design introduced in the following.

Stage-I: Mask3D Encoder Initialization. We initialize the RGB encoder with network weights trained on ImageNet [12] (as pre-training for our pre-training). To maintain a fully self-supervised pre-training paradigm, we initialize with weights obtained by self-supervised ImageNet pre-training [19].

Stage-II: Mask3D Pre-training on ScanNet. Mask3D pre-training leverages 3D priors in RGB-D frame data, for which we use the color and depth maps of ScanNet [10]. Note that this does not use any semantic or reconstruction

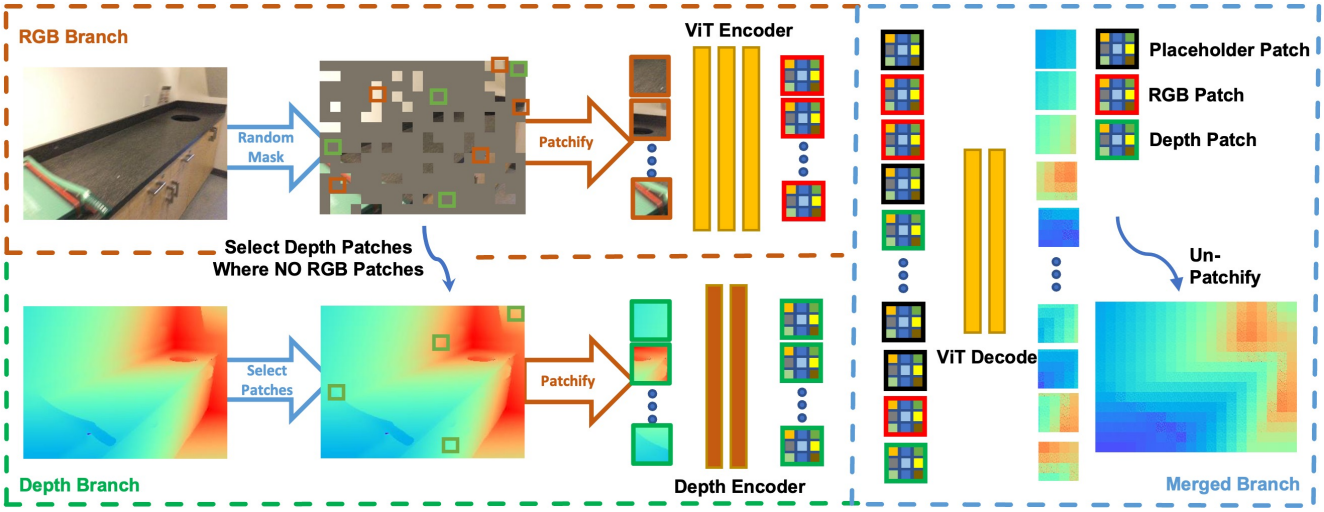

Figure 2. **Overview of Mask3D Pre-training.** As a pretext task, we predict dense depth from color and sparse depth signals. We use masked input by randomly selecting a set of patches from the input color image, which are then mapped to higher dimensional feature vectors; input depth is similarly masked and encoded. The color and depth features are then fused into a bottleneck from which dense depth is reconstructed as a self-supervised loss.

information during pre-training. ScanNet contains 2.5M RGB-D frames from 1513 ScanNet train video sequences. We regularly sample every 25$^{\text{th}}$ frame without any other filtering (e.g., no control on viewpoint variation).

Downstream Fine-tuning. We evaluate our Mask3D pre-training by fine-tuning a variety of downstream image understanding tasks (semantic segmentation, instance segmentation, object detection). We consider in-domain transfers on ScanNet image understanding, and further evaluate the out-of-domain transfer on datasets with different statistical characteristics: the indoor image data of NYUv2 [34], as well as across a strong domain gap to the outdoor image data of Cityscapes [8]. For semantic segmentation tasks, we use both encoder and decoder pre-trained with Mask3D, and for instance segmentation and detection, only the backbone encoder is pre-trained.

Baselines. To evaluate the effectiveness of our learned masked 3D priors for 2D representations, we benchmark our method against relevant baselines:

Supervised ImageNet Pre-training (supIN). This uses the pre-trained weights from ImageNet, provided by torchvision, as is commonly used for image understanding tasks. Here, only ImageNet data is used, and no ScanNet data is involved in the pre-training phase.

2-Stage MoCoV2 (MoCoV2-supIN→SN). Supervised ImageNet pre-trained (supIN) weights are used as network initialization for pre-training. MoCoV2 [20] is used for pre-training with randomly shuffled ScanNet images. Here, both ImageNet and ScanNet image data are used without any geometric priors.

2-Stage MAE (MAE-unsupIN→SN). Self-supervised ImageNet pre-trained weights are used as network initialization for pre-training. MAE [19] is used for pre-training with randomly shuffled ScanNet images. Here, both ImageNet and ScanNet image data are used without any geometric priors.

Pri3D [25]. Supervised ImageNet pre-trained are used to initialize Pri3D pre-training, which leverages multi-view and reconstruction constraints from ScanNet data under a contrastive loss. Here, both ImageNet and ScanNet data are used, incorporating 3D priors from reconstructed RGB-D video sequences for pre-training.

Implementation Details. We use a ViT-B backbone to train our approach. For pre-training, we use an SGD optimizer with a learning rate of 0.1 and an effective batch size of 128 (accumulated gradients from an actual batch size of 64). The learning rate is decreased by a factor of 0.99 every 1000 steps, and our method is trained for 100 epochs. Fine-tuning on semantic segmentation is trained with a batch size of 8 for 80 epochs. The initial learning rate is 0.01, with polynomial decay with a power of 0.9. Fine-tuning on detection and instance segmentation is trained using Detectron2 [38] with the 1x schedule. Pre-training experiments are conducted on a single NVIDIA A6000 GPU, or 2 NVIDIA RTX3090 GPUs, or 4 NVIDIA RTX2080Ti GPUs; semantic segmentation experiments are conducted on a single NVIDIA A6000 GPU; instance segmentation and detection experiments are conducted on 8 V100 GPUs.

4.2. ScanNet Downstream Tasks

We demonstrate the effectiveness of representation learning with 3D priors via downstream tasks on ScanNet [10] images. For fine-tuning, we follow the standard

Pre-training Method	Backbone	Pre-training Data	mIoU
Scratch	ResNet-50	None	39.1
ImageNet Pre-training (supIN)	ResNet-50	ImageNet	55.7
Supervised Pre-training	ViT	ImageNet+ScanNet	65.9 (+10.2)
MoCoV2-supIN→SN [20]	ResNet-50	ImageNet+ScanNet	56.6 (+0.9)
Pri3D [25]	ResNet-50	ImageNet+ScanNet	60.2 (+4.5)
Pri3D [25]	ViT	ImageNet+ScanNet	59.3 (+3.6)
DINO [3]	ViT	ImageNet+ScanNet	58.1 (+3.6)
MAE-unsupIN→SN [19]	ViT	ImageNet+ScanNet	63.3 (+7.6)
Ours – Mask3D (DINO)	ViT	ImageNet+ScanNet	60.5 (+4.8)
Ours – Mask3D (MAE)	ViT	ImageNet+ScanNet	**66.7** (+11.0)

Table 1. **ScanNet 2D Semantic Segmentation.** Mask3D significantly outperforms Pri3D as well as other state-of-the-art pre-training approaches that leverage both ImageNet and ScanNet data.

Pre-training Method	AP@0.5	AP@0.75	AP
Scratch	32.7	17.7	16.9
ImageNet Pretrain (supIN)	41.7	25.9	25.1
MoCoV2-supIN→SN [20]	43.5 (+1.8)	26.8 (+0.9)	25.8 (+0.7)
Pri3D [25]	43.7 (+2.0)	27.0 (+1.1)	26.3 (+1.2)
MAE-unsupIN→SN [19]	46.1 (+4.4)	32.7 (+6.8)	30.5 (+5.4)
Mask3D (Ours)	**50.4** (+8.7)	**35.3** (+9.4)	**32.7** (+7.6)

Table 2. **ScanNet 2D Object Detection.** Fine-tuning with Mask3D pre-trained models leads to improved object detection results across different metrics, in comparison to ImageNet pre-training, MoCo-style pre-training, and a strong MAE-style pre-training method.

Pre-training Method	AP@0.5	AP@0.75	AP
Scratch	25.8	13.1	12.2
ImageNet Pretrain (supIN)	32.6	17.8	17.6
MoCoV2-supIN→SN [20]	33.9 (+1.3)	18.1 (+0.3)	18.3 (+0.7)
Pri3D [25]	34.3 (+1.7)	18.7 (+0.9)	18.3 (+0.7)
MAE-unsupIN→SN [19]	37.4 (+4.8)	20.3 (+2.5)	20.7 (+3.1)
Mask3D (Ours)	**41.2** (+8.6)	**22.7** (+4.9)	**22.8** (+5.2)

Table 3. **ScanNet 2D Instance Segmentation.** Fine-tuning with Mask3D pre-trained models leads to improved instance segmentation results across different metrics compared to ImageNet pre-training, MoCo-style pre-training, and a strong MAE-style pre-training method.

Pre-training Method	AP@0.5	AP@0.75	AP
Scratch	17.2	9.2	8.8
ImageNet Pretrain (supIN)	25.1	13.9	13.4
MoCoV2-supIN→SN [20]	27.2 (+2.1)	14.7 (+0.2)	14.8 (+1.4)
Pri3D [25]	28.1 (+3.0)	15.7 (+1.8)	15.7 (+2.3)
MAE-unsupIN→SN [19]	33.6 (+8.5)	19.0 (+5.1)	19.0 (+5.6)
Mask3D (Ours)	**37.0** (+11.9)	**21.6** (+7.7)	**21.3** (+7.9)

Table 4. **NYUv2 2D Instance Segmentation.** Fine-tuning with Mask3D pre-trained models leads to improved instance segmentation results across different metrics compared to previous methods, demonstrating the cross-dataset transfer ability of Mask3D.

protocol of the ScanNet benchmark [10] and sample every 100^{th} frame, resulting in 20,000 train images and 5,000 validation images.

2D Semantic Segmentation. Tab. 1 shows the fine-tuning for semantic segmentation, in comparison with baseline pre-training approaches. All pre-training methods significantly improve performance over training the semantic segmentation model from scratch. In particular, Mask3D provides substantially better representation quality leading to a much stronger improvement over supervised ImageNet pre-training (+11 mIoU), and notably improving over MAE-unsupIN→SN with ImageNet and ScanNet (+3.4 mIoU) and the 3D-based pre-training of Pri3D (+6.5 mIoU). We note that the multi-view 3D pre-training of Pri3D does not effectively embed informative 3D priors to ViT backbones, rather suffering from performance degradation from a ResNet-50 backbone. In contrast, our Mask3D pre-training can notably improve performance with a ViT backbone, indicating the effectiveness of our learned 3D priors.

2D Object Detection and Instance Segmentation We show that Mask3D provides effective general 3D priors for a variety of image-based tasks, by evaluating downstream object detection and instance segmentation in Tab. 2 and Tab. 3, respectively. Across all tasks, various pre-training approaches yield substantial improvement over training from scratch. Our masked 3D prior learning transfers effectively learned representations for object detection and instance segmentation, notably improving over the best-performing MAE-unsupIN→SN (+4.3 AP@0.5 and +3.8 AP@0.5, respectively).

Data-Efficient Scenarios. We additionally show that our single-view RGB-D pre-training to embed 3D priors in limited data scenarios for downstream ScanNet semantic segmentation in Fig. 5. Mask3D shows consistent improvements across a range of limited data; even with only 20% of the training data, we recover 80% performance with 100% training data available and improving +15.2 mIoU over Pri3D pre-training on a ViT backbone.

4.3. NYUv2 Downstream Tasks

We demonstrate the generalizability of our 3D-imbued learned feature representations across datasets, us-

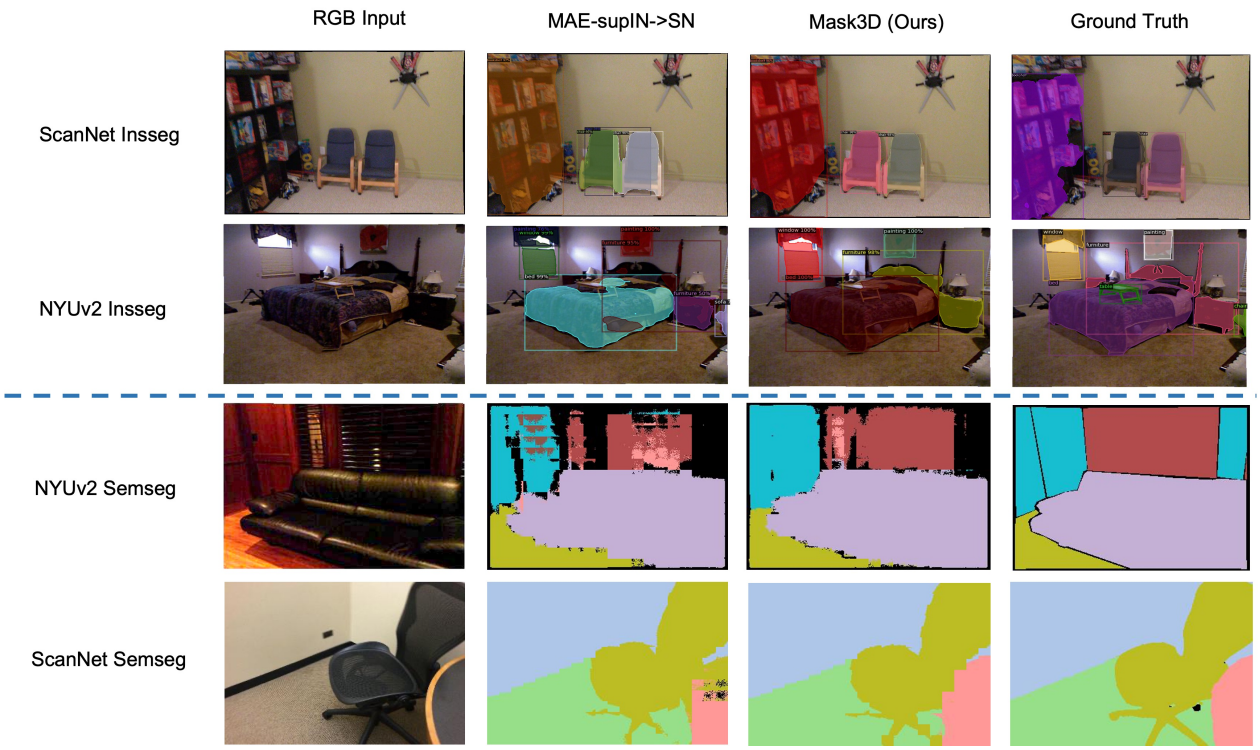

Figure 3. **Qualitative Results on Various Tasks across Different Benchmarks.** We visualize predictions on different tasks across various scene understanding benchmarks. From top to bottom rows: instance segmentation on ScanNet, instance segmentation on NYUv2, semantic segmentation on NYUv2, and semantic segmentation results in ScanNet.

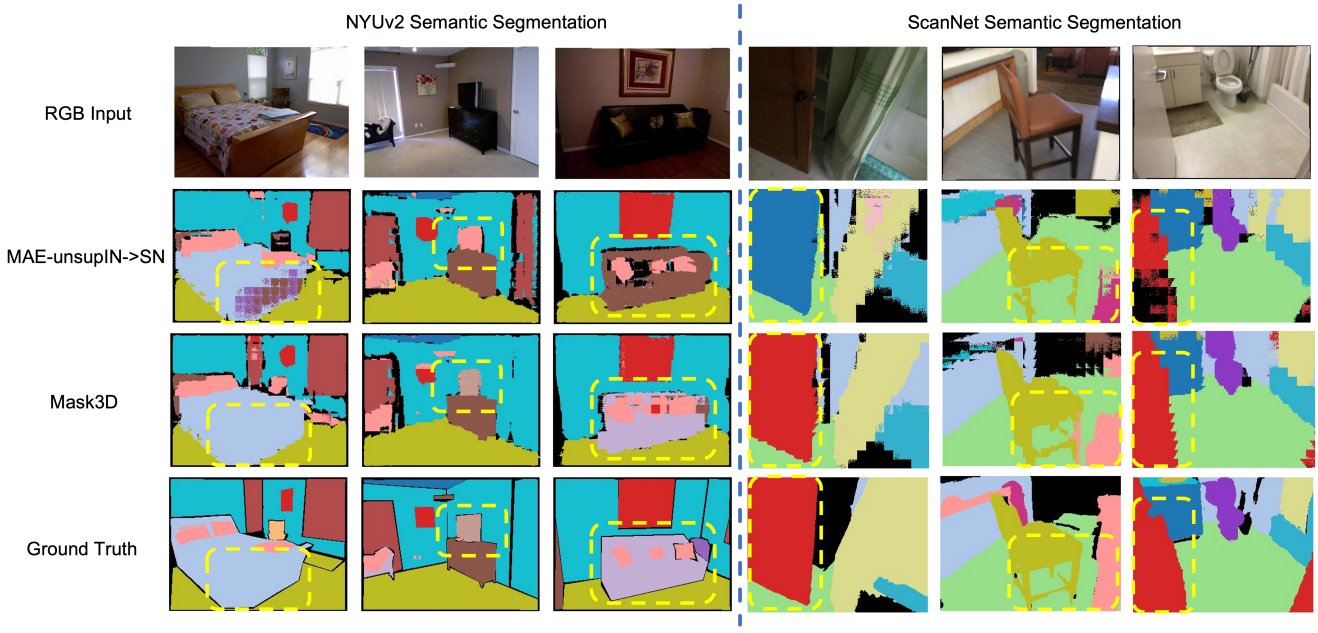

Figure 4. **More Qualitative Results on Semantic Segmentation.** We visualize semantic segmentation predictions on various scene understanding benchmarks including ScanNet and NYUv2.

ing Mask3D pre-trained on ScanNet and fine-tuned on NYUv2 [34] following the same fine-tuning setup as before. NYUv2 contains Microsoft Kinect RGB-D video sequences of indoor scenes, comprising 1449 densely labeled RGB images. We use the official 795/654 train/val split. Tables 5, 6, and 4 evaluate the downstream tasks

Pre-training Method	Backbone	Pre-training Data	mIoU
Scratch	ResNet-50	None	24.8
ImageNet Pre-training (supIN)	ResNet-50	ImageNet	50.0
Supervised Pre-training	ViT	ImageNet+ScanNet	55.5 (+5.5)
MoCoV2-supIN→SN [20]	ResNet-50	ImageNet+ScanNet	47.6 (−2.4)
Pri3D [25]	ResNet-50	ImageNet+ScanNet	54.2 (+4.2)
Pri3D [25]	ViT	ImageNet+ScanNet	53.2 (+3.2)
MAE-unsupIN→SN [19]	ViT	ImageNet+ScanNet	54.9 (+4.9)
Mask3D (Ours)	ViT	ImageNet+ScanNet	**56.9** (+6.9)

Table 5. **NYUv2 2D Semantic Segmentation.** Mask3D significantly outperforms state-of-the-art pre-training approaches, demonstrating its effectiveness in transferring to different dataset characteristics.

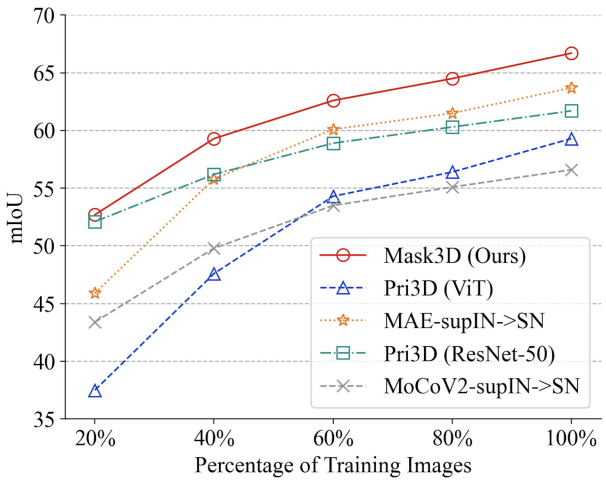

Figure 5. **Data-Efficient Results.** Compared to previous methods, Mask3D demonstrates consistent improvements on ScanNet 2D semantic segmentation across a range of limited data scenarios. Mask3D is particularly effective for ViT pre-training, improving +15.2% mIoU over state-of-the-art Pri3D [25] on a ViT backbone at 20% of the training data.

Pre-training Method	AP@0.5	AP@0.75	AP
Scratch	21.3	10.3	9.0
ImageNet Pretrain (supIN)	29.9	17.3	16.8
MoCoV2-supIN→SN [20]	30.1 (+0.20)	18.1 (+0.80)	17.3 (+0.50)
Pri3D [25]	33.0 (+2.10)	19.8 (+2.60)	18.9 (+2.10)
MAE-unsupIN→SN [19]	40.3 (+10.4)	24.5 (+7.20)	23.2 (+6.40)
Mask3D (Ours)	**44.0** (+14.1)	**28.3** (+6.40)	**25.9** (+9.10)

Table 6. **NYUv2 2D Object Detection.** Fine-tuning with Mask3D pre-trained models leads to improved object detection results across different metrics, showing an effective transfer across dataset characteristics.

of 2D semantic segmentation, object detection, and instance segmentation, respectively. Across all three tasks on NYUv2 data, our Mask3D pre-training achieves notably improved performance than training from scratch as well as the various baseline pre-training methods. In particular, we achieve an improvement of +6.9 mIoU, +14.1 AP@0.5, and +11.9 AP@ 0.5 over the common supervised ImageNet pre-training on semantic segmentation, object detection, and instance segmentation.

Pre-training Method	Backbone	mIoU
ImageNet Pre-training (supIN)	ResNet-50	54.1
Pri3D [25]	ResNet-50	55.1 (+1.00)
MAE-unsupIN→SN [19]	ViT	64.7 (+10.6)
Mask3D (Ours)	ViT	**66.4** (+12.3)

Table 7. **Cityscapes 2D Semantic Segmentation.** Mask3D significantly outperforms state-of-the-art Pri3D as well as a strong MAE-style pre-training. This demonstrates the effectiveness of the transferability of Mask3D, even under a significant domain gap.

4.4. Out-of-domain Transfer

While Mask3D concentrates on pre-training for improving indoor scene understanding, we further demonstrate the effectiveness of our Mask3D pre-training for the out-of-domain transfer on outdoor data, such as Cityscapes [8]. We use the official data split of 3000 images for training and 500 images for the test. To evaluate the transferability in such a large domain gap scenario, we fine-tune the pre-trained models for the 2D semantic segmentation task in Tab. 7. Our approach maintains performance improvement over baseline pre-training methods such as Pri3D (+11.3 mIoU) and MAE-unsupIN→SN (+1.7 mIoU). This indicates an encouraging transferability of our learned representations and their applicability to a variety of scenarios. Please refer to the supplemental material for more out-of-domain transfer results on more generally distributed data, such as ADE20K [46].

4.5. Ablation Studies

Does the pre-training masking ratio matter? We show how different masking ratios influence downstream task results in Tab. 8 on ScanNet semantic segmentation. We found a performance gain when masking more RGB values (keeping 20%), which in combination with the heavy depth masking (keeping 20%) leads to the best performance.

What about other ViT variants? In our experiments, we use ViT-B as the meta-architecture. We show Mask3D also works in other ViT variants, such as ViT-L (see Tab. 12), which exhibits a similar trend of improvements.

Does the normalization in the reconstruction loss help? We normalize the features when computing the reconstruc-

RGB Ratio	Depth Ratio	mIoU
20.0%	0.0%	65.2
20.0%	20.0%	**66.7**
20.0%	80.0%	65.5
50.0%	20.0%	65.9
50.0%	50.0%	64.7
80.0%	20.0%	64.8
80.0%	50.0%	64.8
100.0%	0.0%	64.6
100.0%	20.0%	64.8
100.0%	100.0%	64.5

Table 8. **Ablation Study of Masking Ratios.** on ScanNet 2D semantic segmentation. We mask out different ratios of RGB and depth patches, where the ratio indicates the percentage of kept patches. Refer to supplemental material for a full list.

tion loss and observe an improvement of +0.8% mIoU in the semantic segmentation task on ScanNet.

Method	Backbone	mIoU
Train from Scratch	ViT	32.6
MAE	ViT	37.1
Mask3D	ViT	42.2

Table 9. **Results on ScanNet Semantic Segmentation without ImageNet pre-training.** Similar trend is seen as ImageNet pre-training. The Gap gets larger compared to ImageNet pre-training.

Compared to a pure depth prediction baseline. In Tab. 8, we demonstrate a superior performance with a 20% kept patches of RGB and depth, compared to a pure depth prediction method (66.7 vs. 64.6). Note in the table, pre-training with 100% RGB ratio and 0% depth ratio is equivalent to a pure depth prediction from a RGB image.

Color + depth reconstruction? We found that having joint losses on color and depth during pre-training does not benefit performance (see the following Tab. 10). The RGB reconstruction loss potentially makes pre-training easier, as we already have additional depth priors as guidance.

Method	Reconstruction	mIoU
Mask3D	RGB+Depth	65.6
Mask3D	Depth	66.7

Table 10. **ScanNet Semantic Segmentation.** RGB as an additional signal does not bring a significant improvement.

No Stage-I pre-training. We observe a performance drop without ImageNet pre-training model as initialization for our pre-training. Since ImageNet pre-training is readily available and ScanNet has a relatively small amount of indoor data, we make ImageNet pre-training initialization as default, similar to Pri3D. Meanwhile, we conduct experiments without ImageNet pre-training in Tab. 9, and observe similar trends as when using ImageNet pre-training.

RGB + semantic segmentation as pre-training. Using RGB and semantic segmentation for pre-training rather than depth completion achieved competitive results on ScanNet semantic segmentation, although this requires the use of semantic labels for the pre-training dataset, and is likely to be less transferable across domains than using depth completion. As shown in the following Tab. 11, the gap becomes larger when transferring to both NYUv2 and Cityscapes.

Datasets	Mask3D - Semantics	Mask3D
ScanNet	65.9	66.7
NYUv2	55.5	56.9
CityScapes	63.0	66.4

Table 11. Semantic segmentation results (mIoU). "Mask3D - Semantics" denotes pre-training using RGB+Semantics.

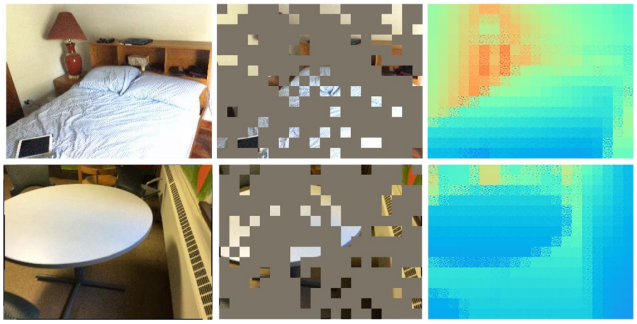

Figure 6. **Pre-trained ViT learns 3D structual priors.** Our proposed pre-training method learns spatial structures from heavily masked RGB images.

Pre-training Method	Backbone	mIoU
Pri3D [25]	ResNet-50	60.2
Pri3D [25]	ViT-B	59.3 (-0.9)
MAE-unsupIN→SN [19]	ViT-B	63.3 (+3.1)
Mask3D (Ours)	ViT-B	66.7 (+6.5)
Pri3D [25]	ViT-L	64.3 (+4.1)
MAE-unsupIN→SN [19]	ViT-L	68.2 (+8.0)
Mask3D (Ours)	ViT-L	**70.7** (+10.5)

Table 12. **ViT Variants on ScanNet 2D Semantic Segmentation.** Mask3D yields consistent improvements for both ViT-B and ViT-L backbone architectures.

5. Conclusion

In this paper, we present Mask3D, a new self-supervised approach to embed 3D priors into learned 2D representations for image scene understanding. We leverage existing large-scale RGB-D data to learn 3D priors without requiring any camera pose or multi-view correspondence information, instead learning geometric and structural cues through a pre-text reconstruction task from masked color and depth. We show that Mask3D is particularly effective in pre-training the modern, powerful ViT backbones, with notable improvements across a variety of image-based tasks and datasets. We believe this shows the strong potential in effectively learning 3D priors and provides new avenues for such 3D-grounded representation learning.

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *ICCV*, 2016. 1, 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer, 2022. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 5

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2

[6] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. *arXiv preprint arXiv:2112.02990*, 2021. 2

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3, 4, 7

[9] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[10] 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. 1, 2, 3, 4, 5

[11] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[15] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6433–6442, 2021. 2

[16] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *CVPR*, 2020. 2

[17] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 2

[18] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. OccuSeg: Occupancy-aware 3D instance segmentation. In *CVPR*, 2020. 2

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 3, 4, 5, 7, 8

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4, 5, 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[22] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *CVPR*, 2019. 2

[23] Ji Hou, Angela Dai, and Matthias Nießner. RevealNet: Seeing Behind Objects in RGB-D Scans. In *CVPR*, 2020. 2

[24] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 2

[25] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *ICCV*, 2021. 2, 3, 4, 5, 7, 8

[26] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *CVPR*, 2020. 2

[27] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *ICCV*, 2019. 2

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[29] Yinyu Nie, Ji Hou, Xiaoguang Han, and Matthias Nießner. Rfd-net: Point scene understanding by semantic instance reconstruction. In *CVPR*, 2021. 2

[30] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3D object detection in point clouds. *ICCV*, 2019. 2

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. *CVPR*, 2017. 2

[32] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 2

[33] 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. 2

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. *ECCV*, 2012. 1, 2, 3, 4, 6

[35] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *CVPR*, 2015. 1

[36] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KPConv: Flexible and deformable convolution for point clouds. In *CVPR*, 2019. 2

[37] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 2

[38] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[39] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas J Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3D point cloud understanding. *ECCV*, 2020. 2

[40] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3D instance segmentation on point clouds. In *NeurIPS*, 2019. 2

[41] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas Guibas. GSPN: Generative shape proposal network for 3D instance segmentation in point cloud. In *CVPR*, 2019. 2

[42] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. *arXiv preprint arXiv:2303.08231*, 2023. 2

[43] Yu Zhang, Junle Yu, Xiaolin Huang, Wenhui Zhou, and Ji Hou. Pcr-cg: Point cloud registration via deep explicit color and geometry. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 443–459. Springer, 2022. 2

[44] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2

[45] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 2

[46] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 7