Semi-Supervised Single-View 3D Reconstruction via Prototype Shape Priors

Zhen Xing[1,2], Hengduo Li[3], Zuxuan Wu[1,2†], and Yu-Gang Jiang[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3] University of Maryland

Abstract. The performance of existing single-view 3D reconstruction methods heavily relies on large-scale 3D annotations. However, such annotations are tedious and expensive to collect. Semi-supervised learning serves as an alternative way to mitigate the need for manual labels, but remains unexplored in 3D reconstruction. Inspired by the recent success of semi-supervised image classification tasks, we propose SSP3D, a semi-supervised framework for 3D reconstruction. In particular, we introduce an attention-guided prototype shape prior module for guiding realistic object reconstruction. We further introduce a discriminator-guided module to incentivize better shape generation, as well as a regularizer to tolerate noisy training samples. On the ShapeNet benchmark, the proposed approach outperforms previous supervised methods by clear margins under various labeling ratios, (*i.e.*, 1%, 5% , 10% and 20%). Moreover, our approach also performs well when transferring to real-world Pix3D datasets under labeling ratios of 10%. We also demonstrate our method could transfer to novel categories with few novel supervised data. Experiments on the popular ShapeNet dataset show that our method outperforms the zero-shot baseline by over 12% and we also perform rigorous ablations and analysis to validate our approach. Code is available at https://github.com/ChenHsing/SSP3D.

Keywords: Semi-supervised learning, 3D Reconstruction, Shape priors

1 Introduction

Reconstructing 3D shape from RGB images plays an important role in many applications, such as 3D printing, virtual reality and 3D scene understanding. Human can easily infer 3D shape and scene object from single-view images mainly because of the powerful shape priors of human visual systems, yet it remains challenging to model such strong priors for accurate single-view 3D reconstruction. While Structure From Motion(SFM) [24] and Simultaneous Localization and Mapping (SLAM) [3] are feasible solutions, they require abundant data annotations and inferring camera parameters.

† Corresponding author.

Recently, with the growing interest in deep learning, great success has been achieved in predicting 3D shape from a single image with deep Convolutional Neural Networks (CNNs) [6,34,39]. But there are still limitations of these methods: (i) The astounding performance comes at the cost of massive amount of labeled images with fine-grained 3D shape, which is time-consuming and labour-intensive to obtain. (ii) Inferring 3D shape from a single image is an ill-posed problem because there are multiple plausible shapes given a 2D image.

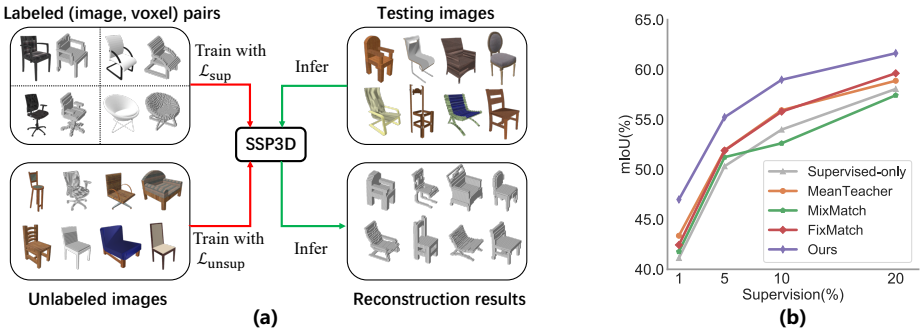

Fig. 1. (a) Illustration of semi-supervised single-view 3D reconstruction. Our SSP3D can predict 3D shape for an unlabeled image after training with a mixture of labeled data and unlabeled data. (b) Our proposed model can efficient leverage the unlabeled data and outperform supervised-only method and state-of-the-art semi-supervised image recognition extended methods.

Semi-Supervised learning (SSL) is a popular strategy to learn in the low-data regime by leveraging the readily available unlabeled data, which has demonstrated great success for image classification [29,1,35] and object detection [19]. Generalizing best practices [29,27] that work well in the 2D domain to 3D reconstruction, while appealing, is challenging. On one hand, it remains unclear how to evaluate the quality of 3D shape pseudo labels, which are the core for SSL. On the other hand, inferring the actual 3D shape of an object from a single image requires strong shape priors, yet existing single-view 3D reconstruction methods [6,39] require a large amount of annotated data to learn the shape priors implicitly with the model parameters. As a result, the 3D reconstruction network trained with limited annotations will likely produce low-quality reconstruction results, especially for the images with heavy occlusion.

To tackle these challenges, we propose a semi-supervised learning framework with several components specially designed for single-view 3D reconstruction as shown in Figure 1. Inspired by the recent advances in SSL for image classification [29,27], we use the teacher-student pseudo labeling method as the training paradigm of our framework. In order to generate more reliable pseudo labels for the unlabeled images, we use a Prototype Attentive Module for providing shape priors explicitly. In particular, we first obtain 3D prototype shape as candidate

shape priors through clustering algorithms (*e.g.*, KMeans). For a given image, we extract the image feature through a 2D encoder. The relationship of image feature and 3D prototype is captured with the help of the attention mechanism to obtain the shape priors, which serve as a bridge to encourage perceptually realistic reconstruction and prevent mode collapses [6,39].

In addition, we introduce a module named Shape Naturalness Module that serves as a discriminator distinguishing predicted 3D shapes from ground-truth 3D shapes. During training, an additional loss is used to penalize unnatural reconstruction results from the model in a generative adversarial learning manner such that the model is incentivized to generate more realistic 3D shapes. Meanwhile, the output of the discriminator can be directly used as an approximation of the quality of pseudo labels such that the inaccurate pseudo labels can be ignored or down-weighted accordingly when training the student model.
In conclusion, the main contributions of this paper is summarized as follows:

- We propose a semi-supervised prototype 3D reconstruction network (SSP3D) to reconstruct 3D shapes from a single RGB image. Our work is the first attempt to reconstruct 3D volume in semi-supervised learning with only 1% labeled data of train set.
- Without additional information, an effective yet lightweight shape prior fusion module is proposed, which can be easily incorporated into 3D reconstruction networks with similar architecture. In addition, the discriminator module we proposed guides the generation of natural shapes and serves as a scorer to filter out noisy training samples for the student model.
- We are the first to establish a semi-supervised benchmark to measure the single-view 3D reconstruction network. Experiments show that our model achieves the state-of-the-art on two datasets and settings under various labeling ratios. We hope that our results serve a strong baseline to encourage future research in more robust semi-supervised 3D reconstruction methods.

2 Related Work

Deep Learning for 3D Reconstruction Recently, deep learning techniques have been widely used for 3D reconstruction. 3D-R2N2 [6] is among the earliest work exploring the 3D reconstruction based on Recurrent Neural Network. It establishes a benchmark for 3D reconstruction with a synthetic ShapeNet dataset. 3D-VAE-GAN [37] builds upon Variational Autoencoders (VAE) and Generative Adversarial Networks (GANs) to reconstruct 3D shapes. OGN [30] and Matryoshka Networks [22] use octree and nested shape layers to represent 3D volumes of objects, respectively. Marrnet [36], ShapeHD [38] and GenRe [46] adopt 2.5D information such as depth, silhouette and surface normal of RGBs as intermediate shape priors to reconstruct 3D shapes. Pix2Vox [39] and Pix2Vox++ [40] build robust backbones for 3D volume reconstruction and achieve state-of-the-art results with encoder-decoder architectures. Mem3D [43] requires a great extra storage space to provide shape priors, which limits its applicability. EVolT [33] and 3D-RETR [25] leverage transformers as backbone

networks to reconstruct 3D shapes. Unlike most existing work that are trained in a supervised manner, we explore semi-supervised learning for 3D reconstruction.

Deep Semi-Supervised Learning The overall purpose of semi-supervised learning (SSL) is to effectively use unlabeled data without relying on any manual supervision to expand supervised learning when the labeled training data is scarce. Recent semi-supervised methods mainly contain two principles: data augmentations and consistency regularization. The model is expected to be consistent and robust to data augmentations—producing consistent outputs for the original and augmented inputs. Many methods use different data augmentations [1,16,23] or dropout [29] of models to generate images of different transformations. Researchers also use multiple networks to generate different views of the same input data [21], or mix input data to generate training data and labels [45,44,10,13]. In single-view 3D reconstruction, Semi-supervised Soft Rasterizer (SSR) [17] and [42] try to reconstruct 3D objects with few amount of annotation data, but they all rely on the annotations of additional camera pose or silhouette. To the best of knowledge, the settings of SSL with only single-view image have not been studied in 3D reconstruction, a complex and challenging task that depends on fine-grained human annotations.

3 Method

Problem Definition For a single-view image x of any object, the goal is to reconstruct the 3D shape y of the object. As discussed earlier, current methods for single-view 3D reconstruction typically require large amount of annotations that are time-consuming and labour-intensive to obtain. We thus explore developing a semi-supervised learning framework for the task to alleviate the need of annotated data during training. Suppose we have N training samples, including N_L labeled image-3D pairs $(x_l, y_l) \in D_L$ and N_U unlabeled image data $(x_u) \in D_U$. As in prior work, D_L and D_U are sampled from the same data distribution (*e.g.*, either synthetic or real-world). Our purpose is to leverage D_L and D_U together to train the model for an improved performance on reconstructing the 3D shapes of objects.

Overview As shown in Figure 2, our framework SSP3D contains two training stages: Warm-up stage and Teacher-student mutual learning stage. In the Warm-up stage, the available labeled set D_L is used to train a "teacher" model; in the Teacher-student mutual learning stage, the teacher model first generates pseudo labels (*i.e.*, predicted 3D shapes) for the unlabeled set D_U, and then a "student" model – initialized from the pre-trained teacher model – is trained on D_L and D_U for an improved performance. For effective distillation, strong data augmentation is applied on the input to student model. The teacher model also temporally aggregates the weights from the student model to produce more refined pseudo labels.

While appealing, directly extending existing SSL methods like Mean-Teacher [29] and FixMatch [27] for single-view 3D reconstruction is challenging

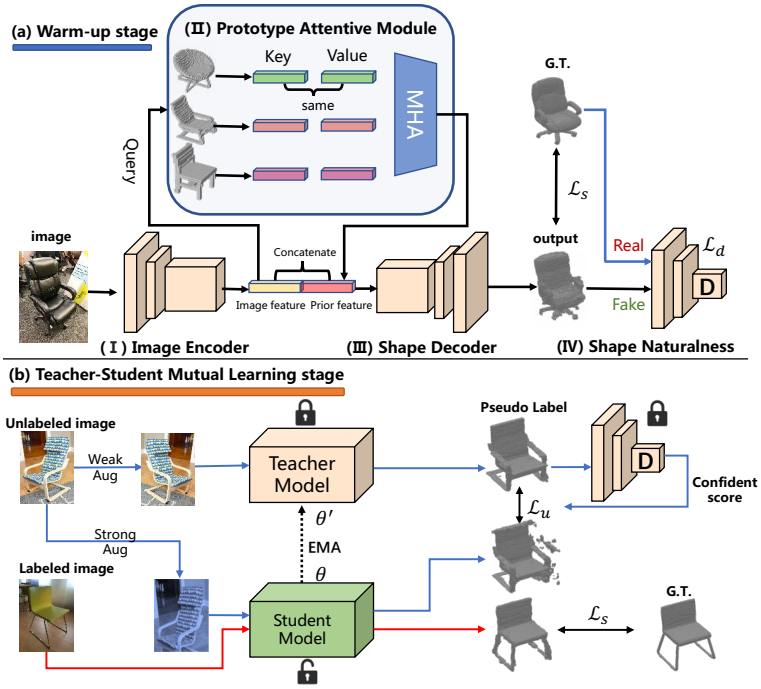

Fig. 2. Overview of Our SSP3D. SSP3D consists of two stages. **Warm-up**: we use available supervised data to train 3D reconstruction network. **Teacher-Student mutual learning stage**: for unsupervised data, *Teacher* with fixed parameters generate pseudo-labels to train *Student*. At the same time, *Teacher* and *Student* are given weakly and strongly augmented inputs respectively. In order to avoid the interference of pseudo-labels noise generated by the *Teacher*, we give a confidence score weight to the unsupervised loss by discriminator. The knowledge learned by *Student* online is slowly transferred to the weight replication mode of *Teacher* through exponential moving average (EMA). When the reconstruction network is trained and converged in the Warm-up stage, we switch to the Teacher-Student mutual learning stage.

since the pseudo labels from the teacher model can be quite noisy for two main reasons: 1) inferring accurate 3D shape from single-view image requires strong prior that is difficult to learn without massive annotated data; 2) it is unclear how to evaluate the quality of the predicted pseudo 3D shapes to filter out inaccurate predictions. To this end, we propose two modules namely Prototype Attention Module and Shape Naturalness Module to address these challenges.

In the following text, we first introduce our proposed model components in Warm-up stage (Sec. 3.1), and then we show how the Teacher-student mutual learning stage works with pseudo labelling and teacher refinement methods (Sec. 3.2). Finally, we elaborate the optimization of our framework in Sec. 3.3.

3.1 Warm-up stage

As shown in Fig. 2, SSP3D consists of four modules, among which image encoder and shape decoder are consistent with the state-of-the-art method Pix2Vox [39], whereas the proposed prototype attentive module and shape naturalness module will be presented below. At this stage, the teacher model is trained on labeled set D_L in standard supervised learning manner.

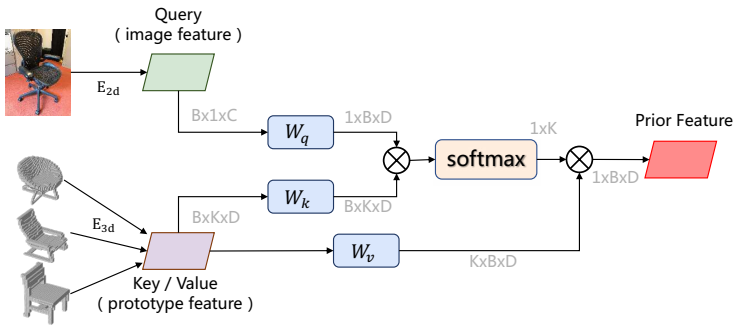

Fig. 3. Overview of Our Prototype Attentive Module.

Prototype Attentive Module In most existing work on single-view 3D reconstruction, the shape priors are learned implicitly with the parameters of the model [6,39,33], which may lead to poor performance for some noisy or occluded images [6], especially when annotated training data is not abundant. Therefore, standard 3D reconstruction models are likely to produce noisy and inaccurate pseudo labels when used as the teacher model directly, resulting in poor performance when training the student model under such semi-supervised learning setting. To tackle this problem, we propose to augment the image features with learned category-specific shape priors explicitly, so that the strong priors learned from labeled data could help infer more realistic and natural object shapes.

For supervised data, we obtain the shape prototype ${P_i}^k$ of the specified categories by doing K-Means clustering on the features learned by a 3D Autoencoder [1], and we take the clustering center of these categories as the prototype shape priors. The designed attention-based shape priors acquisition mechanism is shown in Fig. 3. Firstly, image encoder extracts the 2D feature of the image as query in Eq. 1. Secondly, we extract feature of the prototype through 3D encoder to obtain the prototype feature in Eq. 2, which is used as the key and value in the attention mechanism [31]. We then use three separate linear layers parameterized by W_q, W_k and W_v to extract query, key, value embedding Q, K and V in Eq. 3. Formally, the shape prior feature can be obtained by multi-head attention (MHA) [31] in Eq. 4.

[1] Please refer to Appendix for more details.

$$\text{Image features: } Query = \texttt{Encoder2d}(I_q), \tag{1}$$

$$\text{Prototype features: } Key, Value = \texttt{Encoder3d}(P_i), \tag{2}$$

$$Q = Query \cdot W_q, \quad K = Key \cdot W_k, \quad V = Value \cdot W_v, \tag{3}$$

$$\text{Prior features} = \texttt{MHA}(Q, K, V). \tag{4}$$

Here, I_q is the query image, P_i indicates the prototype 3D voxel, $W_q \in \mathbb{R}^{C\times D}$, W_k, $W_v \in \mathbb{R}^{D\times D}$ are learnable matrices. In the previous work using shape priors [32,43], 3D voxel can be directly used as shape priors in the form of additional inputs, however they can not capture the correlations between the images and multiple prototype shape priors. In contrast, we use the attention-based module to extract the shape priors by exploring the association between image features and 3D prototypes.

Shape Naturalness Module The shape reconstruction network typically uses only one supervised loss during training, yet the inherent uncertainty of the loss will lead to unrealistic and inaccurate prediction of object shapes especially on object surface.

Inspired by [38], we develop a shape naturalness module that servers as a discriminator distinguishing predicted shape and the corresponding ground-truth shape, and penalizes the network in an adversarial learning manner when unnatural shapes are generated.

Unlike [37] and [38] which use a pre-trained 3D-GAN as a discriminator to judge whether a shape is real, our framework is learned in an end-to-end generative adversarial training manner. In particular, we take parts (I)-(III) in Fig. 2 as the generator, and the shape naturalness module is used to distinguish the generated shape from the real shape. The optimization is achieved by minimizing the following loss $\mathcal{L}_d$:

$$\mathcal{L}_d = \mathbb{E}_{y_p\sim D_p}\log D(\mathrm{y_p}) + \mathbb{E}_{\mathrm{y_g}\sim \mathrm{D_g}}\log(1 - D(\mathrm{y_g})), \tag{5}$$

where D_p and D_g are predicted and groundtruth distributions, y_p and y_g are samples in D_p and D_g respectively, and D is the discriminator here.

3.2 Teacher-Student Mutual Learning stage

Overview After the teacher model converges in the Warm-up stage, it is used to produce pseudo labels on unlabeled images to supervise the student model. For effective and efficient distillation, we initialize the student model with the weights of the teacher model and apply strong data augmentations on input images to the student following the common SSL paradigm [27]. On the other hand, the teacher takes weakly augmented images as inputs and aggregates the weights of

the student temporally throughout the Teacher-Student Mutual Learning stage to generate more reliable pseudo labels.

Student Learning To utilize the readily available unlabeled images D_u, we use the pseudo-labeling method to generate labels for D_u to train the student model, which has been shown effective for semi-supervised image classification [29,27] and object detection [19,18].

Formally, for unsupervised data, the teacher model first generates the soft label $\hat{I}$ in voxel, in which each voxel entry belongs to $[0, 1]$. We first binarize it into hard labels, where each entry in the 3D voxel is binarized as follows:

$$I(i,j,k) = \begin{cases} 1, & \hat{I}(i,j,k) > \delta \quad (6) \\ 0, & \text{otherwise} \quad (7) \end{cases}$$

We then train the student model by taking the binarized pseudo label I as the ground truth. In addition, we jointly train students with the same amount of unsupervised and supervised data in each mini-batch to ensure that the model is not biased by pseudo labels.

Confidence Scores for Pseudo Label The predictions from the teacher model are more or less inaccurate compared with the ground-truth shapes. Therefore, a filtering mechanism is desired to keep only the mostly accurate predictions as pseudo labels to train the student model. Existing semi-supervised classification methods often use the confidence scores predicted by the network as a proxy and only keep the confident predictions as pseudo labels through applying predefined thresholds [41] or using Top-k selection [27]. However, such confidence scores are missing in 3D reconstruction, and a new solution is needed to measure the quality of the generated 3D pseudo labels.

To this end, the shape naturalness module is also designed to serve as a naturalness "scorer" directly. In particular, the sigmoid-normalized output of the discriminator naturally indicates the possibility that an output sample is real or fake since the discriminator is optimized by a binary cross entropy loss using label 1 for real ground truth, and 0 for generated shape. We therefore use this output as the confidence score to measure the quality of generated pseudo label. The confidence score can be used to reweight the unsupervised loss, which will be described in detail in Section 3.3.

Teacher Refinement In order to obtain more refined pseudo labels, we use exponential moving average (EMA) to gradually update the teacher model with the weights of the student model. The slow updating process of teacher model can be considered as an ensemble of student models at different training time stamps. The update rule is defined below:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \quad (8)$$

where α is momentum coefficient. In order to make the training process more stable, we slowly increase α to 1 through cosine design as in [7]. This method has been proved to be effective in many existing works, such as self-supervised learning [11,9], SSL image classification [29] and SSL object detection method [19,18].

Here, we are the first to introduce it and validate its effectiveness in semi-supervised 3D reconstruction to the best of our knowledge.

3.3 Training paradigm

The training process is completed in two stages. In the Warm-up stage, we adopt reconstruction loss and GAN loss jointly and train the teacher model on D_L. In the Teacher-Student mutual learning stage, the generator part is duplicated as two models (Teacher and Student). The parameters of teacher and discriminator are fixed in this stage. We only optimize students through supervised and unsupervised losses.

Reconstruction Loss For the 3D reconstructions network, both the reconstruction prediction and the ground truth are in the form of voxels. We follow previous works [32,20,39,40] that adopt binary cross entropy loss as the reconstruction loss function:

$$\mathcal{L}_{rec} = \frac{1}{r_v^3} \sum_{i=1}^{r_v^3} [gt_i \log(pr_i) + (1 - gt_i) \log(1 - pr_i)], \tag{9}$$

where r_v represents the resolution of the voxel space, pr and gt represent the predict and the ground truth volume.

Warm-up Loss In the Warm-up stage, all parts of the models are end-to-end trained on labeled set D_L. The objective function is:

$$\min_{\theta_f} \max_{\theta_d} \mathcal{L}_{rec}(\theta_f) + \lambda_d \mathcal{L}_d(\theta_d). \tag{10}$$

Where θ_f and θ_d are the parameter of generator and discriminator, respectively. λ_d is the balance parameter of loss terms. We set λ_d to 1e-3 here.

Teacher-Student Mutual Loss At the second stage, for supervised data, we use the BCE loss function as in Eq. 9. For unlabeled data, we use the loss function below:

$$\mathcal{L}_{unsup} = \sum_{i=1}^{n} \text{score}_i (\hat{y}_i - y_i)^2, \tag{11}$$

where y_i and $\hat{y}_i$ are the target and predicted shapes, respectively. $score_i$ denotes the confidence score of $\hat{y}_i$ output by the discriminator. Note that we used squared L2 loss or the Birer score [2] instead of binary cross entry loss in the optimization of unsupervised data. The Brier score is widely used in semi-supervised literature because it is bounded and does not severely penalize the probability of being far away from the ground truth. Our initial experiments show that square L2 loss results in slightly better performance than binary cross entropy.

The loss function for training the student model is shown below:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_u \mathcal{L}_{unsup}. \tag{12}$$

where λ_u is the balance parameter of loss terms, which is set as 5 here. Through the joint training of supervised loss and unsupervised loss, we can make full use of labeled and unlabeled data to achieve better performance.

4 Experiments

4.1 Experimental Setup

Datasets We use ShapeNet [4] and Pix3D [28] in our experiments. The ShapeNet [4] is described in 3D-R2N2 [6], which has 13 categories and 43,783 3D models. Following the split defined in Pix2Vox [39], we randomly divide the training set into supervised data and unlabeled data based on the ratio of labeled samples, *i.e.*, 1%, 5%, 10% and 20%. The voxel resolution of ShapeNet is 32^3. Pix3D [28] is a large-scale benchmark with image-shape pairs and pixel level 2D-3D alignment containing 9 categories. We follow the standard S1-split, which contains 7,539 train images and 2,530 test images as in Mesh R-CNN [8]. Because Pix3D is loosely annotated (*i.e.*, an image may contain more than one object but only one object is labeled), we use ground-truth bounding boxes to cut all the images as [39,40]. Similarly, we randomly sample 10% of the training set as labeled data and use the remaining samples as unlabeled data. The voxel resolution of Pix3D is 128^3 and we have also changed the network parameters accordingly following common practice [40].
Evaluation Metric We used Intersection over Union (IoU) for the evaluation metric as in [6,39]. It is defined as follows:

$$\text{IoU} = \frac{\sum_{i,j,k} \mathcal{F}(\hat{p}_{(i,j,k)} > t)\mathcal{F}(p_{(i,j,k)})}{\sum_{i,j,k} \mathcal{F}[\mathcal{F}(\hat{p}_{(i,j,k)} > t) + \mathcal{F}(p_{(i,j,k)})]}, \tag{13}$$

where $\hat{p}_{(i,j,k)}$ and $p_{(i,j,k)}$ represent the predicted possibility and the value of ground truth at voxel entry (i, j, k), respectively. $\mathcal{F}$ is a shifted unit step function and t represents the threshold, which is set to 0.3 in our experiments.
Implementation details In both stages, the batch size is set to 32, and the learning rate decays from $1e-3$ to $1e-4$. We use Adam [15] as the optimizer. We set α to 0.9996, the number of clusters for prototypes to 3, and the number of multi-head of attention to 2. The δ is set to 0.3. We train the network for 250 epochs in the Warm-up stage and 100 epochs in the Teacher-Student mutual learning stage.

4.2 Main Results

Baseline We compare our approach with various baselines and direct extensions of popular semi-supervised approaches for 2D image classification. Firstly, we consider the encoder-decoder architecture of Pix2Vox [39] as our supervised baseline. Note that we change the backbone from VGG19 [26] to ResNet-50 [12] for decreasing parameters following Pix2Vox++ [40]. Secondly, we extend state-of-the-art SSL methods for image classification such as MeanTeacher [29], MixMatch [1] and FixMatch [27], to the task of 3D reconstruction, which serve as strong semi-supervised baselines. We use the same backbone and experimental settings for all the baselines and our approach for fair comparisons. More details of implementation could be found in Appendix.

Table 1. Comparisons of single-view 3D object reconstruction on ShapeNet at 32^3 resolution with different labeling ratios. We report the mean IoU (%) of all categories. The best number for each category is highlighted in bold.

Approach\split	1% 301 labels 30596 images	5% 1527 labels 30596 images	10% 3060 labels 30596 images	20% 6125 labels 30596 images
Supervised (ICCV'19)	41.13	50.32	53.99	58.06
Mean-Teacher (NeurIPS'17)	43.36 (↑2.23)	51.92 (↑1.60)	55.93 (↑1.94)	58.88 (↑0.82)
MixMatch (NeurIPS'19)	41.77 (↑0.64)	51.23 (↑0.91)	52.62 (↓1.37)	57.43 (↓0.63)
FixMatch (NeurIPS'20)	42.44 (↑1.31)	51.89 (↑1.57)	55.79 (↑1.80)	59.63 (↑1.57)
SSP3D (ours)	**46.99 (↑5.86)**	**55.23 (↑4.91)**	**58.98 (↑4.99)**	**61.64 (↑3.58)**

Results on ShapeNet and Pix3D As shown in Table 1, we compare our method with the supervised-only models under the settings of 1%, 5%, 10% and 20% labeled data. The experimental results show that our model outperforms supervised baselines by clear margins, especially under the setting with only 1% labels where our model outperforms the supervised model by 5.86%. Notably, our model outperforms the latest SOTA method FixMatch [27] by 4.55% with only 1% labeled data, demonstrating that directly extending existing SSL methods is sub-optimal for the task of single-view 3D reconstruction and that the proposed prototype attentive module and shape naturalness module are effective.

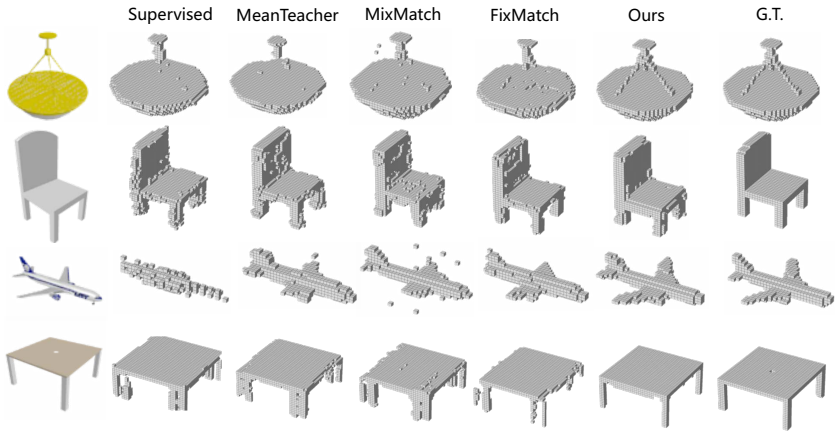

Fig. 4. Examples of single-view 3D Reconstruction on ShapeNet with 5% labels.

We further conduct experiments on Pix3D [28]. The experimental results are shown in Table 2. Considering that the 3D voxel resolution of Pix3D is 128^3, which increases the model complexity, we compare it with the supervised method and the two state-of-the-art methods of MeanTeacher [29] and FixMatch [27]

Table 2. Comparisons of single-view 3D object reconstruction on Pix3D at 128^3 resolution with 10% supervised data. We report the mean IoU (%) of all categories. The best number for each category is highlighted in bold.

Approach\split	chair 267/2672	bed 78/781	bookcase 28/282	desk 54/546	misc 4/48	sofa 153/1532	table 145/1451	tool 3/36	wardrobe 18/189	**Mean**
Supervised [39]	19.27	32.10	23.99	25.32	18.37	62.29	22.77	11.32	81.88	29.80
MeanTeacher [29]	21.66 (↑2.39)	35.04 (↑2.94)	18.88 (↓5.11)	26.17 (↑0.85)	22.37 (↑4.00)	64.19 (↑1.90)	24.03 (↑1.26)	9.18 (↓2.14)	84.34 (↑2.46)	31.40 (↑1.60)
FixMatch [27]	21.95 (↑2.68)	26.69 (↓5.41)	16.06 (↓7.93)	22.12 (↓3.20)	17.87 (↓0.50)	63.74 (↑1.45)	20.64 (↓2.13)	6.89 (↓4.43)	84.45 (↑2.57)	30.35 (↑0.55)
SSP3D (ours)	**23.97 (↑3.04)**	**46.33 (↑13.36)**	**32.77 (↑7.01)**	**32.89 (↑4.76)**	**24.35 (↑2.96)**	**68.32 (↑5.76)**	**23.84 (↑2.13)**	**39.06 (↑27.58)**	**89.59 (↑6.88)**	**35.39 (↑5.59)**

only under the setting of 10% labeled data due to limited GPU resources. We report the reconstruction performance of each category. For some categories with few labeled data (*e.g.*, tool and bed), our model outperforms supervised models by 27.58% and 13.36%. Due to the scarcity of annotated training data and that the other two methods (MeanTeacher and FixMatch) do not have the guidance of strong shape priors, they do not perform as well as supervised methods. Overall, our method outperforms supervised methods by 5.59% measured by on the mean IoU of all categories. Compared with MeanTeacher [29], it is also better by 4.99%.

We further provide qualitative results on both datasets in Fig. 4 and Fig. 5. As can be seen from Fig. 4, for images with clean background, our method produces a smoother object surface than baseline methods. For data with complex background and heavy occlusions in Fig. 5, shapes generated by our method are much better than alternative methods.

Transferring to Novel Category Results Wallace *et al.* [32] propose a few-shot setting for single-view 3D reconstruction via shape priors. We also train the model with seven base categories and finetune the model with only 10 labeled data in the novel categories. During inference, we report the performance on novel categories. We also compare our method with CGCE [20] and PADMix [5], as shown in Table 3. Under the 10-shot setting, our method outperforms the zero-shot baseline by 12%. We hypothesize that the improvement is mainly due to our more reasonable shape prior module design as well as the usage of a large number of unlabeled data.

4.3 Ablation Study

In this section, we evaluate the effectiveness of proposed modules and the impact of hyper-parameters. The experiments are under the 1% ShapeNet setting and 10% Pix3D setting if not mentioned elsewhere.

Prototype Attentive Module Here we analyze the effectiveness of shape prior module in 3D reconstruction. To do this, we compare our method with various prior aggregation methods including totally removing the prototype attention module (w/o PAM), fusing class-specific prototypes through averaging (w. average) and using LSTM [14] fusion for prototype shape priors. As shown in Ta-

Fig. 5. Examples of single-view 3D Reconstruction on Pix3D with 10% labels.

ble 4, removing the prototype attentive module results in a large drop of 3.55% and 4.17% in performance on both datasets , demonstrating the effectiveness of using class-specific shape priors for single-view 3D reconstruction. Our prior module also outperforms all other prior aggregation methods, indicating that self-attention mechanism is better at capturing the relationships between input images and class-specific prototype shape priors.

Shape Naturalness Module In order to demonstrate the effectiveness of the shape naturalness module, we remove the module (w/o SNM), that is, remove the GAN loss $\mathcal{L}_d$ and only use $\mathcal{L}_{rec}$ to optimize the network in the warm-up stage. In addition, we also verify the effectiveness of the confidence scores generated by the discriminator through replacing all the confidence scores as 1 (w/o score) and check the performance. Experiments show that the performance drops 1.12% and 0.67% on ShapeNet without SNM and scorer respectively, indicating that SNM plays an important role in our framework, which may avoid unnatural 3D shape generation, and the confidence score could avoid the negativeness of noisy or biased labels.

Table 3. Comparison of single-view 3D object reconstruction on novel categories of ShapeNet at 32^3 resolution under 10-shot setting. We report the mean IoU(%) per category. The best number for each category is highlighted in bold.

	cabinet	sofa	bench	watercraft	lamp	firearm	**Mean**
Zero-shot	69	52	37	28	19	13	36
Wallace (ICCV'19) [32]	69 (↑0)	54 (↑2)	36 (↓1)	36 (↑8)	19 (↑0)	24 (↑11)	39 (↑3)
CGCE (ECCV'20) [20]	71 (↑2)	54 (↑2)	37 (↑0)	41 (↑13)	20 (↑1)	23 (↑10)	41 (↑5)
PADMiX (AAAI'22) [5]	66 (↓3)	57 (↑5)	41 (↑4)	46 (↑18)	31 (↑12)	39 (↑26)	47 (↑11)
SSP3D(ours)	**72 (↑3)**	**61 (↑9)**	**43 (↑6)**	**49 (↑21)**	**31 (↑12)**	**34 (↑21)**	**48 (↑12)**

Table 4. Ablation study of different modules and losses. We report the mean IoU(%) of both datasets.

	PAM	average	LSTM	SNM	$\mathcal{L}_{unsup}$	$\mathcal{L}_{BCE}$	$\mathcal{L}_{rec}$	score	ShapeNet	Pix3D
baseline							✓		41.13 (↓5.86)	29.80 (↓5.59)
w/o PAM				✓	✓		✓	✓	43.44 (↓3.55)	31.32 (↓4.17)
w average		✓		✓	✓		✓	✓	43.80 (↓3.19)	32.64 (↓2.75)
w LSTM			✓	✓	✓		✓	✓	44.61 (↓2.38)	33.42 (↓1.97)
w/o SNM	✓				✓		✓	-	45.87 (↓1.12)	33.90 (↓1.49)
w/o score	✓			✓	✓		✓		46.32 (↓0.67)	34.62 (↓0.77)
w $\mathcal{L}_{BCE}$	✓			✓		✓	✓	✓	45.85 (↓1.14)	34.02 (↓1.37)
SSP3D(ours)	✓			✓	✓		✓	✓	**46.99**	**35.39**

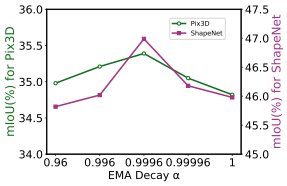

Fig. 6. Ablation study of different EMA decay α.

EMA and Loss We also verify the effect of EMA. In our experiments, we find that EMA decay coefficient $\alpha = 0.9996$ gives the best validation performance. As shown in Fig. 6, the performance slightly drops at different decay rates. For the unsupervised loss function in the Teacher-student mutual learning stage, if the unsupervised squared L2 loss is replaced by binary cross entropy (w $\mathcal{L}_{BCE}$), the performance of the model will also drop 1.14% and 1.37% in performance on two datasets shown in Table 4.

5 Conclusion

We introduced SSP3D, which is the first semi-supervised approach for single-view 3D reconstruction. We presented an effective prototype attentive module for semi-supervised setting to cope with limited annotation data. We also used a discriminator to evaluate the quality of pseudo-labels so as to generate better shapes. We conducted extensive experiments on multiple benchmarks and the results demonstrate the the effectiveness of the proposed approach. In future work, we would like to explore the semi-supervised setting on other 3D representation, such as mesh or implicit function.

Acknowledgement Y.-G. Jiang was sponsored in part by "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 20SG01). Z. Wu was supported by NSFC under Grant No. 62102092.

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: NeurIPS (2019) 2, 4, 10
2. Brier, G.W., et al.: Verification of forecasts expressed in terms of probability. Monthly weather review (1950) 9
3. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Trans. Robotics (2016) 1
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 10
5. Cheng, T.Y., Yang, H.R., Trigoni, N., Chen, H.T., Liu, T.L.: Pose adaptive dual mixup for few-shot single-view 3d reconstruction. In: AAAI (2022) 12, 14
6. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016) 2, 3, 6, 10
7. Ge, C., Liang, Y., Song, Y., Jiao, J., Wang, J., Luo, P.: Revitalizing cnn attention via transformers in self-supervised visual representation learning. In: NeurIPS (2021) 8
8. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. In: ICCV (2019) 10
9. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020) 8
10. Guo, H., Mao, Y., Zhang, R.: Mixup as locally linear out-of-manifold regularization. In: AAAI (2019) 4
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 8
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10
13. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. In: ICLR (2020) 4
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997) 12
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 10
16. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017) 4
17. Laradji, I., Rodríguez, P., Vazquez, D., Nowrouzezahrai, D.: Ssr: Semi-supervised soft rasterizer for single-view 2d to 3d reconstruction. In: ICCVW (2021) 4
18. Li, H., Wu, Z., Shrivastava, A., Davis, L.S.: Rethinking pseudo labels for semi-supervised object detection. In: AAAI (2022) 8
19. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: ICLR (2021) 2, 8
20. Michalkiewicz, M., Parisot, S., Tsogkas, S., Baktashmotlagh, M., Eriksson, A., Belilovsky, E.: Few-shot single-view 3-d object reconstruction with compositional priors. In: ECCV (2020) 9, 12, 14

21. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: ECCV (2018) 4
22. Richter, S.R., Roth, S.: Matryoshka networks: Predicting 3d geometry via nested shape layers. In: CVPR (2018) 3
23. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: NeurIPS (2016) 4
24. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) 1
25. Shi, Z., Meng, Z., Xing, Y., Ma, Y., Wattenhofer, R.: 3d-retr: End-to-end single and multi-view 3d reconstruction with transformers. In: BMVC (2021) 3
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 10
27. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020) 2, 4, 7, 8, 10, 11, 12
28. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: CVPR (2018) 10, 11
29. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017) 2, 4, 8, 10, 11, 12
30. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: ICCV (2017) 3
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 6
32. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors. In: ICCV (2019) 7, 9, 12, 14
33. Wang, D., Cui, X., Chen, X., Zou, Z., Shi, T., Salcudean, S., Wang, Z.J., Ward, R.: Multi-view 3d reconstruction with transformers. In: ICCV (2021) 3, 6
34. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: ECCV (2018) 2
35. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y.G.: Semi-supervised vision transformers. In: ECCV (2022) 2
36. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: Marrnet: 3d shape reconstruction via 2.5 d sketches. In: NeurIPS (2017) 3
37. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: NeurIPS (2016) 3, 7
38. Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W.T., Tenenbaum, J.B.: Learning shape priors for single-view 3d completion and reconstruction. In: ECCV (2018) 3, 7
39. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: ICCV (2019) 2, 3, 6, 9, 10, 12
40. Xie, H., Yao, H., Zhang, S., Zhou, S., Sun, W.: Pix2vox++: multi-scale context-aware 3d object reconstruction from single and multiple images. IJCV (2020) 3, 9, 10
41. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019) 8

42. Yang, G., Cui, Y., Belongie, S., Hariharan, B.: Learning single-view 3d reconstruction with limited pose supervision. In: ECCV (2018) 4
43. Yang, S., Xu, M., Xie, H., Perry, S., Xia, J.: Single-view 3d object reconstruction from shape priors in memory. In: CVPR (2021) 3, 7
44. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019) 4
45. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018) 4
46. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J., Freeman, B., Wu, J.: Learning to reconstruct shapes from unseen classes. In: NeurIPS (2018) 3