Pri3D: Can 3D Priors Help 2D Representation Learning?

Ji Hou[1] Saining Xie[2] Benjamin Graham[2] Angela Dai[1] Matthias Nießner[1]
[1]Technical University of Munich [2]Facebook AI Research

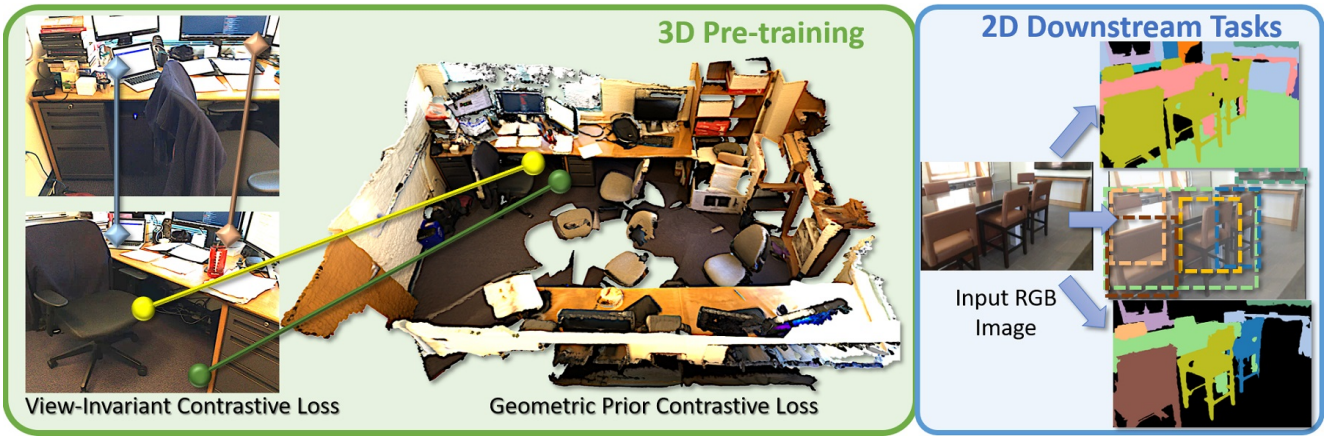

Figure 1: Pri3D leverages 3D priors for downstream 2D image understanding tasks: during pre-training, we incorporate view-invariant and geometric priors from color-geometry information given by RGB-D datasets, imbuing geometric priors into learned features. We show that these 3D-imbued learned features can effectively transfer to improved performance on 2D tasks such as semantic segmentation, object detection, and instance segmentation.

Abstract

Recent advances in 3D perception have shown impressive progress in understanding geometric structures of 3D shapes and even scenes. Inspired by these advances in geometric understanding, we aim to imbue image-based perception with representations learned under geometric constraints. We introduce an approach to learn view-invariant, geometry-aware representations for network pre-training, based on multi-view RGB-D data, that can then be effectively transferred to downstream 2D tasks. We propose to employ contrastive learning under both multi-view image constraints and image-geometry constraints to encode 3D priors into learned 2D representations. This results not only in improvement over 2D-only representation learning on the image-based tasks of semantic segmentation, instance segmentation and object detection on real-world indoor datasets, but moreover, provides significant improvement in the low data regime. We show significant improvement of 6.0% on semantic segmentation on full data as well as 11.9% on 20% data against baselines on ScanNet. Our code is open sourced at https://github.com/Sekunde/Pri3D.

1. Introduction

In recent years, we have seen rapid progress in learning-based approaches for semantic understanding of 3D scenes, particularly in the tasks of 3D semantic segmentation, 3D object detection, and 3D semantic instance segmentation [40, 8, 51, 28, 22, 17, 12, 29, 38]. Such approaches leverage geometric observations, exploiting the representation of points [40, 41], voxels [8, 22], or meshes [28] to obtain accurate 3D semantics. These have shown significant promise towards realizing applications such as depth-based scene understanding for robotics, as well as augmented or virtual reality. In parallel to the development of such methods, the availability of large-scale RGB-D datasets [46, 27, 3, 7], has further accelerated the research in this area.

One advantage of learning directly in 3D in contrast to learning solely from 2D images is that methods operate in metric 3D space; hence, it is not necessary to learn view-dependent effects and/or projective mappings. This allows training 3D neural networks from scratch in a relatively short time frame and typically requires a (relatively) small number of training samples; e.g., state-of-the-art 3D neural networks can be trained with around 1000 scenes from ScanNet. Our main idea is to leverage these advantages in

the form of 3D priors for image-based scene understanding.

Simultaneously, we have seen tremendous progress on representation learning in the image domain, mostly powered by the success of recent contrastive learning based methods [54, 18, 4, 15, 2]. The exploration in 2D representation learning heavily relies on the paradigm of instance discrimination, where different augmented copies of the same instance are drawn closer. Different invariances can be encoded from those low-level augmentations such as random cropping, flipping and scaling, as well as color jittering. However, despite the common belief that 3D view-invariance is an essential property for a capable visual system [33], there remains little study linking the 3D priors and 2D representation learning. The goal of our work is to explore the combination of contrastive representation learning with 3D priors, and offer some preliminary evidence towards answering an important question: can 3D priors help 2D representation learning?

To this end, we introduce Pri3D, which aims to learn with 3D priors in a pre-training stage and subsequently use them as initialization for fine-tuning on image-based downstream tasks such as semantic segmentation, detection, and instance segmentation. More specifically, we introduce geometric constraints to a contrastive learning scheme, which are enabled by multi-view RGB-D data that is readily available. We propose to exploit geometric correlations through implicit multi-view constraints between different images through the correspondence of pixels which correspond to the same geometry, as well as explicit correspondence of geometric patches which correspond to image regions. This imbues geometric knowledge into the learned representations of the image inputs which can then be leveraged as pre-trained features for various image-based vision tasks, particularly in the low training data regime.

We demonstrate our approach by pre-training on ScanNet [7] under these geometric constraints for representation learning, and show that such self-supervised pre-training (i.e., no semantic labels are used) results in improved performance on 2D semantic segmentation, instance segmentation and detection tasks. We demonstrate this not only on ScanNet data, but also generalizing to improved performance on NYUv2 [46] semantic segmentation, instance segmentation and detection tasks. Moreover, leveraging such geometric priors for pre-training provides robust features which can consistently improve performance under a wide range of amount of training data available. While we focus on indoor scene understanding in this paper, we believe our results can shed light on the the paradigm of representation learning with 3D priors and open new opportunities towards more general 3D-aware image understanding.

In summary, our contributions are:

- A first exploration of the effect of 3D priors for 2D image understanding tasks, where we demonstrate the benefit of 3D geometric pre-training towards complex 2D perception such as semantic segmentation, object detection, and instance segmentation.
- A new pre-training approach based on 3D-guided view-invariant constraints and geometric priors from color-geometry correspondence, which learns features that can be transferred to 2D representations, complementing and improving image understanding across multiple datasets.

2. Related Work

3D Scene Understanding. Research in 3D scene understanding has recently been spurred forward with the introduction of larger-scale, real-world 3D scanned scene datasets [1, 7, 3, 13]. We have seen notable progress in development of methods for semantic segmentation [40, 41, 51, 52, 8, 26, 31, 56, 28, 59], object detection [48, 49, 38, 39, 37, 61, 34], and instance segmentation [22, 58, 57, 30, 23, 12, 17, 29] in 3D. In particular, the introduction of sparse convolutional neural networks [14, 6] have presented a computationally-efficient paradigm producing state-of-the-art results in such tasks. Inspired by the developments in 3D scene understanding, we introduce learned geometric priors to representation learning for image-based vision tasks, leveraging a sparse convolutional backbone for 3D features used during pre-training.

In the past year, we have also seen new developments in 3D representation learning. PointContrast [55] first showed that unsupervised, contrastive-based pre-training improves performance across various 3D semantic understanding tasks. Hou et al. [24] introduces spatial context into 3D contrastive pre-training, resulting in improved performance in 3D limited annotation and data scenarios. Zhang et al. [60] introduces a instance-discrimination-style pre-training approach that directly operates on depth frames. Our approach bridges these concepts into feature learning that can be transferred to 2D image understanding tasks.

2D Contrastive Representation Learning. Representation learning has driven significant efforts in deep learning; on the image domain, pre-training a network on a rich set of data has been shown to improve performance in fine-tuning for a smaller target dataset for various applications. In particular, the contrastive learning framework [16] to learn representations from similar/dissimilar pairs of data has been demonstrated to show incredible promise [36, 21, 54, 18, 4, 5, 15, 2]. Notably, using an instance discrimination task in which positive pairs are created with data augmentation, MoCo [18] shows that unsupervised pre-training can surpass various supervised counterparts in detection and segmentation tasks, and SimCLR [4] further reduces the gap to supervised pre-training in linear classifier performance. Our approach leverages multi-view geomet-

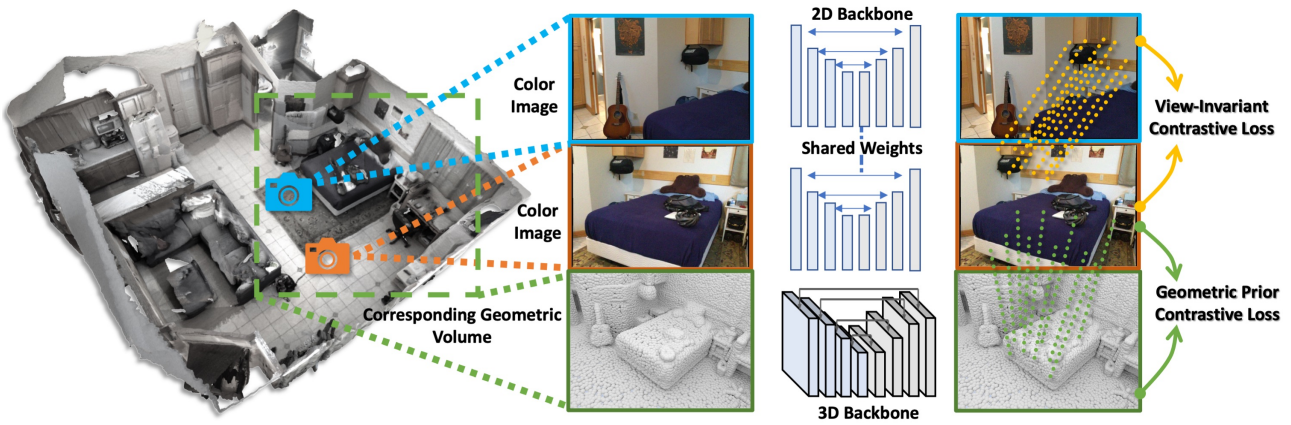

Figure 2: **Method Overview.** During pre-training, we use geometric constraints from RGB-D reconstructions to learn 3D priors for image-based representations. Specifically, we propose a contrastive learning formulation that models multi-view correspondences (View-Invariant Contrastive Loss) as well as geometry-to-image alignments (Geometric Prior Contrastive Loss). Our Pri3D pre-training strategy embeds geometric priors into the learned representations (in a form of pre-trained 2D convolutional network weights) that can be further leveraged for downstream 2D-only image understanding tasks.

ric information to augment contrastive learning and imbue robust geometric priors into learned feature representations.

Multi-Modality Learning CLIP [42] firstly proposes to train on images but with natural language supervision, and achieves significant results on zero-shot learning. BPNet [25] proposes a bidirectional projection module to mutually leverage 2D and 3D information for semantic segmentation task. 3D-to-2D Distillation [32] introduces additional 3D network in the training phase to embed 3D features for 2D semantic segmentation task. Existing works need to modify networks or add fusion modules in the training and/or inference phases. To this end, our method is more flexible as our pre-trained weights can be directly used like the ImageNet pre-trained model without any further modules or 3D/NLP data in the downstream tasks.

Correspondences Matching Schmidt et al. [44] advocates a new approach to learning visual descriptors for dense correspondence estimation for the re-localization purpose, e.g., in the SLAM context. Schuster et al. [45] presents a robust, unified descriptor network leveraging stacked dilated convolutions (SDC) for larger receptive field to better estimate dense pixel matching. HumanGPS [50] estimates dense correspondences between human images under arbitrary camera viewpoints and body poses. Existing works focus on 2D-2D correspondences matching problem itself. Our approach uses 2D-3D as well as 2D-2D view-invariant correspondences matching as pretext task to embed 3D priors for 2D downstream tasks.

3. Learning Representations from 3D Priors

In this section, we introduce Pri3D; our key idea is to leverage constraints from RGB-D reconstructions, now readily available in various datasets [13, 47, 7, 3], to embed 3D priors in image-based representations. From a dataset of RGB-D sequences, each sequence consists of depth and color frames, $\{D_i\}$ and $\{C_i\}$, respectively, as well as automatically-computed 6-DoF camera pose alignments $\{T_i\}$ (mapping from each camera space to world space) from state-of-the-art SLAM, all resulting in a reconstructed 3D surface geometry S. Specifically, we observe that multi-view constraints can be exploited in order to learn view-invariance without the need of costly semantic labels. In addition, we learn features through geometric representations given by the obtained geometry in RGB-D scans, again, without the need of human annotations. For both, we use state-of-the-art contrastive learning in order to constrain the multi-modal input for training. We show that these priors can be embedded in the image-based representations such that the learned features can be used as pre-trained features for purely image-based perception tasks; i.e., we can perform tasks such as image segmentation or instance segmentation on a single RGB image. An overview of our approach is shown in Figure 2.

3.1. View-Invariant Learning

In 2D constrative pre-training algorithms, a variety of data augmentations are used for finding positive matching pairs, such as MoCo [18] and SimCLR [4]. For instance, they use random crops as self-supervised constraints within the same image for positive pairs, and correspondences to crops from other images as negative pairs. Our key idea is that with the availability of 3D data for training, we can leverage geometric knowledge to provide matching constraints between multiple images that see the same points. To this end, we use the ScanNet RGB-D dataset [7] which provides a sequence of RGB-D images with camera poses

computed by a state-of-the-art SLAM method [9], and reconstructed surface geometry S [35]. Note that both the pose alignments and the 3D reconstructions were obtained in a fully-automated fashion without any user input.

For a given RGB-D sequence in the train set, our method then leverages the 3D data to finding pixel-level correspondences between 2D frames. We consider all pairs of frames (i, j) from the RGB-D sequence. We then back-project frame i's depth map D_i to camera space, and transform the points into world space by T_i. The depth values of frame j are similarly transformed into world space. Pixel correspondences between the two frames are then determined as those whose 3D world locations lie within 2cm of each other (see Figure 3). We use the pairs of frames which have at least 30% pixel overlap, with overlap computed as number of corresponding pixels in both frames divided by total number points in the two frames. In total, we sample around 840k pairs of images from the ScanNet training data.

In the training phase, a pair of sampled images is input to a shared 2D network backbone. In our experiments, we use a UNet-style [43] backbone with ResNet [20] architecture as an encoder, but note that our method is agnostic to the underlying encoder backbone. We then consider the feature map from decoder of the 2D backbone, where its size is half of the input resolution. For each image in the pair, we use the aforementioned pixel-to-pixel correspondences which refer to the same physical 3D point. Note that these correspondences may have different color values due to view-dependent lighting effects but represent the same 3D world location; additionally, the regions surrounding the correspondences appear different due to different viewing angles. In this fashion, we treat these pairs of correspondences as positive samples in contrastive learning; we use all non-matching pixels as negatives. Non-matching pixels are also defined within the set of correspondences. For a pair of frames with n pairs of correspondences as positive samples, we use all $n(n-1)$ negative pairs (each of n pixels from the first frame with each $n-1$ non-matching pixel from the second). Non-matching pixel-voxels are defined similarly but from a pair of frame and 3D chunk.

Between the features of matching and non-matching pixel locations, we then compute a PointInfoNCE loss [55], which is defined as:

$$\mathcal{L}_p = -\sum_{(a,b)\in M} \log \frac{\exp(\mathbf{f}_a \cdot \mathbf{f}_b/\tau)}{\sum_{(\cdot,k)\in M} \exp(\mathbf{f}_a \cdot \mathbf{f}_k/\tau)}, \quad (1)$$

where M is the set of pairs of pixel correspondences, and f represents the associated feature vector of a pixel in the feature map. By leveraging multi-view correspondences, we apply implicit 3D priors – without any explicit 3D learning, we imbue view-invariance in the learned image-based features.

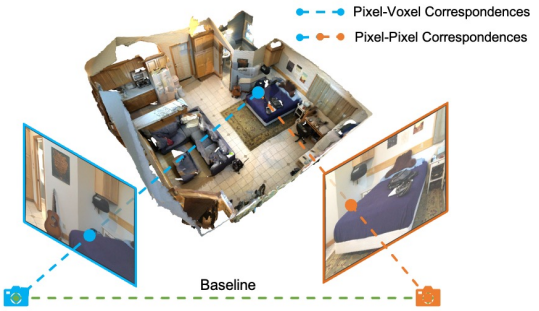

Figure 3: Illustration of finding correspondences between frames via epipolar geometry; world space as intermediary.

3.2. Geometric Prior

In addition to multi-view constraints, we also leverage explicit geometry-color correspondences inherent to the RGB-D data during training. For an RGB-D train sequence, the geometry-color correspondences are given by associating the surface reconstruction S with the RGB frames of the sequence. For each frame i, we compute its view frustum in the world space. A volumetric chunk V_i of S is then cropped from the axis-aligned bounding box of the view frustum. We represent V_i as a 2cm resolution volumetric occupancy grid from the surface. We thus consider pairs of color frames and geometric chunks (C_i, V_i).

From the color-geometry pairs (C_i, V_i), we compute pixel-voxel correspondences by projecting the depth values for each pixel in the corresponding frame D_i into world space to find an associated occupied voxel in V_i that lies within 2cm of the 3D location of the pixel.

During training, we leverage the color-geometry correspondences with a 2D network backbone and a 3D network backbone. We use a UNet-style [43] architecture with ResNet [20] encoder for the 2D network backbone, and a UNet-style sparse convolutional [14, 6] 3D network backbone. Similarly to view-invariant training, we also take the output from the decoder of 2D network backbone where its output size is half of the input resolution. We then use the pixel-voxel correspondences in (C_i, V_i) for contrastive learning, with positives as all matching pixel-voxel pairs and negatives as all non-matching pixel-voxel pairs. We apply the PointInfoNCE loss (Equation 1) with f_i as the 2D features of a pixel, and f_j is the feature vector from its 3D correspondence, and M the set of 2D-3D pixel-voxel correspondence pairs.

3.3. Joint Learning

We can leverage not only the view-invariant constraints and geometric priors during training, but also learn jointly from the combination of both constraints. We can thus employ a shared 2D network backbone and a 3D network backbone, with the 2D network backbone constrained by both view-invariant constraints and as the 2D part of the geomet-

ric prior constraint.

During training, we consider (C_i, C_j, V_i, V_j) of overlapping color frames C_i and C_j as well as V_i and V_j which have geometric correspondence with C_i, C_j respectively. The shared 2D network backbone then processes C_i, C_j and computes the view-invariant loss from Section 3.1. At the same time, V_i and V_j are processed by the 3D sparse convolutional backbone, with the loss (discussed in 3.2) relative to the features of C_i and C_j respectively. This embeds both constraints into the learned 2D representations.

4. Experimental Setup

Our approach aims to embed 3D priors into the learned 2D representation by leveraging our view-invariant and geometric prior constraints. In this section, we introduce our detailed experimental setup for pre-training with an RGB-D dataset and fine-tuning on downstream 2D scene understanding tasks.

Architecture for Pre-training. As described in the previous section, our pre-training method leverages the pixel-to-pixel and geometry-to-color correspondences for view-invariant contrastive learning. The specific form of our pre-training objective requires a feature extractor capable of providing per-pixel or per-3D-point features for the backbone architecture, as the positive and negative matches are defined over 2D pixels or 3D locations.

Our meta-architectures for both view-invariant constraints and geometric priors are U-Nets [43] with residual connections. The encoder part of the U-Net is a standard ResNet. For view-invariant learning with 2D image inputs, we use ResNet18 or ResNet50 as encoders. The decoder part of the U-Net architecture consists of convolutional layers and bi-linear interpolation layers. For learning geometric priors from 3D volumetric occupancy input, we use sparse convolutions [14], specifically a Residual U-Net-32 backbone implemented with MinkowskiEngine [6], using a 2cm voxel size.

Stage I: Pri3D encoder initialization. We empirically found that for the pre-training phase, good initialization of the encoder network is critical to make learning robust. Instead of starting with random initialization, we initialize the encoder with network weights trained on ImageNet (*i.e*. we pre-train the network for pre-training). The whole pipeline can be seen as a two-stage framework. We note that our method aims to improve the *general* representation learning, thus is not tied to a specific learning paradigm (*e.g*. supervised pre-training or self-supervised pre-training). From this perspective, we can leverage supervised pre-training of ResNet [20] encoders with ImageNet [10] data for encoder initialization for pre-training. We name this model **Pri3D**.

Although the use of a supervised ImageNet pre-trained initialization is a common practice, for completeness we also evaluate Pri3D in an unsupervised pipeline without using ImageNet labels. Results suggest that Pri3D does not rely on any semantic supervision (*e.g*. ImageNet labels) to succeed, and still is able to achieve a substantial gain in this setup. We name this variant **Unsupervised Pri3D**. Further results of Unsupervised Pri3D are demonstrated in supplementary materials.

Stage II: Pri3D pre-training on ScanNet. Our pre-training method is enabled by the inherent geometry and color information present in the RGB-D data sequences. For pre-training, we leverage the color image and geometric reconstructions provided by the automatic reconstruction pipeline of ScanNet [7]; note that we do not use the semantic annotations during pre-training. ScanNet contains 2.5M images from 1513 ScanNet train video sequences. We regularly sample every 25^{th} frame without any other filtering (e.g., no control on viewpoint variation), and compute the set of overlapping pairs of frames that have $> 30\%$ pixel overlap, resulting in $\approx$ 840k frame pairs for which we compute their corresponding geometric chunks for each image, in order to apply both our view-invariant and geometric prior constraints.

Downstream Fine-tuning. We evaluate our Pri3D models by fine-tuning them on a suite of downstream image-based scene understanding tasks. We use two datasets, ScanNet [7] and NYUv2 [46], and the three tasks of semantic segmentation, object detection, and instance segmentation. As our pre-training dataset is ImageNet and ScanNet, fine-tuning on ScanNet represents a scenario of in-domain transfer—it would be interesting to know if the 3D priors can help with 2D representations for image-based tasks on *the same dataset*. We further evaluate the performance of Pri3D on the NYUv2 dataset which maintains different statistics. This represents a out-of-domain transfer scenario. For semantic segmentation tasks, we directly use the U-Net architecture for dense prediction. The encoder and decoder networks are both pre-trained with Pri3D. For instance segmentation and detection tasks, we use Mask-RCNN [19] framework implemented in Detectron2 [53]. Only the backbone encoder part is pre-trained.

Implementation details. For pre-training, we use an SGD optimizer with learning rate 0.1 and batch-size of 64. The learning rate is decreased by a factor of 0.99 every 1000 steps, and our method is trained for 60,000 iterations. For MoCoV2 [5], we use the official PyTorch implementation. MoCoV2 is trained for 100 epochs with batch size 256. The fine-tuning experiments on semantic segmentation are trained with a batch size of 64 for 80 epochs. The initial learning rate is 0.01, with polynomial decay with power 0.9. All experiments are conducted on 8 NVIDIA V100 GPUs.

Baselines. As we are using additional RGB-D data from ScanNet, it is important to benchmark our method against relevant baselines in order to answer the question: are 3D priors useful for 2D representation learning?

- **Supervised ImageNet Pre-training (IN)**. We use the ImageNet pre-trained weights provided in torchvision; this represents a widely adopted paradigm for image-based tasks. *No ScanNet data is involved.*
- **1-Stage MoCoV2 (MoCoV2-IN+SN)**. We train MoCoV2 on an expanded dataset that combines ImageNet with ScanNet. We explore two strategies: 1) Directly combining the two datasets with shuffled images and 2) mixing minibatches (sampling half images from ImageNet and the other half from ScanNet). In this case, *we use ScanNet data but no 3D priors are considered.*
- **2-Stage MoCoV2 (MoCoV2-supIN→SN)**. As we use supervised pre-training (IN) in our method as encoder initialization, for fair comparison, we also try one version with (supervised) IN as the encoder initialization, then add another stage to fine-tune MoCoV2 with randomly shuffled ScanNet images. In this case, *we use ScanNet data but no 3D priors are used.*
- **Trivial Correspondences.** We use our framework but instead of learning from multi-view correspondences, we take one single-view image and create two copies by applying color space augmentations including: RGB jittering, random color dropping and Gaussian blur. Positive matches are defined on pixels at the same location. In this case, *we use ScanNet data but no 3D priors are considered.*
- **Depth Prediction** We use single frame depth prediction as a pretext task. Our approach can leverage depth prediction as proxy loss. In this regard, *we use ScanNet data and a simple 3D prior is considered.*

Through above baselines, we aim to justify that Pri3D learns to embed 3D priors in 2D representations that lead to an improved downstream performance; it is nontrivial to achieve the goal, given the auxiliary RGB-D dataset.

5. Results

In this section, we present the results of our downstream fine-tuning results as well as relevant baselines mentioned in the previous section.

5.1. ScanNet

We use our pre-trained network weights learned with Pri3D, and fine-tune for 2D semantic segmentation, object detection, and instance segmentation tasks on ScanNet [7] images, demonstrating the effectiveness of representation learning with 3D geometric priors. For fine-tuning, following the standard protocol in the ScanNet benchmark [7]: we sample every 100 2D frames, resulting in 20,000 train images and 5,000 validation images.

2D Semantic Segmentation. We first show fine-tuning for semantic segmentation results in Table 1, in comparison with several baselines that also use ScanNet RGB-D data. We show the applicability of our approach with a standard ResNet50 backbone and a smaller ResNet18 backbone.

Comparing to just training the semantic segmentation model from scratch on downstream dataset (39.1% with ResNet50), all pre-training methods help significantly, even just using the ImageNet pre-training. This confirms the common belief in computer vision that a good 2D representation is essential for good performance on the target task. Several baselines, when adding the ScanNet RGB-D data, also works reasonably well, but *not* much better than the naive ImageNet Pre-training baseline. This suggests that simply adding the ScanNet data into the representation learning pipeline does not necessarily lead to better results. Our Pri3D variants, including the view-invariant contrastive learning, geometry-color correspondence based contrastive learning and the combination of the two, provides substantially better representation quality that leads to improved semantic segmentation performance. We note that our method has a major performance boost ($+6.0\%$ absolute mIoU) even compared with the ImageNet Pre-training results. We believe this is an encouraging result and represents a practical use case as ImageNet pre-trained networks are often readily available.

Moreover, we evaluate our approach under limited data scenarios in Figure 5. Our Pri3D pre-training shows an even larger gap when using a small subset of the training images, again compared to the strong ImageNet pre-training baseline. With only 20% of the training data, we are able to recover 84% and 80% of the finetuning performance when using 100% training data, with ResNet50 and ResNet18 backbone respectively.

2D Object Detection and Instance Segmentation. To demonstrate that Pri3D is generalizable for different image-based tasks, we show results on fine-tuning for object detection in Table 2 and instance segmentation in Table 3. For both tasks, we observe similar behavior to the semantic segmentation counterpart. All pre-training methods bring substantial improvement over training from scratch, but Pri3D models stand out and yield more gain compared to ImageNet Pre-training alone (+3.2% and +2.8% AP@0.5 for instance segmentation and detection, respectively). We note that for this set of experiments, we only transfer the encoder weights, discarding the decoder weights in the U-Net architecture for pre-training. This resembles similar practice in language domains (*e.g.* BERT [11]) and shows that the main gain of Pri3D is better encoder representations.

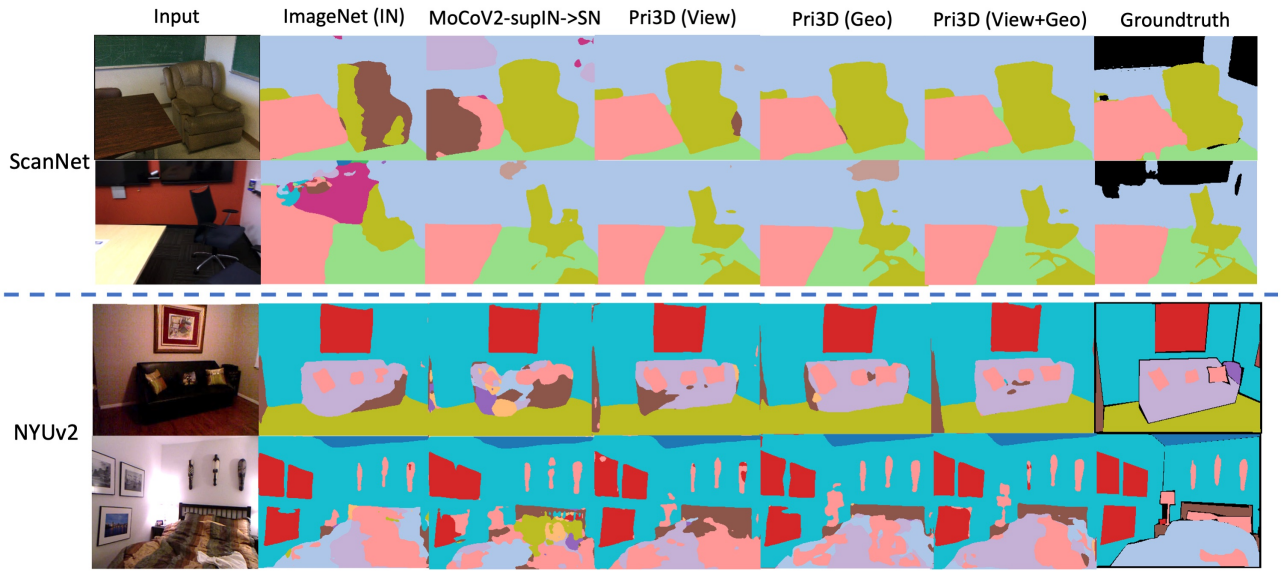

Figure 4: We show qualitative results on 2D semantic segmentation of ScanNet [7] and NYUv2 [46]. By encoding 3D priors, we obtain better segmentation results, in particular where when there are appearance variations over objects.

Method	ResNet50	ResNet18
Scratch	39.1	37.5
ImageNet Pre-training (IN)	55.7	51.0
MoCoV2-supIN→SN	56.6 (+0.9)	52.9 (+1.9)
MoCoV2-IN+SN$_{(combine)}$	54.9 (-0.8)	-
MoCoV2-IN+SN$_{(mixing\ batch)}$	54.5 (-1.2)	-
Trivial Correspondences	56.4 (+0.7)	52.1 (+1.1)
Depth Prediction	58.4 (+2.7)	-
Pri3D (View)	61.3 (+5.6)	54.4 (+3.4)
Pri3D (Geo)	61.1 (+5.4)	55.3 (+4.3)
Pri3D (View + Geo)	**61.7** (+6.0)	**55.7** (+4.7)

Table 1: **2D Semantic Segmentation on ScanNet.** Fine-tuning with Pri3D pre-trained models leads to significantly improved results compared to ImageNet pre-training. Pri3D learns better representations with 3D priors and compares favorably with other baselines that also uses auxiliary RGB-D data. Please refer to Sec. 4 for the detailed setup for those baselines. Metric is mean intersection-over-union (mIoU).

SOTA Segmentation Network. To demonstrate our method is agnostic to semantic segmentation backbones, we further show results with PSPNet and DeepLabV3/DeepLabV3+ in Table 4. Pri3D (Ours) consistently outperforms the baseline across different backbone choices.

5.2. NYUv2

We show that our method learns transferable features across datasets. With Pri3D pre-trained on ScanNet RGB-D data, we explore fine-tuning on NYUv2 [46] for downstream 2D tasks. The NYU-Depth V2 dataset is comprised of video sequences from a variety of indoor scenes, recorded by Microsoft Kinect RGB-D sensors. It contains 1449 densely labeled pairs of aligned RGB and depth images. We use the official split: 795 images for training, 654 images for test. Similar to ScanNet, we also evaluate on 3 popular downstream tasks, 2D semantic segmentation, ob-

Method	AP@0.5	AP@0.75	AP
Scratch	32.7	17.7	16.9
ImageNet (IN)	41.7	25.9	25.1
MoCoV2-supIN→SN	43.5 (+1.8)	26.8 (+0.9)	25.8 (+0.7)
Pri3D (View)	43.7 (+2.0)	27.0 (+1.1)	26.3 (+1.2)
Pri3D (Geo)	44.2 (+2.5)	**27.6** (+1.7)	26.6 (+1.5)
Pri3D (View+Geo)	**44.5** (+2.8)	27.4 (+1.5)	**26.6** (+1.5)

Table 2: **2D Detection on ScanNet.** Fine-tuning with Pri3D pre-trained models leads to improved object detection results across different metrics compared to ImageNet pre-training and a strong MoCo-style pre-training method.

Method	AP@0.5	AP@0.75	AP
Scratch	25.8	13.1	12.2
ImageNet (IN)	32.6	17.8	17.6
MoCoV2-supIN→SN	33.9 (+1.3)	18.1 (+0.3)	18.3 (+0.7)
Pri3D (view)	34.3 (+1.7)	18.7 (+0.9)	18.3 (+0.7)
Pri3D (geo).	34.4 (+1.8)	18.7 (+0.9)	18.3 (+0.7)
Pri3D (view+geo)	**35.8** (+3.2)	**19.3** (+1.5)	**18.7** (+1.1)

Table 3: **Instance Segmentation on ScanNet.** Fine-tuning with Pri3D pre-trained models leads to improved instance segmentation results compared to ImageNet pre-training and a strong MoCo-style pre-training method.

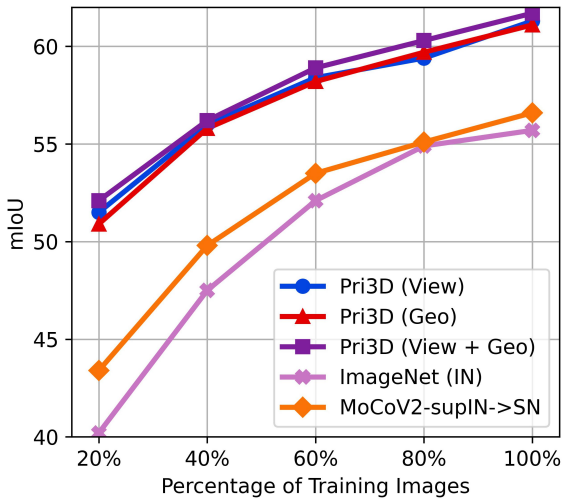

Figure 5: **Data Efficient Learning on ScanNet (ResNet50 Backbone).** Using only 40% of the training data, our pre-training can outperform supervised ImageNet pretraining when fine-tuned with 100% data available for semantic segmentation. We see similar trends with a ResNet18 backbone, which is included in the appendix.

Method	ResNet50
DeepLabV3 (ImageNet)	57.0
DeepLabV3 (Pri3D)	**61.3** (+4.3)
DeepLabV3+ (ImageNet)	57.8
DeepLabV3+ (Pri3D)	**61.6** (+3.8)
PSPNet (ImageNet)	59.7
PSPNet (Pri3D)	**62.8** (+3.1)

Table 4: 2D Semantic Segmentation on ScanNet (mIoU).

ject detection, and instance segmentation. Table 5 shows the semantic segmentation performance on NYUv2.

We show the semantic segmentation fine-tuning performance on NYUv2 in Table 5; the object detection fine-tuning results in Table 6; and the instance segmentation fine-tuning results in Table 7. The experimental setup is similar to the ScanNet downstream fine-tuning counterpart, and we use supervised ImageNet pre-trained weights for encoder initialization of all methods. For all three tasks, we observe improved performance over different baselines such as training from scratch, training with ImageNet pre-trained weights, and MoCoV2-style pre-training on additional ScanNet data. Compared to the ImageNet pre-training baseline, we achieve a margin of $+4.4\%$ AP@0.5 for instance segmentation, $+4.8\%$ mIoU for semantic segmentation (ResNet50 backbone) and $+4.1\%$ AP@0.5 for object detection.

6. Conclusion

We have introduced Pri3D, a new method for representation learning for image-based scene understanding tasks.

Method	ResNet50	ResNet18
Scratch	24.8	22.5
ImageNet Pre-training (IN)	50.0	44.7
MoCoV2-supIN→SN	47.6 (-2.4)	45.1 (+0.4)
Pri3D (View)	54.2 (+4.2)	48.2 (+3.5)
Pri3D (Geo)	**54.8** (+4.8)	**48.6** (+3.9)
Pri3D (View+Geo)	54.7 (+4.7)	48.1 (+3.4)

Table 5: **2D Semantic Segmentation on NYUv2**. Fine-tuning with Pri3D pre-trained models leads to improved semantic segmentation results compared to ImageNet pre-training and a strong MoCo-style pre-training method. Metric is Mean Intersection-Over-Union (mIoU).

Method	AP@0.5	AP@0.75	AP
Scratch	21.3	10.3	9.0
ImageNet (IN)	29.9	17.3	16.8
MoCoV2-supIN→SN	30.1 (+0.2)	18.1 (+0.8)	17.3 (+0.5)
Pri3D (View)	33.0 (+2.1)	19.8 (+2.6)	18.9 (+2.1)
Pri3D (Geo)	33.8 (+2.9)	20.2 (+2.9)	19.1 (+2.3)
Pri3D (View+Geo)	**34.0** (+4.1)	**20.4** (+3.1)	**19.4** (+2.6)

Table 6: **2D Object Detection on NYUv2.** Better object detection AP can be obtained with Pri3D fine-tuning.

Method	AP@0.5	AP@0.75	AP
Scratch	17.2	9.2	8.8
ImageNet (IN)	25.1	13.9	13.4
MoCoV2-supIN→SN	27.2 (+2.1)	14.7 (+0.2)	14.8 (+1.4)
Pri3D (View)	28.1 (+3.0)	15.7 (+1.8)	15.7 (+2.3)
Pri3D (Geo)	29.0 (+3.9)	15.9 (+2.0)	15.2 (+1.8)
Pri3D (View+Geo)	**29.5** (+4.4)	**16.3** (+2.4)	**15.8** (+2.4)

Table 7: **2D Instance Segmentation on NYUv2.** Better instance segmentation AP can be obtained with Pri3D.

Our core idea is to incorporate 3D priors in a pre-training process whose constraints are applied under a contrastive loss formulation. We learn view-invariant and geometry-aware representations by leveraging multi-view and image-geometry correspondence from existing RGB-D dataset. We show that this results in significant improvement compared to 2D-only pre-training. With limited training data available, we outperform the semantic segmentation baselines by 11.9% on ScanNet. We hope our results can shed light on the the general paradigm of representation learning with 3D priors and open up new opportunities towards 3D-aware image understanding.

Acknowledgments This work was supported by a TUM-IAS Rudolf Moßbauer Fellowship, the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Grant Making Machine Learning on Static and Dynamic 3D Data Practical, a Google Research Grant, and the Bavarian State Ministry of Science and the Arts as coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

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. 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2

[3] 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, 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020. 2, 3

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2, 4, 5

[7] 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, 5, 6, 7

[8] 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. 1, 2

[9] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 4

[10] 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. 5

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 6

[12] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *CVPR*, 2020. 1, 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 3

[14] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 2, 4, 5

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 2

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[17] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *CVPR*, 2020. 1, 2

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019. 2

[22] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *CVPR*, 2019. 1, 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] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *CVPR*, 2021. 3

[26] Zeyu Hu, Mingmin Zhen, Xuyang Bai, Hongbo Fu, and Chiew-lan Tai. Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds. *arXiv preprint arXiv:2007.06888*, 2020. 2

[27] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 92–101. IEEE, 2016. 1

[28] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4440–4449, 2019. 1, 2

[29] 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. 1, 2

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

[31] Yiqun Lin, Zizheng Yan, Haibin Huang, Dong Du, Ligang Liu, Shuguang Cui, and Xiaoguang Han. Fpconv: Learning local flattening for point convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2020. 2

[32] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3d-to-2d distillation for indoor scene parsing. In *CVPR*, 2021. 3

[33] David Marr and Tomaso Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1156):301–328, 1979. 2

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

[35] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM TOG*, 32(6):169, 2013. 4

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[37] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3D object detection in point clouds with image votes. In *CVPR*, 2020. 2

[38] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. *ICCV*, 2019. 1, 2

[39] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. 2

[40] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 2017. 1, 2

[41] 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. 1, 2

[42] 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. *arXiv preprint arXiv:2103.00020*, 2021. 3

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 5

[44] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. In *ICRA*, 2017. 3

[45] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *CVPR*, 2019. 3

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. *ECCV*, 2012. 1, 2, 5, 7

[47] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *CVPR*, 2015. 3

[48] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *ECCV*. 2014. 2

[49] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *CVPR*, 2016. 2

[50] Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, et al. Humangps: Geodesic preserving feature for dense human correspondences. *CVPR*, 2021. 3

[51] 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. 1, 2

[52] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. 2

[53] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[54] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[55] 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, 4

[56] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2020. 2

[57] 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

[58] 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

[59] Jiazhao Zhang, Chenyang Zhu, Lintao Zheng, and Kai Xu. Fusion-aware point convolution for online semantic 3d scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4534–4543, 2020. 2

[60] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. *arXiv preprint arXiv:2101.02691*, 2021. 2

[61] 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