**Point-Syn2Real: Semi-Supervised Synthetic-to-Real Cross-Domain Learning for Object Classification in 3D Point Clouds**

Ziwei Wang*, Reza Arablouei*, Jiajun Liu*, Paulo Borges*, Greg Bishop-Hurley†, and Nicholas Heaney‡

*CSIRO Data61 †CSIRO Agriculture & Food ‡Evolve Group

{ziwei.wang, reza.arablouei, jiajun.liu, paulo.borges, greg.bishop-hurley}@csiro.au, nicheaney@me.com

**Abstract**—Object classification using LiDAR 3D point cloud data is critical for modern applications such as autonomous driving. However, labeling point cloud data is labor-intensive as it requires human annotators to visualize and inspect the 3D data from different perspectives. In this paper, we propose a semi-supervised cross-domain learning approach that does not rely on manual annotations of point clouds and performs similar to fully-supervised approaches. We utilize available 3D object models to train classifiers that can generalize to real-world point clouds. We simulate the acquisition of point clouds by sampling 3D object models from multiple viewpoints and with arbitrary partial occlusions. We then augment the resulting set of point clouds through random rotations and adding Gaussian noise to better emulate the real-world scenarios. We then train point cloud encoding models on the synthesized and augmented datasets and evaluate their cross-domain classification performance on corresponding real-world datasets. We also introduce PointSyn2Real, a new benchmark dataset for cross-domain learning on point clouds. The results of our extensive experiments with this dataset demonstrate that the proposed cross-domain learning approach for point clouds outperforms the related baseline and state-of-the-art approaches in both indoor and outdoor settings in terms of cross-domain generalizability.

**Index Terms**—3D object detection, point cloud, deep learning, semi-supervised

I. INTRODUCTION

Detecting and identifying the objects present in a scene is a crucial but challenging task in machine learning. Substantial efforts have been made to develop algorithms that can accurately recognize objects [1], [2], detect objects [3], [4], track objects, and recognize human-object interactions [5] using images or videos captured by cameras. Although object detection in computer vision has significantly advanced in recent years, there exist fundamental limitations. For example, most cameras have difficulty capturing clear images in low-light or excessive-light conditions, or determining the exact distance of objects in 2D images is challenging. In addition, privacy concerns often arise around images as they may contain private information easily perceivable by humans.

Light detection and ranging (LiDAR) sensors use laser beams to scan their surroundings and construct 3D representations of the objects within. The scanned 3D snapshots are stored as the so-called point clouds. Detecting objects from 3D point clouds can help resolve some of the challenges associated with image-based object detection. A LiDAR scanner can obtain precise information of object positions and shapes regardless of lighting conditions. Moreover, as point clouds are less perceivable by humans, they can help enhance privacy preservation. Given the above advantages, computer vision applications can benefit from 3D point clouds. For example, in autonomous driving, when the visibility is poor, LiDAR sensors can help detect obstacles.

Recognizing objects in point clouds using machine learning has been studied by many researchers. Conventional approaches [6], [7] use carefully designed features to represent various shapes in point clouds. More recently, deep learning (DL) models have been used to learn point-level and object-level features in an end-to-end manner. A notable example is PointNet [8] that uses end-to-end DL. It has led to significant improvement in point cloud classification and segmentation performance. Ensuing methods, including PointNet++ [9] and DGCNN [10], consider local neighborhood information for refined feature extraction. Nonetheless, training DL models...
We provide the results of extensive experiments using data generated via the approach described in Section II. Our key contributions in this paper are:

1) We introduce a novel semi-supervised cross-domain learning approach that can generalize the knowledge learned from synthetic 3D point clouds to real-world data collected by LiDAR scanners. We use random rotation/noise addition augmentation, multi-view simulation, and entropy minimization to enhance the robustness and performance of the learned models.

2) We create a comprehensive synthetic-to-real cross-domain 3D point cloud dataset as a benchmark, which includes indoor and outdoor scenarios.

3) We provide the results of extensive experiments using data of both indoor and outdoor settings, and demonstrate the effectiveness of the proposed approach.

II. RELATED WORK

PointNet [8] is one of the first DL-based end-to-end models that can directly process raw point cloud data. It calculates point-level features, which can be aggregated via max-pooling to produce global features. PointNet delivers promising results in point cloud classification and segmentation tasks. However, its functionality is limited as it only considers global features and pays less attention to local geometric features. As an improvement to PointNet, PointNet++ [9] introduces additional sampling and grouping layers to leverage local information. DGCNN [10] is another DL-based point cloud encoder that builds a nearest-neighbor graph to incorporate the local and global geometric information. These models exhibit good performance when trained and evaluated on data from the same domain. However, they usually do not perform well when they are trained on data from one domain and evaluated on data from another domain. Therefore, models trained on synthetic 3D datasets such as ModelNet [12] and ShapeNet [13] may not perform well on real-world datasets such as ScanNet [14].

Obtaining labeled real-world point cloud data for training is challenging due to the associated labor costs or time constraints.
Cross-domain learning utilizes readily-available synthetic data (source domain) for training and adapts the learned model to perform inference on real-world data (target domain) with limited annotations. There are a few existing works that address cross-domain learning with point cloud data. PointDAN [15] aligns the local and global features to mitigate distribution shift between the source and target domains. DefRec [16] learns a representation model by reconstructing point clouds with induced deformations. These methods are useful for adapting models from a data distribution perspective. However, they do not explicitly address challenges involved in real-world data acquisition such as partial occlusions and viewpoint variations. Other cross-domain learning methods such as [17]–[19] deal with point-cloud-related tasks other than object classification.

III. Proposed Approach

We provide a visual overview of the proposed approach in Figure 2. During training, we use a 3D computer-aided design (CAD) tool to generate a set of partially-occluded point clouds taken from multiple viewpoints for each considered object. We denote the generated point cloud set associated with the ith object as $P_i = \{P_{ij1}, P_{ij2}, \ldots, P_{ijM}\}$ where $M$ is the number of viewpoints. Each point cloud is a set of points in the Euclidean space, i.e., $P_{ij} = \{p_{ij1}, p_{ij2}, \ldots, p_{ijN_{ij}}\}$ where $N_{ij}$ is the number of points in $P_{ij}$ and each point $p_{ijk}$ has three coordinate values $(x_{ijk}, y_{ijk}, z_{ijk})$. We utilize unlabeled real-world point clouds from the target domain to realize semi-supervise learning. Thus, we denote the set of real point clouds associated with the ith object as $S_i$. We augment the set of synthetic point clouds by applying multiple random rotations and adding Gaussian noise. We then feed the augmented synthetic data into a DL-based point cloud encoder to extract point-level features. We aggregate these features via max-pooling before forwarding to a fully-connected (FC) classification layers, which output predicted posterior probabilities for each object class using the softmax function. To jointly train the neural networks of the encoder and the classifier, and describe the unified learning objective that consists of cross entropy loss for labeled synthetic data, and entropy loss for unlabeled real data.

$A. \textbf{Multi-view Point Cloud Simulation}$

In LiDAR scans, the objects of interest may be occluded by other objects or even themselves. In our cross-domain learning approach, we simulate occlusions in synthesizing the training set similar to RotationNet [20], which takes snapshots from multiple viewpoints to create multi-view 2D images, one can synthesize multi-view 3D point clouds. In particular, given a 3D object model, one can simulate realistic LiDAR scans and generate multiple point clouds of the object from different viewpoints. In this work, we utilize the open-source software Blender to create synthetic point clouds for training. The procedure has two major steps, i.e., depth map simulation and back projection. In Figure 3, we illustrate an example synthetic point cloud dataset with six objects and four viewpoints.

$B. \textbf{Data Augmentation}$

We augment the synthesized partially-occluded point clouds by applying random rotations and adding Gaussian noise to improve the robustness and accuracy of the learned models.

1) Random Rotation: We rotate each point cloud around the z-axis by a uniformly-distributed random angle, i.e., $\varphi \in [0, 2\pi]$. The rotation of every point of the point cloud $P_{ij}$, i.e., $p_{ijk} = (x_{ijk}, y_{ijk}, z_{ijk})$, around the z-axis by $\varphi$ is expressed via the following linear transformation

\[
\begin{bmatrix}
    x'_{ijk} \\
y'_{ijk} \\
z'_{ijk}
\end{bmatrix} = \begin{bmatrix}
    \cos(\varphi) & \sin(\varphi) & 0 \\
    -\sin(\varphi) & \cos(\varphi) & 0 \\
    0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
    x_{ijk} \\
y_{ijk} \\
z_{ijk}
\end{bmatrix}
\]

We denote the rotated point cloud as $P'_{ij}$.

2) Gaussian Noise: When collecting data in real world, sensing imperfections due to, e.g., measurement noise or error, may corrupt the data. Therefore, to make our synthetic point cloud data more realistic, we add noise to the values of each synthetic point as

\[
p'_{ijk} = p_{ijk} + \nu_{ijk},
\]

where $\nu_{ijk} = \nu_{x_{ijk}}, \nu_{y_{ijk}}, \nu_{z_{ijk}}$ is the additive noise with $\nu_{x_{ijk}}, \nu_{y_{ijk}}, \nu_{z_{ijk}} \in \mathbb{C}, (x, y, z)$, being independently drawn from a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.01$.

$C. \textbf{Cross-domain Point Cloud Encoder}$

In this section, we elaborate the point cloud encoder and classifier, and describe the unified learning objective that consists of cross entropy loss for labeled synthetic data, and entropy loss for unlabeled real data.

Given a partially-occluded and randomly-augmented point cloud $P_{ij}'$, the point cloud encoder, denoted by the function $f(\cdot)$, takes the point cloud as the input and outputs the point-level feature vector of dimension $D$, i.e., $f_{ij} = f(P_{ij}')$. The point features are then aggregated using an effective symmetric aggregation function, i.e., max-pooling denoted by max-pool(·), to produce the pooled global feature vector, i.e., $g_{ij} = \text{max-pool}(f_{ij}) \in \mathbb{R}^D$. The global feature vector is then passed through a classifier, $h(\cdot)$, that is a multilayer fully-connected neural network (perceptron) and outputs the logits for each class. The softmax(·) function is applied to the logits to produce the class-wise posterior probabilities, denoted by $q_{ij} \in \mathbb{C}^C$ where $C$ is the number of classes, i.e.,

\[
q_{ij} = \text{softmax}(h(g_{ij})).
\]

We calculate the categorical cross-entropy loss that evaluates the divergence between the predicted posterior probabilities and the ground-truth label as

\[
l_{ij} = -y_{ij}^T \log(q_{ij})
\]

where $y_{ij} \in \mathbb{C}^C$ is the one-hot vector for the ground-truth label corresponding to $P_{ij}$. We use a weighted version of the cross-entropy loss to mitigate the impact of class imbalance.
To exploit the information available through the unlabeled real point clouds from the target domain, we utilize the entropy loss function calculated as

$$\ell_{it} = -s_{it}^T \log(s_{it})$$

(4)

where $s_{it}$ is the vector of posterior probabilities predicted by the model for the $i$th real point cloud of the $t$th object available for training. Minimizing the entropy loss for unlabeled data encourages the learned model to make more confident predictions, which can in turn improve its performance. The unified objective function that we minimize during training is the weighted average of the cross-entropy and entropy losses for all available synthetic and real point clouds.

The specifications of the encoder is not central to our approach. Hence, any suitable point cloud encoder can be used. In this work, we primarily use DGCNN [10] as the point cloud encoder since it is efficient and leads to good performance. We also consider using PointNet++ [9] as our point cloud encoder and demonstrate consistent result. The detailed study can be found in the supplementary material.

IV. Evaluation

A. Point-Syn2Real Dataset

We compile a new benchmark dataset, called Point-Syn2Real, by gathering data from multiple sources. The dataset can be used to evaluate the performance of cross-domain learning methods that involve transferring knowledge from synthetic 3D data to real-world point cloud data. Point-Syn2Real covers both indoor and outdoor settings as shown in Table I. The multiview synthesis considerably increases the number of instances available for training. The detailed dataset statistics can be found in supplementary materials.

B. Experimental Details

We use DGCNN [10] with the neighborhood size of $k = 20$ as the point cloud encoder. For training, we use a batch size of 32 and a maximum epoch number of 80. We use the Adam optimization algorithm with the learning rate set to $0.001$ and the weight decay to $5 \times 10^{-5}$. The values of other hyperparameters can be found in the provided code.

In our evaluations, we use common classification metrics including overall accuracy and weighted average F1-score [15]. In addition, we calculate the Matthews correlation coefficient (MCC) to measure the performance of multi-class classification, especially given that the considered cross-domain datasets are imbalanced. For semi-supervised cross-domain learning, we use labeled synthetic data from the source domain and unlabeled real point clouds from the target domain to train the model. We use a small set of labeled real point clouds from the target domain to validate the model fit and tune the hyper-parameters. We evaluate the eventual learned model on a held-out target-domain test set that is unseen during the training.

We compare the performance of the proposed approach with a number of existing baseline and state-of-the-art approaches as listed bellow. **Supervised**: The model trained on labeled real-world data from the target domain, which sets an upper bound. **Baseline**: The model trained only on the synthetic data of the source domain with no domain adaptation, multi-view simulation, or random augmentation. **Point-Syn2Real**: The proposed approach. And a number of state-of-the-art methods: **PointDAN** [15], **MMD** [21], **DANN** [22], **DefRec+PCM** [16].

We denote the multiview point cloud simulation described in section III-A by S, the data augmentation described in section III-B by A, and the inclusion of entropy loss for semi-supervised learning by E.

C. Indoor Object Classification

The results presented in Table II for the ModelNet to ScanNet case show that when augmented via random rotations and additive Gaussian noise, the proposed approach outperforms the earlier domain adaptation method PointDAN, which aligns the distribution of the features learned in the source and target domains. There is a similar observation for the ShapeNet to ScanNet case where the accuracy is improved from 33.90% (for PointDAN) to 50.37%.

However, the augmentation alone does not represent the real world, since objects may face different directions in the point cloud coordinate system when they are scanned in the real world. To account for this, we generate simulated training data from multiple view points which results in slightly varied samples of the same object. Benefiting from both augmentation
(A) and multiview simulation (S), Point-Syn2Real A+S, further improves the performances. We incorporate the knowledge of the unlabeled target data into the training to further regularize the model using the entropy loss for the unlabeled data (E) and adapt it to the target domain. The full model, Point-Syn2Real A+S+E, outperforms the state-of-the-art approach DefRec+PCM by 7.33% and 8.98% in the ModelNet to ScanNet and ShapeNet to ScanNet cases, respectively. MCC score is also improved significantly compared to all the existing methods. Overall, it is evident that the proposed approach offers significant performance improvement in the indoor settings. In addition, our experiments demonstrate that our semi-supervised learning approach through the use of information entropy loss for unlabeled data outperforms more complex domain adaptation methods. We provide the detailed ablation study and comparison with the existing domain-adaptation-based methods, e.g., MMD and DANN, in Section IV-E and within the supplementary material.

D. Outdoor Object Classification

For evaluation on outdoor objects, we extract the real object scans from the SemanticKitti [23] autonomous driving dataset. We train the model on labeled synthetic source domain, i.e., 3D City, and adapt the model to target domain during the training with unlabeled real point cloud from SemKitti_Obj training split. A held-out test split from target domain is used for testing. In Table III, we present the performance evaluation results for the considered outdoor setting.

Both Point-Syn2Real A and Point-Syn2Real A+S perform better than the Baseline approach attesting to the effectiveness of the utilized random augmentation and multiview simulation. MMD has high accuracy and F-1 score, close to those of the Supervised approach. However, its MCC value is substantially lower than that of “Supervised”. This is mainly because MMD is able to classify the more common classes with good accuracy but it fails with the classes that have low frequency. The proposed Point-Syn2Real A+S+E approach offers significant improvements over other considered approaches in terms of all three metrics and draws close to the upper bounds set by the Supervised approach. This means that the combination of random augmentation, multiview simulation, and semi-supervised learning appreciably enhances the ability of cross-domain object classification models to generalize to the target domain with minimal supervision.

E. Ablation Study

We conduct an ablation study to better understand the relative contribution of each component in the proposed Point-Syn2Real approach. In Tables II and III, we examine the benefits of including augmentation (A), multiview simulation (S), and entropy loss (E). For the considered indoor settings, as shown in Table II, including the random augmentation alone increases the accuracy significantly, i.e., from 31.09% to 51.33% in the ModelNet to ScanNet case and from 24.02% to 50.37% in the ShapeNet to ScanNet case, compared to Baseline. This suggests that augmentation is an effective way of enhancing the generalization capacity with relatively small training datasets such as ModelNet.

Semi-supervised learning through the use of the entropy loss improves the performance in both indoor and outdoor settings. Especially, in the considered outdoor setting, it increases the MCC value from 0.50 to 0.57. Minimizing the entropy of
the posterior class probabilities predicted by the classifier for the unlabeled target training data encourages the classifier to make more confident predictions. This helps the learned model better generalize to unseen data from the target domain. The inclusion of the entropy loss can also be perceived as a form of regularization that prevents the learned model from overfitting to the source domain without relying on any labeled data from the target domain. Figure 4 shows the class-wise accuracy of the Baseline, MMD, and Point-Syn2Real A+S+E approaches for the ModelNet to ScanNet case. The results indicate that Point-Syn2Real has the best accuracy for most classes. Especially, the accuracy for the Chair class is about 70% with Point-Syn2Real while it is around 30% with Baseline and MMD. The accuracy of Point-Syn2Real is lower than that of MMD for only three classes of Lamp, Monitor, and Plant. It is also interesting to observe that MMD is less accurate than Baseline for five classes. In general, there appears ample room for further improvement considering the class-wise accuracy values, although our proposed approach achieves appreciable improvement over the state-of-the-art.

V. CONCLUSION

We introduced a synthetic-to-real semi-supervised cross-domain learning approach, named Point-Syn2Real, to learn 3D point cloud classification models that can generalize from synthetic domain to real world. Point-Syn2Real produces synthetic point clouds by simulating their LiDAR scans from multiple viewpoints while inducing partial occlusions that may occur in real-world 3D scans. It then augments the simulated point clouds by applying random rotations and Gaussian noise. The synthesized point clouds are then used to train the object classifier that includes a suitable point cloud encoder. To mitigate the likelihood of overfitting to the synthetic data of the source domain and hence improve the performance, we incorporate the entropy associated with the available unlabeled real data from the target domain into the training loss. Through extensive experiments with synthetic and real data in both indoor and outdoor settings on our proposed new point cloud benchmark dataset, we showed that Point-Syn2Real outperforms several relevant existing state-of-the-art approaches.

This work is supported by CSIRO’s Research Plus Science Leader Project R-91559.